PH.D. SCHOOL
UNIVERSITY OF MILANO-BICOCCA

DEPARTMENT OF STATISTICS AND QUANTITATIVE METHODS
Ph.D.: IN STATISTICS AND MATHEMATICS FOR FINANCE - XXXI Cycle

Curriculum: Statistics

Bayesian Variable Selection in Multinomial Logistic Regression: a Conditional Latent Approach

Ph.D. Dissertation of: Alexios Polymeropoulos

Supervisor: Prof. Guido Consonni
Co-Supervisor: Prof. Ioannis Ntzoufras
Tutor: Prof. Guido Consonni
Ph.D. Coordinator: Prof. Giorgio Vittadini

Academic Year 2019-2020

Acknowledgements

I would like to express my gratitude and respect for Prof. Guido Consonni for enlighting my future here in italy. I would like to thank him for being my tutor from the early beginning of this phd program and providing me important consulting in order to be deligent for the demanding prerequisites of course-exams. Moreover, i thank him for his guidance and patience during this difficult stage of phd and thesis. Finally, i thank him for trusting and believing in me, when the things got difficult.

I would like also to thank Prof. Ioannis Ntzoufras, for his encouragement during the hard times of this phd program and for his creative and interactive discussions for the accomplishment of this phd thesis. I thank him for introducing me in this fantastic journey of Bayesian inference and MCMC methods during my master studies in Athens.

I would like to thank all the phd-teaching committee in order to provide us insightful comments and suggestions during the first year of hard exams.

I want to thank my colleagues for all the collaboration, time spent of joking and exchanging skills of knowledge.

I cannot forget also all my family for the moral and financial support. I would like to thank my parents who helped me to continue my studies and overcome many difficulties arose during the whole phd period.

i would like to thank especially my girlfriend Karen for being next to me all the time, in good and difficult moments, of this phd dissertation.

Abstract

Mixtures of g-priors are well established in linear regression models by Liang et al. (2008) and generalized linear models by Bové and Held (2011) and Li and Clyde (2013) for variable selection. This approach enables us to overcome the problem of specifying the dispersion parameter by imposing a hyper-prior on it. By this way we allow for our model to "learn" about the shrinkage from the data. In this work, we implement Bayesian variable selection methods based on g-priors and their mixtures in multinomial logistic regression models. More precisely, we follow two approaches: (a) the traditional implementation by extending the approach of Bové and Held (2011) for multinomial models, and (b) an augmented implementation of Polson et al. (2013) based on latent structure. We will study and compare the two approaches. Furthermore, we will focus on handling class imbalance and sparsity issues appearing when the number of covariates is moderate and the need of specifying different covariate selection across different pairwise logit structures. All proposed methods will be presented in simulation and real datasets. Extensive comparisons and results are also presented for logistic regression in real and simulated settings.

Table of contents

List of figures **vii**

List of tables **xi**

1 Introduction **1**

1.1 Thesis Overview 2

1.2 Statistical Models for Bayesian Model Selection 4

1.2.1 Definition of Statistical and Bayesian Models 4

1.2.2 Bayesian Model Selection 5

1.2.2.1 Posterior Measures of Evidence 5

1.2.2.2 Model Set Approaches 8

1.2.2.3 Jeffrey's-Lindley's Paradox 11

1.3 Objective Bayesian Approach 11

1.3.1 Prior Elicitation 12

1.3.2 Objective Prior Specification 13

1.3.3 Objective Model Selection 15

2 Bayesian Variable Selection in Linear Regression **17**

2.1 The problem of Bayesian Variable Selection in Linear Models 18

2.1.1 Prior Elicitation 20

2.1.2 Prior Choice 20

2.1.2.1 Base Model Considerations and Jeffrey's Approach . . 21

2.1.2.2 Default Prior Choice 22

2.1.2.3 Prior Choice for Model Space 26

2.1.3 Model Choice 27

2.1.4 Model Choice with Mixtures of g-Priors 30

2.1.4.1 Hierachical Prior Specification 31

2.1.4.2 Model Choice with Zellner-Siow Priors 32

2.1.4.3 Model Choice with Hyper-g-Priors 34
2.2 MCMC for Bayesian variable selection in Linear Models 36
2.2.1 Hierarchical Prior Specification for MCMC 38
2.2.2 Stochastic Search Variable Selection 39
2.2.3 Gibbs Variable Selection 42
2.2.4 Posterior Exploration of Model Space 45
2.3 Closing Remarks 46

3 Bayesian Variable Selection in Generalized Linear Models 48
3.1 The problem of Bayesian variable selection in Generalized Linear Models 50
3.1.1 Prior Elicitation 51
3.1.2 Prior Choice 52
3.1.2.1 Default Prior Choice 53
3.1.2.2 Prior and Model Choice of Bove and Held Approach . 55
3.1.2.3 Prior and Model Choice of Li and Clyde Approach . . 59
3.1.2.4 Confluent Hypergeometric Function 62
3.2 Closing Remarks 64

4 Bayesian Variable Selection in Multinomial Logistic Regression 66
4.1 The problem of Bayesian Variable Selection in Multinomial Logistic Regression Models 73
4.1.1 Prior Elicitation 75
4.1.2 Default Prior Choice 76
4.1.3 Prior Choice of Bove and Held Approach 79
4.2 MCMC for Bayesian Variable Selection in Multinomial Logistic Models 83
4.3 SSVS in Typical Multinomial Logistic Setup 86
4.4 GVS in Typical Multinomial Logistic Setup 93
4.5 SSVS vs GVS Within Typical Logistic Multinomial Setup 99
4.6 Polya-Gamma Data Augmentation 100
4.7 SSVS in Augmented Multinomial Logistic Setup 104
4.8 GVS in Augmented Multinomial Logistic Setup 111
4.9 SSVS vs GVS Within Augmented Multinomial setup 118
4.10 Simulated Experiments 119
4.11 Real Application 139
4.12 Closing Remarks 150

5 Conclusions and Future Research 153

A Bayesian Variable Selection in Linear Regression 156

A.1 Posterior Interpretation of g-Prior 156

A.2 Deriving the Joint Posterior of a, $\boldsymbol{\beta}_\gamma$ and σ^2 157

A.3 Laplace Approximation 160

A.4 Laplace Approximation for Zellner-Siow Priors 161

A.5 Calculation of Bayes Factor and Posterior Moments of g and $\frac{g}{g+1}$ 163

A.6 Laplace Approximation for Hyper-g Priors 167

A.7 Implementation of Stochastic Search Variable Selection 169

A.8 Implementation of Gibbs Variable Selection 171

A.9 Simulated Experiment 174

A.10 Real Dataset 182

B Bayesian Variable Selection in Generalized Linear Models 196

B.1 Laplace Approximation of Bove and Held 196

B.2 Gaussian Quadrature Approximation with Mixtures Of g-priors 197

B.3 Laplace Approximation of Li and Clyde 200

B.4 Hyper-g-prior Laplace Approximation of Li and Clyde 202

B.5 Confluent Hypergeometric Laplace Approximation of Li and Clyde . . . 203

C Bayesian Variable Selection in Multinomial Logistic Regression 205

C.1 Proof of SSVS in Augmented Multinomial Logistic Setup 205

C.2 Proof of GVS in Augmented Multinomial Logistic Setup 208

C.3 Bayesian Variable Selection in Logistic Regression 212

C.4 Simulated Experiments 212

C.5 Real Dataset 231

References 251

List of figures

4.1 Posterior inclusion probabilities for 100 repetitions of no-overlapping scenario regarding specific class 2 versus baseline class 1 128
4.2 Posterior inclusion probabilities for 100 repetitions of no-overlapping scenario regarding specific class 3 versus baseline class 1 128
4.3 Posterior inclusion probabilities for 100 repetitions of overlapping scenario regarding specific class 2 versus baseline class 1 129
4.4 Posterior inclusion probabilities for 100 repetitions of overlapping scenario regarding specific class 3 versus baseline class 1 129
4.5 Posterior inclusion probabilities for 100 repetitions of no-overlapping scenario regarding specific class 2 versus baseline class 1 (n=500) 134
4.6 Posterior inclusion probabilities for 100 repetitions of no-overlapping scenario regarding specific class 3 versus baseline class 1 (n=500) (rowwise largest value in bold) 134
4.7 Posterior inclusion probabilities for 100 repetitions of overlapping scenario regarding specific class 2 versus baseline class 1 (n=500) 135
4.8 Posterior inclusion probabilities for 100 repetitions of overlapping scenario regarding specific class 2 versus baseline class 1 (n=500) 135
4.9 Distributions of $\frac{g}{g+1}$ for 100 repetitions of no-overlapping scenario . . . 136
4.10 Distributions of $\frac{g}{g+1}$ for 100 repetitions of no-overlapping scenario . . . 136
4.11 Distributions of $\frac{g}{g+1}$ for 100 repetitions of overlapping scenario 137
4.12 Distributions of $\frac{g}{g+1}$ for 100 repetitions of overlapping scenario 137
4.13 Model selection consistency of SSVS and GVS in typical vs augmented multinomial logistic setup respectively. 138
4.14 Traceplots of shrinkage factor $\frac{g}{g+1}$. 147
4.15 Ergodic means of shrinkage factor $\frac{g}{g+1}$. 147
4.16 Posteriors of accuracy $\widehat{ACC}$. 148
4.17 Posteriors of Coehn's kappa $\widehat{\kappa}_c$. 149

A.1 Multiple scatterplots of the dependent variable Y versus each covariate $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, $\boldsymbol{X_3}$, $\boldsymbol{X_4}$, $\boldsymbol{X_5}$. 174
A.2 Convergence diagnostics of shrinkage factor $\frac{g}{g+1}$ for ssvs$^{\mathrm{R}}$.ZS. 180
A.3 Convergence diagnostics of shrinkage factor $\frac{g}{g+1}$ for gvs$^{\mathrm{R}}$.ZS. 181
A.4 Convergence diagnostics of shrinkage factor $\frac{g}{g+1}$ for ssvs$^{\mathrm{R}}$.hyp. 181
A.5 Convergence diagnostics of shrinkage factor $\frac{g}{g+1}$ for gvs$^{\mathrm{R}}$.hyp. 182
A.6 Convergence diagnostics of shrinkage factor $\frac{g}{g+1}$ for ssvs$^{\mathrm{R}}$.ZS. 188
A.7 Convergence diagnostics of shrinkage factor $\frac{g}{g+1}$ for gvs$^{\mathrm{R}}$.ZS. 188
A.8 Convergence diagnostics of shrinkage factor $\frac{g}{g+1}$ for ssvs$^{\mathrm{R}}$.hyp. 189
A.9 Convergence diagnostics of shrinkage factor $\frac{g}{g+1}$ for gvs$^{\mathrm{R}}$.hyp. 189
A.10 Convergence diagnostics of mean squared error MSE for ssvs$^{\mathrm{R}}$.g. 190
A.11 Convergence diagnostics of mean squared error MSE for gvs$^{\mathrm{R}}$.g. 191
A.12 Convergence diagnostics of mean squared error MSE for ssvs$^{\mathrm{R}}$.ZS. . . . 192
A.13 Convergence diagnostics of mean squared error MSE for gvs$^{\mathrm{R}}$.ZS. . . . 193
A.14 Convergence diagnostics of mean squared error MSE for ssvs$^{\mathrm{R}}$.hyp. . . 194
A.15 Convergence diagnostics of mean squared error MSE for gvs$^{\mathrm{R}}$.hyp. . . . 195

C.1 Posterior inclusion probabilities under the various methods from 100 repetitions of independent ($r = 0$) covariates for different scenarios. . . 220
C.2 Posterior inclusion probabilities under the various methods from 100 repetitions of correlated ($r = 0.75$) covariates for different scenarios. . . 221
C.3 Posterior regression coefficients under the various methods from 100 repetitions of independent ($r = 0$) covariates for different scenarios. . . 227
C.4 Posterior regression coefficients under the various methods from 100 repetitions of correlated ($r = 0.75$) covariates for different scenarios. . . 228
C.5 Posterior inclusion probabilities under the various methods from 100 repetitions of independent ($r = 0$) covariates for different scenarios ($n = 500$). 229
C.6 Posterior inclusion probabilities under the various methods from 100 repetitions of correlated ($r = 0.75$) covariates for different scenarios ($n = 500$). 230
C.7 Posterior density plots, ergodic mean plots, autocorrelation plots and traceplots of the shrinkage factor $\frac{g}{g+1}$ for typical and augmented logistic regression methods ssvs$^{\mathrm{R}}$.hyp.typ and ssvs$^{\mathrm{R}}$.hyp.aug with hyper-g prior. 240
C.8 Posterior density plots, ergodic mean plots, autocorrelation plots and traceplots of the shrinkage factor $\frac{g}{g+1}$ for typical and augmented logistic regression methods gvs$^{\mathrm{R}}$.hyp.typ and gvs$^{\mathrm{R}}$.hyp.aug with hyper-g prior. 240

C.9 Posterior density plots, ergodic mean plots, autocorrelation plots and traceplots of the shrinkage factor $\frac{g}{g+1}$ for typical and augmented logistic regression methods ssvs$^{\mathrm{R}}$.ZS.typ and ssvs$^{\mathrm{R}}$.ZS.aug with Zellner-Siow prior. 241

C.10 Posterior density plots, ergodic mean plots, autocorrelation plots and traceplots of the shrinkage factor $\frac{g}{g+1}$ for typical and augmented logistic regression methods gvs$^{\mathrm{R}}$.ZS.typ and gvs$^{\mathrm{R}}$.ZS.aug with Zellner-Siow prior. 241

C.11 Posterior density plots, ergodic mean plots, autocorrelation plots and traceplots of false negative $\widehat{FN}$, false positive $\widehat{FP}$, accuracy $\widehat{ACC}$ and missclassification error $\widehat{ERR}$ for typical and augmented logistic regression methods ssvs$^{\mathrm{R}}$.hyp.typ and ssvs$^{\mathrm{R}}$.hyp.aug with hyper-g priors under the MAP and MPM. 243

C.12 Posterior density plots, ergodic mean plots, autocorrelation plots and traceplots of false negative $\widehat{FN}$, false positive $\widehat{FP}$, accuracy $\widehat{ACC}$ and missclassification error $\widehat{ERR}$ for typical and augmented logistic regression methods gvs$^{\mathrm{R}}$.hyp.typ and gvs$^{\mathrm{R}}$.hyp.aug with hyper-g priors under the MAP. 244

C.13 Posterior density plots, ergodic mean plots, autocorrelation plots and traceplots of false negative $\widehat{FN}$, false positive $\widehat{FP}$, accuracy $\widehat{ACC}$ and missclassification error $\widehat{ERR}$ for typical and augmented logistic regression methods gvs$^{\mathrm{R}}$.hyp.typ and gvs$^{\mathrm{R}}$.hyp.aug with hyper-g priors under the MPM. 245

C.14 Posterior density plots, ergodic mean plots, autocorrelation plots and traceplots of false negative $\widehat{FN}$, false positive $\widehat{FP}$, accuracy $\widehat{ACC}$ and missclassification error $\widehat{ERR}$ for typical and augmented logistic regression methods ssvs$^{\mathrm{R}}$.ZS.typ and ssvs$^{\mathrm{R}}$.ZS.aug with Zellner-Siow priors under the MAP and MPM. 246

C.15 Posterior density plots, ergodic mean plots, autocorrelation plots and traceplots of false negative $\widehat{FN}$, false positive $\widehat{FP}$, accuracy $\widehat{ACC}$ and missclassification error $\widehat{ERR}$ for typical and augmented logistic regression methods gvs$^{\mathrm{R}}$.ZS.typ and gvs$^{\mathrm{R}}$.ZS.aug with Zellner-Siow priors under the MAP and MPM. 247

C.16 Posterior density plots, ergodic mean plots, autocorrelation plots and traceplots of false negative $\widehat{FN}$, false positive $\widehat{FP}$, accuracy $\widehat{ACC}$ and missclassification error $\widehat{ERR}$ for typical and augmented logistic regression methods gvs$^{\mathrm{R}}$.ZS.typ and gvs$^{\mathrm{R}}$.ZS.aug with Zellner-Siow priors under the MAP and MPM. 248

C.17 Posterior density plots, ergodic mean plots, autocorrelation plots and traceplots of false negative $\widehat{FN}$, false positive $\widehat{FP}$, accuracy $\widehat{ACC}$ and missclassification error $\widehat{ERR}$ for typical and augmented logistic regression methods gvs$^{\mathrm{R}}$.g.typ and gvs$^{\mathrm{R}}$.g.aug with g-priors under the MAP. . 249

C.18 Posterior density plots, ergodic mean plots, autocorrelation plots and traceplots of false negative $\widehat{FN}$, false positive $\widehat{FP}$, accuracy $\widehat{ACC}$ and missclassification error $\widehat{ERR}$ for typical and augmented logistic regression methods gvs$^{\mathrm{R}}$.g.typ and gvs$^{\mathrm{R}}$.g.aug with g-priors under the MPM. 250

List of tables

1.1 Bayes factor interpretation of $log_{10} BF_{[M_1:M_0]}$. 7

1.2 Bayes factor interpretation of $log_e BF_{[M_1:M_0]}$. 7

4.1 Multinomial logistic regression sparse scenarios using independent variables 119

4.2 Prior-inputs and initial values 120

4.3 Acronyms of Bayesian variable selection methods with MCMC for multinomial logistic regression 121

4.4 Number of 100 simulated samples that the MAP coincides with the true generating model of Table (4.1) for no-overlapping scenario for specific class 2 and 3 versus baseline class 1 (row-wise largest value in bold) . . 128

4.5 Number of 100 simulated samples that the MAP coincides with the true generating model of Table (4.1) for overlapping scenario for specific class 2 and 3 versus baseline class 1 (row-wise largest value in bold) 129

4.6 Effective sample size comparison (in bold lowest value). 130

4.7 Monte Carlo error comparison (in bold lowest value). 130

4.8 Effective sample size comparison (in bold lowest value). 131

4.9 Monte Carlo error comparison (in bold lowest value). 131

4.10 Effective sample size comparison (in bold lowest value). 132

4.11 Monte Carlo error comparison (in bold lowest value). 132

4.12 Effective sample size comparison (in bold lowest value). 133

4.13 Monte Carlo error comparison (in bold lowest value). 133

4.14 Number of 100 simulated samples that the MAP coincides with the true generating model of Table (4.1) for no-overlapping scenario for specific class 2 and 3 versus baseline class 1 (n=500) (row-wise largest value in bold) 134

4.15 Number of 100 simulated samples that the MAP coincides with the true generating model of Table (4.1) for overlapping scenario for specific class 2 and 3 versus baseline class 1 (n=500) (row-wise largest value in bold) 135

4.16 Class distribution of cardiotography's recordings 139
4.17 Largerst pair-wise correlation among covariates 140
4.18 Confusion matrix at s-th MCMC iteration of the predictive response set $\boldsymbol{y}^{*(s)}$ versus the respective response of test set $\boldsymbol{y}^{te}$ 142
4.19 Results of acceptance rates for parameters a_2, a_3, $\boldsymbol{\beta}$ and g of each Bayesian variable selection methods for typical and augmented multinomial logistic regression model with mixtures of g-priors over 40000 of MCMC values. 145
4.20 Results of all Bayesian variable selection methods under the MAP for each specific class 2 and 3 given baseline class 1 145
4.21 Results of all Bayesian variable selection methods under the MPM for each specific class 2 and 3 given baseline class 1 146
4.22 Results of all Bayesian variable selection methods under the BMA for each specific class 2 and 3 given baseline class 1 146

A.1 Prior-inputs and initial values 175
A.2 Acronyms of Bayesian variable selection methods. 176
A.3 Posterior model probabilities of top 5 models. 177
A.4 Posterior model probabilities of top 5 models. 177
A.5 Posterior model probabilities of top 5 models. 177
A.6 Marginal posterior inclusion probabilities for each independent variable $\boldsymbol{X}_j$ regarding Bayesian variable selection methods. 179
A.7 Results of posterior summary statistics regarding each Bayesian variable selection methods for mixtures of g-priors. 182
A.8 Prior-inputs and initial values of 1st Analysis 183
A.9 Prior-inputs and initial values of 2nd Analysis 183
A.10 Results of posterior model probabilities regarding Bayesian variable selection methods. 185
A.11 Marginal posterior inclusion probabilities for each independent variable X_j regarding each Bayesian variable selection method. 186
A.12 Posterior summary statistics of mean squared error MSE for ssvs$^{\mathrm{R}}$.g. . 190
A.13 Posterior summary statistics of mean squared error MSE for gvs$^{\mathrm{R}}$.g. . . 191
A.14 Posterior summary statistics of mean squared error MSE for ssvs$^{\mathrm{R}}$.ZS. . 192
A.15 Posterior summary statistics of mean squared error MSE for gvs$^{\mathrm{R}}$.ZS. . 193
A.16 Posterior summary statistics of mean squared error MSE for ssvs$^{\mathrm{R}}$.hyp. 194
A.17 Posterior summary statistics of mean squared error MSE for gvs$^{\mathrm{R}}$.hyp. 195

C.1 Logistic regression scenarios using independent ($r = 0$) and correlated predictors ($r = 0.75$). 212
C.2 Prior-inputs and initial values 213
C.3 Acronyms of Bayesian variable selection methods with MCMC for logistic regression. 214
C.4 Number of 100 simulated samples that the MAP coincides with the true generating model of Table (C.1) under various scenarios for independent and correlated covariates (row-wise largest value in bold). 215
C.5 Number of 100 simulated samples that the MAP coincides with the true generating model of Table (C.1) under various scenarios for independent and correlated covariates (row-wise largest value in bold). 217
C.6 Effective sample size comparison (in bold lowest value). 222
C.7 Monte Carlo error comparison (in bold lowest value). 222
C.8 Effective sample size comparison (in bold lowest value). 223
C.9 Monte Carlo error comparison (in bold lowest value). 223
C.10 Effective sample size comparison (in bold lowest value). 224
C.11 Monte Carlo error comparison (in bold lowest value). 224
C.12 Effective sample size comparison (in bold lowest value). 225
C.13 Monte Carlo error comparison (in bold lowest value). 225
C.14 Effective sample size comparison (in bold lowest value). 226
C.15 Monte Carlo error comparison (in bold lowest value). 226
C.16 Acronyms of Bayesian variable selection methods with MCMC for logistic regression. 232
C.17 Marginal posterior inclusion probabilities for each independent variable X_j regarding each Bayesian variable selection method for 40000 MCMC iterations. 233
C.18 Prior-inputs and initial values of 1st Analysis 234
C.19 Prior-inputs and initial values of 2nd Analysis 234
C.20 Results of acceptance rates for parameters a, $\boldsymbol{\beta}$ and g of each Bayesian variable selection methods for typical and augmented logistic regression model with mixtures of g-priors. 236
C.21 Confusion matrix at s-th MCMC iteration of the predictive response set $\boldsymbol{y}^{*(s)}$ versus the respective response test set $\boldsymbol{y}^{te}$. 237
C.22 Results of false negative $\widehat{FN}$, false positive $\widehat{FP}$, accuracy $\widehat{ACC}$ and miss-classification error $\widehat{ERR}$ among all Bayesian variable selection methods under the MAP (column-wise largest value in bold). 242

C.23 Results of false negative $\widehat{FN}$, false positive $\widehat{FP}$, accuracy $\widehat{ACC}$ and miss-classification error $\widehat{ERR}$ among all Bayesian variable selection methods under the MPM (column-wise largest value in bold). 242

Chapter 1

Introduction

In daily basis, large scale applications are encountered and produced routinely with emphasis in genomics and marketing sectors for classification purposes, where usually only a small number of covariates affect the response. While the development for classification with only two classes is standard, developing more sophisticated methods with 3 or more categories is not trivial, due to the additional model complexity and class imbalance. The need to extract the intrinsic information in the layers of response associated univoquely to a distinct subset of variables, requires advanced methods appropriately described by probabilistic reasoning on face of Bayesian methods, since in reality these subsets are considered unknown. Moreover, the no available or scarce information regarding each subset of variables suggests the use of objective Bayesian methodology. The profit of applying these methods stand for separating the noise variables from the relevant, ensuring accurate predictions, when spike-slab priors are met with mixtures of g-priors.

Although, mixtures of g-priors have been well recognised in practice for linear regression and generalized linear models, their use was restricted in multiclassification and especially in multinomial logistic regression mainly due to posterior intractability, computation of posterior model probabilities and Fisher information matrix. The ability to infer on the dispersion parameter, by simply adopting a hyper-prior on it, ranked mixtures of g-priors the most common prior choice for data-driven methods in the objective Bayesian catalogue.

This thesis investigates the Bayesian variable selection problem for multinomial logistic regression models with mixtures of g-priors. Under this setting, the problem is outlined as the selection of variables that varies according to each class of a polychotomous response given a baseline class, namely "class-specific predictor selection". Thus, this problem is regarded as a contemporaneous variable selection procedure, where there is a distinct set of variables that is related separately to each level of multicategorical target variable. By this way, each selected subset of variables shares a unique profile related exclusively to each categorical class of the response, meaning that the distribution of each response outcome varies conditional on different subsets.

As formal Bayesian variable selection methods in the multinomial logistic regression, cannot adequately deal with the computation of posterior model probabilities mainly due to large model spaces and posterior intractability resulting in the excessive model uncertainty both at the level of class and covariates that grows exponentially, model-based inference via MCMC methods consists of a flexible alternative.

In this thesis, we tackle the problem using model search algorithms based on the approaches of George and McCulloch (1993) and Dellaportas et al. (2002) that explore sufficiently the model space and provide handy solutions to the difficulties stated above, summarizing important aspects central to Bayesian variable selection for a traditional implementation of multinomial logistic regression based on Bové and Held (2011). Furthermore, we extend the latter Bayesian variable selection methods constructed with Polya-Gamma latent variables that mimic the behaviour of the original one, sharing identical properties to those of linear models owing to the data augmentation strategy of Polson et al. (2013), surpassing the most difficult aspects of MCMC methods. In this framework, these methods can be alternatively seen as conditional latent approaches, where the latent variables are entering in our disposal in order to convert the "incompatible" likelihood into a likelihood of "convenience", where model-based inference is facilitated due to posterior tractability of regression parameters amenable to Gibbs sampling. Both approaches are compared and assessed in simulation and real settings.

1.1 Thesis Overview

The overall context of this thesis is summarized as follows:

Chapter 1 introduces the reader to the essentials of Bayesian model selection and objective Bayesian methodology within common specific prior choices. The most popular objective Bayesian approaches are mentioned, although we will not enter in details through this chapter, because these approaches are beyond the scopes of this thesis.

Chapter 2 introduces the problem of Bayesian variable selection for linear regression models and reviews in details the main aspects of objective prior and model selection based on mixtures of g-priors with full enumeration and MCMC methods. During the

presentation of this chapter priority is given also to the notion of centering and its influence on the resulting inference. In addition, formal Bayesian variable selection and MCMC methods are assessed in simulation settings compared with standard methods in order to show effectiveness of MCMC, and then we describe a real application with priority to MCMC procedures. In both cases small model spaces are used in order to illustrate the attractive comparisons. Appendix A sections are intended to provide further demonstrated quantities or additional results (Appendix section A.9 and A.10) to support and justify the presented material.
Chapter 3 focuses on Bayesian variable selection problem for generalized linear models regarding the basic concepts of objective prior and model selection based on the approaches of Bové and Held (2011) and Li and Clyde (2013) with mixtures of g-priors, which are reviewed in detail, emphasizing more on Bové and Held (2011), which will serve as basis for the next chapter. Appendix B provides additional material with demonstrations and useful expressions.
Chapter 4 is the main topic of this thesis, where we illustrate the problem of variable selection in multinomial logistic regression from a fully Bayesian perspective. Next, we describe and adopt the prior specification of Bové and Held (2011) adapted in multinomial logistic regression based on mixtures of g-priors, representing the main computational requirement in order to initialize the MCMC methods. Afterwards, we introduce the MCMC methods as the main computational tools for Bayesian variable selection, where they are developed initially for a traditional implementation and then extended under the conditional latent approach based on Polya-Gamma data augmentation of Polson et al. (2013). Both approaches are highlighted in details regarding the additional computational steps with additional material in the Appendix C (Appendix sections C.1 and C.2 respectively) especially for the latter one, in order to demonstrate our ideas. Finally, this chapter ends with the applications of MCMC methods for typical and augmented multinomial logistic regression in simulation and real datasets. To conclude, Chapter 5 includes the discussion and the main directions towards future developments, while Appendix C ends with sections C.3, C.4 and C.5 respectively for the problem of Bayesian variable selection in logistic models emerging as a special case of the multinomial logistic regression.

1.2 Statistical Models for Bayesian Model Selection

1.2.1 Definition of Statistical and Bayesian Models

Let $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ be a set of i.i.d. observables that are realizations of a real-valued random variable $\boldsymbol{Y}$. A statistical model M is defined as a probabilistic law $f_M(.)$ assigned to $\boldsymbol{Y}$ whose behaviour can be determined through a set of parameters $\boldsymbol{\theta}_M$. The values of these parameters generate a set of distributions related to each other whose differences depend only on the chosen values. This set is called *parametric family of distributions* and is formally described as $f_M(\boldsymbol{y}|\boldsymbol{\theta}_M)$. Usually, a statistical model is characterized by the joint distribution of its observed values $\boldsymbol{y}$

$$f_M(\boldsymbol{y}|\boldsymbol{\theta}_M) = \prod_{i=1}^{n} f_M(y_i|\boldsymbol{\theta}_M),$$

this joint distribution is called *likelihood* and it contains all appropriate information for the sample and the structural dependencies connected with parameters of population. Often models are usually constructed in order to assess or interpret causal or dependency relationships among observed values of a response variable and parameters of interest. For instance, a regression model can be seen as a probabilistic structure that contains a deterministic component that links covariates (attributes of the population) with observed values $\boldsymbol{y}$. In addition, Bayesian theory treats the parameters of a model $\boldsymbol{\theta}_M$ and the model itself M as random variables assigning them (pure) probabilistic distributions. These distributions $\pi(\boldsymbol{\theta}_M)$, $\pi(M)$ are called *priors* and represent the prior beliefs over the model parameters and the model respectively denoted. In this way, these prior distributions allow for additional uncertainty both at the level of model parameters and the model itself.

Moreover, a Bayesian model can be expressed throughout the joint posterior of model parameter vector $\boldsymbol{\theta}_M$ and model M

$$f(\boldsymbol{\theta}_M, M|\boldsymbol{y}) \propto f_M(\boldsymbol{y}|\boldsymbol{\theta}_M)\pi(\boldsymbol{\theta}_M)\pi(M),$$

which combines the information contained in sampling distribution $f_M(\boldsymbol{y}|\boldsymbol{\theta}_M)$ alongside with prior beliefs $\pi(\boldsymbol{\theta}_M)$ and $\pi(M)$.

The former factorization is relevant for model selection since it allows to update initial uncertainty of model parameters and model itself into a new knowledge represented by the joint posterior distribution.

1.2.2 Bayesian Model Selection

Model choice is one of the most important aspects of statistical inference and fetches the final part of decisions. Generally, in model selection researchers represent models as “claims" rephrased into probability statements and then they adopt a criterion that will decide which model will be the best in the domain under consideration. Standard model choice routines avoid completely model uncertainty and they apply criteria whose distribution is not identifiable like Bayesian models, thus the Bayesian version of model selection is essential in situations of these type. In this section, we give the reader a “first taste" of the fundamentals of Bayesian model choice based on calculations of posterior model probabilities and Bayes factors. The bibliography of research contributions is vast surrounding Bayes factors, henceforth we mention only the most distinguished. These include the research works of Smith and Spiegelhalter (1980), Spiegelhalter and Smith (1982) Kass and Raftery (1995) and Hoeting et al. (1999).

Mostly, marginal likelihoods and posterior model probabilities are a ramification of high complexed integral which inspired researchers to publish many scrutiny achievements utilizing numerical and Monte Carlo Approximations; see Gelfand and Dey (1994), Kass and Raftery (1995), Kass and Wasserman (1995), Verdinelli and Wasserman (1995), Chib (1995) and DiCiccio et al. (2006). On the other side, the latest progress of computer technology enabled the usage of intricated MCMC methods in order to obtain a more flexible and efficient model selection procedures towards unsolved problems of last years. Although there were numerous attempts for MCMC methods construction in various publications, we will focus only in model search algorithms of George and McCulloch (1993) and Dellaportas et al. (2000) for the next chapters and these represent the basis of this thesis.

1.2.2.1 Posterior Measures of Evidence

The formal approach to Bayesian model selection is based on the original work of Kass and Raftery (1995). The main core of this approach is based on the calculation of posterior model probabilities and posterior odds. Let K be a collection of competing models $M_1, \dots, M_k$ and $\boldsymbol{y}$ the available data generated probably by one of these models. Each model $M_k \in \mathcal{M}$ specifies a different sampling density $f_{M_k}(\boldsymbol{y}|\boldsymbol{\theta}_{M_k})$ of the data $\boldsymbol{y}$. If model selection is priority, then applying the Bayes theorem

$$\pi(M_{k'}|\boldsymbol{y}) = \frac{m(\boldsymbol{y}|M_{k'})\pi(M_{k'})}{\sum_{k=1}^{K} m(\boldsymbol{y}|M_k)\pi(M_k)} \propto m(\boldsymbol{y}|M_{k'})\pi(M_{k'}), \tag{1.1}$$

the posterior probability of a model $M_{k'}$, $1 \leq k^{'} \leq K$ is computed, which is interpreted as the probability that the true generating mechanism of the data $\boldsymbol{y}$ was model $M_{k'}$. Also, notice that expression (1.1) is up to a proportionality constant of product of two terms. The first term is called marginal distribution of available data $\boldsymbol{y}$ under model $M_{k'}$ or marginal or integrated likelihood of model $M_{k'}$ if we refer to model $M_{k'}$ and is calculated

$$m(\boldsymbol{y}|M_{k'}) = \int\limits_{\boldsymbol{\theta}_{M_{k'}}} f_{M_{k'}}(\boldsymbol{y}|\boldsymbol{\theta}_{M_{k'}})\pi(\boldsymbol{\theta}_{M_{k'}})d\boldsymbol{\theta}_{M_{k'}},$$

which integrates out the uncertainty of model parameters $\pi(\boldsymbol{\theta}_{M_{k'}})$ together with sampling distribution $f_{M_{k'}}(\boldsymbol{y}|\boldsymbol{\theta}_{M_{k'}})$.

The Marginal likelihood is a key quantity in model selection for calculation of posterior model probabilities. However, it is a demanding quantity due to the computational cost that requires the evaluation of integrals. Furthermore, (Kass and Raftery, 1995) pointed out the importance of marginal likelihood $m(\boldsymbol{y}|M_{k'})$ as "the predictive distribution of the data $\boldsymbol{y}$ given the model $M_{k'}$", that is interpreted as the probability of observing the data before any data were seen under the assumption the model $M_{k'}$ holds.

The second term is the prior distribution of model $M_{k'}$ and represents the prior knowledge of model $M_{k'}$ before observing the data $\boldsymbol{y}$. In case of “prior ignorance", a simple and popular prior specification is the uniform which assigns same probability to all models of consideration. This prior takes the form of

$$\pi(M_{k'}) = \frac{1}{|K|}, \tag{1.2}$$

where $|K|$ denotes the cardinality of the set of models. Despite its conventional utility, there is a major drawback with this prior choice discussed in the work of Chipman et al. (2001). Although, someone might expect the same amount of probability to each model, the latter situation will not hold. The authors mentioned this prior as “deceptive" due to its preference of models of moderate size. For this reason, in variable selection for regression models this prior is substituted with other types of priors that we will see in next chapters. On the contrary, if we are interested in comparing two different models, M_1 and M_0, we can calculate their ratio of posterior model probabilities

$$PO_{[M_1:M_0]} = \frac{\pi(M_1|\boldsymbol{y})}{\pi(M_0|\boldsymbol{y})} = \frac{m(\boldsymbol{y}|M_1)}{m(\boldsymbol{y}|M_0)}\frac{\pi(M_1)}{\pi(M_0)}, \tag{1.3}$$

where M_0, M_1 are complement elements of the model space $\mathcal{M}$, implying that $\pi(M_1) = 1 - \pi(M_0)$. The ratio of posterior model probabilities is also called posterior odds $PO_{[.]}$

of model M_1 versus model M_0 and depends on the products of respective ratios of marginal likelihoods and prior model probabilities of model M_1 versus M_0. Posterior odds “rests firmly" on the update of prior model uncertainty through the ratio of marginal likelihoods. Posterior odds $PO_{[M_1:M_0]}$ with values larger than 1 indicate that model M_1 is favoured against model M_0, while for values lesser than 1 there are claims against model M_1. If we consider the uniform prior (1.2) across models M_0, M_1, then the ratio of prior model probabilities M_1, M_0 vanishes in (1.3) and posterior odds $PO_{[M_1:M_0]}$ turns into

$$BF_{[M_1:M_0]} = PO_{[M_1:M_0]} = \frac{\pi(M_1|\boldsymbol{y})}{\pi(M_0|\boldsymbol{y})} = \frac{m(\boldsymbol{y}|M_1)}{m(\boldsymbol{y}|M_0)},$$

which depends exsclusively on the ratio of marginal likelihood called “Bayes factor of model M_1 versus M_0" denoted as $BF_{[M_1:M_0]}$. The Bayes factor, as stated by (Kass and Raftery, 1995), measures “the accumulated evidence provided by the data of model M_1 against M_0", in other words the way it favours model M_0 or model M_1. The term “evidence" rests on the notion of marginal likelihood which is usually expressed in log-scale. If M_0 denotes the null model, a brief manual of Bayes factor’s quantification $BF_{[M_1:M_0]}$ is provided in Tables (1.1) and (1.2) by Kass and Raftery (1995), which contain a straightforward interpretation of Bayes factor both in log-scales $log_{10}(.)$, $log_e(.)$.

$log_{10}BF_{[M_1:M_0]}$	$BF_{[M_1:M_0]}$	**Evidence against** M_0
0.0-0.5	1.0-3.2	Not worth than a bare mention
0.5-1.0	3.2-10	Substantial
1.0-2.0	10-100	Strong
> 2	> 100	Decisive

Table 1.1 Bayes factor interpretation of $log_{10}BF_{[M_1:M_0]}$.

$log_eBF_{[M_1:M_0]}$	$BF_{[M_1:M_0]}$	**Evidence against** M_0
0-2	1-3	Not worth than a bare mention
2-5	3-12	Substantial
5-10	12-150	Strong
> 10	> 150	Decisive

Table 1.2 Bayes factor interpretation of $log_eBF_{[M_1:M_0]}$.

The null model M_0 is considered as reference base model for model selection purposes and in variable selection in regression is expressed as a model containing only a constant term called intercept without covariates. If we now consider the previous case of the set of models $M_1, \ldots, M_k$ and include the null model M_0 into this set, an alternative definition of posterior model probabilities using Bayes theorem is the following

$$\begin{aligned}\pi(M_{k'}|\boldsymbol{y}) =& \frac{m(\boldsymbol{y}|M_{k'})\pi(M_{k'})}{\sum_{k=1}^{K} m(\boldsymbol{y}|M_k)\pi(M_k)}\\ =& \frac{m(\boldsymbol{y}|M_{k'})\pi(M_{k'})/m(\boldsymbol{y}|M_0)\pi(M_0)}{\sum_{k=1}^{K} m(\boldsymbol{y}|M_k)\pi(M_k)/m(\boldsymbol{y}|M_0)\pi(M_0)}\\ =& \frac{PO_{[M_{k'}:M_0]}}{\sum_{k=1}^{K} PO_{[M_k:M_0]}},\end{aligned}$$

which is based on the comparison of posterior odds $PO_{[M_{k'}:M_0]}$ of each model $M_{k'}$ versus null model M_0. In case of loss of information regarding which model is more plausible, the use of uniform prior (1.2) reduces the previous expression to

$$\begin{aligned}\pi(M_{k'}|\boldsymbol{y}) =& \frac{m(\boldsymbol{y}|M_{k'})\pi(M_{k'})}{\sum_{k=1}^{K} m(\boldsymbol{y}|M_k)\pi(M_k)}\\ =& \frac{PO_{[M_{k'}:M_0]}}{\sum_{k=1}^{K} PO_{[M_k:M_0]}}\\ =& \frac{BF_{[M_{k'}:M_0]}}{\sum_{k=1}^{K} BF_{[M_k:M_0]}},\end{aligned}$$

which depends on Bayes factor $BF_{[M_{k'}:M_0]}$ of model $M_{k'}$ versus null model M_0.
To conclude, posterior model probabilities, posterior odds and Bayes factors depend on marginal likelihood which is a high dimensional integral obtained in closed form only for particular instances, whereas for most of the other cases analytic approximations such as Laplace (normal) or Monte Carlo methods are used; the Laplace approach Tierney et al. (1989) is presented in the next chapters, while Monte Carlo methods are avoided, since they are not of applied in this thesis.

1.2.2.2 Model Set Approaches

Model selection techniques distract the attention of many researchers regarding the interpretation of posterior model measures, from which including posterior model probabilities, Bayes factors and marginal likelihoods; see for more Bernardo (1979). The posterior calculation of model probabilities reflects the way how the model uncertainty is adjusted through the normalization over the models (see denominator of Bayes theorem). When the goal is model determination, all computed probabilities are compared and only the model with the highest posterior probability is selected as the best model in consideration, commonly named as the maximum aposteriori model (MAP). In this way, Bernardo (1979) proposed useful ideas for the set of Models $\mathcal{M}$

1. $\mathcal{M}$-closed view: Assume the true model M_T is an unknown element of the model space $\mathcal{M}$.

2. $\mathcal{M}$-completed view: Assume that the set of different models, “to be evaluated in the light of the individuals separate actual belief model".

3. $\mathcal{M}$-open view: Assume that the model space is the same set of competing models under consideration.

$\mathcal{M}$-closed view is often used due to direct perception of posterior model probability, that is the probability that data were generated under the true model, however this assumption is valid only in case of a simulation study design where we know the true form of generating mechanism. In most other cases, especially in real world applications where $\mathcal{M}$-closed view seems too pragmatic, we approximate the underlying phenomenon providing a representative set of “competing" models with similar behaviour with the real one. Usually, we assess the performance of model either based on model fit of data in total sample or in out of sample predictions. Model fit refers to the applied criterion that describes adequately the data and out of sample predictions refers to the ability of testing performance of model to new data. Both rely on the way of "learning" from data. For example, in linear regression problems we asses performance of model based on mean squared error, while in classification problem we calculate misclassification errors.

On the other hand, a major drawback of posterior model probabilities in comparison to posterior odds and Bayes Factor is their reduction in similar models; we invite the reader to give a look to Consonni et al. (2018) as the authors present nice review research contributions that covers important aspects of Bayesian model and prior specification. As long as the size of models is large, the respective posterior model probabilities are decreasing in magnitude, even for the MAP (Consonni et al., 2018). This is related to the dilution of posterior mass probability shared to common areas of model space, as stated in Consonni et al. (2018); see also George and McCulloch (1993). Thus, one has to think carefully for alternative measures of posterior model selection. The authors provide very nice suggestions regarding prior choice and variable selection.

More precisely in variable selection, assume a collection of different models $M_1, \dots, M_\Xi$. If any collection of this form $\mathcal{M}_j = \{M_\xi \in \mathcal{M} : \gamma_j = 1\} \subset \mathcal{M}$ is the set of all models containing the independent variable X_j as subsets of model space $\mathcal{M}$, we can additionally compute the posterior inclusion probabilities of a respective covariate X_j

$$\pi(\gamma_j = 1|\boldsymbol{y}) = \sum_{\xi=1}^{\Xi} \pi(M_\xi|\boldsymbol{y}) = \frac{\sum_{\xi=1}^{\Xi} PO_{[M_\xi:M_0]}}{\sum_{k=1}^{K} PO_{[M_k:M_0]}}, \tag{1.4}$$

where γ_j is the binary indicator of covariate X_j, which indicates the inclusion or exclusion depending on its value. The binary indicator vector $\boldsymbol{\gamma}$ includes each γ_j respectively and usually substitutes the model indicator M_k, indexing the 2^p possible subsets of $\boldsymbol{X}_1, \dots, \boldsymbol{X}_p$, where p is the number of independent variables, see George and McCulloch (1993). We will see in next chapters that is a quantity of central interest when we will start talking about highly complexed methods called MCMC usually adopted for model search algorithms in variable selection problem.

Posterior inclusion probabilities like Bayes factors and posterior odds, indicate the accumulated evidence of favouring the inclusion of covariate X_j provided by clues of data. They represent a useful and reliable tool of evidence in variable selection problem. Most of their success is devoted to their capabilities to ensure better predictive optimality than even the MAP highlighted by Barbieri and Berger (2004), when we consider the median probability model (MPM) which contains only covariates posterior inclusion probabilities greater than 0.5. According to Consonni et al. (2018), posterior inclusion probabilities are not vulnerable unlike posterior model probabilities because they can be rewritten as

$$\pi(\gamma_j = 1|\boldsymbol{y}) = \frac{1}{1 + O_j}, \tag{1.5}$$

$$O_j = \frac{\sum_{\xi=1}^{\Xi} PO_{[M_\xi:M_0]}}{\sum_{\bar{\xi}=1}^{\bar{\Xi}} PO_{[M_{\bar{\xi}}:M_0]}}, \tag{1.6}$$

where $\overline{\mathcal{M}}_j = \left\{M_{\bar{\xi}} \in \mathcal{M} : \gamma_j = 0\right\} \subset \mathcal{M}$ is the complement set of $\mathcal{M}_j$, that is the set of all models as elements of model space $\mathcal{M}$ excluding the independent variable X_j. In the above expression, notice that O_j is simply a fraction of sums up to 2^{p-1} terms which confirms robustness and flexibility in large model sizes overcoming "dilution" stated before. Furthermore, quantities O_j are also used in steps of a model search algorithm that we will further exploit in next chapter where we will discussing for variable selection in regression models. Thus, posterior inclusion probabilities are more trusted and faster estimated than individual posterior model probabilities due to large sizes of models that contain little but different from zero probability mass in normalisation constant of Bayes theorem for model selection.

1.2.2.3 Jeffrey's-Lindley's Paradox

(Lindley, 1957) observed an odd phenomenon of Bayes factor when he experimented with a simple Bayesian model comparison for the mean and he called it "paradox". This deportment is delineated: "as the sample size n goes to infinity, the posterior odds tends to zero supporting parsimonious models for every level of significance. In frequentist stastistics the level of significance as long as the sample n becomes larger tend to reject simpler models. Then, in order to conduct a Bayesian or frequentist method doesn't matter at all due to the fact that different practitioners will end up "fairly" supporting completely different models, that is a completely contradiction. (Bartlett, 1957) later discovered that for large values of prior variance the posterior odds supports the parsimonious model. Given this discovery, a practitioner must be careful and select with "care" priors with large variances. Thus, prior with large variances may exhibit undesired features even in Bayesian model selection which can be seen also as non-informative priors. Concluding, non-informative priors may result problematic even for model choice, but objective Bayes methodology, presented in the next sections, introducing a completely different point of view through unorthodox Bayes factors promises alternative solutions.

1.3 Objective Bayesian Approach

Bayesian model selection usually starts by assigning prior distributions to the model parameters and the model itself. However, a common burden within prior specification is always transformed into two frequently asked questions that a researcher has to justify, "how define prior elicitation?" and "what type of prior distribution should be used?". Often there are two ways to elicit prior knowledge, the subjective and the objective, while we restrict prior specification only for objective approach. We will illustrate their difference in this section, while emphasis is given only to objective priors, where Bayesian model selection is briefly introduced because is not of interest in this thesi's topic. We give a short overview of these different approaches based on Consonni et al. (2018), while the most distinguished objective model selection approaches are mentioned in the final part of this chapter.

1.3.1 Prior Elicitation

A Bayesian approach for model selection requires the specification of prior distributions for each model parameter and model itself. There are two usual manners of expressing prior knowledge regarding unknown quantities, *subjective* and *objective*. If a subjective approach is considered, priors should rely their information from past studies or from expertisers knowledge, in this way one elicits prior information quantifying his prior beliefs for parameters of interest. Subjective analysis is considered as a "great source of information", especially in applications where there is available knowledge of background for the case under study; see for additional information Morris (1983), Consonni et al. (2018).

The term prior elicitation usually refers in the way that prior beliefs are expressed and its characteristics. A widespread known method initially starts with noticing how many values of the parameter are more likely to appear and then to assign a probability mass or density function, which respectively sums up or integrates to one, possibly referring to a discrete or continuous parameter space. Of course, graphical representation of prior distribution may be helpful, but for multiparameter space things get difficult. An intuitive idea, is to assume a prior distribution belonging to a parametric family of distributions that fits our prior beliefs as much as possible. When someone uses a known probabilistic rule for the unknown quantities, he wonders if there are better alternatives for posterior inference than others. More precisely, researchers adopted prior distributions devoted to their functional adaptability with likelihood which lead to posterior conclusions belonging in the same family like the prior. These priors were introduced by Morris (1983) and the basic result states that any likelihood sharing characteristics of exponential family will always have a conjugate prior. Most of the cases, belong to exponential family and therefore they are broadly used. Additional nice statements on the ground of conjugate properties under exponential family are found in Consonni and Veronese (1992).

On the other side, objective approach is adopted in situations where there is lack or loss of information regarding parameters. Objectivity refers to the way of providing prior information as minimal as possible without influencing data, so letting data decide. These approaches are known respectively as subjective Bayesian and objective Bayesian approaches, whereas emphasis is only given to methods and prior specification of objective Bayes.

Objective Bayesian analysis has a long history and it was the main reason for intense debate among Bayesians due its concepts, Berger underlined the main philosophical parts of objective Bayes. Although, it seems reasonable for a practitioner to express his opinion into a prior distribution, in real world applications of high dimensions, an authentic prior elicitation turns out problematic. A useful justification by Consonni et al. (2018) alongside Berger in the previous statement, is that as long as there is complex parameter space, it will be very difficult to capture all dependence connections among parameters within prior specification.

For example, variable selection in multinomial logistic regression is a problem of this type with a complex parameter and model space.

On the other hand, for many applications there will not be available information with respect to unknown parameters before observing data, thus a researcher has to find a way to depict the loss of information into a prior distribution. For this reason, we will focus on objective Bayesian methods. This section is described as follows: first we discuss specific objective prior specification, secondly we introduce the reader to basic concepts related to objective model selection and finally reviewing the most important objective Bayesian methods.

1.3.2 Objective Prior Specification

Specification of prior distributions has an important impact in Bayesian statistics. As we mentioned above, prior information are usually expressed through past studies or from opinions of an expertiser's area. When there is little or no information available for the respective parameters, *non-informative* priors are used; Carlin and Louis (1996). According to (Consonni et al., 2018), "any kind of this prior was used for many decades in an attempt to prepare the Bayesian omelette without breaking the Bayesian eggs", based on the original work of Savage (1954). The main idea behind words of this kind, is the connection between frequentist and Bayesian modeling that someone can summon probabilistic likelihood inference, avoiding subjective prior distributions. In many situations there will not be trusty prior information related to unknown parameters, thus objective inference based purely on the data is desired. In these situations, the prior distribution of $\boldsymbol{\theta}$ shouldn't contain any relevant information such that no value of $\boldsymbol{\theta}$ observed in the parameter space $\boldsymbol{\Theta}$ should be preferred over others, notice for simplicity we dropped the subscript of M from parameter $\boldsymbol{\theta}$. When the parameter space is a finite set with discrete values such that $\boldsymbol{\Theta} = \{\theta_1, \ldots, \theta_D\}$ the probability mass function

$$\pi(\theta_d) = \frac{1}{D}, \;\; d = 1, \ldots, D,$$

is non-informative because it gives same probability amount to each value of θ, notice also that in this example θ is only a scalar compared to before. If the parameter space is continuous and bounded $\boldsymbol{\Theta} = (-\infty, \infty)$, with $-\infty < \tilde{\delta} < \tilde{\epsilon} < \infty$, the prior density distribution is

$$\pi(\theta) = \frac{1}{\tilde{\epsilon} - \tilde{\delta}}, \;\; \tilde{\delta} < 0 < \tilde{\epsilon}.$$

Moreover, if the parameter space turns into an unbounded interval, the probability density function of θ becomes

$$\pi(\theta) = \tilde{c}, \;\; \tilde{c} > 0,$$

which is called improper because holds, $\int\limits_{\boldsymbol{\Theta}} \pi(\theta) d\theta = \infty.$

Even if it seems strange to use this prior, the integral involving the product of likelihood and prior over θ may result a finite number and so a proper posterior provides us regular conclusions.

The two examples analyzed before for finite parameter spaces either in the discrete either in continuous are called uniform distributions respectively, like the uniform that we presented in the section of Bayesian model selection previously for prior beliefs of models.

A disadvantage of uniform prior is its invariance in transformations. This property states that if a non-informative prior distribution $\pi(\theta)$ is used, any reparametrizations of form $G(\theta)$ will not maintain the non-informativeness property, $G(.)$ is usually a "1-1" real-valued function. A convenient solution to this issue is the adoption of *Jeffreys'* non-informative priors that are invariant to any kind of transformations and based on expected Fisher information matrix. If we consider multiparameter $\boldsymbol{\theta}$, a Jeffreys' prior takes the following form

$$\pi^J(\boldsymbol{\theta}) \propto det(\mathcal{I}(\boldsymbol{\theta}))^{1/2}, \tag{1.7}$$

where $\mathcal{I}(\boldsymbol{\theta})$ denotes the expected Fisher information matrix, whose generic element $\mathcal{I}_{ij}(\boldsymbol{\theta})$ under regularity conditions and in the continuous case, is given by

$$\mathcal{I}_{ij}(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{\theta}} \left(\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f(\boldsymbol{Y}|\boldsymbol{\theta}) \right),$$

where $\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{\theta}}(.)$ is the expected value over the observed random variable $\boldsymbol{Y}$ given $\boldsymbol{\theta}$. In case of scalar parameter θ, Jeffreys' prior reduces to

$$\pi^J(\theta) \propto \sqrt{\mathcal{I}(\theta)}, \tag{1.8}$$

which is simpler. Fisher information matrix $\mathcal{I}(\boldsymbol{\theta})$ requires intensive computational cost especially in high dimensional settings, thus a useful recommendation is to derive its

Jeffreys' prior for each element of the parameter vector $\boldsymbol{\theta}$ and then taking the joint prior which is formed from the product of each individual Jeffreys' prior. (Consonni et al., 2018) underlines that Jeffreys' prior enjoys many important properties besides its invariance property. It is considered an automatic prior setup which maximizes asymptotic divergence among prior and posterior for $\boldsymbol{\theta}$, increasing the optimality conditions in absence of nuisance parameters.

Although, it is viewed as one of the most popular objective prior choice, it has many limitations surrounded by incoherences and paradoxes which are further discussed in Consonni et al. (2018). Other common objective prior choices not focuses on this thesis include, the reference prior Bernardo (1979), Datta et al. (1995) and Datta and Ghosh (1996) in multidimensional parameter spaces, the matching prior which shares properties like frequentist confidence intervals and the maximum entropy prior of Jaynes (2003), see Consonni et al. (2018) for more details.

To conclude, we summarized the most used prior specifications in objective Bayesian analysis and in the next section we will describe the most important objective Bayesian methods based on the current prior specification of non-informative priors.

1.3.3 Objective Model Selection

In this section, we introduce the main concepts of objective Bayesian model selection. Let $M_1, \ldots, M_K$ be a collection of K different models and observables $\boldsymbol{y}$ come possibly from one of these models. In case of substantial lack of information, we represent this ignorance with (objective) non-informative priors $\pi^N(\boldsymbol{\theta}_{M_k})$. Note that the superscript in $\pi^N(.)$ refers to class of non-informative priors.

Model determination begins with marginal distribution of data $\boldsymbol{y}$ under model $M_{k'}$, $m^N(\boldsymbol{y}|M_{k'})$, which depends on the form of the non-informative prior $\pi^N(\boldsymbol{\theta}_{M_{k'}})$. This suggests that the estimation is not affected, whereas model selection it does. The issues of arbitrary constants influence strongly model selection and this is better illustrated in the works of O'Hagan (1995), Berger and Pericchi (1996) and Berger and Pericchi (2001). Hence, they cannot be used naively in objective model selection to calculate marginal likelihoods and Bayes factors Berger and Pericchi (2001).

In last years many objective methods were established in order to surpass the indeterminacy of Bayes factors based on the use of non-informative priors. Among them, we distinct the posterior Bayes factor of Aitkin (1991), the fractional Bayes factor of O'Hagan (1995), the intrinsic Bayes factor of (Berger and Pericchi, 1996), the power prior of Ibrahim and Chen (2000) and the expected posterior prior Pérez and Berger (2002). These methods despite they are very interesting in theory, in practice may be not trivial and share the use of a thought experiment that will allow to convert the improper prior to a proper one and then use this for model selection purposes. Often some of them methods such as Aitkin (1991) and O'Hagan (1995) are inconsistent because they using the data twice, violating the likelihood principle. To conclude, a hybrid approach which couples the power priors and expected posterior priors, namely power expected posterior priors, is found in Fouskakis and Ntzoufras (2012) and Fouskakis et al. (2018) in variable selection for linear and generalized models settings respectively. Next chapter introduces the problem of Bayesian variable selection in linear regression from an objective point of view with mixtures of g-priors.

Chapter 2

Bayesian Variable Selection in Linear Regression

Regression is an approved and attractive tool in different fields of science for the investigation of linear dependencies among independent variables and a response variable. Usually, in order to build an ideal regression model, a common question is "how many relevant pieces of information need to be added", which in the regression context means including only the important predictor variables which are related to Y, namely *variable selection* Marin and Robert (2014). Including variables that are not related to the target variable Y may mask important attributes of the population and increase the error of predictions. Thus, variable selection is regarded as one of the most important aspects of model selection where each model corresponds to a different subset of covariates that balances the predictive performance with good estimation properties. Nowadays, variable selection remains an open area of research.

As traditional variable selection cannot adequately deal with the issue of model uncertainty, research has proved Bayesian methods more efficient in terms of probabilistic reasoning. Especially, when little or no information is available regarding the covariates, objective Bayesian methods are proposed as an alternative tool.

Thus, it is often necessary to resort to some formal prior elicitations based especially on objective Bayesian approach.

Generally, improper priors are prohibited in model selection procedure, as their distributional forms depend on proportionality constants, which are defined arbitrarily. However, the latter doesn't influence the estimation process since these constants are added as products in the numerator and the denominator of the posterior, causing them to cancel out. Moreover, these constants will not vanish in the model selection procedure, as the marginal likelihood will contain them leading to "ill" posterior measures of evidence like Bayes factors and posterior model probabilities. Hence, to avoid undetermined model choice assessments, the use of conventional proper priors is strongly suggested. For these reasons, in this chapter, we focus only on Bayesian variable selection for linear regression models with mixtures of g priors Liang et al. (2008), namely g-prior; an idea developed originally by Zellner (1986). These Bayesian variable selection methods seem promising in terms of objective consistency and are characterized by optimality properties such as sparsity and parsimony. The first research developments of Bayesian variable selection have occurred in linear regression because of its simplicity and approximation to complicate relationships, such as wavelets or splines which are expressed as linear functions of the covariates Chipman et al. (2001). Research publications in Bayesian variable selection abound regarding regression, especially for Gaussian linear models, making Bayesian variable selection alone an area of study, which exceeds the limit of this publication. Furthermore, many statistical models in Bayesian variable selection replace the whole model indicator M_k with $\boldsymbol{\gamma}$, such that $M_k \in \mathcal{M} \equiv \boldsymbol{\gamma} \in \{0,1\}^p$, where $\boldsymbol{\gamma}$ is a latent vector parameter that indexes each subset of independent variables to the initial set of variables $X_1, \ldots, X_p$. On the other hand, an evident disadvantage of variable selection is commonly known as "curse of dimensionality"; this means that when the number of parameters grows serious problems are caused even in the context of a "formal" Bayesian model selection. These problems are notably related to the computation of the posterior model probabilities and the size of the model space. Additionally, latest technology progress gave the opportunity to deal with the problem using highly complex algorithms called MCMC which successfully perform the variable selection process. The MCMC methods George and McCulloch (1993) and Dellaportas et al. (2002) are the main computational tools of this thesis and consist of model search algorithms that provide analytic summary of Bayesian variable selection. Even if regression was highly regarded with favor for its simple computational form and its adaptability, in terms of MCMC, with Gibbs sampling, the latter is not trivial when mixtures of g-priors are adopted. This topic is presented in the second part of this chapter.

2.1 The problem of Bayesian Variable Selection in Linear Models

The variable selection in linear regression models is a widely used methodology in various applications and remains a hot topic for many research publications in the present. Variable selection is regarded as one of the major encountered problems of statistical modelling in practice which aims to find important covariates $X_1, \ldots, X_p$ that explain or predict a phenomenon measured through a response variable Y . However, one must pay attention to the importance of explanatory variables because irrelevant covariates may harm the predictive ability and create noise in the produced estimates. Thus, it is essential to select which variables-among a large pool of explanatory variables should be included in the model that ensures a good predictive accuracy Marin and Robert (2014). A linear model, as stated in previous chapter, is a probabilistic structure that embodies a systematic component among the observables and the population parameters containing the covariates usually expressed in terms of a possible parameter vector $\boldsymbol{\beta}$. In addition, model choice in linear model formulation begins with a probability density distribution of observed values $\boldsymbol{y} = (y_1, \ldots, y_n)^{\boldsymbol{T}}$ for the response variable $\boldsymbol{Y}$

$$\boldsymbol{Y}|a, \boldsymbol{\beta}, \sigma^2 \sim N_n\left(a\mathbf{1}_{\boldsymbol{n}} + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_{\boldsymbol{n}}\right), \tag{2.1}$$

where p denotes the number of covariates in the design matrix $\boldsymbol{X} = [\boldsymbol{X}_{\mathbf{1}}, \ldots, \boldsymbol{X}_{\boldsymbol{p}}]$, a denotes the intercept as a scalar quantity multiplied by a column of n ones, $\boldsymbol{X}$ denotes the design matrix of dimension $n \times p$, $\boldsymbol{\beta}$ is a vector of dimension $p \times 1$ which denotes the corresponding effects of each covariate, σ^2 represents the variance related to the measuring error of the linear model and $\boldsymbol{I}_{\boldsymbol{n}}$ is the identity matrix of dimension $n \times n$. Notice that the relationship in (2.1) is expressed as a linear function of a and $\boldsymbol{\beta}$ which is due to the centering of the design matrix $\boldsymbol{X}$; this step plays a decisive role in prior specification and model selection which will be highlighted in next sections.

The next step is to identify promising subsets of variables related to the linear model (2.1) which are represented by the binary vector. This notation is quite handy, since it indicates which covariates are included or not. According to (Liang et al., 2008) variable selection can be seen alternatively as the restriction in each subspace of the linear model (2.1) conditional on $\boldsymbol{\gamma}$; thus each subspace is a reduced linear model or subset of variables. Moreover, the parameter $\boldsymbol{\gamma}$ serves as model index and it maps each corresponding model $\boldsymbol{\gamma}$ to the model space 2^p. Under these considerations, for each candidate model $\boldsymbol{\gamma} \in \{0, 1\}^p$, the variable selection problem in linear regression models can be defined for the random variable $\boldsymbol{Y}$ as

$$\boldsymbol{Y}|a, \sigma^2, \boldsymbol{\beta}_\gamma, \boldsymbol{\gamma} \sim N_n\left(a\mathbf{1}_{\boldsymbol{n}} + \boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma, \sigma^2\boldsymbol{I}_{\boldsymbol{n}}\right), \tag{2.2}$$

where a denotes again the intercept and remains the same across all models $\boldsymbol{\gamma}$, p_γ denotes the selected number of covariates in the design matrix $\boldsymbol{X}_\gamma$ of each model $\boldsymbol{\gamma}$, $\boldsymbol{X}_\gamma = [\boldsymbol{X}_{\gamma\mathbf{1}}, \ldots, \boldsymbol{X}_{\gamma\boldsymbol{p}_\gamma}]$ denotes the design matrix of dimension $n \times p_\gamma$ for model $\boldsymbol{\gamma}$, $\boldsymbol{\beta}_\gamma$

is a vector of dimension $p_\gamma \times 1$ which denotes the corresponding effects of selected covariates and σ^2 is the variance related to the measuring error of model $\boldsymbol{\gamma}$. Practically speaking, the design matrix $\boldsymbol{X}_\gamma$ represents a sub-matrix of the full matrix $\boldsymbol{X}$ where each column of it is included in model $\boldsymbol{\gamma}$; the same applies to $\boldsymbol{\beta}_\gamma$ respectively to $\boldsymbol{\beta}$.
In other words, the variable selection problem becomes a decision problem where candidate models are ranked using a corresponding criterion which determines the model selection. Thus, there are 2^p competitive models that enumerate the model space and we search the best subset of $\boldsymbol{X_1}, \ldots, \boldsymbol{X_p}$ explanatory variables that describe or explain at most the variability of the response variable Y.
To conclude, in the next section, we will present the essentials of prior choice for Bayesian Variable Selection in linear regression models.

2.1.1 Prior Elicitation

Although there are many reasons behind a researcher's plan to express his subjective opinion through a prior distribution for a quantity of interest, in general it results in a futile attempt, especially in real problems. The former situation emerges in multi-parameter problems, where it is impossible to envelop all prior parameter features even if an expert's knowledge is available. Thus, objectivity becomes crucial in situations where subjective information is non available. This justification becomes evident in variable selection for regression models where explanatory variables $X_1, \ldots, X_p$ enter in competition in terms of the possible 2^p subsets, as their representation will correspond to a specific couple of prior parameters sharing completely distinct values conditional on each model, yielding obviously problematic subjective prior information. For this reason, it will be impossible to encapsulate individual model characteristics reflected in each distinct prior internal feature and hence objective Bayesian methods are preferred.
In the setting of Bayesian variable selection we follow the approach of Jeffreys (1961) which is presented in the next section afterwards the prior choice.

2.1.2 Prior Choice

In Bayesian variable selection, a special application of statistical modeling widely accepted, it is a common practice to consider any subset of independent variables as the probability density functions and then to compute the posterior distribution of all these subsets in order to figure out the uncertainty on the model space 2^p Consonni et al. (2018). Thus, the model choice even for the variable selection, is summarized through posterior model probabilities for all 2^p models using the Bayes theorem compared to a base model $\boldsymbol{\gamma}_0$

$$\pi(\boldsymbol{\gamma}^{'}|\boldsymbol{y}) = \frac{m(\boldsymbol{y}|\boldsymbol{\gamma}^{'})\pi(\boldsymbol{\gamma}^{'})}{\sum_{\boldsymbol{\gamma}}^{2p} m(\boldsymbol{y}|\boldsymbol{\gamma})\pi(\boldsymbol{\gamma})} = \frac{PO_{[\gamma^{'}:\gamma_0]}}{\sum_{\boldsymbol{\gamma}}^{2p} PO_{[\gamma:\gamma_0]}},$$

where $\pi(\boldsymbol{\gamma})$ are the prior beliefs for each model $\boldsymbol{\gamma}$ and $m(\boldsymbol{y}|\boldsymbol{\gamma})$ is the marginal distribution of data $\boldsymbol{y}$ under model $\boldsymbol{\gamma}$. Furthermore, posterior model probabilities often result from the computation of marginal likelihoods for each model $\boldsymbol{\gamma}$ in the model space 2^p

$$m(\boldsymbol{y}|\boldsymbol{\gamma}) = \quad \int_{a}\int_{\sigma^2}\int_{\boldsymbol{\beta}_\gamma} f(\boldsymbol{y}|a, \sigma^2, \boldsymbol{\beta}_\gamma, \gamma)\pi(a, \boldsymbol{\beta}_\gamma, \sigma^2|\gamma) da d\sigma^2 d\boldsymbol{\beta}_\gamma,$$

where $\pi(a, \boldsymbol{\beta}_\gamma, \sigma^2|\boldsymbol{\gamma})$ denotes the joint prior distribution of parameters given model $\boldsymbol{\gamma}$. Apart from $\pi(\boldsymbol{\gamma})$, apparently, for a given set of models the effectiveness of the Bayesian approach depends exclusively on the prior $\pi(a, \boldsymbol{\beta}_\gamma, \sigma^2|\boldsymbol{\gamma})$ usually requiring the evaluation of a highly complex integral which leads to closed forms only for some specific prior forms Liang et al. (2008); whereas for the other cases Laplace approximation Tierney et al. (1989) is used. Thus, the specification of a prior distribution needs to be handled with utmost care. Prior elicitation for $\pi(a, \boldsymbol{\beta}_\gamma, \sigma^2|\boldsymbol{\gamma})$ will be based on the original work of Jeffreys (1961) on "conventional" priors, which provide reasonable comparisons of Bayes factors across the model space Forte (2014). On the other hand, prior model probabilities $\pi(\boldsymbol{\gamma})$ regarding the model space are not of interest in this section and will be further discussed in the second part of this chapter, including the most popular by Scott and Berger (2010).

2.1.2.1 Base Model Considerations and Jeffrey's Approach

The Bayesian approach to model selection and consequently to variable selection is regarded as simultaneous hypothesis testing through the calculation of Bayes factors and, consequently, posterior model probabilities. In equation (1.1), posterior probabilities result from the simultaneous comparison of Bayes factor of each model $\boldsymbol{\gamma}$ with respect to a fixed model $\boldsymbol{\gamma}_0$, called *base model.* The choice of the base model is crucial and affects the model selection from an objective point of view, whereas from a subjective stance it can be ommited. Consequently, the prior distribution $\pi(a, \sigma^2, \boldsymbol{\beta}_\gamma|\boldsymbol{\gamma})$ needs a careful specification because it takes part in every model comparison; hence its construction depends exclusively on the characteristics of the model we are comparing $\boldsymbol{\gamma}$ it with. The main intuition behind the base model is better understood in the next example based on (Consonni et al., 2018) which originates from the work of Jeffreys

(1961). Suppose we are interested in comparing two models, let's say, $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}_0$ for simplicity, assumed as sampling densities that govern the observables $\boldsymbol{y}$

$$\begin{aligned} \text{Model } \boldsymbol{\gamma}_0 &: f(\boldsymbol{y}|\boldsymbol{\theta}_{\boldsymbol{\gamma_0}}, \gamma_0), \;\; \boldsymbol{\theta}_{\boldsymbol{\gamma_0}} = (a, \sigma^2)^T \in \mathbb{R}^{d_{\gamma_0}}, \\ \text{Model } \boldsymbol{\gamma} &: f(\boldsymbol{y}|\boldsymbol{\theta}_{\gamma}, \gamma), \;\; \boldsymbol{\theta}_{\gamma} = (a, \sigma^2, \boldsymbol{\beta}_{\gamma}^T)^T \in \mathbb{R}^{d_{\gamma}}, \end{aligned}$$

where model parameters $\boldsymbol{\theta}_{\boldsymbol{\gamma_0}}$, $\boldsymbol{\theta}_{\gamma}$ have different dimensions d_{γ_0}, d_{γ_0} respectively. Note the model dimensions correspond to $d_{\gamma_0} = 2$ and $d_{\gamma} = 2 + p_{\gamma}$ respectively for $\boldsymbol{\gamma}_0$ and $\boldsymbol{\gamma}$. When model $\boldsymbol{\gamma}_0$ is nested in model $\boldsymbol{\gamma}$, which means that after additional restrictions on the parameter space of model $\boldsymbol{\gamma}$ model $\boldsymbol{\gamma}_0$ is retrieved, we conventionally consider as "common" the model parameters $(a, \sigma^2)^T$ between the two models, whereas parameter $\boldsymbol{\beta}_{\gamma}$ is model "specific". The use of a "common" parameter $(a, \sigma^2)^T$ in nested model comparisons often permits the employment of the same improper prior for $(a, \sigma^2)^T$ across the model space. Despite its broad use due to the above simplification, it is not appropriate in the intrinsic methodology as stated by Consonni et al. (2018). A "plausible" choice of the base model $\boldsymbol{\gamma}_0$ in the example above is null model which is nested in each respective model $\boldsymbol{\gamma}$ in consideration. This option is considered to be the most plausible since model evaluation compares each Bayes factor of nested models $\boldsymbol{\gamma}_0$ and $\boldsymbol{\gamma}$ with common model parameters $(a, \sigma^2)^T$ for $\boldsymbol{\gamma}_0$ and $\boldsymbol{\beta}_{\gamma}$ specific for each $\boldsymbol{\gamma}$, called *null-based* approach Liang, (2008). Hence, each model $\boldsymbol{\gamma}$ univoquely captures different features in the specific prior distribution $\pi(a, \sigma^2, \boldsymbol{\beta}_{\gamma}|\boldsymbol{\gamma})$. In this way, let $\pi(a, \sigma^2|\boldsymbol{\gamma}_0)$ be prior under the null model $\boldsymbol{\gamma}_0$ and without loss of generality let the prior under model $\boldsymbol{\gamma}$, have the following hierarchical distribution form

$$\pi(a, \sigma^2, \boldsymbol{\beta}_{\gamma}|\boldsymbol{\gamma}) = \pi(\boldsymbol{\beta}_{\gamma}|a, \sigma^2, \boldsymbol{\gamma})\pi(a, \sigma^2|\boldsymbol{\gamma}),$$

From the above prior formulation notice that for any other option of base model $\boldsymbol{\gamma}_0$ the common and specific model parameters will vary respectively for the Bayes factor comparisons of each model $\boldsymbol{\gamma}$ with respect to $\boldsymbol{\gamma}_0$ which will also directly affect the specific prior itself. To conclude, other prior distributions related to other optional base model comparisons are found in Casella and Moreno (2006); Liang et al. (2008).

2.1.2.2 Default Prior Choice

Model choice remains indisputably one of the most important domains in statistics and is considered the final stage of conclusions. From a Bayesian perspective, before taking final decisions, any model is initialized after a researcher assigns prior distributions for the unknowns which are of central interest. Thus, any prior represents the so called "kick off" of a Bayesian model and their choice is a delicate issue even for variable selection which will be further explored throught this section. In this section, we present a detailed prior specification as stated previously following the guidelines of Jeffreys (1961) and Zellner and Siow (1980) based on the approach of Zellner (1986). Although a subjective view is useful, in real life applications with moderate model space subjectivity is rejected even for variable selection. Thus, objectivity is still helpful if there is little information at hands for the problem under study. However, several objective Bayesian methods are based on the convenient use of non-informative priors, usually improper or Jeffreys' priors, which typically lead to ill-determined comparisons of Bayes factors due to arbitrary constants Berger and Pericchi (1996) and Berger and Pericchi (2001). Only in some instances one can search for solutions regarding the indeterminacy of Bayes factors; see Aitkin (1991), O'Hagan (1995), Pérez and Berger (2002) and Casella and Moreno (2006). Moreover, researchers often prefer to act in the framework of conjugate priors because of their computational tractability in the marginal likelihood, nonetheless conjugate priors exhibit undesired behaviour in the resulting posterior measures of evidence as the sample size grows. Likewise, although vague priors depict prior ignorance suitably; there is still debate arguing to avoid them. Therefore, the use of conjugate priors in Bayesian variable selection for linear regression models has limited potentials and a more elaborate strategy must be used; see Marin and Robert (2014). The core beyond this strategy borrows strength by the original idea proposed by Zellner (1986)in the linear regression.

Hence, researchers resort to suitable methods of objective Bayesian model selection which are insensitive to the matters stated above and then use priors on the semi-conjugate sketch of Zellner (1986). The basic approach behind this idea is to allow practitioners to introduce (possibly weak) information about the vector of regression coefficients and to settle the matter of the prior specification of the scale hyperparameters, by comprising the variance-covariance matrix through the expected Fisher information matrix. The original "pre-processed" g-prior, suggested by Zellner (1986) alongside Jeffrey's conventional directions, is based on a combined prior, expressed through a Jeffreys' prior for the common model parameter σ^2 and a proper conjugate prior for the specific model parameter $\boldsymbol{\beta}_{\gamma+1}$ which takes in practice the following form

$$\pi(a, \boldsymbol{\beta}_{\gamma}, \sigma^2|\boldsymbol{\gamma}+1) = \pi(\sigma^2|\boldsymbol{\gamma}+1)\pi(\boldsymbol{\beta}_{\gamma+1}|g, \sigma^2, \boldsymbol{\gamma}+1), \tag{2.3}$$

$$\boldsymbol{\beta}_{\gamma+1}|g, \sigma^2, \boldsymbol{\gamma}+1 \sim N_{p_{\gamma+1}}\left(\mathbf{0}_{\boldsymbol{p}_{\gamma+1}}, g\sigma^2(\boldsymbol{X}^{\boldsymbol{T}}_{\gamma+\mathbf{1}}\boldsymbol{X}_{\gamma+\mathbf{1}})^{-1}\right), \tag{2.4}$$

where $\pi(\sigma^2|\boldsymbol{\gamma}+1) \propto \frac{1}{\sigma^2}$, $\boldsymbol{\beta}_{\gamma+1} = (a, {\boldsymbol{\beta}_{\gamma}}^{T})^{T}$ denotes the joint parameter vector $\boldsymbol{\beta}_{\gamma}$ including the intercept a. Its mean is centered to zero because we assume there is an a-priori negligible effect of selected covariates and g is a positive scalar multiplied by a variance-covariance matrix which depends on the observed data of the design matrix $\boldsymbol{X}_{\gamma+1}$. Nonetheless, this is not a crucial issue since the whole model $\boldsymbol{\gamma}+1$ is conditional on $\boldsymbol{X}_{\gamma+1}$ from construction. While the improper prior for σ^2 is not intended to provide information, some researchers prefer to use this prior in an informative sense specifying a combined normal-inverse-gamma modification of (2.4). Furthermore, the prior scale of (2.4) up to the scalar g is the expected Fisher information matrix which is equivalent to the variance-covariance matrix for a maximum likelihood estimator $\widehat{\boldsymbol{\beta}}_{\gamma+1}$ conditional on a model $\boldsymbol{\gamma}+1$. This prior form offers a major advantage in the unification of objective Bayesian and frequentist analysis enabling us to consider straightforward prior specifications without activating the Jeffrey's-Lindley's paradox Lindley (1957) and Bartlett (1957). According to Zellner's first ideas, dated in the 80s and mainly inspired by Jeffrey's work of conventional prior specification, the g-prior was initially introduced in his first attempt to compare a simple hypothesis testing for linear regression even if he didn't consider explicitly nested scenarios. At least his earlier exposition was satisfied by the orthogonality of Fisher information matrix which was a prerequisite for adopting Jeffrey's pioneering ideas. In this work, orthogonality plays an important role for any objective prior specification in the Fisherian sense, that implies a block diagonal matrix of parameters in developing objective Bayesian methods. For instance, the main body of prior specification (2.3) for the full parameter vector $\boldsymbol{\theta}_{\boldsymbol{\gamma+1}} = (\boldsymbol{\beta}_{\gamma+1}, \sigma^2)$ is due to Fisher information matrix defined as

$$\mathcal{I}(\boldsymbol{\theta}_{\gamma+1}) = -\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{\theta}_{\gamma+1}} \begin{pmatrix} \frac{\partial^2 \log\big((f(\boldsymbol{Y}|\boldsymbol{\beta}_{\gamma+1},\sigma^2,\gamma+1)\big)}{\partial\sigma^4} & \frac{\partial^2 \log\big(f(\boldsymbol{Y}|\boldsymbol{\beta}_{\gamma+1},\sigma^2,\gamma+1)\big)}{\partial\sigma^2\partial\boldsymbol{\beta}_{\boldsymbol{\gamma+1}}} \\ \frac{\partial^2 \log\big(f(\boldsymbol{Y}|\boldsymbol{\beta}_{\gamma+1},\sigma^2,\gamma+1)\big)^T}{\partial\boldsymbol{\beta}_{\boldsymbol{\gamma+1}}\partial\sigma^2} & \frac{\partial^2 \log\big(f(\boldsymbol{Y}|\boldsymbol{\beta}_{\gamma+1},\sigma^2,\gamma+1)\big)}{\partial\boldsymbol{\beta}^{\mathbf{2}}_{\boldsymbol{\gamma+1}}} \end{pmatrix},$$

where the $\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{\theta}_{\gamma+1}}(.)$ is taken with respect to the random variable $\boldsymbol{Y}$ given the parameter $\boldsymbol{\theta}_{\gamma+1}$ which after simple algebraic steps becomes

$$\mathcal{I}(\boldsymbol{\theta}_{\gamma+1}) = \begin{pmatrix} \frac{n}{\sigma^4} & \mathbf{0}_{p_{\gamma+1}} \\ \mathbf{0}^T_{p_{\gamma+1}} & \frac{\boldsymbol{X}^{\boldsymbol{T}}_{\boldsymbol{\gamma+1}}\boldsymbol{X}_{\boldsymbol{\gamma+1}}}{\sigma^2} \end{pmatrix} = \begin{pmatrix} \mathcal{I}(\sigma^2) & \mathcal{I}(\sigma^2, \boldsymbol{\beta}_{\gamma+1}) \\ \mathcal{I}(\boldsymbol{\beta}_{\gamma+1}, \sigma^2)^T & \mathcal{I}(\boldsymbol{\beta}_{\gamma+1}) \end{pmatrix},$$

where the block diagonal part $\mathcal{I}(\boldsymbol{\beta}_{\gamma+1})^{-1}$ is used in the main part of the scale specification of the g-prior. Note that block diagonal element $\mathcal{I}(\sigma^2) = \frac{n}{\sigma^4}$ can be used also for the prior of σ^2 which is also a Jeffreys' prior due to the invariance principle under

transformations.
In addition, another extension of the initial g-prior of Zellner, was introduced by Liang et al. (2008) which became popular for its computational advantages in Bayesian variable selection in linear models. This version is simply a modified version based on the orthogonality (centering) of the design matrix $\boldsymbol{X}_\gamma$ which allows to express distinctly model parameters a, $\boldsymbol{\beta}_\gamma$ in the following joint prior specification

$$\pi(a, \boldsymbol{\beta}_\gamma, \sigma^2|\boldsymbol{\gamma}) = \pi(a, \sigma^2|\boldsymbol{\gamma})\pi(\boldsymbol{\beta}_\gamma|g, \sigma^2, \boldsymbol{\gamma}), \tag{2.5}$$

where in the above equation we define

$$\pi(a, \sigma^2|\boldsymbol{\gamma}) \propto \frac{1}{\sigma^2}, \tag{2.6}$$

$$\boldsymbol{\beta}_\gamma|g, \sigma^2, \boldsymbol{\gamma} \sim N_{p_\gamma}\left(\mathbf{0}_{\boldsymbol{p}_\gamma}, g\sigma^2(\boldsymbol{X}_\gamma^{\boldsymbol{T}}\boldsymbol{X}_\gamma)^{-1}\right), \tag{2.7}$$

where the joint prior (2.5) reflects the main ideas of Jeffreys (1961) although the intercept is no more conditional on the coefficients of the g-prior (2.7). Moreover, apart from Jeffreys', orthogonality, usually known as centering, one of the main computational steps in the aspects of prior specification which coerces to adopt the same improper prior for a, σ^2 along with the prior independence of a, $\boldsymbol{\beta}_\gamma$ is related to Fisher information matrix of a, $\boldsymbol{\beta}_\gamma$. Although, Liang et al. (2008) took advantage of the centering step, subtracting the mean of the design matrix from the own matrix, they miss the chance to justify the employment of prior specification (2.5) in the Fisherian sense and this important feature is taken into account in this thesis. Unlike with the original g-prior, the Fisher information matrix of a, $\boldsymbol{\beta}_\gamma$ is sufficient

$$\mathcal{I}(a, \boldsymbol{\beta}_\gamma) = -\mathbb{E}_{\boldsymbol{Y}|a,\boldsymbol{\beta}_\gamma,\sigma^2}\begin{pmatrix} \frac{\partial^2 \log\left(\left(f(\boldsymbol{Y}|a,\boldsymbol{\beta}_\gamma,\sigma^2,\gamma)\right)\right.}{\partial a^2} & \frac{\partial^2 \log\left(f(\boldsymbol{Y}|a,\boldsymbol{\beta}_\gamma,\sigma^2,\gamma)\right)}{\partial a \partial \boldsymbol{\beta}_\gamma} \\ \frac{\partial^2 \log\left(f(\boldsymbol{Y}|a,\boldsymbol{\beta}_\gamma,\sigma,\gamma)\right)^T}{\partial \boldsymbol{\beta}_\gamma \partial a} & \frac{\partial^2 \log\left(f(\boldsymbol{Y}|a,\boldsymbol{\beta}_\gamma,\sigma^2,\gamma)\right)}{\partial \boldsymbol{\beta}_\gamma^{\boldsymbol{2}}} \end{pmatrix}, \tag{2.8}$$

which is computed after performing some mathematical steps

$$\mathcal{I}(a, \boldsymbol{\beta}_\gamma) = \begin{pmatrix} \frac{n}{\sigma^2} & \mathbf{0}_{p_\gamma} \\ \mathbf{0}_{p_\gamma}^T & \frac{\boldsymbol{X}_\gamma^{\boldsymbol{T}}\boldsymbol{X}_\gamma}{\sigma^2} \end{pmatrix} = \begin{pmatrix} \mathcal{I}(a) & \mathcal{I}(a, \boldsymbol{\beta}_\gamma) \\ \mathcal{I}(\boldsymbol{\beta}_\gamma, a)^T & \mathcal{I}(\boldsymbol{\beta}_\gamma) \end{pmatrix},$$

where the block diagonality of the above matrix is due to the centering assumption ,$\boldsymbol{X}_\gamma^T\mathbf{1}_n = \mathbf{0}_{p_\gamma}$, which justifies the plausible a-priori independence of a, $\boldsymbol{\beta}_\gamma$. The block diagonal part $\mathcal{I}(\boldsymbol{\beta}_\gamma)^{-1}$ is used as variance-covariance matrix of Liang's g-prior (2.7). Another important feature is the interpretability of the g-prior approach represented in both equations (2.4) and (2.7). In the first equation, the original g-prior is the posterior distribution of an imaginary sample $\boldsymbol{y_0} = (y_{01}, \ldots, y_{0n^*})^T$, that can be used for a response $\boldsymbol{Y_0}$, whose values are a vector of zeros of dimension $n^* \times 1$ for a linear model with known variance. Denoted as $\boldsymbol{Y_0}|\boldsymbol{\beta}, \sigma^2, g \sim N_{n^*}\left((\mathbf{1}_{n^*}\boldsymbol{X_0})\boldsymbol{\beta}, g\sigma^2\right)$, coupled with a joint improper prior, $\pi(\boldsymbol{\beta}) \propto 1$, where $\boldsymbol{X_0}$ is any centered design matrix of dimension $n^* \times p$ for which holds $(\mathbf{1}_{n^*}\boldsymbol{X_0})^T(\mathbf{1}_{n^*}\boldsymbol{X_0}) = (\mathbf{1}_{n^*}\boldsymbol{X})^T(\mathbf{1}_{n^*}\boldsymbol{X})$; see Bové and Held (2011). Similarly the same hold for Liang's g-prior; for more information see Appendix section A.1.

To conclude, in the next section we will see how the prior knowledge is incorporated in the posterior and we discuss the fundamentals of model selection in the context of Bayesian variable selection for linear regression models. In the next subsection, we review the main prior specifications regarding the model space composed by the possible subsets of independent variables.

2.1.2.3 Prior Choice for Model Space

Prior model probabilities $\pi(\boldsymbol{\gamma})$ are essential in order to complete the model selection procedure using the Bayes theorem through posterior model probabilities and Bayes factors. If a subjective point of view is adopted, the additional uncertainty that arises from the priors $\pi(\boldsymbol{\gamma})$ would be updated properly and incorporated in the post summary after we have seen the data $\boldsymbol{y}$; see (Chipman et al., 2001). Although it is a great source of information, subjective opinion is forbidden due to practical limitations related to the increased number of parameters and consequently the risen complexity and high uncertainty Chipman et al. (2001). Perhaps the most important reason for rejecting the subjective approach is that we cannot "naturally" describe this uncertainty (Chipman et al., 2001). Variable selection is a problem of this type where it is impossible to manage the possible elicitations of 2^p subsets both at the level of model parameter and the model itself, although in this section emphasis will be given only to the latter one. As we mentioned in previous sections, we will deal with Bayesian approaches that use prior specifications based on non-informative and semi-automatic formulations, using objective methods. The main purpose of using an objective approach is to specify priors that allow the posterior probability to accumulate near the true generating mechanism that generated the data $\boldsymbol{y}$. When $\mathcal{M}$-closed view is used, a common choice to express indifference between two or more competing subsets was for many years the uniform distribution $\pi(\boldsymbol{\gamma}) = 1/2^p$ which is used in the case of prior ignorance and it favours equally each model $\boldsymbol{\gamma}$ (Consonni et al., 2018). Equivalently, in variable selection terms, beyond those covariates which must be present in each subset, the uniform distribution

is obtained when all covariates are equally probable to enter the subset.
The uniform prior simplifies the calculation of the posterior model probabilites since they are expressed as a proportionality constant to the marginal likelihood $\pi(\boldsymbol{\gamma}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\gamma})$ and the prior model probabilites are cancelled out leading to a straightforward comparison between Bayes factors. According to (Chipman et al., 2001); this non-informative prior is characterized as deceptive due to the sensitivity to the model complexity. The idea of assigning equally prior probability mass to each model, is not preferable because the model size has a large impact on the resulting posterior causing it to reallocate the posterior mass probability away from the true model probabilities. Thus, the selected model based on the posterior model probabilites will not coincide with the true generating mechanism that generated the data $\boldsymbol{y}$.
In the next sections, when we will discuss Bayesian variable selection using MCMC, we will introduce a more general class of priors on model space which spreads the probability mass around plausible model neighborhoods without being affected by the model complexity.

2.1.3 Model Choice

Model selection is one of the most important aspects in statistical modeling since it decides which mathematical structure is more appropriate to describe the genuine mechanism that produced the data $\boldsymbol{y}$ and then uses a criterion or a measure of evidence to single out a unique representative of the model uncertainty. In Bayesian model selection, since we are treating the unknown parameters under a probabilistic framework, these measures that decide which model is better or not are based on posterior probabilities and Bayes factors. Model selection in the context of Bayesian variable selection is of particular interest when g-prior formulation of (2.5) is adopted. Its advantages are the analytical tractability of marginal likelihood, the closed forms for posterior estimation and the automatic elicitation based only on the specification of g, which are presented in details throughout this section Liang et al. (2008) and Held et al. (2015). With regard to the latter, some methods based on arbitrary assignment of values and empirical Bayes are also mentioned instead of only Bayesian methods, since g takes place both in parameter estimates and posterior measures. This might create unexpected surprises and should be avoided in research. This is further discussed in Liang et al. (2008), Lindley (1957) and Bartlett (1957). Despite most authors start with estimation and then conclude their work with model selection, we would like to start from the marginal likelihood expression and then proceed to the explanation based on computed posterior closed forms. Afterwards this section finishes with the description of the possible choices of g. To begin with, marginal likelihood is a key quantity since it is involved in the subsequent calculations of the posterior model probabilities and the Bayes factors. In addition, based on the g-prior under equations (2.6) and (2.7) combined with the Gaussian sampling distribution (2.2) the marginal likelihood closed expression is obtained as

$$m(\boldsymbol{y}|\boldsymbol{\gamma}, g) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\pi^{\frac{n-1}{2}}} ||\boldsymbol{y} - \bar{y}\mathbf{1}_{\boldsymbol{n}}||^{-(n-1)} \frac{(1+g)^{\frac{n-1-p_{\gamma}}{2}}}{[1+g(1-R_{\gamma}^2)]^{\frac{n-1}{2}}},$$

where R_{γ}^2 denotes the usual coefficient of determination of the regression model $\boldsymbol{\gamma}$ and we notice that the marginal likelihood is expressed as a function of the coefficient of determination and the parameter g induced by the g-prior; see Liang et al. (2008). This expression derives from the fact that the marginal likelihood specified under the g-prior, is a normal-inverse-gamma distribution, leading to posterior distributions retrieved also in closed forms which belong to the same family as the prior distributions for each respective parameter $\boldsymbol{\beta}_{\gamma}$, a, σ^2; see Marin and Robert (2014); for more details see also equations (A.2), (A.3) and (A.4) respectively found in Appendix section A.2. However, the mean and variance-covariance structure of the posterior distribution $\boldsymbol{\beta}_{\gamma}|g, \boldsymbol{y}, \sigma^2, \boldsymbol{\gamma}$ depend on $\frac{g}{g+1}$, called shrinkage factor, which preserves nice properties and shrinks the maximum likelihood estimator $\widehat{\boldsymbol{\beta}}_{\gamma}$ to prior zero mean. The issue of the shrinkage factor $\frac{g}{g+1}$ will be more discussed in the next sections, which refer to fully Bayesian variable selection methods and how it takes part in the computation of the marginal likelihood. In particular, the posterior mean enables the so-called data *linear* dependent shrinkage, which is an adaptive shrinkage provided by the data. The posterior mean of the parameter $\boldsymbol{\beta}_{\gamma}$ of a given model $\boldsymbol{\gamma}$ is expressed as

$$\mathbb{E}\left(\boldsymbol{\beta}_{\gamma}|g, \boldsymbol{y}, \sigma^2, \boldsymbol{\gamma}\right) = \frac{\widehat{\boldsymbol{\beta}}_{\gamma}}{1+\frac{1}{g}},$$

where the above expression shows that the posterior mean $\mathbb{E}\left(\boldsymbol{\beta}_{\gamma}|g, \boldsymbol{y}, \sigma^2, \boldsymbol{\gamma}\right)$ of $\boldsymbol{\beta}_{\gamma}$ is a weighted average of the data and the prior mean, with weights 1, $\frac{1}{g}$ respectively.

The whole uncertainty regarding the model parameters can be integrated (averaged) in the marginal likelihood leading to recognizable kernels of these distributions with some mathematical steps which result in the calculation of the above marginal likelihood.

In addition, when model choice is of interest and a convenient base model is chosen (in our case always the null model) Bayesian variable selection selection proceeds with comparing each Bayes factor of model space with respect to the null model. In an alternative way, Bayes factors comparison is regarded as a simultaneous hypothesis testing, where the null hypothesis restricts the values of parameter $\boldsymbol{\beta}_\gamma$. More precisely, if there is no prior knowledge regarding model choice, often a uniform prior is used to express the ignorance across model space, posterior model probabilities are up to a proportionality constant of each model's $\boldsymbol{\gamma}$ Bayes factor versus the null model $\boldsymbol{\gamma}_0$ as

$$\pi(\boldsymbol{\gamma}|\boldsymbol{y}, g) \propto BF_{[\gamma:\gamma_0]}(g),$$

where the Bayes factors $BF_{[\gamma:\gamma_0]}(.)$ is a function of g hyperparameter and the above is a result of Jeffrey's ideas that took into account in the original work of Liang et al. (2008). Moreover, each Bayes factor of model $\boldsymbol{\gamma}$ versus $\boldsymbol{\gamma}_0$ is equal to

$$BF_{[\gamma:\gamma_0]}(g) = (1+g)^{\frac{n-1-p_\gamma}{2}}[1+g(1-R_\gamma^2)]^{-\frac{n-1}{2}}.$$

The above expression is reduced due to the marginal likelihood of null model $\boldsymbol{\gamma}_0$ which holds for $p_{\gamma_0} = 0$, $R^2_{\gamma_0} = 0$; see for more Liang et al., (2008).
On the other hand, even though default prior specification of g-prior leads to automatic model selection procedure, the hyperparameter g remains an open discussion for many research publications. After performing the model selection, its influence is most notably evident in the estimation and model selection measures. This influence is related to the specification of the g hyperparameter, which acts as a dimensionality penalty in the model choice problem and causes strange behaviours. This dimensionality penalty emerges from the fact that if the specification of g is large, the model selection procedure enforces the selection of more parsimonious models, whereas it is small, saturated models are favoured. This fact is deduced from Jeffreys paradox which was mentioned above and is strongly related to the sensitivity of the large g specification which favors sparse models, forcing the Bayes factor $BF_{[\gamma:\gamma_0]} \to 0$; see Bartlett (1957) and Lindley (1957). In the literature there are approaches which deal with the specification of g either with prefixed values or by estimating a value from the corresponding data $\boldsymbol{y}$. Although, the specification of g was subject to intense debate, the use of a fixed value has been less criticized for some cases, whereas empirical Bayes procedures provide reasonable estimates over the corresponding marginal likelihood of g using the data. These approaches are summarized as follows *unit information prior* Kass and Wasserman (1995), *risk inflation criterion* Foster and George (1994), *benchmark priors* Fernandez et al. (2001), *local empirical Bayes* and *Global empirical Bayes* George and Foster (2000). Except for empirical Bayes methods, prefixed Bayesian methods including those of unit information prior, risk inflation criterion and benchmark prior

don't resolve paradoxes related to information consistency. Liang mentioned in her original paper that under prespecified g, the information paradox is encountered since these values do not include information from the data and thus seem odd. This paradox appears when a large quantity of evidence is accumulated for a particular model $\boldsymbol{\gamma}$, but instead of forcing the Bayes factor $BF_{[\boldsymbol{\gamma}:\boldsymbol{\gamma}_0]} \to \infty$, it forces it to diverge to the constant term $(1+g)^{\frac{n-1-p_{\gamma}}{2}}$; and this is unusual. Furthermore, empirical Bayes procedures exhibit even stronger non linear shrinkage and completely ignore the model uncertainty reflected in the estimated standard errors of g. Although, the approaches described above are considered very useful in terms of dealing with the specification of hyperparameter g, in general an arbitrary value to the unknown quantities may cause bias in selecting the correct model or even providing estimators using the data twice; see Liang et al. (2008). Thus, the need to use pragmatic methods to eliminate the dependence on g and the need to account for the additional uncertainty of g hyperparameter lead to fully Bayesian approaches using mixtures of g-priors which will be analyzed in the context of Bayesian variable selection. These will represent the main core of the present thesis completing the objective approach and are presented in the next section.

2.1.4 Model Choice with Mixtures of g-Priors

The need for creating default fully Bayesian variable selection methods fascinated researchers working intensively on extensions of g-priors. They surpassed the most difficult aspects related to g, which led to the development of the so-called mixtures of g-priors Liang et al. (2008). In this context, a prior distribution $\pi(g)$ is assigned to g to account for the additional uncertainty in a realistic sense rather than using an odd or plug-in value for g. Priors of this type are born from the desire to create automatic Bayesian variable selection procedures which assure consistency in model selection providing shrinkage and sparsity in covariate terms. These notions will be further discussed in the second part of this chapter when we will describe model search algorithms for the variable selection problem George and McCulloch (1993) and Dellaportas et al. (2002). In addition, the literature of Bayesian variable selection is vastly centered on prior distributions on g but these approaches are beyond the scope of the present thesis. Especially in linear regression, most of the research contributions came into light due to the simplified version of prior covariance matrix of g-prior, which did not depend on model coefficients. Some distinctive research contributions includes the approaches of Zellner and Siow (1980), Liang et al. (2008), Cui and George (2008), Maruyama and George (2011), Ley and Steel (2012) and Bayarri et al. (2012). The work of Liang et al. (2008) is of major interest along with that of Zellner and Siow (1980). This thesis covers the topic related to Zellner and Siow (1980) prior and Liang et al. (2008) hyper-g prior. Despite the enormous success of mixtures of g-priors in Bayesian variable selection for linear models, in the generalized linear model settings the case was just the opposite and this will be presented in the next chapter, Bové and Held (2011) and Li and Clyde (2018). In this section, the reader is introduced to the main points of mixtures of g-priors based on the work of Zellner and Siow (1980), namely the Zellner-Siow prior. Then the hyper-g-prior will be presented.

2.1.4.1 Hierachical Prior Specification

A natural Bayesian determinant to the uncertainty regarding the choice of g, is a hyper-prior on g to allow the data decide in an automatic manner. This will make actually the analysis more robust sharing the ideas of objective Bayesian methods with respect to the assumptions on g extending the prior specification (2.5). In this section, we review the basic ideas regarding the g-prior combined with a proper prior $\pi(g)$ distribution for g based on Jeffreys (1961) and Zellner and Siow (1980) seminal works. Appealing to Jeffrey's ingenious ideas for a simple hypothesis testing for a normal mean, (Zellner and Siow, 1980) proposed the use of multivariate Cauchy priors for Bayesian variable selection. As is commonly accepted and highlighted in Liang's research paper, Cauchy priors inherit heavy tails like a t-student distribution and can be easily expressed as a scale mixtures of normal priors. Thus, the Zellner-Siow proposal prior for the variable selection problem results just as a modification of (2.5)

$$\pi(a, \boldsymbol{\beta}_{\gamma}, \sigma^2, g|\boldsymbol{\gamma}) = \pi(a, \sigma^2|\boldsymbol{\gamma})\pi(\boldsymbol{\beta}_{\gamma}|\sigma^2, \boldsymbol{\gamma}), \tag{2.9}$$

with

$$\pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\sigma^2, \boldsymbol{\gamma}) = \int_0^{\infty} N_{p_{\gamma}}\left(\boldsymbol{\beta}_{\gamma}\middle|\mathbf{0}_{\boldsymbol{p}_{\gamma}}, g\sigma^2(\boldsymbol{X}_{\gamma}^{\boldsymbol{T}}\boldsymbol{X}_{\gamma})^{-1}\right)\pi(g)dg,$$

where the previous represents the cauchy prior as integrated version of Liang's g-prior (2.7) over the hyper-prior $\pi(g)$ for g. Distributions of such form are known from probability theory as scale mixtures of normals and the corresponding $\pi(g)$ is called mixing function. An equivalent way of seeing mixing of distributions in (**??**), is like a hierachical prior specification for constructing (2.5) starting from the coefficients vector $\boldsymbol{\beta}_{\gamma}$ given $\boldsymbol{\gamma}$ and then employing a proper prior for g. On the other hand, as we saw previously g appears not only in posterior model probabilities and Bayes factors, but even in the resulting posterior measures for estimation, as for example, in the posterior mean and matrix covariance of $\boldsymbol{\beta}_{\gamma}$. Thus the prior specification must be selected carefully to allow tractable computation both in model selection and estimation terms. When the goal is model determination, after adopting a base model comparison, the calculation of the posterior model probabilities and Bayes factors of each model $\boldsymbol{\gamma}$ based on the mixture representation are modified as

$$\pi(\boldsymbol{\gamma}|\boldsymbol{y}) \propto BF_{[\gamma:\gamma_0]},$$

where,

$$BF_{[\gamma:\gamma_0]} = \int_0^\infty (1+g)^{\frac{n-1-p_\gamma}{2}} [1+g(1-R^2_\gamma)]^{-\frac{n-1}{2}} \pi(g) dg.$$

Notice that the previous posterior measures in are refined versions cleaned from the dependence of hyperparameter g enabling a straightforward model comparison in variable selection. While tractable marginal likelihood and prediction are essentials for model selection, researchers often resort to prior specifications that lead to consistent Bayesian variable selection methods. For this reason, in the next section we will present in detail two default Bayesian methods, the so-called Zellner-Siow prior based on the previous mixture representation through an inverse-gamma and the hyper g as special mixing functions.

2.1.4.2 Model Choice with Zellner-Siow Priors

In the context of model selection, Jeffreys (1961) avoided the use of normal distributions for testing a simple normal mean due to the paradoxes of the Bayes factors. Inspired by Jeffrey's, the first mixtures of g-priors was the Zellner-Siow's prior, which can be described through a multivariate Cauchy distribution satisfying consistency in terms of Bayesian variable selection for regression model; see for more Zellner and Siow (1980). However, the popularity of this family of mixture priors was restricted due to non tractable forms in terms of model selection and posterior estimation. Consequently, one has to apply numerical methods. The prior formulation (2.9) starts by assigning for model parameters σ^2, a a common Jeffreys' prior as introduced by (2.6) and for regression coefficients $\boldsymbol{\beta}_\gamma$ of each model $\boldsymbol{\gamma}$ a multivariate Cauchy of the following form

$$\pi(\boldsymbol{\beta}_\gamma|\sigma^2, \boldsymbol{\gamma}) \propto \left(1 + \frac{\boldsymbol{\beta}_\gamma^{\boldsymbol{T}} \boldsymbol{X}_\gamma^{\boldsymbol{T}} \boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma}{n\sigma^2}\right)^{-\frac{p_\gamma}{2}}.$$

However, Liang stated in her original paper that the computation of marginal likelihood using Cauchy distribution doesn't lead to closed form representation and when the model dimension increases this computation becomes clearly unfeasible. Later, (Zellner and Siow, 1980) took advantage of mixture representation of Cauchy prior in expressing it, as a mixture of normals over inverse-gamma distributions as

$$\pi(\boldsymbol{\beta}_\gamma|\sigma^2, \boldsymbol{\gamma}) = \int_0^\infty N_{p_\gamma}\left(\boldsymbol{\beta}_\gamma \middle| \mathbf{0}_{\boldsymbol{p}_\gamma}, g\sigma^2(\boldsymbol{X}_\gamma^{\boldsymbol{T}}\boldsymbol{X}_\gamma)^{-1}\right)\pi^{ZS}(g)dg,$$

where $\pi^{ZS}(g)$ is a proper prior distribution on the hyperparameter g denoted as

$$\pi^{ZS}(g) = \frac{n^{\frac{1}{2}}}{\Gamma\left(\frac{1}{2}\right)} g^{-\frac{3}{2}} \exp\left\{-\frac{n}{2g}\right\}, \tag{2.10}$$

where the above implies $g \sim IG\left(\frac{1}{2}, \frac{n}{2}\right)$; see Liang et al. (2008). Note that for notation reasons, $\pi^{ZS}(.)$ denotes the hyper-prior of Zellner-Siow.

If we want to calculate the posterior model probabilities, one has to calculate first each compared Bayes factor of a given model $\boldsymbol{\gamma}$ versus the null model $\boldsymbol{\gamma}_0$ expressed under Zellner-Siow prior formulation

$$BF^{ZS}_{[\gamma:\gamma_0]} = \frac{n^{\frac{1}{2}}}{\Gamma\left(\frac{1}{2}\right)} \int_0^\infty (1+g)^{\frac{n-1-p_\gamma}{2}}[1+g(1-R^2_\gamma)]^{-\frac{n-1}{2}} g^{-\frac{3}{2}} \exp\left\{-\frac{n}{2g}\right\} dg, \tag{2.11}$$

where the above Bayes factor is a unidimensional integral over the g hyperparameter for which there are no available closed mathematical forms.

In order to handle such integrals of the form $\int_0^\infty h(g)dg$, where $h(.)$ is a real-valued function, it is recommended to use the Laplace approximation Tierney et al. (1989). This approximation is valid only under certain regularity conditions and consists of expanding a smooth unimodal function twice differentiable $H(g) = \log h(g)$ in a Taylor series expansion of second order around $\hat{g}$, the mode of $H(g)$. The Laplace approximation can be implemented as the following

$$\int_0^\infty \exp\{H(g)\}\, dg \approx \sqrt{2\pi}\hat{\sigma}_H h(\hat{g}),$$

where $\widehat{\sigma}_H \approx \left[\frac{-d^2H(g)}{dg^2}\Big|_{g=\hat{g}}\right]^{-\frac{1}{2}}$; for more details see Appendix section A.3. The mode $\hat{g}$ will result as the solution of the equation $\frac{dH(g)}{dg} = 0$.

In addition, Bayes factor of (2.11) under the Zellner-Siow prior (2.10) is approximated by the Laplace approximation

$$\widehat{BF}^{ZS}_{[\gamma:\gamma_0]} \approx \frac{n^{\frac{1}{2}}}{\Gamma\left(\frac{1}{2}\right)} \sqrt{2\pi}\hat{\sigma}_{H^{ZS}} h(\widehat{g}^{ZS}),$$

where $\widehat{\sigma}_{H^{ZS}}$ is the sampling error of posterior mode $\widehat{g}^{ZS}$ and the mode $\widehat{g}^{ZS}$. The quantities $\widehat{g}^{ZS}$ and the mode $\widehat{g}^{ZS}$ are based on Abramowitz and Stegun (1970) and Liang et al. (2008); see also for more details on equations (A.9) and (A.10) Appendix section A.4. Further details are described by Liang et al. (2008), where the mode $\hat{g}$ of Bayes factor (2.11) is provided as the solution of a cubic equation; see for more Liang et al. (2008) Appendix section A.1.

2.1.4.3 Model Choice with Hyper-g-Priors

As an alternative to Zellner-Siow prior for the model choice problem, another family of mixture of g-priors is the hyper-g-prior which is mainly preferred for its reasonable analytical tractability of marginal likelihood, Bayes factors and its closed expressions for all posterior statistics of interest. In addition to the original work of Liang et al. (2008), we went a step further providing closed mathematical expressions for the first and second posterior moments that represent important means of location and scale. Similarly to Zellner-Siow, the mixture representation presented in the previous section may be extended to a hierachical prior specification, but this time using the hyper-g-prior. It is important to point out that this mixture of g-priors must not be confused with that of Cauchy prior in the sense that it does not lead to Cauchy prior. The hyper-g-prior borrows its name from the Gaussian hypergeometric function included in posterior measures of evidence and estimation. The Gaussian hypergeometric function ${}_2F_1(.)$, is defined for a generic real variable x

$$ {}_2F_1(a,b;c;z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \int_0^1 \frac{x^{b-1}(1-x)^{c-b-1}}{(1-xz)^a} dx, $$

where the integral is convergent for real $|z| < 1$ with $c > b > 0$; see Abramowitz and Stegun (1970) and Liang et al. (2008). In addition, the mixture representation provided initially by Jeffreys (1961) and Zellner and Siow (1980) regarding Bayesian variable selection, allows to express the joint prior (2.5) implying the same improper prior (2.6) and including the additional model uncertainty for g through a multivariate distribution for $\boldsymbol{\beta}_\gamma$ with a mixing distribution

$$ \pi(\boldsymbol{\beta}_\gamma|\sigma^2,\boldsymbol{\gamma}) = \int_0^\infty N_{p_\gamma}\left(\boldsymbol{\beta}_\gamma\middle|\mathbf{0}_{\boldsymbol{p}_\gamma}, g\sigma^2(\boldsymbol{X}_\gamma^{\boldsymbol{T}}\boldsymbol{X}_\gamma)^{-1}\right)\pi^{hy}(g)dg, $$

where $\pi^{hy}(g)$ is the hyper-g-prior on g denoted as

$$ \pi^{hy}(g) = \frac{\alpha-2}{2}(1+g)^{-\frac{\alpha}{2}}, \ \ \alpha > 2, \tag{2.12} $$

where the prior elicitation of parameter g depends on the choice of the hyperparameter α. For values $\alpha > 2$, this is a proper distribution and for values $\alpha \leq 2$ this is an improper distribution. Notice for $\alpha = 2$ it corresponds to the usual Jeffreys' prior; see Strawderman (1971) and Cui and George (2008). In general, any choice of $\alpha \in (2, 4]$ lead to robust performance, apart from values near 2 that tend to activate Jeffrey's-Lindley's Paradox. Notice again we denote as $\pi^{hy}(.)$ the prior indexed by the hyper-g. Alternatively, an equivalent fully Bayesian analysis is obtained by translating the hyper-g-prior for g into a Beta prior on the shrinkage factor $\frac{g}{g+1}$ common for all models $\boldsymbol{\gamma}$

$$\frac{g}{g+1} \sim Beta\left(1, \frac{\alpha}{2} - 1\right), \tag{2.13}$$

which leads to a prior mean equal to $\frac{2}{\alpha}$. In a similar way, the elicitation of g is determined by the values of α in the range $[2, +\infty)$. The choice of $\alpha = 4$, results non-informative in the sense that it turns the prior of the shrinkage factor $\frac{g}{g+1}$ into uniform. A value close to 2, concentrates the probability mass on the shrinkage factor close to 1. Conversely, any $\alpha > 4$ concentrates probability mass near 0. Furthermore, the authors derived the model specific posterior distribution of g and some posterior statistics by relying on the integral representation of Gaussian hypergeometric function ${}_2F_1(a, b; c; z)$ which is computed in exact form

$$\pi(g|\boldsymbol{\gamma}, \boldsymbol{y}) = \frac{p_\gamma + \alpha - 2}{2\,{}_2F_1\left(\frac{n-1}{2}, 1; \frac{p_\gamma + \delta}{2}; R^2_\gamma\right)}(1+g)^{\frac{n-1-p_\gamma-\alpha}{2}}[1 + g(1 - R^2_\gamma)]^{-\frac{n-1}{2}}.$$

In addition, when model selection is mandatory, the computation of posterior model probabilities via the Bayes theorem include the comparison of Bayes factor of each model in the model space versus the null model after adopting the hyper-g-prior

$$BF^{hy}_{[\gamma:\gamma_0]} = \frac{\alpha - 2}{2}\int_0^\infty (1+g)^{\frac{n-1-p_\gamma-\alpha}{2}}[1 + g(1 - R^2_\gamma)]^{-\frac{n-1}{2}}dg, \tag{2.14}$$

which can be found in closed form based on previous considerations. The corresponding Bayes factor (2.14) using either hyper-g-prior (2.12) either the hyper-g-prior (2.13) for the shrinkage factor is calculated after recognizing the normalization constant in posterior of g as

$$BF^{hy}_{[\gamma:\gamma_0]} = \frac{\alpha - 2}{p_\gamma + \alpha - 2}\,{}_2F_1\left(\frac{n-1}{2}, 1; \frac{p_\gamma + \alpha}{2}; R^2_\gamma\right), \tag{2.15}$$

which is found in tractable form in terms of Gaussian hypergeometric function; see for more Appendix section A.5. As it was mentioned before, the posterior distribution of

g allows to derive important posterior summary statistics in closed form such as first and second moments both for g and $\frac{g}{g+1}$ in comparison with Liang et al. (2008) who rest only on the posterior mean of shrinkage factor, while we derived also the first and second posterior moments of g and $\frac{g}{g+1}$; see for more equations (A.15), (A.17), (A.16) and (A.18).

On the contrary, most notably, posterior expectations given in Liang et al., (2008) involve ratios of Gaussian hypergeometric functions $2F_1(.)$. The authors propose to compute via Laplace approximations when the problem of numerical overflows appears. Their approach, in order to avoid problems with the boundary of Zellner's g parameter space, uses a Laplace approximation after applying the transformation $z = \log(g)$ in the corresponding Bayes factor (2.14)

$$\widehat{BF}^{hy}_{[\gamma:\gamma_0]} \approx \frac{\alpha - 2}{2} \sqrt{2\pi} \widehat{\sigma}_{H^{hy}} h(\widehat{z}^{hy}),$$

where $\widehat{\sigma}_{H^{hy}}$ is the standard error of the mode $\widehat{z}^{hy}$ and the mode $\widehat{z}^{hy}$ is resulting as the solution to a quadratic equation, all the above results are described further by Liang et al. (2008) and Abramowitz and Stegun (1970); see also for additional details of the respective calculations equations (A.21) and (A.22) in Appendix section A.6.

2.2 MCMC for Bayesian variable selection in Linear Models

Bayesian model selection is regarded as a dominant procedure in daily practice that successfully deals with the model and parameter uncertainty. Model uncertainty is translated into covariate uncertainty in Bayesian variable selection problem where there are not sufficient guidelines on how to select an "ideal" subset. This problem is encountered in many scientific areas where an automatic procedure sharing probabilistic nature, like objective Bayesian methods, is always essential. More precisely, Bayesian variable selection in linear regression models is considered one of the most important aspects of model selection, useful for predictions and causal relationships between the response variable and the explanatory variables, while at the same time it decides which variables should be included or not in the model. In other words, the variable selection problem arises when there is an unknown subset of explanatory variables with regression coefficients too small that it would be preferable to ignore them Chipman et al. (2001). In this way, variable selection in a regression problem is viewed as one of inducing sparsity or parsimony and shrinkage inspired by the ingenious ideas of *spike-*

slab priors. Moreover, parsimony and shrinkage, are two terms that are completing each other in the sense that one implies the other and this refers in the way that models are produced in the most notably simple form forcing noise variables to be omitted. Equivalently, ignoring variables that are noise to the data produces simpler models. Under this consideration, spike-slab priors gained much popularity for their adaptability to the data both in prediction and in model selection.

In addition, the use of computers and technology progress reawakened new paths that were clearly infeasible in the last decades and within the advent of Markov Chain Monte Carlo (MCMC), the Bayesian core of variable selection received great attention promising innovative solutions and ideas for complex models in high dimensional settings. The MCMC methods provide a powerful tool regarding the model selection avoiding the computational burden of the solid Bayesian oracle. We invite the reader to give a short look to the main manual for MCMC provided by Gilks et al. (1996).

Two common problems that we encounter within the Bayesian variable selection are the computation of the posterior model probability and the enumeration of the model space. The Bayesian variable selection with MCMC methods was quite promising to overcome these limitations. The MCMC methods in Bayesian variable selection problem produce flexible inference by generating a simulated sample from the joint posterior distribution of the model parameters and the model itself through a Markov chain that operates on the model space converging to the target distribution. The joint distribution generated by the sampler of MCMC, represents the posterior distribution by imposing a hierarchical mixture prior on the regression coefficients; we will discuss this hierarchical prior in the next section. Despite, these methods show an exclusive way to calculate the posterior, they provide a substantial exploration of the model space summarizing relevant information related to variable selection. The MCMC methods in Bayesian variable selection for linear models is covered immensely in the literature, as it gained much attention due to a variety of research accomplishments. The most important research contributions were described by George and McCulloch (1993), Madigan and Raftery (1994), Carlin and Chib (1995), Green (1995), Smith and Kohn (1996), Raftery et al. (1997), Hoeting et al. (1999) , Kuo and Mallick (1998), Dellaportas et al. (2002) and Ročková and George (2014) each of them describing a special contribution. However, we will restrict our attention to the Stochastic Search Variable Selection (SSVS) of George and McCulloch (1993), the Gibbs Variable Selection (GVS) of Dellaportas et al. (2002), which are the main tools applied in the present thesis and then we will summarize important features related to each variable selection procedure. As a consequence, in this second part of this chapter we want to extend the above Bayesian variable selection for these methods using mixtures of g-priors introduced by Liang et al. (2008), including Zellner-Siow prior and hyper-g-prior and enlighting up some additional formulations which will be used within the frames of a simulated and a real data-set in order to assess their performance. Before that, let us introduce some basic concepts of a more general hierachical Bayesian variable selection scheme in the next section.

2.2.1 Hierarchical Prior Specification for MCMC

Prior specification is regarded one of the most crucial steps to initialize a Bayesian model selection method, while its choice deserves great attention. When, variable selection takes place as a special case of model choice, prior elicitation becomes virtually impossible as the number of independent variables grows, thus researchers have to resort to alternative priors to deal with this issue, especially in high dimensional problems. The prior specification to solve this problem, namely spike-slab, was introduced by Mitchell and Beauchamp (1988) and then generalized by George and McCulloch (1993) and Chipman et al. (2001). The main intuition behind these priors rests firmly on the context of variable selection, through a hierachical construction conditional on the values of a binary latent vector $\boldsymbol{\gamma}$. Incorporating the binary indicator variable $\boldsymbol{\gamma}$ in the analysis, implicitly indicates that the vector of coefficients $\boldsymbol{\beta}$ is an example of a spike-slab prior distribution. More specifically, a spike-slab prior is a mixture of two distributions, of a spike and of a slab respectively, where the spike is highly peaked in a region of zero with small variance or precisely at zero capturing the coefficients that are not significant towards zero and the slab prior is a more spread distribution covering plausible values moving away from zero. In other words, we simply assign less uncertainty reflected in the small variance of the spike component just to capture the non significant effects and more uncertainty expressed in the large variance of the slab component to allow the possibility for capturing important effects. In addition, Mitchell and Beauchamp (1988) introduced these priors in order to facilitate the variable selection problem putting additional constraints to the corresponding effects of the independent variables being zero or not. Prior distributions of this form are formulated component-wise for each element β_j as

$$\pi(\beta_j|\gamma_j) = \gamma_j \pi^{slab}(\beta_j) + (1-\gamma_j)\pi^{spike}(\beta_j),$$

where $\pi^{slab}(.)$, $\pi^{slab}(.)$ denote the slab and spike components respectively and the above expression implies prior independence between the individual elements of $\boldsymbol{\beta}$, for $j = 1, \dots, p$. In their original approach Mitchell and Beauchamp (1988); used a spike highly peaked at zero and a uniform slab large spread around plausible values. Based on the current value of γ_j, the effect of β_j will belong either to the spike or to the slab. According to (Ročková and George, 2014); if there is strong evidence provided by the data against the inclusion of the respective effect of the covariate, the spike component will dominate the posterior for which it will effectively shrink the posterior mean towards zero. Thus, the spike-slab distributions clarify which regression coefficients are significant and which are not. Alternatively, the spike part will be substituted by a continuous distribution with zero mean and small variance; see George and McCulloch (1993); Dellaportas et al. (2002).

2.2.2 Stochastic Search Variable Selection

The stochastic search variable selection (SSVS) is one of the great fundamentals of Bayesian variable selection and was introduced by George and McCulloch (1993) in their attempt to extend the spike-slab, provided an efficient model exploration algorithm using a Gibbs sampler. In parallel with the outbreak of MCMC methods, SSVS was the main motivation that inspired many researcher to develop efficient Bayesian variable selection methods for the exploration of model space and it is considered as the predecessor of many methods. Even though it was introduced for variable selection in linear regression, many extensions surrounding SSVS were also used for instance in factor analysis and time series in many different areas of science. The SSVS approach in linear regression begins describing the relationship between the response variable Y and the set of predictors $X_1, \dots, X_p$ using a linear model of the form (2.1) where the above is the usual regression setup. The latent binary vector $\boldsymbol{\gamma}$ was first introduced by George and McCulloch (1993) and arose naturally for the interpretation of the inclusion or exclusion of the respective covariates. In addition, for each covariate and its respective effect, a mixture of two normal distributions with the one having most of its mass concentrated around zero and the other one spread over plausible values, tuned by possible additional hyperparameters, is assigned. Their choice ensures the good mixing of the underlying MCMC method. This is achieved through a prior distribution for each element β_j conditional on the values of γ_j

$$\pi(\beta_j|\gamma_j) = \gamma_j N(0, c_j^2\tau_j^2) + (1-\gamma_j)N(0, \tau_j^2),$$

where the parameters τ_j, c_j are large and set respectively so that the distribution $N(0, c_j^2\tau_j^2)$ and the $N(0, \tau_j^2)$ is diffuse. The tuning of parameters τ_j, c_j determine the performance of the procedure, while the magnitude of the shrinkage of the non important covariates depends on τ_j. The authors in their original paper, mention that the option of τ_j, c_j is based on practical significance rather than statistical which appears as a solution of the intersection points between the two densities $N(0, c_j^2\tau_j^2)$ and the $N(0, \tau_j^2)$; see for details George and McCulloch (1993) and Chipman et al. (2001). The intuitive idea behind the above prior specification implies that, when $\gamma_j = 1$, the corresponding effect β_j is included in the model with a prior that is sharp in reasonable values away from zero in order to let the data decide through the posterior. On the contrary, when $\gamma_j = 0$, β_j is absent from the model with an informative prior around zero; then the effect is shrunk to zero. Although, the initial prior construction of George and McCulloch (1993) was developed for different τ_j, c_j, we would like to propose something slightly different but more flexible and automatic from the previous consideration by assuming the same prior inputs $\tau_j = \tau$, $c_j = c$ for each regression coefficients β_j such that to obtain results near objective Bayesian approaches. The above prior for each element β_j, can be expressed also under a more general multivariate normal $\boldsymbol{\beta}$, which is described in George and McCulloch (1993) and Chipman et al. (2001). Conditional on the binary vector $\boldsymbol{\gamma}$, the prior for $\boldsymbol{\beta}$ is expressed as

$$\boldsymbol{\beta}|\boldsymbol{\gamma} \sim N_p(\mathbf{0}_p, \boldsymbol{DRD}),$$

where $\boldsymbol{R}$ is the prior correlation matrix and $\boldsymbol{D}$ is a diagonal matrix with j-th entry equal to $\gamma_j c_j \tau_j + (1-\gamma_j)\tau_j$ that arranges the scale of prior-covariance. Although, the choice of $\boldsymbol{R} = \boldsymbol{I_p}$ yields in equivalence this prior with the above, $\boldsymbol{R} = g\sigma^2\boldsymbol{D}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{D}$ is adopted as the main objective choice based on g-prior. The SSVS ends with the joint prior specification of parameters $a, \sigma^2, g, \boldsymbol{\gamma}$ based on Jeffrey's objective approach for the linear model

$$\pi^{SSVS}(a, \beta, \sigma^2, g, \boldsymbol{\gamma}) = \pi(a, \sigma^2)\pi^{SSVS}(\boldsymbol{\beta}|g, \sigma^2, \boldsymbol{\gamma})\pi(g)\pi(\boldsymbol{\gamma}), \tag{2.16}$$

where denotes $\pi(a, \sigma^2)$ the same improper Jeffreys' prior as (2.6), but with the only difference that the model indicator $\boldsymbol{\gamma}$ is ommited for simplicification purposes. Also, regarding the prior for model space

$$\pi(\boldsymbol{\gamma}) = \prod_{j=1}^{p} w_j^{\gamma_j}(1-w_j)^{1-\gamma_j},$$

where the above prior is an independent Bernoulli with prior probability weights of inclusion w_j and exclusion $1 - w_j$ for each respective covariate. This prior is a realistic choice which implies that the inclusion of one covariate is independent from the inclusion of any one other; the same assumption holds also for exclusion. Prior reduces to uniform prior which puts equal prior probability mass to each covariate, in other words, the inclusion of each covariate is likely the same. Moreover, if we consider $w = w_j$ the prior choice of $\boldsymbol{\gamma}$ may be extended to a more general class of hierarchical prior which deals with the specification of prior weights w_j as the following $w|\tilde{p}_1, \tilde{p}_2 \sim \; Beta(\tilde{p}_1, \tilde{p}_2)$. If hyperparameters $\tilde{p}_1$, $\tilde{p}_2$ are set to one, prior becomes non-informative, whereas there is an expert's opinion stating that it can be used also as an informative prior. Scott and Berger (2010) showed that the above hierarchical prior is much preferable than that of Bernoulli distribution, with specified prior weights w_j as it preserves sparsity in high dimensional settings.

To conclude, regarding the prior for g, it is either adopted with a Zellner-Siow prior (2.10) or with a hyper-g (2.12).

One important feature of the SSVS prior (2.16) , is that by construction it keeps dimensionality constant across all models $\boldsymbol{\gamma}$, which accomodates the incorporation of the binary latent vector $\boldsymbol{\gamma}$ in main prior parameter specification, in contradiction with formal Bayesian model selection which specifies prior model probabilities separately after marginalizing the model parameters. The joint posterior distribution of the model parameters and the model space is expressed as the product of the sampling density (2.1) and the SSVS prior (2.16)

$$\pi^{SSVS}(a, \beta, \sigma^2, g, \boldsymbol{\gamma}|\boldsymbol{y}) \propto f(\boldsymbol{y}|a, \boldsymbol{\beta}, \sigma^2)\pi(a, \sigma^2)\pi^{SSVS}(\boldsymbol{\beta}|g, \sigma^2, \boldsymbol{\gamma})\pi(g)\pi(\boldsymbol{\gamma}).$$

Notice in the above that the likelihood remains the same across model space in contradiction with the joint posterior that changes over each model $\boldsymbol{\gamma}$ and the joint posterior remains in an unrecognised form. In this way, although the joint posterior distribution is intractable in closed form, MCMC methods facilitate exploration of the joint posterior of the model. Model indicator parameters identify only the important areas of model space with high posterior probability avoiding exhaustive exact calculation of marginal likelihood and full enumeration of the model space. The corresponding MCMC procedure is applied through a Gibbs sampler which samples indirectly from the joint posterior distribution, using the full conditionals of each parameter a, $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{\gamma}$ resulting as a sample from joint posterior distribution $\pi^{SSVS}(a, \beta, \sigma^2, g, \boldsymbol{\gamma}|\boldsymbol{y})$. To conclude, analytical details and computations of full conditionals with respect to model specific parameters are avoided because are out of the scope of this thesis and hence we provide an analytic description of the implementation of SSVS in order to familiarize the interesting reader; see for more details Appendix section A.7.

2.2.3 Gibbs Variable Selection

The Gibbs variable selection (GVS) of Ntzoufras (1999) and Dellaportas et al. (2002) can be considered a variant of SSVS. It was first proposed for Bayesian variable selection in linear regression models and uses a similar but modified prior specification allowing to safely jump from one model to another when different sizes are concerned. This prior schema extends the idea of SSVS through a flexible prior mixture component for the non present effects ensuring the dimensionality across different size of models. In this approach, the linear relationship of the response Y and the set of predictors X_1,. . . , X_p is affected by the binary vector $\boldsymbol{\gamma}$ using a linear model of the form

$$\boldsymbol{Y}|a, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma} \sim N_n\left(a\mathbf{1}_{\boldsymbol{n}} + \boldsymbol{X}\boldsymbol{\Gamma}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_{\boldsymbol{n}}\right), \tag{2.17}$$

where $\boldsymbol{\Gamma} = \text{diag}(\boldsymbol{\gamma})$ is a matrix of dimension $p \times p$. In comparison with SSVS of George and McCulloch (1993), each effect β_j of $\boldsymbol{\beta}$ is now assigned with a modified prior scale mixture of normal distributions. Alike SSVS, this prior mixture incorporates the binary indicators γ_j of β_j as

$$\beta_j|\gamma_j \sim \gamma_j N(0, c_j^2\tau_j^2) + (1-\gamma_j)N(\bar{\mu}_j, \bar{s}_j^2),$$

where parameters τ_j, c_j are set small and large respectively, whereas $\bar{\mu}_j$, $\bar{s}_j$ are means and standard deviations of β_j obtained by a pilot study for a saturated model. Thus, the corresponding effect β_j will be classified to diffuse either prior component $N(0, c_j^2\tau_j^2)$ or prior component $N(\bar{\mu}_j, \bar{s}_j^2)$. Unlike SSVS, the mixture prior of GVS is distinguised because of the mixture component $N(\bar{\mu}_j, \bar{s}_j^2)$, commonly known as pseudo-prior, which acts as a "passive" prior incorporating no additional knowledge for β_j and hence does not affect the data and the resulting posterior. In order to familiarize the reader with the concept of pseudo-prior, we illustrate a simple example of the main prior specification as adopted in Dellaportas et al. (2002). Let the above univariate mixture prior be expressed in a more fashionable way like a mixture prior of SSVS with minor differentiations as a multivariate distribution for the vector of effects $\boldsymbol{\beta}$ as the following

$$\boldsymbol{\beta}|\boldsymbol{\gamma} \sim N_p\left(\boldsymbol{\mu}, \widetilde{\boldsymbol{D}}^2\right),$$

where we denote as $\boldsymbol{\mu} = (1-\boldsymbol{\gamma})\bar{\boldsymbol{\mu}}$ with $\bar{\boldsymbol{\mu}} = (\mu_1, \ldots, \mu_p)^{\boldsymbol{T}}$ and $\widetilde{\boldsymbol{D}}$ a diagonal matrix with j-th diagonal element equal to $\gamma_j c_j \tau_j + (1-\gamma_j)\bar{s}_j$ implying prior independence given latent vector $\boldsymbol{\gamma}$. Both $\boldsymbol{\mu}$, $\widetilde{\boldsymbol{D}}$ determine the prior mean and variance-covariance structure of the prior which are the main ingredients that distinguish the GVS from the SSVS method. If we consider in our disposal a fixed value $\boldsymbol{\gamma}^* = (1,0,1)$ for a given model $\boldsymbol{\gamma}$ the full conditional of $\boldsymbol{\beta}$ is computed

$$\begin{aligned}
&\pi(\beta_1,\beta_2,\beta_3|a,\sigma^2\boldsymbol{\gamma}^*,\boldsymbol{y}) \propto f(\boldsymbol{y}|a,\beta_1,\beta_2,\beta_3,\sigma^2,\boldsymbol{\gamma}^*)\pi(\beta_1,\beta_2,\beta_3|\boldsymbol{\gamma}^*)\\
&\propto \exp\left(-\frac{1}{2\sigma^2}[\boldsymbol{y}-a\mathbf{1}_{\boldsymbol{n}}-X_1\beta_1-X_3\beta_3)^T(\boldsymbol{y}-a\mathbf{1}_{\boldsymbol{n}}-X_1\beta_1-X_3\beta_3]\right),\\
&\exp\left(-\frac{\beta_1^2}{2c_1^2\tau_1^2}\right)\exp\left(-\frac{(\beta_2-\bar{\mu}_2)^2}{2\bar{s}_2^2}\right)\exp\left(-\frac{\beta_3^2}{2c_3^2\tau_3^2}\right).
\end{aligned}$$

Notice above that the priors for the included coefficients β_1,β_3 are independent of the pseudo-prior β_2 and the likelihood doesn't depend on β_2 in contradiction with β_1, β_3 for which $\gamma_1^* = 1, \gamma_3^* = 1$. This suggests that the actual posterior will be based on β_1, β_3 conditional on β_2, whereas β_2 will be updated only through the pseudo-prior as the following full conditionals imply

$$\begin{aligned}
&\pi(\beta_1,\beta_3|a,\sigma^2\gamma_1^*=1,\gamma_3^*=1,\boldsymbol{y}) \propto\\
&\exp\left(-\frac{1}{2\sigma^2}[\boldsymbol{y}-a\mathbf{1}_{\boldsymbol{n}}-X_1\beta_1-X_3\beta_3)^T(\boldsymbol{y}-a\mathbf{1}_{\boldsymbol{n}}-X_1\beta_1-X_3\beta_3]\right)\\
&\exp\left(-\frac{\beta_1^2}{2c_1^2\tau_1^2}\right)\exp\left(-\frac{\beta_3^2}{2c_3^2\tau_3^2}\right),
\end{aligned}$$

$$\pi(\beta_2|\gamma_2^*=0,\boldsymbol{y}) \propto \exp\left(-\frac{(\beta_2-\bar{\mu}_2)^2}{2\bar{s}_2^2}\right).$$

If we consider the previous example, in a more general framework, the full conditional of $\boldsymbol{\beta}$ in the previous example is computed

$$\begin{aligned}
&\boldsymbol{\beta}|\sigma^2,\boldsymbol{\gamma},\boldsymbol{y} \sim\\
&N_p\left(\left(\boldsymbol{\Gamma X^T X\Gamma}+\sigma^2\widetilde{\boldsymbol{D}}^{-2}\right)^{-1}\left(\boldsymbol{\Gamma X^T y}+\sigma^2\widetilde{\boldsymbol{D}}^{-2}\boldsymbol{\mu}\right),\sigma^2\left(\boldsymbol{\Gamma X^T X\Gamma}+\sigma^2\widetilde{\boldsymbol{D}}^{-2}\right)^{-1}\right),
\end{aligned}$$

under the guidelines of Dellaportas et al. (2002), who shows that if we consider the partition of $\boldsymbol{\beta} = (\boldsymbol{\beta}_\gamma, \boldsymbol{\beta}_{-\gamma})^T$ denoting the inclusion and exclusion vectors of coefficient vectors, the full conditional posterior distribution of the included effects is $\boldsymbol{\beta}_\gamma$

$$\pi(\boldsymbol{\beta}_\gamma|\sigma^2,\boldsymbol{y}) \propto \begin{cases} N_{p_\gamma}\left(\left(\boldsymbol{X}_\gamma^T\boldsymbol{X}_\gamma+\sigma^2\widetilde{\boldsymbol{D}}_\gamma^{-2}\right)^{-1}\widetilde{\boldsymbol{\mu}}_\gamma,\sigma^2\left(\boldsymbol{X}_\gamma^T\boldsymbol{X}_\gamma+\sigma^2\widetilde{\boldsymbol{D}}_\gamma^{-2}\right)^{-1}\right), & \boldsymbol{\gamma}=\mathbf{1}_{p_\gamma},\\ N_{p_{-\gamma}}\left(\boldsymbol{\mu}_{-\gamma},\widetilde{\boldsymbol{D}}_{-\gamma}^{-2}\right), \quad \boldsymbol{\gamma}=\mathbf{0}_{p_{-\gamma}}, \end{cases}$$

where $\widetilde{\boldsymbol{\mu}}_\gamma = \left(\boldsymbol{X}_\gamma^T\boldsymbol{y} + \sigma^2\widetilde{\boldsymbol{D}}_\gamma^{-2}\boldsymbol{\mu}_\gamma\right)$, $\widetilde{\boldsymbol{D}}_\gamma, \widetilde{\boldsymbol{D}}_{-\gamma}$ and $\boldsymbol{\mu}_\gamma, \boldsymbol{\mu}_{-\gamma}$ are the partitions of $\widetilde{\boldsymbol{D}}$ and $\boldsymbol{\mu}$ respectively. The above shows that the full conditional of $\boldsymbol{\beta}$ is equivalent to the full conditional of $\boldsymbol{\beta}_\gamma$ as a product of the actual posterior times the pseudo-prior for $\boldsymbol{\beta}_{-\gamma}$ which doesn't affect the posterior. Thus, if one either considered to use the joint full conditional for $\boldsymbol{\beta}$ or the partitioned full conditional based on the included effects $\boldsymbol{\beta}_\gamma$ he would reach the same conclusions as shown in Paroli and Spezia (2006). In our analysis, we follow this approach instead of the partitioned vector.

The above multivariate mixture prior was extended by Ntzoufras et al. (2002) and Perrakis and Ntzoufras (2018) for adopting g-priors in a more fashionable way unlike mixture prior of SSVS with minor differentiations into a multivariate distribution for the vector of effects $\boldsymbol{\beta}$ with prior precision matrix $\widetilde{\boldsymbol{D}}$ defined as

$$\widetilde{\boldsymbol{D}} = \left(\frac{\boldsymbol{\Gamma X^T X \Gamma}}{g} + \text{diag}(1-\boldsymbol{\gamma})\frac{1}{\bar{\boldsymbol{s}}^2}\right)^{-1},$$

where $\bar{\boldsymbol{s}} = (\bar{s}_1, \ldots, \bar{s}_p)^{\boldsymbol{T}}$ denotes the standard error vector obtained from a pilot run of a saturated model. To illustrate the adoption of this prior specification through pseudo-prior, consider the joint prior distribution of the partitioned vector $\boldsymbol{\beta}_\gamma, \boldsymbol{\beta}_{-\gamma}$

$$\pi(\boldsymbol{\beta}_\gamma, \boldsymbol{\beta}_{-\gamma}|\sigma^2, g) \propto \begin{cases} N_{p_\gamma}\left(\boldsymbol{\beta}_\gamma \middle| \boldsymbol{\mu}_\gamma, g\sigma^2(\boldsymbol{X}_\gamma^{\boldsymbol{T}}\boldsymbol{X}_\gamma)^{-1}\right), & \boldsymbol{\gamma} = \mathbf{1}_{p_\gamma}, \\ N_{p_{-\gamma}}\left(\boldsymbol{\beta}_{-\gamma} \middle| \boldsymbol{\mu}_{-\gamma}, \widetilde{\boldsymbol{D}}_{-\gamma}^{-1}\right), & \boldsymbol{\gamma} = \mathbf{0}_{p_{-\gamma}} \end{cases},$$

where the above prior suggests that the actual prior of included effects $\boldsymbol{\beta}_\gamma$ is generated from the g-prior independently again of the pseudo-prior which is not affected in case we update the whole parameter $\boldsymbol{\beta}$. To conclude, the joint prior specification of parameters $a, \boldsymbol{\beta}, \sigma^2, g, \boldsymbol{\gamma}$ is according to Jeffreys' ideas likewise the joint prior for SSVS previously

$$\pi^{GVS}(a, \beta, \sigma^2, g, \boldsymbol{\gamma}) = \pi(a, \sigma^2)\pi^{GVS}(\beta|g, \sigma^2, \boldsymbol{\gamma})\pi(g)\pi(\boldsymbol{\gamma}). \tag{2.18}$$

On the other side, unlike SSVS, GVS maintains similar properties since the prior for latent vector $\boldsymbol{\gamma}$ is embedded in joint prior specification (2.18). Under these settings, GVS method produces a Markov chain which covers a substantial set of most likely models and approximates sufficiently joint posterior distribution. This joint posterior distribution of specific model parameters and model itself is expressed as the product of the sampling density (2.17) and prior (2.18)

$$\pi^{GVS}(a, \beta, \sigma^2, g, \boldsymbol{\gamma}|\boldsymbol{y}) \propto f(\boldsymbol{y}|a, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma})\pi(a, \sigma^2)\pi^{GVS}(\boldsymbol{\beta}|g, \sigma^2, \boldsymbol{\gamma})\pi(g)\pi(\boldsymbol{\gamma}), \tag{2.19}$$

where it should be noted that the likelihood changes dimension given $\boldsymbol{\gamma}$ and the joint posterior remains in an intractable form. According to (Dellaportas et al., 2002), most of the success of the GVS method is due to the pseudo-prior that speeds up the convergence of the MCMC in order to target the distribution and maintains the dimensionality difference which is balanced in conjuction with the likelihood. In order to sample from the joint posterior $\pi^{GVS}(a, \beta, \sigma^2, g, \boldsymbol{\gamma}|\boldsymbol{y})$, an MCMC method is used based on Gibbs algorithm that samples sequentially from full conditionals of each parameter a, $\boldsymbol{\beta}$, σ^2, g, $\boldsymbol{\gamma}$ apart from the conditional of a which remains the same as (**??**). Finally, we omit subsequent steps of how to recover the full conditionals of each parameter for the same reasons as in SSVS; the kind reader may find also an analytic implementation of GVS in Appendix section A.8.

2.2.4 Posterior Exploration of Model Space

The implementation of model search algorithms for variable selection usually consist of two stages. The first stage entails setting the prior inputs to the hierarchical model, for example c_j, τ_j in case of SSVS, $\bar{\mu}_j$, $\bar{s}_j^2$ in GVS and the prior weights $\pi(\boldsymbol{\gamma})$ so that $\boldsymbol{\gamma}$ values corresponding to promising models are assigned higher posterior probability under $\pi(\boldsymbol{\gamma}|\boldsymbol{y})$ Gilks et al. (1996). The second stage includes the identification of these high posterior model probabilities via a Metropolis-Hastings step within Gibbs sampling Gilks et al. (1996). These approaches avoid the overwhelmed computational cost of computing all 2^p posterior model probabilities by numerical or analytical methods. Both variable selection methods use Gibbs sampler with Metropolis-Hastings step to generate a sequence

$$\boldsymbol{\gamma}^{(0)}, \ldots, \boldsymbol{\gamma}^{(S)},$$

which converges to the posterior distribution of model space $\pi(\boldsymbol{\gamma}|\boldsymbol{y})$. The relative frequency of each model $\boldsymbol{\gamma}$ converges to its probability $\pi(\boldsymbol{\gamma}|\boldsymbol{y})$. In particular, those values of $\boldsymbol{\gamma}$ with highest posterior model probability can be identified as those which appear most frequently in this sequence. It should be noted that, in contrast to many application of the Gibbs sampler, the goal here is not the evaluation of the entire posterior distribution of model space $\pi(\boldsymbol{\gamma}|\boldsymbol{y})$. In practice, most of the 2^p values of $\boldsymbol{\gamma}$ will result with very low probability, in the sense that their appearance will be rare in the sequence and so they can be ignored. In effect, SSVS and GVS use the Gibbs sampler to explore the model space rather than evaluating the whole posterior distribution $\pi(\boldsymbol{\gamma}|\boldsymbol{y})$. Consequently, the length of the sequence of $\boldsymbol{\gamma}$ is much smaller than the actual of 2^p possible subsets, and that's why it serves to identify only the areas of model space with high posterior model probabilities. Both MCMC algorithms generate a sequence by appending the Gibbs with Metropolis-Hastings step to the respective joint posteriors $\pi^{SSVS}(a, \boldsymbol{\beta}, \sigma^2, g, \boldsymbol{\gamma}|\boldsymbol{y})$ and $\pi^{GVS}(a, \boldsymbol{\beta}, \sigma^2, g, \boldsymbol{\gamma}|\boldsymbol{y})$. Then, these produce a complete sequence of couples of model specific parameters and model given the iteration

$$a^{(0)}, \boldsymbol{\beta}^{(0)}, \sigma^{2(0)}, g^{(0)}\boldsymbol{\gamma}^{(0)} \ldots, a^{(S)}, \boldsymbol{\beta}^{(S)}, \sigma^{2(S)}, g^{(S)}\boldsymbol{\gamma}^{(S)},$$

where a Markov chain operates on the model space converging to the joint posterior in which the above sequence is embedded. These MCMC methods are implemented as follows: the model specific parameters are initialized at some reasonable guess $a^{(0)}, \boldsymbol{\beta}^{(0)}, \sigma^{2(0)}, g^{(0)}\boldsymbol{\gamma}^{(0)}$ and are such that they might be obtained by generation from the respective priors or frequentist estimators. To conclude, subsequent values are generated in both algorithms iterating over the full conditional of model parameters and the model itself as it was shown in the respective sections. In this way, a sample containing all the appropriate information is provided by the MCMC methods.

2.3 Closing Remarks

In this chapter we revisited the variable selection problem for linear regression from a fully Bayesian perspective. We reviewed step by step the main aspects of objective prior specification and model selection based on Zellner's g-prior and its mixtures for Bayesian variable selection methods with full enumeration and MCMC methods. In the first part of this chapter, we provided the notion of centering and its importance in the Bayesian variable selection procedure. Closed form expressions of g and $\frac{g}{g+1}$ were derived in detail of the first and second posterior moments which are absent in the bibliography and in Liang et al. (2008) in Appendix section A.5. Furthermore, in the second half of this chapter, we outlined the Bayesian variable selection algorithms of SSVS and GVS extended in the framework of mixtures of g-priors, which require highly complex MCMC methods for their implementation (Appendix section A.7 and A.8 respectively) in order to prepare the interesting reader for the next chapters. These Bayesian variable selection methods, are not trivial especially in the case of hyper-g, since the treatment of g, which is responsible for the shrinkage of covariates from the data when the model is trained, requires a smart Metropolis-Hastings sampling in order to complete the procedures. In addition, the performance of MCMC and formal methods of full enumeration is assessed on both the simulated and the real data-set which are compared to the Bayesian adaptive sampling of the R programming language and GVS implemented in WINBUGS Appendix section A.9 and A.10. Regarding the simulation, the results were characterized by similarity and showed that the mixtures of g-priors work well in practice towards the goal of identifying the correct model. Similarly, regarding the real data-set, and whether the method applied is Zellner-Siow or the hyper-g-prior the results led to models of different dimensionality. Specifically, the use of the hyper-g-prior produces more complex models in comparison to the Zellner-Siow, as was expected according to the bibliography. Concluding, the behavior of the fixed g-priors is sparser due to the presence of the moderate sample size. Our preference to assess the performance of Bayesian variable selection methods of full enumeration and MCMC on small model space is justified as an attractive case for comparisons.

Chapter 3

Bayesian Variable Selection in Generalized Linear Models

The variable selection problem has been well recognised in daily practice for its intensive theoretical and computational challenge, whereas it still remains an open "quest" for extensive research nowadays due to issues regarding a) prior elicitation and b) the exact computation of all posterior model probabilities. The need to move beyond Gaussian responses occupied the attention of many researchers who considered a broader family of probabilistic structures, namely *generalized linear models* (GLMs) McCullagh and Nelder (1989). The Bayesian approach to this subject is quite challenging and well documented, Dey et al. (1999). Despite the popularity of generalized linear models, Bayesian inference in this domain of statistical modelling is always a difficult task due to the inconvenient form of the likelihood, which means that analytical approximations Tierney et al. (1989) or MCMC methods Gilks et al. (1996) are needed. Under this setup, variable selection applications emerge especially in classification problems where a moderate number of covariates enters in disposal and the challenge is to find only a small subset of the initial set that will "truely" affect the binary response variable.

From an objective Bayesian point of view, a probabilistic procedure is essential for the variable selection problem when information regarding the inclusion or exclusion of covariates is scarce or not available, whereas if a subjective approach is adopted, it is immediately rejected due to the impracticability of considering prior elicitations that control the prior structure of all 2^p subsets of variables.

Hence, researchers resort to elicitations of objective Bayesian methods introduced by Jeffreys in order to express in an automatic manner all possible prior features within model and model specific parameters across the possible 2^p subsets of covariates.

On the other hand, default objective priors based on improper priors are avoided in model determination procedure due to the known limitations related to indetermined constants in posterior measures. In those measures semi-automatic priors based on Zellner's g-prior Zellner (1986) design coupled with mixtures of g Liang et al. (2008) are preferred ensuring flexible performance across the model space. Although, Zellner's g-prior brought enormous success in Bayesian variable selection for linear models, limited versions were established for generalized linear model settings due to the fact that the expected Fisher information matrix depends on the regression coefficients. Thus this was one of the main reasons that inspired practitioners to work hard over this problem in order to enrich the research bibliography.

Chen and Ibrahim (2003) introduced for the first time conjugate priors for generalized linear models based essentially on historical data (equivalent on the notion of imaginary data sample) related with a scalar precision parameter which is similar to g-prior sketch.

However, this prior was difficult to handle due to the intractable form of generalized linear models apart from normal regression models. Therefore, researchers proposed MCMC methods to deal with this issue. In particular, they proposed a unit information g-prior based on Kass and Wasserman (1995) for variable selection and link identification in logistic regression models through reversible-jump MCMC. Bové and Held (2011) considered the asymptotic posterior distribution based on the original prior construct of Chen and Ibrahim (2003), which coincides with the same g-prior introduced by Ntzoufras et al. (2003). Their proposed method consisted of an integrated Laplace approximation, based on Gauss-Hermite approximation after a log-transformation of the initial g parameter, which allows the implementation of full enumeration or MCMC in variable selection for small or moderate model spaces. Alternative prior formulations for generalized linear models coupled with g-prior mixtures, were established also under the empirical Bayes fashion as it was expected alike linear models, through the observed or expected information matrix evaluated at the maximum likelihood estimates described in the research works of Hansen and Yu (2003), Wang and George (2007) and Li and Clyde (2018). A computational advantage of empirical Bayesian methods is that the integrated Laplace approximation provides closed form expressions as functions of the maximum likelihood estimators amenable for model selection like exact functions. While Hansen and Yu (2003) evaluated Fisher's information matrix at maximum likelihood estimate likewise Ntzoufras et al. (2003) only for canonical link functions, Wang and George (2007) used maximum likelihood estimates for the evaluation of the observed information matrix instead of the expected Fisher's information matrix. On the other hand, Gupta and Ibrahim (2009) preferred to avoid the choice of maximum likelihood estimates and kept the expected Fisher information matrix depending on model parameters, leading to undetermined Bayes factors.

Recently, Fouskakis et al. (2018) introduced the power expected posterior prior methodology coupled with mixtures of g-priors in the GLMs settings.

However, in this chapter, we focus on the main aspects of generalized linear models emphasized in variable selection regarding the objective Bayesian methodology of prior and model choice based on the seminal works of Bové and Held (2011) and Li and Clyde (2013) as the main ingredients in order to illustrate step by step the evolution of g-prior together with their extension for mixtures of g-priors Liang et al. (2008). Both incorporate the innovative idea of mixtures of g-priors in generalized liner models framework which was absent for many years due to the difficulties stated above resulting in a big gap from the original work of Liang et al. (2008), where they provided extensions to a more general class of hyperpriors, namely *incomplete gamma* and *compound confluent hypergeometric distribution.* To conclude, the approach of (Bové and Held (2011)) will serve as basis later for the introduction of last chapter, and more precisely of Bayesian variable selection in multinomial logistic regression with MCMC methods.

3.1 The problem of Bayesian variable selection in Generalized Linear Models

Variable selection problem in linear regression models is very challenging and promises very interesting applications nowadays. However, there are many instances where both the assumptions of linearity and normality are violated for specific data, especially when the support of the response variable is restricted to $\mathbb{R}^+$ or $\mathbb{N}$, thus GLMs are often required in such situations; see McCullagh and Nelder (1989) and Marin and Robert (2014). The research bibliography of Bayesian variable selection in GLMs is vast, with the most distinguished approaches of Hansen and Yu (2003), Chen and Ibrahim (2003), Ntzoufras et al. (2003), Wang and George (2007), Chen et al. (2008), Bové and Held (2011), Li and Clyde (2013) and Li and Clyde (2018). Usually, a GLM's set-up involves specifying for the observed values $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ of random variable $\boldsymbol{Y}$ a sampling density with probabilistic nature as

$$f(\boldsymbol{y}|a, \boldsymbol{\beta}, \boldsymbol{\phi}) = \exp\left(\boldsymbol{y}^T \boldsymbol{\Phi}^{-1} \boldsymbol{\psi} - b^T(\boldsymbol{\psi})\mathbf{1}_n + c^T(\boldsymbol{y}, \boldsymbol{\phi})\mathbf{1}_n\right), \tag{3.1}$$

where $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_n)^T$, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)^T$ are unknown parameters that may depend on p observed independent variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$ (equivalently on a, $\boldsymbol{\beta}$) representing location and scale parameters respectively, $\Phi = \text{diag}(\boldsymbol{\phi})$ the dispersion scale matrix and together with the functions $b(.)$, $c(.)$, a separate distribution of the form (3.1) is characterized each time sharing the features of exponential family.

More precisely, a GLM is usually equipped with a deterministic function $\boldsymbol{\eta}(.)$, called linear predictor which encapsulates the covariates as $\boldsymbol{\eta}(a, \boldsymbol{\beta}) = a\mathbf{1}_n + \boldsymbol{X}\boldsymbol{\beta}$ (possibly through the parameters a, $\boldsymbol{\beta}$), which is often expressed as function of the mean response $g(E(\boldsymbol{Y}))$ via a link function $g(.)$. The link function is of major importance since it has the ability to induce equality between the natural parameter $\boldsymbol{\psi}$ and linear predictor $\boldsymbol{\eta}$ known as canonical link, denoted in mathematical terms $\boldsymbol{\psi} = a + \boldsymbol{X}\boldsymbol{\beta}$; see Bové and Held (2011). Although the dispersion parameter $\boldsymbol{\phi}$ is unknown only for few cases incorporating weights $\widetilde{\boldsymbol{w}} = (\widetilde{w}_1, \ldots, \widetilde{w}_n)^T$ through $\phi_i = \frac{\phi}{\widetilde{w}_i}$, where ϕ is a scalar parameter denoting the dispersion, will be assumed known through the work of this thesis and hence it can be omitted. Moreover, the variance of the i-th response variable $Var(Y_i) = b^{''}\left((b^{'})^{-1}(E(Y_i))\right)\frac{\phi_i}{w_i}$ results from expressing the natural parameter $\boldsymbol{\psi}$ as function of the mean response $E(\boldsymbol{Y})$; see Bové and Held (2011).

In addition, variable selection problem in GLMs can be stated as follows: let $\boldsymbol{\gamma}$ be the binary latent vector representing all the 2^p possible subsets of independent variables and assume the $\boldsymbol{\gamma}$-th subset be of size $p_{\boldsymbol{\gamma}}$, in this way candidate models $\boldsymbol{\gamma}$ are entering in competition where we consider to select only the "best" model of the form

$$g(\mathbb{E}_{\gamma}(\boldsymbol{Y})) = a\mathbf{1}_n + \boldsymbol{X}_{\gamma}\boldsymbol{\beta}_{\gamma}, \tag{3.2}$$

where the above shows that there is not only uncertainty related to model parameters a, $\boldsymbol{\beta}_{\gamma}$ but also uncertainty lying from the choice of each respective subset of each respective model $\boldsymbol{\gamma}$.

3.1.1 Prior Elicitation

One of the most hard topics that caught the attention of the Bayesian community for many years was that of prior elicitation in variable selection which is crucial for the computation of posterior model probabilities and the enumeration of model space. More precisely, the same situation is more challenging in the GLMs framework where a researcher's design has not only to face the prior features of all possible 2^p subsets, but also the intractable product of prior with likelihood in the resulting posterior. In the previous case, we shall remark that even standard conjugate priors are insufficient to overcome the problem, thus a researcher must select with care convenient priors that allow flexible variable selection methods. As pointed in the previous chapter of linear regression, the same problem of eliciting all prior structures across all 2^p elements of the model space results also impractical in GLMs settings even for with expert's opinion, whereas an objective point of view is often desired on behalf of the fact that a researcher's plan will always be rarely confident about the choice of the best subset, hence objective stance is strongly suggested to elicit manually the whole prior dependences. To conclude, we follow again the fundamentals of prior elicitation based on Jeffreys (1961) and Zellner's work Zellner (1986) in details in the next subsection and we present the main ideas regarding prior formulation in GLMs framework.

3.1.2 Prior Choice

A common way to initialize a Bayesian method, is the prior specification which plays a decisive role in between steps of the calculation of posterior model probabilities based on Bayes factors and marginal likelihoods. When model selection is of interest, posterior model probabilities are usually computed in order to find the maximum a posteriori model as unique representative, where the Bayes theorem is an immediate consequence of Jeffreys' ideas through the comparison of each model's $\boldsymbol{\gamma}$ Bayes factor versus the null model $\boldsymbol{\gamma}_0$. Moreover, posterior model probabilities and hence Bayes factors usually depend on the marginal distribution of data $\boldsymbol{y}$ given model $\boldsymbol{\gamma}$ accompanied by the prior model probabilities $\pi(\boldsymbol{\gamma})$

$$m(\boldsymbol{y}|\boldsymbol{\gamma}) = \int\limits_{a}\int\limits_{\boldsymbol{\beta}_\gamma} f(\boldsymbol{y}|a, \boldsymbol{\beta}_\gamma, \boldsymbol{\phi}, \boldsymbol{\gamma})\pi(a, \boldsymbol{\beta}_\gamma|\boldsymbol{\gamma})d\boldsymbol{\beta}_\gamma da, \tag{3.3}$$

which is a key quantity in order to proceed to model selection . However, the prior choice of $\pi(a, \boldsymbol{\beta}_\gamma|\boldsymbol{\gamma})$ is always a delicate issue which limits the broad use in GLMs framework due to the analytic intractability of marginal likelihood $m(\boldsymbol{y}|\boldsymbol{\gamma})$. Furthermore, the use of conjugate priors is often prohibited in order to derive closed form expressions for marginal likelihood $m(\boldsymbol{y}|\boldsymbol{\gamma})$, which turns into a cumbersome "convenience" due to the product of the likelihood (3.1) with the prior $\pi(a, \boldsymbol{\beta}_\gamma|\boldsymbol{\gamma})$. Thus the use of Laplace Tierney et al. (1989) or MCMC approximations Gilks et al. (1996) are usually suggested even in this case. On the other hand, despite the choice of $\pi(a, \boldsymbol{\beta}_\gamma|\boldsymbol{\gamma})$ is always a computational burden to obtain exact inference, at least in can embody special properties of Jeffreys' approach in order to provide a consistent method in the model space. To conclude, in the next section we assume that the interested reader is familiar with Jeffreys' work and the notion of the base model which is outside the scope of the present chapter. We will discuss the form of this prior and review the main prior specification based on the semi-automatic design of Zellner (1986) adopted by Liang et al. (2008).

3.1.2.1 Default Prior Choice

Prior choice has been problematical for statisticians for many years and is still a hot topic in scientific research, especially in variable selection. More precisely, many research publications were motivated from the desire to create automatic Bayesian procedures for eliciting manually all 2^p subsets when no or little information is available for the best subset in consideration, while the intractability of marginal likelihood inspired statisticians to work intensively based on Laplace Tierney et al. (1989) or MCMC approximations Gilks et al. (1996). In this section, we present and review a detailed prior specification based on Jeffreys (1961) surrounding the research establishment of Zellner (1986) g-prior in the GLM framework. The Zellner's g-prior consists one of the standard objective choices that was initially introduced for variable selection in linear regression models with the seminal paper of Liang et al. (2008) which gained too much recognition extending the ideas of Zellner through the centering step of design matrix. It's utility is based on the semi-automatic specification only of g parameter multiplied by the structure of variance covariance matrix of maximum likelihood estimator which represents expected Fisher information matrix maintaining the main bridge in the unification of frequentist and Bayesian approach. Furthermore, the expected Fisher information matrix is free from any model specific parameters without requiring additional specifications.

In addition, similar extensions of g-prior for the family of GLMs reveal to be a harder task in the sense that their Fisher information matrices depend on the model parameters which was further investigated from many authors stated in the introduction, from which we restrict only to Bové and Held (2011) and Li and Clyde (2013). Therefore, under the guidelines of Zellner (1986) in the GLMs framework, Liang's g-prior takes the form

$$\pi(a, \boldsymbol{\beta}_{\gamma} | \boldsymbol{\phi}, \boldsymbol{\gamma}) = \pi(a|\boldsymbol{\gamma})\pi(\boldsymbol{\beta}_{\gamma}|g, \boldsymbol{\gamma}), \tag{3.4}$$

$$\boldsymbol{\beta}_{\gamma}|g, \boldsymbol{\phi}, \boldsymbol{\gamma} \sim N_{p_{\gamma}}\left(\mathbf{0}_{p_{\gamma}}, g(\boldsymbol{X}_{\gamma}^{\boldsymbol{T}}\boldsymbol{H}(\boldsymbol{\psi}_{\gamma})\boldsymbol{X}_{\gamma})^{-1}\right), \tag{3.5}$$

where we clarify that the form of improper prior $\pi(a|\boldsymbol{\gamma})$ depends exclusively on the Bayesian approach that is adopted and the above g-prior (3.5) differs from Liang's

g-prior only by the matrix $\boldsymbol{H}(.)$, which is a function of the natural parameter $\boldsymbol{\psi}_\gamma$ depending on model $\boldsymbol{\gamma}$. The prior variance-covariance matrix of (3.5) originates from the fundamentals of g-prior related to maximum likelihood theory of the sampling density of GLM (3.1), where the score function is obtained by differentiating the log-likelihood function with respect to the natural parameter $\boldsymbol{\psi}_\gamma$ and hence the parameters a, $\boldsymbol{\beta}_\gamma$ as the following

$$\frac{\partial \log\left(f(\boldsymbol{Y}|a,\boldsymbol{\beta}_\gamma,\boldsymbol{\phi},\boldsymbol{\gamma}\right)}{\partial\boldsymbol{\psi}_\gamma} = \left(\begin{array}{c} \frac{\partial\log\left((f(\boldsymbol{Y}|a,\boldsymbol{\beta}_\gamma,\boldsymbol{\phi},\boldsymbol{\gamma})\right)}{\partial a} \\ \frac{\partial^2\log\left(f(\boldsymbol{Y}|a,\boldsymbol{\beta}_\gamma,\boldsymbol{\phi},\boldsymbol{\gamma})\right)}{\partial\boldsymbol{\beta}_\gamma} \end{array}\right),$$

which is reduced after some elementary algebraic step to the following

$$\frac{\partial \log\left(f(\boldsymbol{Y}|a,\boldsymbol{\beta}_\gamma,\boldsymbol{\phi},\boldsymbol{\gamma}\right)}{\partial\boldsymbol{\psi}_\gamma} = \left(\begin{array}{c} \mathbf{1}_n^T\boldsymbol{H}(\boldsymbol{\psi}_\gamma)z_\gamma \\ \boldsymbol{X}_\gamma^{\boldsymbol{T}}\boldsymbol{H}(\boldsymbol{\psi}_\gamma)z_\gamma \end{array}\right),$$

where $\boldsymbol{z}_\gamma = (z_{1\gamma},\ldots,z_{n\gamma})^T$ are the working responses with $z_{i\gamma} = Y_i - \mathbb{E}_\gamma(Y_i)$ and $\boldsymbol{H}(\boldsymbol{\psi}_\gamma) = \text{diag}(h_{1\gamma},\ldots,h_{n\gamma})$ with $h_{i\gamma} = \frac{1}{Var_\gamma(Y_i)g'(E_\gamma(Y_i))^2}$. The next step is to obtain an expression of expected Fisher's information matrix given by

$$\mathcal{I}(\boldsymbol{\psi}_\gamma) = -\mathbb{E}_{\boldsymbol{Y}|a,\boldsymbol{\beta}_\gamma}\left(\begin{array}{cc} \frac{\partial^2\log\left((f(\boldsymbol{Y}|a,\boldsymbol{\beta}_\gamma,\boldsymbol{\phi},\boldsymbol{\gamma})\right)}{\partial a^2} & \frac{\partial^2\log\left(f(\boldsymbol{Y}|a,\boldsymbol{\beta}_\gamma,\boldsymbol{\phi},\boldsymbol{\gamma})\right)}{\partial a\partial\boldsymbol{\beta}_\gamma} \\ \frac{\partial^2\log\left(f(\boldsymbol{Y}|a,\boldsymbol{\beta}_\gamma,\boldsymbol{\phi},\boldsymbol{\gamma})\right)^T}{\partial\boldsymbol{\beta}_\gamma\partial a} & \frac{\partial^2\log\left(f(\boldsymbol{Y}|a,\boldsymbol{\beta}_\gamma,\boldsymbol{\phi},\boldsymbol{\gamma})\right)}{\partial\boldsymbol{\beta}_\gamma^{\boldsymbol{2}}} \end{array}\right),$$

which is reduced after some mathematical steps

$$\mathcal{I}(\boldsymbol{\psi}_\gamma) = \left(\begin{array}{cc} \mathbf{1}_n^T\boldsymbol{H}(\boldsymbol{\psi}_\gamma)\mathbf{1}_n & \mathbf{1}_n^T\boldsymbol{H}(\boldsymbol{\psi}_\gamma)\boldsymbol{X}_\gamma \\ \boldsymbol{X}_\gamma^{\boldsymbol{T}}\boldsymbol{H}(\boldsymbol{\psi}_\gamma)\mathbf{1}_n^T & \boldsymbol{X}_\gamma^{\boldsymbol{T}}\boldsymbol{H}(\boldsymbol{\psi}_\gamma)\boldsymbol{X}_\gamma \end{array}\right) = \left(\begin{array}{cc} \mathcal{I}(a) & \mathcal{I}(a,\boldsymbol{\beta}_\gamma) \\ \mathcal{I}(\boldsymbol{\beta}_\gamma,a)^T & \mathcal{I}(\boldsymbol{\beta}_\gamma) \end{array}\right).$$

As we can see the above Fisher information matrix $\mathcal{I}(\boldsymbol{\psi}_\gamma)$ depends on model parameters a, $\boldsymbol{\beta}_\gamma$ and the non diagonal elements make the whole prior structure very complicated implying additional correlation. Note also that the block diagonal element $\mathcal{I}(\boldsymbol{\beta}_\gamma)$ is used in the scale part of generalized g-prior (3.5). A strong requirement for the employment of generalized g-prior (3.4) rests firmly on a special centring step for matrix $\boldsymbol{H}(\boldsymbol{\psi}_\gamma)$ which is satisfied only in the case of block diagonality for the Fisher information matrix, otherwise the prior specification (3.4) is not valid for the independence among a and $\boldsymbol{\beta}_\gamma$ due to the presence of $\boldsymbol{H}(\boldsymbol{\psi}_\gamma)$. Under canonical link representation, the linear predictor is incorporated in the matrix $\boldsymbol{H}(\boldsymbol{\eta}_\gamma(a,\boldsymbol{\beta}))$ with elements $\frac{w_i b''(\eta_{i\gamma}(a,\boldsymbol{\beta}))}{\phi}$ and specific strategies of handling this matrix are reviewed and discussed in the next sections based on the approach of Bové and Held (2011) and Li and Clyde (2013). To conclude, the approach of Bové and Held (2011) is adopted as the main prior specification throughout this thesis in order to examine out Bayesian variable selection in GLMs settings and evaluate the performance of MCMC model search algorithms in the second part of this chapter.

3.1.2.2 Prior and Model Choice of Bove and Held Approach

The problem of variable selection in GLMs methodology was considered a special matter for many years mainly for i) the intractability of marginal likelihood, ii) the prior specification of the 2^p subsets and iii) the dependence of the expected Fisher information matrix on model specific parameters and still remains one of the most challenging topics for the Bayesian community, which tries to propose quite promising solutions based on the objectivity of g-priors. More precisely, the approach of Bové and Held (2011) was the first research attempt which extended the ideas of g-priors and consequently the mixtures of g-priors for variable selection under the GLM methodology that consisted the main inspiration of a similar approach of Li and Clyde (2013). In this section, we review and describe the major aspects of the prior specification and model choice as presented by the authors in their paper. This approach consists of adopting the authentic interpretation of g-prior as introduced initially by Zellner (1986) and then by Chen and Ibrahim (2003) and Chen et al. (2008) through the use of imaginary sample size for the construction of prior specification, whereas for model selection a Laplace approximation is provided simplifying the computational steps of marginal likelihood. It is worth telling that the actual predecessor of this approach was that of Ntzoufras et al. (2003) who proposed to set up the model parameters equal to the prior mean. To begin with, let an imaginary sample $\boldsymbol{y_0} = g^{-1}(a)\mathbf{1}_{\boldsymbol{n^*}}$ of size n^* for fixed values of the intercept a (assuming that the columns of design matrix $\boldsymbol{X}_\gamma$ have been centred) and for simplicity $w_i = 1$, if an improper joint prior $\pi(a, \boldsymbol{\beta}_\gamma) \propto 1$ is adopted for a GLM (3.1) scaled by $g\phi$, the posterior distribution of $\boldsymbol{\beta}_\gamma$ given the imaginary data $\boldsymbol{y_0}$ is proportional to

$$\pi(\boldsymbol{\beta}_\gamma|\boldsymbol{y}_0, a, g, \boldsymbol{\phi}, \boldsymbol{\gamma}) \propto \exp\left(\boldsymbol{y}_0^T \boldsymbol{\Phi}^{-1} \boldsymbol{\eta}_\gamma(a, \boldsymbol{\beta}_\gamma) - b^T(\boldsymbol{\eta}_\gamma(a, \boldsymbol{\beta}_\gamma))\mathbf{1}_n\right),$$

where the above distribution as long as $n \to +\infty$, converges to a multivariate normal distribution

$$\boldsymbol{\beta}_\gamma|\boldsymbol{y}_0, a, g, \boldsymbol{\phi}, \boldsymbol{\gamma} \sim N_{p_\gamma}(\widehat{\boldsymbol{\mu}}_\gamma(a), g(\boldsymbol{X}_\gamma^{\boldsymbol{T}} \boldsymbol{H}(\boldsymbol{\eta}_\gamma(a, \mathbf{0}_{p_\gamma}))\boldsymbol{X}_\gamma)^{-1}),$$

where $\widehat{\boldsymbol{\mu}}_\gamma(a)$ denotes the posterior mode which results as function of the intercept a after evaluating it to the imaginary sample size $\boldsymbol{y_0}$ and the diagonal matrix $\boldsymbol{H}(\boldsymbol{\eta}_\gamma(a, \mathbf{0}_{p_\gamma}))$ has elements $\frac{b''(\eta_{i\gamma}(a))}{\phi}$. This posterior distribution can be reduced further as the following

$$\boldsymbol{\beta}_\gamma|\boldsymbol{y}_0, a, g, \boldsymbol{\phi}, \delta, \boldsymbol{\gamma} \sim N_{p_\gamma}(\widehat{\boldsymbol{\mu}}_\gamma(a), g\mathcal{I}^{(BH)}(\widehat{\boldsymbol{\mu}}_\gamma(a))^{-1}), \tag{3.6}$$

where $\delta = \frac{1}{b''(a)}$, $\mathcal{I}^{(BH)}(\widehat{\boldsymbol{\mu}}_\gamma(a)) = \frac{\boldsymbol{X}_\gamma^{\boldsymbol{T}}\boldsymbol{X}_\gamma}{\phi\delta}$ and note that the superscript $\mathcal{I}^{(\cdot)}(\widehat{\boldsymbol{\mu}}_\gamma(a))$ will refer to Bové and Held (2011) expected Fisher information matrix. The latter asymptotic result was proposed by Bernardo (1979) and then used by Chen and Ibrahim (2003) and Chen et al. (2008) in their attempt to develop conjugate prior distributions in GLMs, whereas the posterior distribution (3.6) is a consequent step of the previous approaches described in Bové and Held (2011) and its expected Fisher information matrix is evaluated at the mode $\widehat{\boldsymbol{\mu}}_\gamma(a)$ for fixed values of the intercept a as shown in Chen and Ibrahim (2003) and Chen et al. (2008). To end this, we remark an important issue that lies within generalized g-prior of Bové and Held (2011), that is the constant's δ dependence on the fixed values of a, which remains undefined and for this reason Held et al. (2015) suggested to set the value of a equal to zero or to the maximum likelihood of the null or full model to avoid indeterminacies. Both choices work well in practice, but in this thesis we set up the intercept a equal to zero instead of the maximum likelihood estimator in order to vanish any trace of the intercept that may cause correlation in the resulting g-prior, which will reduce the posterior mode to $\widehat{\boldsymbol{\mu}}_\gamma(a) = \mathbf{0}_{p_\gamma}$ and consequently the posterior distribution (3.6) as the following

$$\boldsymbol{\beta}_\gamma|g, \boldsymbol{\phi}, \delta, \boldsymbol{\gamma} \sim N_{p_\gamma}(\mathbf{0}_{p_\gamma}, g\mathcal{I}^{(BH)}(\mathbf{0}_{p_\gamma})^{-1}), \tag{3.7}$$

which can be viewed as the original generalized g-prior of Bové and Held (2011) evaluated at the mode $\mathbf{0}_{p_\gamma}$. In addition, an equivalent way to relate the previous result (3.7) with the joint expected Fisher information matrix of Bové and Held (2011) is based on the approach of Ntzoufras et al. (2003) that proposed to set the regression coefficients equal to their prior means resulting in a block diagonal matrix of the following form

$$\mathcal{I}^{(BH)}(\boldsymbol{\psi}_\gamma) = \begin{pmatrix} \frac{\mathbf{1}_n^T\mathbf{1}_n}{\phi\delta} & \mathbf{0}_{p_\gamma} \\ \mathbf{0}_{p_\gamma}^T & \frac{\boldsymbol{X}_\gamma^{\boldsymbol{T}}\boldsymbol{X}_\gamma}{\phi\delta} \end{pmatrix} = \begin{pmatrix} \mathcal{I}^{(BH)}(a) & \mathcal{I}^{(BH)}(a, \mathbf{0}_{p_\gamma}) \\ \mathcal{I}^{(BH)}(\mathbf{0}_{p_\gamma}, a)^T & \mathcal{I}^{(BH)}(\mathbf{0}_{p_\gamma}) \end{pmatrix},$$

where the block diagonal element $\mathcal{I}^{(BH)}(\mathbf{0}_{p_\gamma})^{-1}$ is used in the g-prior scale (3.7) and the above structure of the expected Fisher's information matrix induces orthogonality between model specific parameters because of the centring step of the design matrix, $\boldsymbol{X}_{\gamma}^{\boldsymbol{T}}\mathbf{1}_n = \mathbf{0}_{p_{\gamma}}$. In this way, the orthogonality allows to consider independence of parameters $a, \boldsymbol{\beta}_{\gamma}$ and adopting g-prior proposed by Zellner (1986), the generalized g-prior formulation of Bové and Held (2011) is expressed as

$$\pi^{(BH)}(a, \boldsymbol{\beta}_{\gamma}|\boldsymbol{\phi}, \delta, \boldsymbol{\gamma}) = \pi^{(BH)}(a|\boldsymbol{\gamma})\pi^{(BH)}(\boldsymbol{\beta}_{\gamma}|g, \boldsymbol{\phi}, \delta, \boldsymbol{\gamma}), \tag{3.8}$$

where in the above equation we define

$$\pi^{(BH)}(\alpha|\boldsymbol{\gamma}) \propto 1, \tag{3.9}$$

and $\pi^{(BH)}(\boldsymbol{\beta}_{\gamma}|g, \boldsymbol{\phi}, \delta, \boldsymbol{\gamma})$ is the generalized g-prior (3.7) and.
On the other hand, if we might be interested in model selection, the marginal likelihood computation is essential also for GLMs in Bayesian variable selection using the generalized g-prior (3.7) of Bové and Held (2011)

$$m^{(BH)}(\boldsymbol{y}|\boldsymbol{\gamma}) = \int\limits_{a}\int\limits_{\boldsymbol{\beta}_{\gamma}} f(\boldsymbol{y}|a, \boldsymbol{\beta}_{\gamma}, \boldsymbol{\phi}, \boldsymbol{\gamma})\pi^{(BH)}(a|\boldsymbol{\gamma})\pi^{(BH)}(\boldsymbol{\beta}_{\gamma}|g, \boldsymbol{\phi}, \delta, \boldsymbol{\gamma})d\boldsymbol{\beta}_{\gamma}da, \tag{3.10}$$

where the above quantity has no closed form and a numerical approximation based on the Laplace method is presented in the work Bové and Held (2011).
This numerical approximation consists of expanding the unnormalized log-posterior of $\pi^{BH}(\boldsymbol{\beta}_{\gamma+1}|\boldsymbol{y}, g, \boldsymbol{\phi}, \delta, \boldsymbol{\gamma})$ around it's posterior mode $\widehat{\boldsymbol{\mu}}_{\gamma+1}$ with precision matrix $\widehat{\boldsymbol{R}}_{\gamma+1}$ evaluated at $\widehat{\boldsymbol{\mu}}_{\gamma+1}$ as the following

$$\begin{aligned}\widehat{m}^{(BH)}(\boldsymbol{y}|g, \boldsymbol{\gamma}) \approx & f(\boldsymbol{y}|\widehat{\boldsymbol{\mu}}_{\gamma+1}, \boldsymbol{\phi}, \boldsymbol{\gamma})(2\pi)^{\frac{p_{\gamma}+1}{2}}\det(\widehat{\boldsymbol{R}}_{\gamma+1})^{-\frac{1}{2}}(2\pi g\delta\phi)^{-\frac{p_{\gamma}}{2}} \\ & \det(\boldsymbol{X}_{\gamma}^{\boldsymbol{T}}\boldsymbol{X}_{\gamma})^{\frac{1}{2}}\exp\left\{-\frac{1}{2g\phi\delta}\widehat{\boldsymbol{\mu}}_{\gamma}^{T}\boldsymbol{X}_{\gamma}^{\boldsymbol{T}}\boldsymbol{X}_{\gamma}\widehat{\boldsymbol{\mu}}_{\gamma}^{T}\right\},\end{aligned} \tag{3.11}$$

where the above calculation steps are simplified due to the assumption of considering the joint parameter vector $\boldsymbol{\beta}_{\gamma+1} = (a, \boldsymbol{\beta}_{\gamma}^{T})^{T}$ in order to express prior (3.8) as a normal kernel of the form $\pi^{(BH)}(\boldsymbol{\beta}_{\gamma+1}|g, \boldsymbol{\phi}, \delta, \boldsymbol{\gamma}) \propto \exp\left\{-\frac{1}{2}\boldsymbol{\beta}_{\gamma+1}^{T}\widetilde{\boldsymbol{R}}_{\gamma+1}\boldsymbol{\beta}_{\gamma+1}\right\}$ with singular precision matrix $\widetilde{\boldsymbol{R}}_{\gamma+1} = \operatorname{diag}\left(0, \frac{1}{g\phi\delta}\boldsymbol{\beta}_{\gamma}^{T}\boldsymbol{X}_{\gamma}^{T}\boldsymbol{X}_{\gamma}\boldsymbol{\beta}_{\gamma}\right)$; see for more details Bové and Held (2011) and Appendix section B.1. In addition, when model selection is of interest, usually a model comparison is made across each 2^p subsets versus a base model, typically the null model $\boldsymbol{\gamma}_0$. After adopting the null model $\boldsymbol{\gamma}_0$ as a reference, computation of posterior model probabilities involves the Bayes factor $BF_{[\gamma:\gamma_0]}$ pairwise comparisons of each model $\boldsymbol{\gamma}$ versus the null model $\boldsymbol{\gamma}_0$, where the latter is of major importance since it indicates the clues provided by the data of supporting or rejecting a model. The approximated Bayes factor based on (3.11) is expressed as function of g hyper-parameter permitting the extension to mixtures of g-priors even in GLMs framework for different choices of hyper-prior $\pi(g)$ as the following

$$\begin{aligned}\widehat{BF}^{(BH)}_{[\gamma:\gamma_0]} \approx & \frac{f(\boldsymbol{y}|\widehat{\boldsymbol{\mu}}_{\gamma+1},\boldsymbol{\phi},\boldsymbol{\gamma})}{f(\boldsymbol{y}|\widehat{\mu}_1,\boldsymbol{\phi},\boldsymbol{\gamma}_0)}(2\pi g\delta\phi)^{-\frac{p_\gamma}{2}}\det(\boldsymbol{X}^{\boldsymbol{T}}_\gamma\boldsymbol{X}_\gamma)^{\frac{1}{2}}\left(\frac{1}{\widehat{\sigma}_1}\right)^{-\frac{1}{2}} \\ & \int_0^\infty \det(\widehat{\boldsymbol{R}}_{\gamma+1})^{-\frac{1}{2}}\exp\left\{-\frac{1}{2g\phi\delta}\widehat{\boldsymbol{\mu}}^T_\gamma\boldsymbol{X}^{\boldsymbol{T}}_\gamma\boldsymbol{X}_\gamma\widehat{\boldsymbol{\mu}}^T_\gamma\right\}\pi(g)dg, \qquad (3.12)\end{aligned}$$

where the log-posterior of $\pi^{BH}(\boldsymbol{\beta}_{\gamma+1}|\boldsymbol{y},g,\boldsymbol{\phi})$ under the null model $\boldsymbol{\gamma}_0$ reduces to log-likelihood of $f(\boldsymbol{y}|\widehat{\mu}_1,\boldsymbol{\phi},\boldsymbol{\gamma}_0)$ evaluated at the maximum likelihood estimate $\widehat{\mu}_1$ since g-prior vanishes under the null model $\boldsymbol{\gamma}_0$ with $\widehat{\sigma}_1$ denoting the standard error of the respective log-likelihood. Moreover, the one dimensional integral involved in Bayes factors (3.12) when mixtures of g-priors are employed has no closed form since g is contained in the determinant $\det(\widehat{\boldsymbol{R}}_{\gamma+1})$ and is carried out in log-scale using Gauss-Hermite quadrature after applying the transformation $z = \log(g)$; see Salzer and Zucker (1952), Naylor and Smith (19), and Bové and Held (2011); for more details see Appendix section B.2. The authors suggested to initially approximate the posterior moments of z, mainly, the mode $\widehat{z}$ and the variance $\widehat{\sigma}_z$ evaluated at the mode $\widehat{z}$, which is derived using the unnormalized posterior distribution of z

$$\pi(z|\boldsymbol{y},\boldsymbol{\gamma}) \propto f(\boldsymbol{y}|z,\boldsymbol{\gamma})\pi(z)J_z,$$

where $J_.$ is the associated Jacobian of z due to the transformation on log-scale and then the Gaussian-quadrature is applied on the resulting marginal likelihood.
Consequently, the step of Gaussian quadrature for the approximation of unidimensional integral (3.12) reduces to Bayes factor

$$\begin{aligned}\widehat{BF}^{(BH)}_{[\gamma:\gamma_0]} \approx & \frac{f(\boldsymbol{y}|\widehat{\boldsymbol{\mu}}_{\gamma+1},\boldsymbol{\phi},\boldsymbol{\gamma})}{f(\boldsymbol{y}|\widehat{\mu}_1,\boldsymbol{\phi},\boldsymbol{\gamma}_0)}(2\pi g\delta\phi)^{-\frac{p_\gamma}{2}}\det(\boldsymbol{X}^{\boldsymbol{T}}_\gamma\boldsymbol{X}_\gamma)^{\frac{1}{2}}\det(\widehat{\boldsymbol{R}}_{\gamma+1})^{-\frac{1}{2}}\left(\frac{1}{\widehat{\sigma}_1}\right)^{-\frac{1}{2}} \\ & \sum_{j=1}^{N} m_j\pi(z_j|\boldsymbol{y},\boldsymbol{\gamma}), \qquad (3.13)\end{aligned}$$

where the weights $m_j = v_j \exp{(t_j^2)}\sqrt{2}\widehat{\sigma}_z$ and nodes $z_j = \widehat{z} + \sqrt{2}\widehat{\sigma}_z t_j$ depend on posterior moments $\widehat{z}$, $\widehat{\sigma}_z$ and consequently on v_j, t_j which are considered as original weights. The approximation of Gaussian quadrature is accurate if and only if the unnormalized posterior $\pi(z|\boldsymbol{y}, \boldsymbol{\gamma})$ is the product of a normal distribution $N(\widehat{z}, \widehat{\sigma}_z)$ with a polynomial of order $2N - 1$. To conclude, Bayes factor (3.12) varies based on the choice of the hyper-prior $\pi(g)$ leading to different approximated forms either Zellner and Siow (1980) either hyper-g-prior Liang et al. (2008). In their work they introduced another form of mixtures g-priors, namely incomplete gamma function which will not be further discussed in this thesis. In the next section we will describe the recent method of Li and Clyde (2013) also for GLMs with mixtures of g-priors.

3.1.2.3 Prior and Model Choice of Li and Clyde Approach

The approach of Li and Clyde (2013) is considered the last development of mixtures of g-priors for Bayesian variable selection in GLMs framework and gathered all the attention in recent years moving in similar grounds with the approach of Bové and Held (2011). This approach consists of a special centring evaluated in maximum likelihood estimators in terms of projection in order to ensure that Fisher's information matrix is block diagonal and free from any arbitrariness of regression coefficients. Then it performs accordingly a Laplace approximation for the log-likelihood retrieving approximated closed forms of Bayes factors. The special centring was initially presented for linear models in Forte (2014) and then followed its extension by Li and Clyde (2013) in the GLMs framework. Moreover, Li and Clyde (2013) introduced a more generic family of g-priors mixtures which encompasses the main characteristics of the well known hyper-g and Zellner-Siow as well as also other mixtures of g priors Bové and Held (2011) and Maruyama and George (2011).

To begin with, consider the familiar GLM framework of model $\boldsymbol{\gamma}$ with design matrix $\widetilde{\boldsymbol{X}}_\gamma$ prior to centring

$$g(\mathbb{E}_\gamma(\boldsymbol{Y})) = \widetilde{a}\mathbf{1}_n + \widetilde{\boldsymbol{X}}_\gamma \boldsymbol{\beta}_\gamma, \tag{3.14}$$

then the main intuition behind Li and Clyde's special centring approach is to use the projection matrix $\widehat{\boldsymbol{P}} = \mathbf{1}_n(\mathbf{1}_n^T \boldsymbol{H}(\boldsymbol{\eta}_\gamma(\widehat{a}, \widehat{\boldsymbol{\beta}}_\gamma))\mathbf{1}_n)^{-1}\mathbf{1}_n^T \boldsymbol{H}(\boldsymbol{\eta}_\gamma(\widehat{a}, \widehat{\boldsymbol{\beta}}_\gamma))$ evaluated at the maximum likelihood estimator $(\widehat{a}, \widehat{\boldsymbol{\beta}}_\gamma)$ that turns a GLM $\boldsymbol{\gamma}$ into (3.2) with intercept a and design matrix $\boldsymbol{X}_\gamma$. The correspondence of $\widetilde{a}$ to a and $\widetilde{\boldsymbol{X}}_\gamma$ to $\boldsymbol{X}_\gamma$ are defined respectively as $a = \widetilde{a} + \mathbf{1}_n(\mathbf{1}_n^T \boldsymbol{H}(\boldsymbol{\eta}_\gamma(\widehat{a}, \widehat{\boldsymbol{\beta}}_\gamma))\mathbf{1}_n)^{-1}$, $\boldsymbol{X}_\gamma = (\boldsymbol{I}_n - \widehat{\boldsymbol{P}})\widetilde{\boldsymbol{X}}_\gamma$. The interpretation of matrix $\widehat{\boldsymbol{P}}$ rests firmly on projecting the matrix $\boldsymbol{X}_\gamma$ on the space created by the span of $\mathbf{1}_n$ with inner product $< \boldsymbol{c}, \boldsymbol{d} >= \boldsymbol{c}^T \boldsymbol{H}(\boldsymbol{\eta}_\gamma(\widehat{a}, \widehat{\boldsymbol{\beta}}_\gamma))\boldsymbol{d}$, where $\boldsymbol{c}, \boldsymbol{d} \in \mathbb{R}^n$.

However, notice that the approach by Bové and Held (2011) corresponds to a special case of Li and Clyde (2013) with projection matrix $\boldsymbol{P}_0 = \mathbf{1}_n(\mathbf{1}_n^T \boldsymbol{H}(\boldsymbol{\eta}_\gamma(a, \mathbf{0}_p, \mathbf{1}_n)^{-1}\mathbf{1}_n^T \boldsymbol{H}(\boldsymbol{\eta}_\gamma(a, \mathbf{0}_p)) = \mathbf{1}_n(\mathbf{1}_n^T \mathbf{1}_n)^{-1}\mathbf{1}_n^T$ evaluated at prior mean $\mathbf{0}_p$ for fixed values of a. The approach of Li and Clyde (2013) reflects an alternative idea of centring which practically facilitates the whole structure of Fisher's information matrix after its evaluation to the maximum likelihood estimator

$$\begin{aligned}\mathcal{I}^{(LC)}(\widehat{\boldsymbol{\psi}}_\gamma) &= \begin{pmatrix} \mathbf{1}_n^T \boldsymbol{H}(\boldsymbol{\eta}_\gamma(\widehat{a}, \widehat{\boldsymbol{\beta}}_\gamma))\mathbf{1}_n & \mathbf{0}_p \\ \mathbf{0}_p^T & \boldsymbol{X}_\gamma^T \boldsymbol{H}(\boldsymbol{\eta}_\gamma(\widehat{a}, \widehat{\boldsymbol{\beta}}_\gamma))\boldsymbol{X}_\gamma \end{pmatrix} \\ &= \begin{pmatrix} \mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma}) & \mathcal{I}^{(LC)}(\widehat{a}, \widehat{\boldsymbol{\beta}}_\gamma) \\ \mathcal{I}^{(LC)}(\widehat{\boldsymbol{\beta}}_\gamma, \widehat{a})^T & \mathcal{I}^{(LC)}(\widehat{\boldsymbol{\beta}}_\gamma) \end{pmatrix},\end{aligned}$$

where the block diagonal part $\mathcal{I}^{(LC)}(\widehat{\boldsymbol{\beta}}_\gamma)$ will be used in generalized g-prior and its diagonal structure implies the model parameters $a, \boldsymbol{\beta}_\gamma$ to be orthogonal due to the invariance of translating the maximum likelihood estimator, otherwise the prior specification (3.4) will not be valid anymore for independence. For notational reasons, the superscript $\mathcal{I}^{(\cdot)}$ shall refer here to Li and Clyde's expected Fisher information matrix and notice that the block diagonal element $\mathcal{I}^{LC}(\widehat{a}|\boldsymbol{\gamma})$ of a depends on model $\boldsymbol{\gamma}$ because it includes the linear predictor evaluated to maximum likelihood estimator $\widehat{\boldsymbol{\beta}}_\gamma$ which varies according to each model $\boldsymbol{\gamma}$. Since, the above centring step facilitates computational issues and implies plausibly a priori independence of $a, \boldsymbol{\beta}_\gamma$ for each model $\boldsymbol{\gamma}$, then the generalized g-prior of Li and Clyde (2013) can be formulated as the following

$$\pi^{(LC)}(a, \boldsymbol{\beta}_\gamma|\boldsymbol{\phi}, \boldsymbol{\gamma}) = \pi^{(LC)}(a|\boldsymbol{\gamma})\pi^{(LC)}(\boldsymbol{\beta}_\gamma|\widehat{a}, \widehat{\boldsymbol{\beta}}_\gamma, g, \boldsymbol{\phi}, \boldsymbol{\gamma}), \tag{3.15}$$

where in the above expression we define

$$a|n, v \sim N(0, nv), \quad v > 0, \tag{3.16}$$

$$\boldsymbol{\beta}_\gamma|\widehat{a}, \widehat{\boldsymbol{\beta}}_\gamma, g, \boldsymbol{\phi}, \boldsymbol{\gamma} \sim N_{p_\gamma}(\mathbf{0}_{p_\gamma}, g\mathcal{I}^{(LC)}(\widehat{\boldsymbol{\beta}}_\gamma)^{-1}), \tag{3.17}$$

where notice the generalized g-prior (3.17) is additionally conditional on the maximum likelihood estimators $\widehat{a}, \widehat{\boldsymbol{\beta}}_\gamma$ and v is a constant set as large as possible and when $v \to \infty$ prior (3.16) degenerates to the usual improper prior $\pi^{(LC)}(a|\boldsymbol{\gamma}) \propto 1$ as suggested by Li and Clyde (2013). Hence, under Li and Clyde's prior setup (3.15), model selection initially proceeds with the calculation of marginal likelihood of a GLM $\boldsymbol{\gamma}$

$$m^{(LC)}(\boldsymbol{y}|\boldsymbol{\gamma}) = \int\limits_a \int\limits_{\boldsymbol{\beta}_\gamma} f(\boldsymbol{y}|a, \boldsymbol{\beta}_\gamma, \boldsymbol{\phi}, \boldsymbol{\gamma})\pi^{(LC)}(a|\boldsymbol{\gamma})\pi^{(LC)}(\boldsymbol{\beta}_\gamma|\widehat{a}, \widehat{\boldsymbol{\beta}}_\gamma, g, \boldsymbol{\phi}, \boldsymbol{\gamma})d\boldsymbol{\beta}_\gamma da, \tag{3.18}$$

where the above marginal likelihood is difficult to manage even if the prior was factorized with (3.15) and so a Laplace approximation is always suggested. More precisely, the authors proposed a Laplace approximation with expanding first the log-likelihood of $f(\boldsymbol{y}|a, \boldsymbol{\beta}_\gamma, \boldsymbol{\phi}, \boldsymbol{\gamma})$ around the maximum likelihood estimators $\widehat{a}$ and $\widehat{\boldsymbol{\beta}}_\gamma$

$$\widehat{m}^{(LC)}(\boldsymbol{y}|g, \boldsymbol{\gamma}) \approx f(\boldsymbol{y}|\widehat{a}, \widehat{\boldsymbol{\beta}}_\gamma, \boldsymbol{\phi}, \boldsymbol{\gamma}) \quad \int_a \exp\left\{-\frac{1}{2}\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})(a-\widehat{a})^2\right\}\pi^{(LC)}(a|\boldsymbol{\gamma})da$$
$$\int_{\boldsymbol{\beta}_\gamma} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}_\gamma - \widehat{\boldsymbol{\beta}}\right)^T \mathcal{I}^{(LC)}(\widehat{\boldsymbol{\beta}}_\gamma)\left(\boldsymbol{\beta}_\gamma - \widehat{\boldsymbol{\beta}}\right)\right\}\pi^{(LC)}(\boldsymbol{\beta}_\gamma|\widehat{a}, \widehat{\boldsymbol{\beta}}_\gamma, g, \boldsymbol{\phi}, \boldsymbol{\gamma})d\boldsymbol{\beta}_\gamma, \qquad (3.19)$$

and then the expanded log-likelihood combined with the factorized priors (3.16), (3.17) allows to provide approximated closed form expressions of marginal likelihood (3.18) as the following

$$\widehat{m}^{(LC)}(\boldsymbol{y}|g, \boldsymbol{\gamma}) \approx f(\boldsymbol{y}|\widehat{a}, \widehat{\boldsymbol{\beta}}_\gamma, \boldsymbol{\phi}, \boldsymbol{\gamma})[1 + nv\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})]^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\frac{\widehat{a}^2\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})}{1 + nv\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})}\right)\right\}$$
$$(1+g)^{-\frac{p_\gamma}{2}} \exp\left\{-\frac{Q_\gamma}{2(g+1)}\right\}, \qquad (3.20)$$

where $Q_\gamma = \widehat{\boldsymbol{\beta}}_\gamma^T \mathcal{I}^{(LC)}(\widehat{\boldsymbol{\beta}}_\gamma)\widehat{\boldsymbol{\beta}}_\gamma$ denotes a sum of squares regression analogue for a GLM; see for more Li and Clyde (2013) and Li and Clyde (2018) and Appendix section B.3. In addition, when the null model $\boldsymbol{\gamma}_0$ is adopted, model choice involves

$$\widehat{BF}^{(LC)}_{[\gamma:\gamma_0]}(g) \approx \frac{f(\boldsymbol{y}|\widehat{a}, \widehat{\boldsymbol{\beta}}_\gamma, \boldsymbol{\phi}, \boldsymbol{\gamma})}{f(\boldsymbol{y}|\widehat{a}, \boldsymbol{\phi}, \boldsymbol{\gamma}_0)}\left[\frac{1 + nv\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})}{1 + nv\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma}_0)}\right]^{-\frac{1}{2}}$$
$$\exp\left\{-\frac{1}{2}\left[\frac{\widehat{a}^2\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})}{1 + nv\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})} - \frac{\widehat{a}^2\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma}_0)}{1 + nv\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma}_0)}\right]\right\}$$
$$(1+g)^{-\frac{p_\gamma}{2}} \exp\left\{-\frac{Q_\gamma}{2(g+1)}\right\}, \qquad (3.21)$$

where the above expression is reduced due to the marginal likelihood of null model $\boldsymbol{\gamma}_0$ for which holds p_{γ_0} and $Q_{\gamma_0} = 0$. The first term in Bayes factor (3.21) consists of classical likelihood ratio test and additional penalty quantities contributed by the intercept a, whereas the second term results from generalized g-prior; see for more Li and Clyde (2013), Li and Clyde (2018). In addition, for large values v under prior ignorance, the prior (3.16) becomes improper and then the Bayes factor (3.21) is expressed

$$\widehat{BF}^{(LC)}_{[\gamma:\gamma_0]}(g) \approx \frac{f(\boldsymbol{y}|\widehat{a}, \widehat{\boldsymbol{\beta}}_\gamma, \boldsymbol{\phi}, \boldsymbol{\gamma})}{f(\boldsymbol{y}|\widehat{a}, \boldsymbol{\phi}, \boldsymbol{\gamma}_0)}\left[\frac{\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})}{\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma}_0)}\right]^{-\frac{1}{2}}(1+g)^{-\frac{p_\gamma}{2}} \exp\left\{-\frac{Q_\gamma}{2(g+1)}\right\}. \qquad (3.22)$$

Furthermore, posterior measures of model selection such as marginal likelihood and Bayes factors usually depend on the parameter g which allows to express the Bayes factor (3.22) with mixtures of g-priors after adopting a hyper-prior $\pi(g)$

$$\widehat{BF}^{(LC)}_{[\gamma:\gamma_0]} \approx \frac{f(\boldsymbol{y}|\widehat{a}, \widehat{\boldsymbol{\beta}}_{\gamma}, \boldsymbol{\phi}, \boldsymbol{\gamma})}{f(\boldsymbol{y}|\widehat{a}, \boldsymbol{\phi}, \boldsymbol{\gamma}_0)} \left[\frac{\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})}{\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma}_0)}\right]^{-\frac{1}{2}} \int_0^{\infty} (1+g)^{-\frac{p_\gamma}{2}} \exp\left\{-\frac{Q_\gamma}{2(g+1)}\right\} \pi(g) dg, \tag{3.23}$$

where each choice $\pi(g)$ implies a different g-prior mixture coupled with the integrated likelihood which enables straightforward closed or numerical expressions. Possible choices of $\pi(g)$ include the Zellner-Siow which needs an additional Laplace approximation, whereas hyper-g provide closed forms in terms of confluent hypergeometric function ${}_1F_1(.)$ based on a more generalized mixture of g-priors called confluent hypergeometric distribution; see Gordy (1998), Li and Clyde (2013) and Li and Clyde (2018). Other mixtures of g-priors Maruyama and George (2011) and Bové and Held (2011) are conjugate to the approximated Bayes factors of Li and Clyde's approach but are beyond the scope of this thesis. In this first part of this chapter we emphasize more on mixtures of g-priors obtained under Li and Clyde (2013) approach due to the attractive approximated closed forms than those obtained by Bové and Held (2011). To conclude, the first part of this chapter finishes with the next subsection and then the section on Bayesian variable selection with MCMC methods for GLMs with mixtures of g-priors begins.

3.1.2.4 Confluent Hypergeometric Function

Mixtures of g-priors are a common choice in objective variable selection for linear regression and have been well recognised for their automatic set-up when the subset of variables is unknown. Moreover, they share adaptability properties that ensure predictive optimalities and learning from the data through the shrinkage parameter $\frac{g}{g+1}$, while they surpass information and Jeffreys Lindleys paradox Liang et al. (2008) Lindley (1957) and Bartlett (1957). Even though, mixtures of g-priors are always a difficult task in GLMs settings due to inconveniences of likelihood incompatibility when coupled with the prior and the dependence on regression coefficients, Li and Clyde (2013) and Li and Clyde (2018) introduced a more general hierarchical prior mixture called *confluent hypergeometric distribution* (CH). This extension includes as special cases popular choices of g-priors mixtures like Zellner-Siow and hyper-g, whereas the same authors proved that other mixtures of g-priors including beta-prime Maruyama and George (2011) and incomplete gamma Bové and Held (2011) are comprised as special cases into a more general family of distributions called compound confluent hypergeometric distribution which will not be further discussed in this thesis. Recall that hyper-g-priors became popular with respect to Zellner-Siow mainly for it's attractive closed form expressions of posterior measures in terms of the Gausian hypergeometric function. However, when approximating marginal likelihood (3.20) or Bayes factor (3.22) with hyper-g-prior a similar family of functions with Gaussian hypergeometric functions called confluent hypergeometric functions are used to integrate out the additional uncertainty related to hyperparameter g. Consider the usual variable selection problem in GLMs framework, then the model selection under hyper$g$-prior $\pi^{(hy)}(g)$ proceeds with the calculation of Bayes factor (3.23)

$$\widehat{BF}^{(LC)}_{[\gamma:\gamma_0]} \approx \frac{f(\boldsymbol{y}|\widehat{a}, \widehat{\boldsymbol{\beta}}_\gamma, \boldsymbol{\phi}, \boldsymbol{\gamma})}{f(\boldsymbol{y}|\widehat{a}, \boldsymbol{\phi}, \boldsymbol{\gamma}_0)} \left[\frac{\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})}{\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma}_0)}\right]^{-\frac{1}{2}} \int_0^\infty (1+g)^{-\frac{p_\gamma}{2}} \exp\left\{-\frac{Q_\gamma}{2(g+1)}\right\} \pi^{hy}(g)dg, \tag{3.24}$$

where the above integral is difficult to handle and only under affine transformation of g a closed from expression is provided. In particular, the change of variable $u = \frac{1}{g+1}$, simplifies the mathematical steps and leads to the general family of confluent hypergeometric distribution introduced by Gordy (1998), a generalization of Beta distribution. In this case, let the random variable $x \sim CH(e, d, r)$, where $CH(., ., .)$ denotes the confluent hypergeometric distribution, then we say that x follows a Confluent hypergeometric distribution with probability density function of x

$$\pi^{CH}(x) = \frac{x^{e-1}(1-x)^{d-1}\exp(-rz)}{B(e,d)\,{}_1F_1(e, e+d, -r)}, \quad x \in [0,1],$$

where $e > 0$, $d > 0$, $s \in \mathbb{R}$, $B(.)$ is the Beta function and ${}_1F_1(.)$ is the confluent hypergeometric function defined as

$${}_1F_1(e, d, r) = \frac{\Gamma(d)}{\Gamma(d-e)\Gamma(e)} \int_0^1 x^{e-1}(1-x)^{e-d-1}\exp(rx)dx.$$

To end this, the Bayes factor representation (3.24) after recognising the normalising constant of posterior confluent hypergeometric distribution $CH\left(\frac{p_\gamma+\alpha}{2} - 1, \frac{p_\gamma+\alpha}{2}, \frac{Q_\gamma}{2}\right)$ in the integrand of g is computed as follows

$$\widehat{BF}^{(LC)}_{[\gamma:\gamma_0]} \approx \frac{\alpha-2}{2}\frac{f(\boldsymbol{y}|\widehat{a},\widehat{\boldsymbol{\beta}}_{\gamma},\boldsymbol{\phi},\boldsymbol{\gamma})}{f(\boldsymbol{y}|\widehat{a},\boldsymbol{\phi},\boldsymbol{\gamma}_0)}\left[\frac{\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})}{\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma}_0)}\right]^{-\frac{1}{2}} B\left(\frac{p_\gamma+\alpha}{2}-1,1\right)\, {}_1F_1\left(\frac{p_\gamma+\alpha}{2}-1,\frac{p_\gamma+\alpha}{2},\frac{Q_\gamma}{2}\right), \tag{3.25}$$

see for more details Appendix section B.4. In addition, motivated by the confluent hypergeometric distribution, Li and Clyde (2013) and Li and Clyde (2018) introduced a novel mixture of g-priors of the form

$$\pi^{LC}(g) = \frac{g^{\frac{e}{2}-1}(1+g)^{-\frac{e+d}{2}}\exp\left(\frac{r}{2}\frac{g}{g+1}\right)}{B(\frac{e}{2},\frac{d}{2})\, {}_1F_1(\frac{e}{2},\frac{e+d}{2},\frac{r}{2})},$$

whose construction is of major importance since the transformation $u=\frac{1}{g+1}$ turns the above prior into confluent hypergeometric distribution of Gordy (1998). Moreover, based on this result, Li and Clyde (2013) and Li and Clyde (2018) approximated Bayes factor (3.23)

$$\widehat{BF}^{(LC)}_{[\gamma:\gamma_0]} \approx \frac{f(\boldsymbol{y}|\widehat{a},\widehat{\boldsymbol{\beta}}_{\gamma},\boldsymbol{\phi},\boldsymbol{\gamma})}{f(\boldsymbol{y}|\widehat{a},\boldsymbol{\phi},\boldsymbol{\gamma}_0)}\left[\frac{\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})}{\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma}_0)}\right]^{-\frac{1}{2}} \int_0^\infty (1+g)^{-\frac{p_\gamma}{2}}\exp\left\{-\frac{Q_\gamma}{2(g+1)}\right\}\pi^{LC}(g)dg, \tag{3.26}$$

which is reduced after applying the transformation $u=\frac{1}{g+1}$ to the following

$$\widehat{BF}^{(LC)}_{[\gamma:\gamma_0]} \approx \frac{f(\boldsymbol{y}|\widehat{a},\widehat{\boldsymbol{\beta}}_{\gamma},\boldsymbol{\phi},\boldsymbol{\gamma})}{f(\boldsymbol{y}|\widehat{a},\boldsymbol{\phi},\boldsymbol{\gamma}_0)}\left[\frac{\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})}{\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma}_0)}\right]^{-\frac{1}{2}} \frac{B\left(\frac{e+d+p_\gamma}{2},\frac{d+p_\gamma}{2}\right)\, {}_1F_1\left(\frac{d+p_\gamma}{2},\frac{e+d+p_\gamma}{2},\frac{r+Q_\gamma}{2}\right)}{B(\frac{e}{2},\frac{d}{2})\, {}_1F_1(\frac{e}{2},\frac{e+d}{2},\frac{r}{2})}, \tag{3.27}$$

where the latter holds to due to the integrated kernel of posterior distribution $CH\left(\frac{d+p_\gamma}{2},\frac{e+d+p_\gamma}{2},\frac{r+Q_\gamma}{2}\right)$ see for more information Appendix section B.5.

3.2 Closing Remarks

In this chapter, the Bayesian variable selection problem for generalized linear models was introduced. At first, the fundamentals of objective Bayesian variable selection concerning the prior specification and model selection for Zellner's g-prior and its mixtures based on the approaches of Li and Clyde (2013) and Bové and Held (2011) were reviewed. The importance of centering for the expected Fisher information matrix in their respective analytical tools was underlined; the Laplace approximation was verified step by step and demonstrated in the respective Appendix sections B.1 and B.3 respectively.

Chapter 4

Bayesian Variable Selection in Multinomial Logistic Regression

Multinomial regression, a broader class of generalized linear models and binary regression, has recently received immense attention regarding the variable selection problem for multiclassification with many applications in household (Allenby et al., 2005) and disease classification data (Aijun and Xinyuan, 2010) and (Aijun et al., 2016), resulting in advancements, especially in marketing and biomedical sectors. Issues related to class imbalance in this domain of research area often occur. Usually in order to build a multinomial regression one has to set up identifiability constraints using as reference (baseline) one of the categories of the response variable. In this framework, the variable selection problem is regarded as simultaneous, intermediate variable selection steps across different pairwise binary regressions to account for the possible dependencies among the classes of response. The interest lies on seeking only the important covariates that vary according to each class of the nominal response variable, namely *class specific predictor selection* (Gustafson and Lefebvre, 2008). In other words, the distribution of each class is said to vary conditional on covariates. Its Bayesian variable selection version seems impractical regarding the computation of posterior model probabilities related to prior elicitation and model complexity, thus researchers search to find solutions by the use of MCMC. In the Bayesian paradigm, the degree of sparseness is affected by imposing hierarchical mixture priors among the different coefficients of each baseline binary regression forcing the redundant variables to be zero or not, conditional on features of the model space Mitchell and Beauchamp (1988), Bae and Mallick (2004), Park and Casella (2008) and Li and Lin (2010). From an objective point of view the little or no information regarding the unknown subsets across the different binary regression models suggest the use of g-priors Zellner (1986) and their mixtures Liang et al. (2008) and Li and Clyde (2013). This is very important since it will allow to create automatic Bayesian variable selection methods for complex models such as multinomial regression.

The first research bibliography establishments of multinomial regression were made upon the probit regression owing to the simple data augmentation strategy of Albert and Chib (1993). Several approaches among them are considered simply variants or extensions of other existing methods. Yau et al. (2003) proposed a semi-parametric approach for Bayesian variable selection with application in multinomial probit regression. This approach initially used a multivariate model for the unobserved latent variables to express the probability of observing the response outcomes as the normal cumulative evaluated at a smooth function of covariates without making any assumption of its functional form and then approximated this smooth function by using a linear radial basis functions Holmes and Mallick (1998) to decompose it into two components for main effects and two-way interactions. Their approach consisted of an ingenious Bayesian variable selection method that implements an MCMC method in order to estimate first the coefficients of the smooth function and then the probabilities of each response outcome belonging to a class through Bayesian model selection and model averaging procedures. However, this approach may bias the true model identification due to the combinations of parametric and non parametric methods. Later Panagiotelis and Smith (2008) considered a similar development for binary probit regression allowing more functional components (splines) of the linear basis of covariates. The only difference was that the binary indicators where also involved in the hierarchical prior specification of the coefficients of functional components similarly to George and McCulloch (1993) which focus more on sparsity rather than shrinkage as Yau et al. (2003).

Sha et al. (2004) extended a multivariate probit regression for Bayesian variable selection to account for the structural dependences of the different outcomes of multicategorical response by constructing a multivariate linear model based on the unobserved latent variables analogous to the approach of Brown and Vannucci (1998), which can be though as a generalization of SSVS George and McCulloch (1993). Their idea was motivated by David's arguments for matrix notations Dawid (1981). In order to avoid the painful computational cost encountered in the Kronecker product if vectorized forms were preferred, the authors enabled the implementation of a flexible MCMC scheme for Bayesian variable selection based on the full conditionals of latent variables and binary indicators, after the convenient marginalization of parameters of interest as nuisance. The current approach is the extension of their previous work Kyeong et al. (2003) for binary probit regression. Zhou et al. (2006) considered multinomial probit regressions for a cancer multi-classification problem. He introduced two different Bayesian variable selection methods to identify the most relevant gene expressions that vary based on the cancer type and to find out the strongest genes among all cancer types. Their approach based on the first model accounts for the heterogeneity among all different genes to explain the differences in each cancer type and responds to the trace of our work. The second model controls the homogeneity discovery of the most relevant genes that are responsible for the cancers. This is outlined respectively by using different binary indicators for each gene expression (covariate) in the first case, whereas in the second the binary vector is considered invariant according to each class of cancer. However, both models leave observed variability in each case and fit well only in practice depending on the authors' research questions. A fast updating Bayesian variable selection procedure with MCMC is outlined by integrating the parameters of interest as nuisance and performing QR decomposition to increase the computational speed, in such way that the algorithm is based only on the full conditionals of the latent variables and binary inclusion indicators. The current approach resembles to Sha et al. (2004) and represents actually the extension of Smith and Kohn (1996) in the multivariate probit regression.

The notion of class specific predictor selection using a multinomial probit regression was introduced for the first time by Gustafson and Lefebvre (2008) in order to model the association of covariates and each class with respect to the baseline using a special hierarchical prior characterized by a hyperparameter that decides the relevance or not of covariates among the different classes that is set under the non informative sense. In that way, they provided a flexible MCMC that produces estimates in terms of model selection and model averaging.

Recently, Aijun et al. (2016) proposed a sparse Bayesian variable selection approach for a multivariate probit regression using a two level hierarchical mixture prior for the regression coefficients Mitchell and Beauchamp (1988) which responds also to the trace of our work specifying the prior variances based on either the inverse-gamma Li and Lin (2010) or the gamma distribution Park and Casella (2008). Based on this prior setup, they developed a fast Bayesian variable selection method using an MCMC method by updating jointly the pairs of latent variables with the prior variances and the regression coefficients with binary inclusion vectors to reduce the strong posterior correlations. Their approach implements a fast updating scheme based on the marginalization of regression coefficients from the full conditionals of the latent variables, binary inclusion indicators and Woodburry-Sherman Morison formula, which is used to reduce computational complexity of the involved inverse. This development can be seen as the extension of the Bayesian lasso Park and Casella (2008) with varying variances and Bayesian elastic net Li and Lin (2010) for multinomial probit regression. The same approach turns out to be a successful generalization of Bae and Mallick (2004) since it focuses further on sparsity and shrinkage rather than only shrinkage, accounting for the uncertainty of the covariates through the incorporation of binary inclusion vector in the prior specification. This approach adopts features from the work of Aijun and Xinyuan (2010) where he introduced a different hierachical prior specification based only on the included regression coefficients and marginalization of the regression coefficients in order to implement efficiently the MCMC for variable selection method. Other applications of Bayesian modelling, especially for multi-classification problems, were based on support vector machines.

Chakraborty (2009) developed a Bayesian hierachical model for Bayesian variable selection based on a reproduced kernel Hilbert space where they initially introduced the latent variables to capture indirectly the true relationship among the observed outcomes and the covariates and then they approximated the unknown function of covariates with reproduction of the kernel functions especially based on Gaussian kernels. A simultaneous variable selection method based on MCMC is implemented efficiently based on the features of the underlying sampling density and hierarchical prior specification resembling to an extension of stochastic search with respect to support vector machines. This approach can be seen as the generalization of the approach of Mallick et al. (2003) as the only difference is that they didn't consider the variable selection uncertainty.

Furthermore, from the above Bayesian variable selection approaches, only Kyeong et al. (2003), Mallick et al. (2003) Zhou et al. (2006) and Aijun et al. (2016) used a fixed g-prior approach for the regression coefficients, whereas mixtures of g-priors remains an uncovered topic for the moment. More precisely the authors in the latter approach introduced a novel g-prior for high dimensional settings when $n << p$ based on the generalized inverse of Moore-Penrose. Despite the vast research bibliography of Bayesian variable selection in multinomial probit regression, multinomial regression with probit link has been criticized for the interpretability of regression coefficients, hence researchers chase alternatives in terms of multinomial logistic regression. Moreover, many variable selection methods are encountered in research bibliography with respect to applications of multi-classification based on the frequentist approach but their full presentation exceed the limits of the present work. These approaches include the weighted voting scheme (Golub et al., 1999), the threshold number of misclassification score (Ben-Dor et al., 2000), the significance analysis of micro-array statistic (Tusher et al., 2001), the mixture model algorithm (Pan, 2002), the ratio of between-groups to within-groups sum of squares (Dudoit et al., 2002), the partial least squares (Nguyen and Rocke, 2002), the Wilcoxon test statistic (Dettling, 2004) and the support vector machines of (Bradley and Mangasarian, 1998) and (Guyon et al., 1992). However, all the above methods share some important disadvantages: (a) lack of probability reasoning since the ignore model uncertainty; (b) they don't account for the multivariate correlations of covariates; (c) identification based on significance such as t or F tests, are not reliable since the distribution of the implemented algorithm is not identifiable, thus Bayesian variable selection methods are preferred in these situations.

On the other hand, the bibliography for Bayesian variable selection in multinomial logistic regression is very scarce and that is the main reason we want to contribute with the present work. More precisely, posterior intractability and model complexity issues prevented multinomial logistic regression from being popularized in the context of the Bayesian variable selection problem. Even for a standard MCMC method, the added strong posterior correlations among regression coefficients of different class-specific slow the convergence and result in poor mixing. Exactly this is the problem we would like to address with the present work. Moreover, applications of g-priors Zellner (1986) and their mixtures have been reduced dramatically in comparison with the standard GLMs due to the severe complexity of the likelihood across the different pairwise baseline-logits where even the approved approaches of Liang et al. (2008), Li and Clyde (2013) cannot give a satisfying answer to the problem.

In the last decade, there was no computational tool that guaranteed flexible solutions over the hard aspects of MCMC methods, such as Metropolis-Hastings. Recent advances of computer technology and MCMC methods changed the expectations as it was possible to approximate complex model structures in Bayesian inference. That was clearly impossible in the early 2000's via smart sampling but became possible with the development of data augmentation schemes of Tanner and Wong (1987) and Albert and Chib (1993). These approaches share the flexible ideas of converting standard families of generalized linear models through the incorporation of latent variables into familiar model results. For instance, Bayesian variable selection is a data augmentation method due to the incorporation of a latent binary vector that quantifies the importance of the inclusion or exclusion of covariates. Furthermore, we note that the idea of data augmentation originates from the frequentist statistics related to expected maximization algorithm Dempster et al. (1977) in handling missing data mechanism. Although, the work of Albert and Chib (1993) was developed initially for the probit regression model, many researchers were inspired to transport similar ideas in logistic regression models

Holmes and Held (2006), Polson et al. (2013) and Frühwirth-Schnatter (2016). These research establishments are direct analogues of Albert and Chib's data augmentation in logistic regression, in the sense that the step of latent variables is substituted by truncated logistic and Polya-Gamma distributions especially for the first two mentioned approaches. Their difference, though lies substantially in content because they are based on scale mixture rather than location. In particular, Polson et al. (2013) took advantage of binomial likelihoods through the mixture of normal densities over Polya-Gamma distributions. Their success was devoted not only to binomial likelihood but also in generalizing the same concepts in other models like negative binomial regression and multinomial logistic regression which is the main topic of this thesis. The topic of data-augmentation strategies are further analysed in details in the second part of this chapter. However, the authors provide clever data augmentations for multinomial logistic regression only for estimation purposes rather than extending it to variable selection uncertainty.

Moreover, it is important also to highlight the work of Ghosh et al. (2011) which brings valuable research information for the reseach topic of this thesis, even if it doesn't focus at all in this research topic but remains a very important application of multinomial logistic regression with data augmentation but under the aspect of latent class models. In particular, the authors considered a two level latent class model, which accounts simultaneously for estimation and variable selection uncertainty based on potential set of covariates that are incorporated in the class probabilities of a multinomial logistic regression arisen within the distribution of latent variables in order to explain the eterogeneity of a univariate continuous response variable. By this way, conditional on prior features of latent class models and spike-slab prior, the authors developed a clever stochastic search Gibbs sampler that allows to explore the both the latent and model space in order to deliver accurate estimates with primary interest in marginal inclusion probabilities of covariates applied within model averaging framework. This approach can be viewed as a hybrid extension of SSVS George and McCulloch (1993) and Holmes and Held (2006) in latent class models, where the data augmentation of Holmes and Held (2006) ensures the computational convenience of a collapsing Gibbs sampler based on the categories of polychotomous latent variable. Even if their approach differs due to the focus on latent class models, is absolutely related to our work, in the sense that if the latent class labels were known, then the latent class model will reduce to the original multinomial logistic regression uncertainty accounted in this thesis. Moreover, in this work we considered similar extension of SSVS within the data augmentation scheme of Polson et al. (2013). In this work, we take advantage of Polya-Gamma data

augmentation and provide two similar sparse Bayesian variable selection methods for multinomial logistic regression based on hierarchical prior specifications of George and McCulloch (1993) and Dellaportas et al. (2002) as adopted in Bové and Held (2011). In particular, we implement SSVS and GVS for multinomial logistic regression by extending the covariate uncertainty in the framework of mixtures of g-priors both for typical and augmented multinomial logistic regression models. This approach allowed us to deal with the additional uncertainty of g in Bayesian variable selection parameter by adopting a hyperprior on it, hence the model learns regarding the adaptive shrinkage of data on the covariates. The success of our methods is encapsulated in the degree of sparseness that is preserved conditional on covariate components that characterize features of the model space, ensuring complete separation of the important covariates from the noise ensuring optimal predictive properties in terms of multi-classification. Although our method looks similar to the approach of Aijun et al. (2016), it differs in the hierachical prior specification only by the spike component which was allowed to lie with small variance in a region of zero rather placing a spike distribution at zero.
In addition, the research bibliography is enriched with novelty, in a way that the present work with Polya-Gamma data augmentation, manages to reproduce the complexity of an authentic multinomial logistic regression in terms of an amenable Gibbs sampling technique with respect to known linear model results. In this way, a $Q-1$ (number of categories without baseline) nested Gibbs sampler is outlined based on Holmes and Held (2006) and Polson et al. (2013) re-expressing the coefficients of each class conditional on the rest and splitting the g-prior to subclass of g-priors according to the coefficients of each class, then a flexible Bayesian variable selection method is implemented in order to improve the mixing of the chain and surpass the difficult aspects of MCMC in variable selection uncertainty.

We think SSVS and GVS prove extremely useful to accommodate the excessive model uncertainty that lies within each baseline-logit for typical and augmented multinomial logistic regression, whereas they are ideally suited for augmented multinomial logistic regression since they were developed initially for linear models as the model structure of augmented multinomial logistic regression is in fact identical to those of a linear model. Furthermore, our approach deals effectively with sparsity issues of the response variables regardless the sample size in contrast to Laplace approximation, which collapses due to the small sample size Bové and Held (2011) and Li and Clyde (2013).
Finally, the discussed methodologies are compared in terms of typical and augmented multinomial logistic regression with emphasis in the mixtures of g-priors and their performance is assessed on simulated and real datasets.

To conclude, the present work is organized as follows: in the first section we outline the problem of Bayesian variable selection for multinomial logistic regression and discuss the g-prior formulation, in the second section we introduce SSVS and GVS for typical and augmented multinomial logistic regression respectively and the last section is dedicated to the simulated and real dataset applications.

4.1 The problem of Bayesian Variable Selection in Multinomial Logistic Regression Models

In the last decades, variable selection for multinomial logistic regression has received great importance with the explosion of large scale datasets in bioinformatics and marketing sectors promising very interesting applications in multi-classification. Most of the real applications focus on diseases multi-classification according to a certain profile of genes expression such as type of cancer either multi-classification of products based on a specific advertising campaign profile. Multinomial logistic regression is used to investigate linear relationships of a nominal polychotomous response and a set of potentially covariates. Often, a multinomial logistic regression is initialized, after the identifiability constraints are set to zero for the regression coefficients and the intercept of the reference category (the first or the last category is chosen), called *baseline.* In this setup, it is of central interest not only to seek the important covariates, but also to describe the variability of each class membership of the nominal response based on a set of covariates for given baseline. This suggests that modulating the effect of k-th covariate on the q-th class membership of the response, is completely different from modulating the effect of the same response on the q'-th class, hence the distribution of each class-specific membership varies each time based on the respective set of covariates given the baseline q^*.

Usually, the multinomial logistic regression involves specifying for the random variable $\boldsymbol{Y} = (\boldsymbol{Y}_{q^*}, \boldsymbol{Y}_1 \dots, \boldsymbol{Y}_{Q-1})^T$ as linear function of covariates a model for i-th observed values $\boldsymbol{y}_i = (y_{i,1}, \dots, y_{i,q})^T$ as the following

$$\boldsymbol{Y}_i | \boldsymbol{a}, \boldsymbol{\beta} \sim \mathcal{MU}\left(1; p_{i,q^*}(a_{q^*}, \boldsymbol{\beta}_{q^*}), p_{i,1}(a_1, \boldsymbol{\beta}_1), \dots, p_{i,Q-1}(a_{Q-1}, \boldsymbol{\beta}_{Q-1})\right),$$

denoting with $p_{i,q}(a_q, \boldsymbol{\beta}_q) = P(y_{i,q} = q | a_q, \boldsymbol{\beta}_q)$, the probability that the i-th obsevation falls into q-th category encapsulating the effects of covariates except for those corresponding to baseline q^*.

$$p_{i,q}(a_q, \boldsymbol{\beta}_q) = \begin{cases} \frac{1}{1+\sum_{q=1}^{Q-1} \exp{(a_q + \boldsymbol{x}_i \boldsymbol{\beta}_q)}} & , q = q^* \\ \frac{\exp{(a_q + \boldsymbol{x}_i \boldsymbol{\beta}_q)}}{1+\sum_{q=1}^{Q-1} \exp{(a_q + \boldsymbol{x}_i \boldsymbol{\beta}_q)}} & , q \neq q^* \end{cases} ,$$

where $\mathcal{MU}\,(1;.)$ denotes the single sample unit multinomial distribution with normalizing assumption $\sum_{q=1}^{Q-1} p_{i,q}(a_q, \boldsymbol{\beta}_q) + p_{i,q^*}(a_{q^*}, \boldsymbol{\beta}_{q^*}) = 1$, q^* indexes the baseline category with intercept and regression coefficients respectively $a_{q^*} = 0$, $\boldsymbol{\beta}_{q^*} = \mathbf{0}_{p_{q^*}}$, $\boldsymbol{a} = (a_1, \ldots, a_{Q-1})^T$ denotes the complete intercept vector of dimension $(Q-1) \times 1$ without the baseline category q^* including each q-th class-specific intercepts, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_{q-1}^T)^T$ denotes the complete regression coefficients vector of dimension $(Q-1)p_q \times 1$ without the baseline category q^* which includes each q-th class-specific regression coefficients $\boldsymbol{\beta}_q = (\beta_{1,q}, \ldots, \beta_{p,q})^T$ of dimension $p_q \times 1$ and we assume that the design matrix $\boldsymbol{X}$ has been centered in order to consider separately the regression coefficients from the respective intercepts of each specific.

If we were interested in variable selection, covariate uncertainty would be accommodated by assuming binary latent vectors for each class membership with respect to the baseline denoting as $(Q-1)\boldsymbol{\gamma}_q \in \mathcal{C}_{q^*} \times 2^{p_q} \equiv \mathcal{C}_{q^*} \times \{0,1\}^{p_q}$, where $\mathcal{C}_{q^*}$ is the joint reduced set of class memberships of the response $\boldsymbol{Y}$ given the baseline class q^*. Equivalently, Bayesian variable selection formulation in multinomial logistic regression is outlined as simultaneous variable selection steps across different pairwise logistic regression models given the baseline class q^* , which involves the binary inclusion indicators as the following for $q = 1, \ldots, Q-1$

$$\log\left(\frac{P(y_{i,q} = q | a_q, \boldsymbol{\beta}_{q|\gamma_q}, \boldsymbol{\gamma}_q)}{P(y_{i,q^*} = q^* | a_{q^*}, \boldsymbol{\beta}_{q^*})}\right) = a_q + \boldsymbol{X}_{\gamma_q} \boldsymbol{\beta}_{q|\gamma_q} = \boldsymbol{\eta}_{\gamma_q}(a_q, \boldsymbol{\beta}_{q|\gamma_q}), \tag{4.1}$$

where $\boldsymbol{\gamma}_q = (\gamma_{1,q}, \ldots, \gamma_{p,q})$ is the binary latent vector of q-th baseline logit or q-th level of response variable Y given baseline class q^*, $\boldsymbol{X}_{\gamma_q} = [\boldsymbol{X}_{\gamma_q,1}, \ldots, \boldsymbol{X}_{\gamma_q,p_{\gamma_q}}]$ denotes the design matrix of dimension $n \times p_{\gamma_q}$ based only on the included components of binary latent vector $\boldsymbol{\gamma}_q$, $\boldsymbol{\beta}_{\gamma} = (\boldsymbol{\beta}_{1|\gamma_1}^T, \ldots, \boldsymbol{\beta}_{Q-1|\gamma_{Q-1}}^T)^T$ denotes the complete included regression coefficients vector of dimension $p_{\gamma} \times 1$ conditional on the complete latent vector $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{Q-1})$ with $p_{\gamma} = \sum_{q=1}^{Q-1} p_{\gamma_q}$ which contains each included q-th class-specific regression coefficients $\boldsymbol{\beta}_{q|\gamma_q} = (\beta_{1,q|\gamma_q}, \ldots, \beta_{p,q|\gamma_q})^T$ of dimension $p_{\gamma_q} \times 1$ whose components only have been entered conditional on binary latent vector of each q-th class-specific $\boldsymbol{\gamma}_q$, and $\boldsymbol{\eta}_{\gamma_q}(a_q, \boldsymbol{\beta}_{q|\gamma_q})$ denotes the linear predictor of the q-th baseline logit. The above variable selection procedure (4.1) is something more intuitive than representative, since it accounts also for the joint dependence among the possible class outcomes of the response and finding the best subset of each 2^{pq} combinations according to each $(Q-1)$ baseline logit comparisons. Under this settings, a researcher has to deal with the excessive uncertainty of parameters and covariate uncertainty in $\boldsymbol{a}$, $\boldsymbol{\beta}_{\gamma}$ and $\boldsymbol{\gamma}$ using appropriate priors to proceed with an accurate Bayesian variable selection procedure.

4.1.1 Prior Elicitation

Prior elicitation has been one of the most argued topics, which provoked debates among Bayesians and is still considered charming among researchers in the context of Bayesian variable selection. Even for a standard linear regression or logistic regression, practice showed that to account for all 2^p prior elicitation is not a trivial task and hence special strategies must be considered. Serious issues emerge related to posterior intractability and computation of posterior model probabilities within an ordinary generalized linear model such as multinomial logistic regression. These problems are notably seen in practice when one has to consider all joint prior features within model and parametrical space for the possible subsets of each baseline logit given baseline resulting in an excessive uncertainty. From a subjective stance, it will be difficult to obtain information from past studies or practitioners in high dimensional settings, but even if there was available it could not be pragmatic and hence is rejected. On the contrary, we shall argue that it will be more plausible to adopt an objective point of view. The latter choice is more intuitive rather than strategical, due to the fact that it will be virtually impossible to envelop all the different key features of $\mathcal{C}_{q^*} \times 2^{p_q}$ possible subsets encapsulating in a prior distribution if a subjective stance was adopted rather than an objective. Even in that case the objective Bayesian methodology outmatches Consonni and Veronese (1992). In fact, the excessive model uncertainty resulting from each $\mathcal{C}_{q^*} \times 2^{p_q}$ possible subsets across each baseline logit and the little or no guidelines regarding which variables to include or not favour the use of the objective approach of Jeffreys (1961) and Zellner (1986) in order to elicit manually all prior features of Bayesian variable selection. To end with this, we present the main idea of prior elicitation based on the seminal papers of Jeffreys (1961) and Zellner (1986) in details in the next subsection for the multinomial logistic regression framework.

4.1.2 Default Prior Choice

Prior choice has been studied extensively for many decades. It is an important area of research promising elegant applications even in Bayesian variable selection. Many research works were motivated by a desire to develop reasonable objective Bayesian methods for complex models without available guidelines regarding the best subset and the difficult prior elicitations of each subset. Meanwhile the intractability of the posterior encouraged the applications of numerical or MCMC approximations.

In the present section, we introduce a novel prior specification based on Zellner (1986) g-prior extending the generalized g-prior methodology of Bové and Held (2011) for Bayesian variable selection in the framework of multinomial logistic regression. The Zellner's g-prior was first popularized in linear regression settings and it represents one of the most common objective tools with emphasis in Bayesian variable selection procedure. This approach consists of a straightforward prior constructed by the expected Fisher information matrix scaled by a scalar g surpassing the difficulties of prior specification. Later Liang et al. (2008), Bové and Held (2011) and Li and Clyde (2013) introduced mixtures of g priors completing the initial work of Zellner. They established Bayesian variable selection procedures which allow the model to be trained and learn about the shrinkage of covariates by simply adopting a hyper-prior for g.

In addition, similar extensions of g-priors and its mixtures are more challenging in GLM settings and more specifically in multinomial logistic regression models since the expected Fisher information matrix depends on the regression coefficients and its structure is more complicated due to the added covariances matrices among different class-specific baseline logits. Hence, a most elaborated strategy must be adopted to deal with these issues.

Hence, adopting Zellner (1986) approach in the style of Liang's g-prior, we define the generalized g-prior for Bayesian variable selection in the framework of multinomial logistic regression as

$$\begin{aligned} \pi(\boldsymbol{a}, \boldsymbol{\beta}_{\boldsymbol{\gamma}} | \boldsymbol{\gamma}) &= \pi(\boldsymbol{a}|\boldsymbol{\gamma})\pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|g, \boldsymbol{\gamma}), \\ \boldsymbol{\beta}_{\boldsymbol{\gamma}}|g, \boldsymbol{\gamma} &\sim N_{p_{\gamma}}\left(\mathbf{0}_{\boldsymbol{p}_{\gamma}}, g\mathcal{I}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})^{-1}\right), \end{aligned} \tag{4.2}$$

where $\mathcal{I}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})$ denotes the expected Fisher information matrix which includes all the variance-covariance matrices belonging to the same baseline logit and the covariance matrices of different class-specific regression coefficients given the baseline class. In particular, the form of Fisher information matrix depends exclusively on each class-specific regression coefficients embedded in each q-th logistic regression linear predictor

given baseline class, suggesting a complex structure.
The above formula suggests prior independence among the parameters $\boldsymbol{a}$, $\boldsymbol{\beta}_\gamma$ which is proved by the maximum likelihood estimation in the Fisher sense based on the block diagonality in case of Bové and Held (2011) prior specification approach. The prior variance covariance matrix results from maximum likelihood estimation for the sampling density of the multinomial logistic regression (4.1) based initially on the score function and then on the curvature of the log-likelihood.
To begin with, the score function is calculated with the differentiation of the log-likelihood with respect to the linear predictor $\boldsymbol{\eta}_\gamma(\boldsymbol{a}, \boldsymbol{\beta}_\gamma)$ as

$$\begin{aligned}\frac{\partial \log\Big(f(\boldsymbol{Y}|\boldsymbol{a}, \boldsymbol{\beta}_\gamma, \boldsymbol{\gamma}\Big)}{\partial \boldsymbol{\eta}_\gamma(\boldsymbol{a}, \boldsymbol{\beta}_\gamma)} &= \begin{pmatrix} \frac{\partial \log\big((f(\boldsymbol{Y}|\boldsymbol{a},\boldsymbol{\beta}_\gamma,\gamma)\big)}{\partial \boldsymbol{a}} \\ \frac{\partial \log\big(f(\boldsymbol{Y}|\boldsymbol{a},\boldsymbol{\beta}_\gamma,\gamma)\big)}{\partial \boldsymbol{\beta}_\gamma} \end{pmatrix} \\ &= \begin{pmatrix} \frac{\partial \log\big((f(\boldsymbol{Y}|\boldsymbol{a},\boldsymbol{\beta}_\gamma,\gamma)\big)}{\partial a_1} \\ \vdots \\ \frac{\partial \log\big((f(\boldsymbol{Y}|\boldsymbol{a},\boldsymbol{\beta}_\gamma,\gamma)\big)}{\partial a_{Q-1}} \\ \frac{\partial \log\big(f(\boldsymbol{Y}|\boldsymbol{a},\boldsymbol{\beta}_\gamma,\gamma)\big)}{\partial \boldsymbol{\beta}_{1|\gamma_1}} \\ \vdots \\ \frac{\partial \log\big(f(\boldsymbol{Y}|\boldsymbol{a},\boldsymbol{\beta}_\gamma,\gamma)\big)}{\partial \boldsymbol{\beta}_{Q-1|\gamma_{Q-1}}} \end{pmatrix},\end{aligned}$$

which reduces after some mathematical steps to the following

$$\begin{aligned}\frac{\partial \log\Big(f(\boldsymbol{Y}|\boldsymbol{a}, \boldsymbol{\beta}_\gamma, \boldsymbol{\gamma}\Big)}{\partial \boldsymbol{\eta}_\gamma(\boldsymbol{a}, \boldsymbol{\beta}_\gamma)} &= \begin{pmatrix} \mathbf{1}^T_{n\times Q-1}\boldsymbol{y} - \mathbf{1}^T_{n\times Q-1}\boldsymbol{p}(\boldsymbol{a}, \boldsymbol{\beta}_\gamma) \\ \boldsymbol{X}^T_\gamma \boldsymbol{y} - \boldsymbol{X}^T_\gamma \boldsymbol{p}(\boldsymbol{a}, \boldsymbol{\beta}_\gamma) \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{1}^T_n \boldsymbol{y}_1 - \mathbf{1}^T_n \boldsymbol{p}_1(a_1, \boldsymbol{\beta}_{1|\gamma_1}) \\ \vdots \\ \mathbf{1}^T_n \boldsymbol{y}_{Q-1} - \mathbf{1}^T_n \boldsymbol{p}_{Q-1}(a_{Q-1}, \boldsymbol{\beta}_{Q-1|\gamma_{Q-1}}) \\ \boldsymbol{X}^T_{\gamma_1}\boldsymbol{y}_1 - \boldsymbol{X}^T_{\gamma_1}\boldsymbol{p}_1(a_1, \boldsymbol{\beta}_{1|\gamma_1}) \\ \vdots \\ \boldsymbol{X}^T_{\gamma_{Q-1}}\boldsymbol{y}_{Q-1} - \boldsymbol{X}^T_{\gamma_{Q-1}}\boldsymbol{p}_{Q-1}(a_{Q-1}, \boldsymbol{\beta}_{Q-1|\gamma_{Q-1}}) \end{pmatrix},\end{aligned}$$

where $\boldsymbol{X}_\gamma = [\boldsymbol{X}_{\gamma_1}, \ldots, \boldsymbol{X}_{\gamma_{Q-1}}]$ denotes the complete design matrix of dimension $n \times p_\gamma$ based on each q-th latent vector $\boldsymbol{\gamma}_q$ or q-th class-specific, $\boldsymbol{p}_q(a_q, \boldsymbol{\beta}_{q|\gamma_q}) = \Big\{p_{i,q}(a_q, \boldsymbol{\beta}_{q|\gamma_q})\Big\}_{i=1}^{n}$ denotes the probabilities of each i-th observation given q-th class-specific, $\boldsymbol{p}(\boldsymbol{a}, \boldsymbol{\beta}_\gamma) = \Big\{p_{i,q}(a_q, \boldsymbol{\beta}_{q|\gamma_q})\Big\}_{i=1,q=1}^{n,Q-1}$ denotes all the probabilities of each i-th observation belonging to each q-th class-specific and $\boldsymbol{y}_q = \{y_{i,q}\}_{i=1}^{n}$, denotes the observed values given q-th class-specific of the response variable $\boldsymbol{Y}$.

The next step, involves expected Fisher information matrix after differentiating the log-likelihood function twice with respect to the unknown parameters $\boldsymbol{a}$, $\boldsymbol{\beta}_\gamma$ as

$$\mathcal{I}(\boldsymbol{\eta}_\gamma(\boldsymbol{a}, \boldsymbol{\beta}_\gamma)) = -\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{a},\boldsymbol{\beta}_\gamma} \left(\begin{array}{cc} \frac{\partial^2 \log\big((f(\boldsymbol{Y}|\boldsymbol{a},\boldsymbol{\beta}_\gamma,\gamma)\big)}{\partial \boldsymbol{\alpha}^2} & \frac{\partial^2 \log\big(f(\boldsymbol{Y}|\boldsymbol{a},\boldsymbol{\beta}_\gamma,\gamma)\big)}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\beta}_\gamma} \\ \frac{\partial^2 \log\big(f(\boldsymbol{Y}|\boldsymbol{a},\boldsymbol{\beta}_\gamma,\gamma)\big)^T}{\partial \boldsymbol{\beta}_\gamma \partial \boldsymbol{\alpha}} & \frac{\partial^2 \log\big(f(\boldsymbol{Y}|\boldsymbol{a},\boldsymbol{\beta}_\gamma,\gamma)\big)}{\partial \boldsymbol{\beta}_\gamma^{\mathbf{2}}} \end{array} \right),$$

where the above Fisher information suggests a complex structure which encapsulates all the between and within cross-correlations of class-specific regression coefficients and intercepts respectively. In particular, the expected Fisher information matrix to account for all the cross-correlations may be reexpressed as

$$\mathcal{I}(\boldsymbol{\eta}_\gamma(\boldsymbol{a}, \boldsymbol{\beta}_\gamma)) = \left(\begin{array}{cc} \mathcal{I}(\boldsymbol{a}) & \mathcal{I}(\boldsymbol{a}, \boldsymbol{\beta}_\gamma) \\ \mathcal{I}(\boldsymbol{\beta}_\gamma, \boldsymbol{a})^T & \mathcal{I}(\boldsymbol{\beta}_\gamma) \end{array} \right)$$

$$= \left(\begin{array}{cccccc} \mathcal{I}(a_1) & \dots & \mathcal{I}(a_1, a_{Q-1}) & \mathcal{I}(a_1, \boldsymbol{\beta}_{1|\gamma_1}) & \dots & \mathcal{I}(a_1, \boldsymbol{\beta}_{Q-1|\gamma_{Q-1}}) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \mathcal{I}(a_{Q-1}, a_1) & \dots & \mathcal{I}(a_{Q-1}) & \mathcal{I}(\boldsymbol{\beta}_{Q-1|\gamma_{Q-1}}, a_1) & \dots & \mathcal{I}(a_{Q-1}, \boldsymbol{\beta}_{1|\gamma_{Q-1}}) \\ \mathcal{I}(\boldsymbol{\beta}_{1|\gamma_1}, a_1) & \dots & \mathcal{I}(\boldsymbol{\beta}_{Q-1|\gamma_{Q-1}}, a_1) & \mathcal{I}(\boldsymbol{\beta}_{1|\gamma_1}) & \dots & \mathcal{I}(\boldsymbol{\beta}_{1|\gamma_1}, \boldsymbol{\beta}_{Q-1|\gamma_{Q-1}}) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \dots \\ \mathcal{I}(a_1, \boldsymbol{\beta}_{Q-1|\gamma_{Q-1}}) & \dots & \mathcal{I}(\boldsymbol{\beta}_{Q-1|\gamma_{Q-1}}, a_{Q-1}) & \mathcal{I}(\boldsymbol{\beta}_{Q-1|\gamma_{Q-1}}, \boldsymbol{\beta}_{1|\gamma_1}) & \dots & \mathcal{I}(\boldsymbol{\beta}_{Q-1|\gamma_{Q-1}}) \end{array} \right)$$

where each $\mathcal{I}(a_q)$, $\mathcal{I}(a_q, a_{q'})$, $\mathcal{I}(a_q, \boldsymbol{\beta}_{q|\gamma_q})$, $\mathcal{I}(a_q, \boldsymbol{\beta}_{q'|\gamma_{q'}})$, $\mathcal{I}(\boldsymbol{\beta}_{q|\gamma_q})$, $\mathcal{I}(\boldsymbol{\beta}_{q|\gamma_q})$ and $\mathcal{I}(\boldsymbol{\beta}_{q|\gamma_q}, \boldsymbol{\beta}_{q'|\gamma_{q'}})$ are defined respectively as

$$\begin{aligned} \mathcal{I}(a_q, a_q) &= \begin{cases} \mathcal{I}(a_q) = \mathbf{1}_n^T \boldsymbol{p}_q(a_q, \boldsymbol{\beta}_{q|\gamma_q})(\mathbf{1}_n - \boldsymbol{p}_q(a_q, \boldsymbol{\beta}_{q|\gamma_q}))^T \mathbf{1}_n & , q = q' \\ -\mathbf{1}_n^T \boldsymbol{p}_q(a_q, \boldsymbol{\beta}_{q|\gamma_q}) \boldsymbol{p}_{q'}(a_{q'}, \boldsymbol{\beta}_{q'|\gamma_{q'}}) \mathbf{1}_n & , q \neq q' \end{cases}, \\ \mathcal{I}(a_q, \boldsymbol{\beta}_{q|\gamma_q}) &= \begin{cases} \mathbf{1}_n^T \boldsymbol{p}_q(a_q, \boldsymbol{\beta}_{q|\gamma_q})(\mathbf{1}_n - \boldsymbol{p}_q(a_q, \boldsymbol{\beta}_{q|\gamma_q}))^T \boldsymbol{X}_{\gamma_q} & , q = q' \\ -\mathbf{1}_n^T \boldsymbol{p}_q(a_q, \boldsymbol{\beta}_{q|\gamma_q}) \boldsymbol{p}_{q'}(a_{q'}, \boldsymbol{\beta}_{q'|\gamma_{q'}}) \boldsymbol{X}_{\gamma_{q'}} & , q \neq q' \end{cases}, \\ \mathcal{I}(\boldsymbol{\beta}_{q|\gamma_q}, \boldsymbol{\beta}_{q|\gamma_q}) &= \begin{cases} \mathcal{I}(\boldsymbol{\beta}_{q|\gamma_q}) = \boldsymbol{X}_{\gamma_q}^T \boldsymbol{p}_q(a_q, \boldsymbol{\beta}_{q|\gamma_q})(\mathbf{1}_n - \boldsymbol{p}_q(a_q, \boldsymbol{\beta}_{q|\gamma_q}))^T \boldsymbol{X}_{\gamma_q} & , q = q' \\ -\boldsymbol{X}_{\gamma_q}^T \boldsymbol{p}_q(a_q, \boldsymbol{\beta}_{q|\gamma_q}) \boldsymbol{p}_{q'}(a_{q'}, \boldsymbol{\beta}_{q'|\gamma_{q'}}) \boldsymbol{X}_{\gamma_{q'}} & , q \neq q' \end{cases}, \end{aligned} \tag{4.3}$$

showing the respective correlations and cross-correlations of the intercepts and the regression coefficients belonging to the same or different classes, the cross-correlations among regression coefficients and intercepts belonging or not to to the same class.

We notice that the joint Fisher information matrix includes in each main block $\mathcal{I}(\boldsymbol{a})$, $\mathcal{I}(\boldsymbol{a}, \boldsymbol{\beta}_\gamma)$, $\mathcal{I}(\boldsymbol{\beta}_\gamma, \boldsymbol{a})^T$ and $\mathcal{I}(\boldsymbol{\beta}_\gamma)$ the model parameters $\boldsymbol{a}, \boldsymbol{\beta}_\gamma$ and the non diagonal blocks constitute the prior independence among them a special matter of attention. Moreover, notice that the block diagonal $\mathcal{I}(\boldsymbol{\beta}_\gamma)$ corresponds to the Fisher information matrix used as prior variance-covariance matrix in the g-prior formulation (4.2) and differs only by the component $\boldsymbol{p}_q(a_q, \boldsymbol{\beta}_{q|\gamma_q})(\mathbf{1}_n - \boldsymbol{p}_q(a_q, \boldsymbol{\beta}_{q|\gamma_q}))^T$, which vanishes under certain prior specification such as Bové and Held (2011) and its expression based on (4.3) is more intuitive rather than representative which should be clarified to the interesting reader. In particular, expression (4.3) shows that the variance-covariance matrix and hence expected Fisher information matrix $\mathcal{I}(\boldsymbol{\beta}_\gamma)$ breaks into additional Fisher information matrices for $q = q'$ and $q \neq q'$ which add a nice interpretation with respect to multinomial logistic model. For instance, the $q = q'$ refers to the variance-covariance matrix of q-th baseline logit part of the multinomial logistic regression model, while the latter only to the covariances across different logistic parts of the multinomial logistic regression model. In ideal situations, we would like to the second part to be equal to zero. By this way, we could decompose the problem to $Q - 1$ logistic regression which is not of course the case here. In addition, a necessary and intuitive condition for the adoption of the prior independence (4.2) among $\boldsymbol{a}$, $\boldsymbol{\beta}_\gamma$ is the block diagonality of the joint expected Fisher information matrix preserved only by specific strategies, whereas in other cases it results too pragmatic and hence must be avoided. This strategies usually involve special centering and other methods to eradicate the dependence of expected Fisher information matrix from the resulting regression coefficients such as Bové and Held (2011) and Li and Clyde (2013).

To conclude, in the next subsection, we introduce and extend a detailed prior specification based on Bové and Held (2011) for Bayesian variable selection in the multinomial logistic regression framework.

4.1.3 Prior Choice of Bove and Held Approach

The problem of variable selection has been pervasive in daily practice for many years in linear regression and GLMs due to common issues related to the intractability of the posterior, computation of posterior model probabilities and arbitrary expected Fisher information matrix based on the regression coefficients. The same problem seems to be challenging in the multinomial logistic regression framework due to the complex nature of the model, as we seek solutions in the framework of objective Bayesian methods. Recently, approaches Liang et al. (2008) and Li and Clyde (2013) have been proposed to deal with the above issues, based mainly on g-priors Zellner (1986) and its mixtures to obtain consistent Bayesian variable selection across the model space, but in this chapter we will focus only on Bové and Held (2011). Consequently, in this section we attempt to give a first taste of detailed extensions of g-priors and its mixtures to the interesting reader based on a novel hierarchical prior specification likewise Bové and Held (2011). In particular, we construct and present a detailed generalized g-prior through the device of an imaginary sample size adopting the same steps adopted by the authors in their original paper based on Chen and Ibrahim (2003) and Chen et al. (2008) showing that is equivalent to Ntzoufras et al. (2003). For instance, let an imaginary sample size, $\boldsymbol{y}_{0_{n^*\times Q-1}} = \left\{y_{0_{i,q}}\right\}_{i=1,q=1}^{n^*,Q-1} = \boldsymbol{g}^{-1}(\boldsymbol{a})\boldsymbol{1}_{n^*\times Q-1}$ for fixed values of $\boldsymbol{a}$, where $\boldsymbol{g}^{-1}(\boldsymbol{a}) = (\boldsymbol{g}_1^{-1}(a_1), \ldots, \boldsymbol{g}_{Q-1}^{-1}(a_{Q-1}))$ with each element $\boldsymbol{g}_q^{-1}(\alpha_q) = \left\{g_{i,q}^{-1}(\alpha_q)\right\}_{i=1}^{n^*}$, denotes all inverse link functions for each i-th observation given q-th class and consider the multinomial logistic regression sampling density for $\boldsymbol{y}_{0_{n^*\times Q-1}}$ expressed as a multivariate GLM scaled by $g\phi$

$$\begin{aligned} f(\boldsymbol{y}_{0_{n^*\times Q-1}}|\boldsymbol{a}, \boldsymbol{\beta}_\gamma, g, \boldsymbol{\phi}, \boldsymbol{\gamma}) = \exp\left(\boldsymbol{1}_n^T \boldsymbol{y}_{0_{n^*\times q-1}} \boldsymbol{\Phi}^{-1} \boldsymbol{\eta}_\gamma(\boldsymbol{a}, \boldsymbol{\beta}_\gamma)\boldsymbol{1}_{Q-1}\right) \\ \exp\left(-\boldsymbol{1}_{Q-1}^T \boldsymbol{b}^T(\boldsymbol{\eta}_\gamma(\boldsymbol{a}, \boldsymbol{\beta}_\gamma))\boldsymbol{1}_{Q-1} + \boldsymbol{1}_{Q-1}^T \boldsymbol{c}^T(\boldsymbol{y}_{\boldsymbol{0}_{\boldsymbol{n^*}\times \boldsymbol{Q-1}}}, \boldsymbol{\phi})\boldsymbol{1}_{n^*}\right), \end{aligned}$$

where $\boldsymbol{b}(\boldsymbol{\eta}_\gamma(\boldsymbol{a}, \boldsymbol{\beta}_\gamma)) = (\boldsymbol{b}(\boldsymbol{\eta}_{\gamma_1}(a_1, \boldsymbol{\beta}_{1|\gamma_1})), \ldots, \boldsymbol{b}(\boldsymbol{\eta}_{\gamma_{Q-1}}(a_{Q-1}, \boldsymbol{\beta}_{Q-1|\gamma_{Q-1}}))),$
$\boldsymbol{c}(\boldsymbol{y}_{\boldsymbol{0}_{\boldsymbol{n^*}\times \boldsymbol{Q-1}}}, \boldsymbol{\phi}) = (\boldsymbol{c}(\boldsymbol{y}_{0_1}, \boldsymbol{\phi}), \ldots, \boldsymbol{c}(\boldsymbol{y}_{0_{Q-1}}, \boldsymbol{\phi}))$
with each element $\boldsymbol{c}(\boldsymbol{y}_{0_q}, \boldsymbol{\phi}) = \left\{c(y_{0_{i,q}}, \phi_i)\right\}_{i=1}^{n^*}$. Moreover, the specific functions $\boldsymbol{g}(.)$, $\boldsymbol{b}(.)$ and $\boldsymbol{c(.)}$ can be recognised according to the multinomial logistic regression model respectively as
$\boldsymbol{g}^{-1}(\boldsymbol{a}) = (\boldsymbol{p}_1(a_1, \boldsymbol{0}_{p_{\gamma_1}}), \ldots, \boldsymbol{p}_{Q-1}(a_{Q-1}, \boldsymbol{0}_{p_{\gamma_{Q-1}}}))$, $\boldsymbol{b}(\boldsymbol{\eta}_\gamma(\boldsymbol{a}, \boldsymbol{\beta}_\gamma)) = \boldsymbol{p}_{q^*}(a_{q^*}, \boldsymbol{\beta}_{q^*})$ and $\boldsymbol{c}(\boldsymbol{y}_{\boldsymbol{0}_{\boldsymbol{n^*}\times \boldsymbol{Q-1}}}, \boldsymbol{\phi}) = \left(\binom{n}{\boldsymbol{y}_{0_1}}, \ldots, \binom{n}{\boldsymbol{y}_{0_{Q-1}}}\right)$, then if an improper joint prior $\pi(\boldsymbol{a}, \boldsymbol{\beta}_\gamma) \propto 1$ has been adopted for the above multinomial logistic regression model, the posterior distribution of $\boldsymbol{\beta}_\gamma$ conditional on imaginary sample size $\boldsymbol{y}_{0_{n^*\times Q-1}}$ is expressed as

$$\begin{aligned} \pi(\boldsymbol{\beta}_\gamma|\boldsymbol{a}, \boldsymbol{\gamma}) \propto \exp\left(\boldsymbol{1}_{Q-1}^T \boldsymbol{y}_{0_{n^*\times Q-1}}^T \boldsymbol{\Phi}^{-1} \boldsymbol{\eta}_\gamma(\boldsymbol{a}, \boldsymbol{\beta}_\gamma)\boldsymbol{1}_{Q-1}\right) \\ \exp\left(-\boldsymbol{1}_{Q-1}^T \boldsymbol{b}^T(\boldsymbol{\eta}_\gamma(\boldsymbol{a}, \boldsymbol{\beta}_\gamma))\boldsymbol{1}_{n^*}\right), \end{aligned}$$

where the above distribution converges as $n \to +\infty$, to the asymptotic multivariate normal distribution Bernardo (1979)

$$\boldsymbol{\beta}_\gamma|\boldsymbol{y}_{0_{n^*\times Q-1}}, \boldsymbol{a}, g, \boldsymbol{\phi}, \boldsymbol{\delta}, \boldsymbol{\gamma} \sim N_{p_\gamma}(\hat{\boldsymbol{\mu}}_{p_\gamma}(\boldsymbol{a}), g\mathcal{I}(\hat{\boldsymbol{\mu}}_{p_\gamma}(\boldsymbol{a}))^{-1}),$$

where $\widehat{\boldsymbol{\mu}}_{\gamma}(a)$ denotes the posterior mode which results as function of class-specific intercepts $\boldsymbol{a}$ after evaluating it to the imaginary sample size $\boldsymbol{y}_{0_{n^* \times Q-1}}$, $\boldsymbol{\delta} = \boldsymbol{\delta}(\boldsymbol{a})$ and $\mathcal{I}(\widehat{\boldsymbol{\mu}}_{p_\gamma}(\boldsymbol{a})) = \frac{\boldsymbol{\delta}(a)\boldsymbol{X}_\gamma^T \boldsymbol{X}_\gamma}{\phi}$ is the expected Fisher information matrix evaluated at the mode $\widehat{\boldsymbol{\mu}}_{p_\gamma}(\boldsymbol{a})$ as it was demonstrated in Chen and Ibrahim (2003), Chen et al. (2008) and Bové and Held (2011). This expected Fisher information matrix is based on (4.3), whether it belongs to the same class-specific $q = q'$ or to two different class-specifics $q \neq q'$ as the following implies

$$\mathcal{I}(\widehat{\boldsymbol{\mu}}_{\gamma_q}(a_q), \widehat{\boldsymbol{\mu}}_{\gamma_q}(a_q)) = \begin{cases} \mathcal{I}(\widehat{\boldsymbol{\mu}}_{\gamma_q}(a_q)) = \delta(a_q)\boldsymbol{X}_{\gamma_q}^T \boldsymbol{X}_{\gamma_{q'}} & , q = q' \\ -\delta(a_q, a_{q'})\boldsymbol{X}_{\gamma_q}^T \boldsymbol{X}_{\gamma_{q'}} & , q \neq q' \end{cases},$$

denoting with $\boldsymbol{\delta}(\boldsymbol{a})$

$$\delta(a_q, a_q) = \begin{cases} \delta(a_q) = \left(\frac{\exp{(a_q)}}{1+\sum_{q=1}^{Q-1} \exp{(a_q)}}\right)\left(1 - \frac{\exp{(a_q)}}{1+\sum_{q=1}^{Q-1} \exp{(a_q)}}\right) & , q = q' \\ \frac{\exp{(a_q + a_{q'})}}{\left(1+\sum_{q=1}^{Q-1} \exp{(a_q)}\right)^2} & , q \neq q' \end{cases}.$$

Furthermore, assume for simplicity that $\boldsymbol{\Phi} = I$ and notice that the expected Fisher information matrix depends on $\boldsymbol{a}$ which can be substituted with zero or the maximum likelihood estimator according to the guidelines Held et al. (2015). In order to avoid its influence, we prefer to set it to zero to avoid any undesired correlation with the class-specific regression coefficients and hence eliminating the dependence of imaginary sample size $\boldsymbol{y}_{0_{n^* \times Q-1}}$, reducing (4.4) to a posterior distribution evaluated at mode $\widehat{\boldsymbol{\mu}}_{\gamma}(\boldsymbol{0}_{p_\gamma}) = \boldsymbol{0}_{p\gamma}$

$$\boldsymbol{\beta}_\gamma | Q, g, \boldsymbol{\delta}, \boldsymbol{\gamma} \sim N_{p_\gamma}(\boldsymbol{0}_{p_\gamma}, gQ^2 \mathcal{I}^{(BH)}(\boldsymbol{0}_{p_\gamma})^{-1}), \tag{4.4}$$

where the above generalized g-prior is a consequent step of simplification of the above expressions to the following respectively

$$\mathcal{I}^{(BH)}(\boldsymbol{0}_{p_{\gamma_q}}, \boldsymbol{0}_{p_{\gamma_q}}) = \begin{cases} \mathcal{I}^{(BH)}(\boldsymbol{0}_{p_{\gamma_q}}) = \delta(0)\boldsymbol{X}_{\gamma_q}^T \boldsymbol{X}_{\gamma_{q'}} & , q = q' \\ -\delta(0,0)\boldsymbol{X}_{\gamma_q}^T \boldsymbol{X}_{\gamma_{q'}} & , q \neq q' \end{cases}, \tag{4.5}$$

denoting with $\delta(0,0)$

$$\delta(0,0) = \begin{cases} \delta(0) = \frac{Q-1}{Q^2} & , q = q' \\ \frac{1}{Q^2} & , q \neq q' \end{cases}, \tag{4.6}$$

differs from the standard g-prior Liang et al. (2008) only by the scalar Q^2 and its structure of this Fisher information matrix is also related to the approach of Ntzoufras et al. (2003) . In addition, the approach of Ntzoufras et al. (2003) proves useful to relate the above generalized g-prior factorization (4.2) in Liang's style, as he proposed to set the regression coefficients to their prior means equal to zero, then based on the resulting joint expected Fisher information matrix of class-specific intercepts and regression coefficients we obtain

$$
\begin{aligned}
\mathcal{I}^{(BH)}(\boldsymbol{\eta}_{\boldsymbol{\gamma}}(\boldsymbol{a},\mathbf{0}_{p_\gamma})) &= \begin{pmatrix} \mathcal{I}^{(BH)}(\boldsymbol{a}) & \mathcal{I}^{(BH)}(\boldsymbol{a},\mathbf{0}_{p_\gamma}) \\ \mathcal{I}^{(BH)}(\mathbf{0}_{p_\gamma},\boldsymbol{a})^T & \mathcal{I}^{(BH)}(\mathbf{0}_{p_\gamma}) \end{pmatrix} \\
&= \begin{pmatrix}
\mathcal{I}^{(BH)}(a_1) & \dots & \mathcal{I}^{(BH)}(a_1,a_{Q-1}) & \mathcal{I}^{(BH)}(a_1,\mathbf{0}_{p_{\gamma_1}}) & \dots & \mathcal{I}^{(BH)}(a_1,\mathbf{0}_{p_{\gamma_{Q-1}}}) \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
\mathcal{I}^{(BH)}(a_{Q-1},a_1) & \dots & \mathcal{I}^{(BH)}(a_{Q-1}) & \mathcal{I}^{(BH)}(\mathbf{0}_{p_{\gamma_{Q-1}}},a_1) & \dots & \mathcal{I}^{(BH)}(a_{Q-1},\mathbf{0}_{p_{\gamma_{Q-1}}}) \\
\mathcal{I}^{(BH)}(\mathbf{0}_{p_{\gamma_1}},a_1) & \dots & \mathcal{I}^{(BH)}(\mathbf{0}_{p_{\gamma_{q-1}}},a_1) & \mathcal{I}^{(BH)}(\mathbf{0}_{p_{\gamma_1}}) & \dots & \mathcal{I}^{(BH)}(\mathbf{0}_{p_{\gamma_1}},\mathbf{0}_{p_{\gamma_{Q-1}}}) \\
\vdots & \ddots & \vdots & \vdots & \ddots & \dots \\
\mathcal{I}^{(BH)}(a_1,\mathbf{0}_{p_{\gamma_{Q-1}}}) & \dots & \mathcal{I}^{(BH)}(\mathbf{0}_{p_{\gamma_{Q-1}}},a_{q-1}) & \mathcal{I}^{(BH)}(\mathbf{0}_{p_{\gamma_{Q-1}}},\mathbf{0}_{p_{\gamma_1}}) & \dots & \mathcal{I}^{(BH)}(\mathbf{0}_{p_{\gamma_{Q-1}}})
\end{pmatrix},
\end{aligned}
$$

where $\mathcal{I}^{(BH)}(a_q)$, $\mathcal{I}^{(BH)}(a_q,a_{q'})$, $\mathcal{I}^{(BH)}(a_q,\mathbf{0}_{p_{\gamma_q}})$, $\mathcal{I}^{(BH)}(a_q,\mathbf{0}_{p_{\gamma_{q'}}})$ are defined as

$$
\mathcal{I}^{(BH)}(a_q,a_q) = \begin{cases} \mathcal{I}^{(BH)}(a_q) = \delta(a_q)\mathbf{1}_n^T\mathbf{1}_n & ,q = q' \\ -\delta(a_q,a_{q'})\mathbf{1}_n^T\mathbf{1}_n & ,q \neq q' \end{cases},
$$

$$
\mathcal{I}^{(BH)}(a_q,\mathbf{0}_{p_{\gamma_q}}) = \begin{cases} \mathbf{0}_{p_{\gamma_q}} & ,q = q' \\ \mathbf{0}_{p_{\gamma_q}} & ,q \neq q' \end{cases},
$$

where notice all the above steps are reduced due to the centering of the design matrix $\boldsymbol{X}_\gamma$ implying $\boldsymbol{X}_{\gamma_q}^T\mathbf{1}_n = \mathbf{0}_{p_\gamma}$ to each q-th level design matrix. The following block diagonal structure allows to express the joint expected Fisher information matrix as

$$
\mathcal{I}^{(BH)}(\boldsymbol{\eta}_{\boldsymbol{\gamma}}(\boldsymbol{a},\mathbf{0}_{p_\gamma})) = \begin{pmatrix} \mathcal{I}^{(BH)}(\boldsymbol{a}) & \mathbf{0}_{p_\gamma} \\ \mathbf{0}_{p_\gamma}^T & \mathcal{I}^{(BH)}(\mathbf{0}_{p_\gamma}) \end{pmatrix},
$$

which suggests prior independence of $\boldsymbol{a}$, $\boldsymbol{\beta}_\gamma$. In this case, the joint prior specification based on generalized g-prior of Bové and Held (2011) is defined as the following

$$
\begin{aligned}
\pi^{(BH)}(\boldsymbol{a},\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma}) &= \pi^{(BH)}(\boldsymbol{a}|\boldsymbol{\gamma})\pi^{(BH)}(\boldsymbol{\beta}_\gamma|Q,g,\boldsymbol{\delta},\boldsymbol{\gamma}) \\
&= \prod_{q=1}^{Q-1}\pi^{(BH)}(a_q|\boldsymbol{\gamma})\pi^{(BH)}(\boldsymbol{\beta}_\gamma|g,Q,\boldsymbol{\gamma}), \qquad (4.7)
\end{aligned}
$$

$$
\pi^{(BH)}(a_q|\boldsymbol{\gamma}) \propto 1, \text{ for } \quad q = 1,\dots,Q-1,
$$

$$
\boldsymbol{\beta}_\gamma|Q,g,\boldsymbol{\delta},\boldsymbol{\gamma} \sim N_{p_\gamma}\left(\mathbf{0}_{\boldsymbol{p}_\gamma}, gQ^2\mathcal{I}^{(BH)}(\mathbf{0}_{p_\gamma})^{-1}\right), \qquad (4.8)
$$

where the joint prior of $\boldsymbol{a}$ can be considered as a product of independent improper priors adopting Jeffreys (1961) approach. The expected Fisher information matrix $\mathcal{I}^{(BH)}(\boldsymbol{a})$ suggests correlation among the intercepts and this dependence can be omitted since the parameters are not of interest and inference is not affected. To conclude, since we highlight the main aspects of Bayesian variable selection and illustrate the generalized g-prior, we can go a step further in the next sections by introducing the Bayesian variable selection methods for multinomial logistic regression with MCMC.

4.2 MCMC for Bayesian Variable Selection in Multinomial Logistic Models

Modern technology progress increased the tendency of classification datasets in the latest century, where the problem of variable selection in multinomial logistic regression models, as a special family of generalized linear models, has been an open topic for a long time in the Bayesian community with respect to i) the irreconcilability of posterior, ii) the prior specification and iii) the enumeration of model space through the calculation of posterior model probabilities when the number of predictors is grows super exponentially with specific-class memberships. This fact led Bayesians to consider alternative variable selection methods based on construction of MCMC Gilks et al. (1996) or MCMC with data augmentation Tanner and Wong (1987).

In this way, the variable selection in multinomial logistic regression turns into a decision problem where there are distinct subsets of the initial variables, that compensate in describing the distribution of each class membership of polychotomous response Y given baseline class, while these identified subsets can be used for prediction purposes in order to test if the observed value of the response's class membership was classified correctly or not.

From a Bayesian perspective, variable selection in multinomial logistic regression is supplied with the probabilistic nature of each class-specific subset of variables which captures successfully the additional uncertainty related to the pair model-model parameter and then this uncertainty is rephrased as a post-summary containing important information to variable selection. When there isn't available information about which subset to prefer, the issue is fairly treated from an objective point of view adopting default prior designs based on Zellner (1986) g-prior and it's mixtures Liang et al. (2008).

In addition, recent advances of computer technology enabled the construction of MCMC methods Gilks et al. (1996) that revived important pathways in complex statistical modelling likewise generalized linear models and led to consistent variable selection solutions especially in high-dimensional settings when spike-slab priors are chosen. The main intuition behind MCMC methods lies on creating automatic objective methods coupled with the notions of g-priors mixtures even in multinomial logistic models framework, where sparsity is preserved forcing noise covariates in an area near zero.

Despite, MCMC methods count an enormous cataloque of variable selection approaches in generalized linear models, for the shake of feasibility we cannot describe every method in detail, thus we emphasize only the most recent approaches related to g-priors and their mixtures. Moreover, these approaches were summarized in details at the third part of this chapter including Chen and Ibrahim (2003), Ntzoufras et al. (2003), Hansen and Yu (2003), Wang and George (2007), Chen et al. (2008), Bové and Held (2011), Li and Clyde (2013) and Li and Clyde (2018). These approaches suggest a different structure of Fisher information matrix based on observed or expected Fisher information matrix accompanied by an MCMC algorithm. Despite their utility, we will pay attention only to Bové and Held (2011) prior specification and we propose model search algorithms for variable selection based explicitly on SSVS and GVS. We think that these algorithms suit ideally with the problem of variable selection for multinomial logistic regression, in identifying the most probable subsets within a model space for each class-specific.

Despite the flexibility of these algorithms to create a Markov chain that operates on the surface of the joint posterior distribution of model parameters and parameters in linear regression models, their implementation is quite challenging in GLMs terms and hence multinomial logistic models, in the sense that the likelihood is not anymore conjugate, thus Metropolis-Hastings stages are added for the class-specifc regression coefficients and intercepts.

On the other hand, the interaction of modern technology with the advent of MCMC gave rise to *data augmentation* pioneering ideas of Tanner and Wong (1987) catalyzing Bayesian variable selection methods especially in binary regression, a special case of multinomial logistic regression. Under this approach, a latent layer of possibly unobserved random variables are incorporated in the model reproducing an equivalent pseudo-generated mechanism (*augmented model*) of the genuine one. This idea proves very useful for computational and simplicity purposes in situations where the true model nature is complex by construction. In this way, the sampling density of under-study generating mechanism can be still retrieved by intergrating or summing the augmented model over the latent variables, depending on the type of latent random variables if they are continuous or not, but in this thesis interest lies only to models with continuous latent components. A simple intuition behind a strategy of this kind, is just to incorporate as many as possible latent data points to reconstruct a similar behaviour of the "unmanageable" true model, where the latent data points may be seen as hidden pieces of information that gave rise to the observed values of the unmanageable model. It should be stressed out that this idea actually originates from frequentist statistics where usually a missing-data strategy is applied through an expectation-maximization algorithm Dempster et al. (1977) for the joint sampling density of the missing and observed data, so data augmentation can be thought of as the Bayesian version of missing data.

In addition, the seminal paper of Albert and Chib (1993) was the first Bayesian model with data augmentation scheme that was introduced for probit regression. Even so, the original idea of frequentist probit regression shares the notion of latent variables as exogenous pertubations that produced the observed binary responses, the Bayesian version implies a flexible MCMC procedure that turns into known results of a standard linear model in order to avoid the painful computational cost. On the contrary, the need of extending data augmentation designs similar to that of Albert and Chib's, was the main motivation of not remaining stable on models without model interpretability like probit regression and to proceed on grounds of logistic regression marking a new revolutionary era with the papers of Holmes and Held (2006), Polson et al. (2013) and Frühwirth-Schnatter (2016). The clever idea of these approaches lie in the same directions of Albert and Chib (1993) data augmentation scheme where the original sampling density may be expressed as an increased density after the incorporation of latent variables which ends with a standard Gaussian density. Thus the posterior for the regression coefficients is reduced also to familiar results with those of a linear regression depending on the latent structure. While in the latent approach of Albert and Chib (1993) the prior latent structure was based on truncated normal distributions, the recent approaches of Holmes and Held (2006) and Polson et al. (2013) use instead truncated logistic and Polya-Gamma distributions, which affect in a different way the speed of convergence of MCMC methods. The latter family of distributions will be examined further in the next sections when presenting the notion of data augmentation.

Despite the absence of Bayesian variable selection in multinomial logistic models, due to the non closed form of posterior for many years, we aim to contribute in this thesis by establishing the main bridge between linear regression and generalized linear models by extending the model search algorithms SSVS of George and McCulloch (1993) and GVS of Dellaportas et al. (2002) based on the clever introduction of the latent scheme of Polson et al. (2013). We think that SSVS and GVS are the most appropriate to apply for these Bayesian models since they were first introduced for linear models. The main motivation behind the construction of these algorithms lie on creating automatic Bayesian procedures that will manage to control the complex nature of the $\mathcal{C}_{q^*} \times 2^{p_q}$ possible subsets of a multinomial logistic regression sharing the objective Bayesian properties when mixtures of g-priors are adopted and to bypass the hard aspects of MCMC methods through the tuning of posterior for regression coefficients.

To conclude, in this second part of this chapter, we will present and compare the typical multinomial logistic model with the augmented model under mixtures of g-priors based on Bové and Held (2011). Prior specification and their performance is assessed based on simulation and real datasets. More precisely, this section is organized as follows: we begin by introducing the model search algorithms SSVS and GVS for an ordinary multinomial logistic model with mixtures of g priors, Zellner-Siow Zellner and Siow (1980) and hyper-g Liang et al. (2008) and its augmented version, then we end up with this chapter by illustrating the main results from both comparisons in simulated and real datasets.

4.3 SSVS in Typical Multinomial Logistic Setup

SSVS has been the cornerstone of early 90's in the Bayesian universe launching the variable selection problem in the world of MCMC procedures promising very interesting applications, while it represents the source of many existing Bayesian variable selection methods. After its establishment, the research bibliography on the Bayesian variable selection problem increased due to the ideas of quantifying and monitoring respectively the uncertainty and importance of covariates via the binary vector $\boldsymbol{\gamma}$ and spike-slab prior representation. In this way, noise covariates are omitted permitting the identification of only important variables that contribute to actual relationship and prediction. Even SSVS was popularized broadly in linear regression settings, its utility was limited in GLMs framework due to strains of i) prior elicitation, ii) posterior intractability and iii) computation of posterior model probabilities and especially in the case of multinomial logistic regression due to the complexity of expected Fisher information matrix. These are becoming more and more evident and intensive due to the additional model complexity in the resulting $(Q-1)2^{p_q}$ possible subsets.

Here in this section, in order to deal with these issues, we present a novel detailed Bayesian variable selection method based on SSVS of George and McCulloch (1993) with mixtures of g-priors for Bayesian variable selection in multinomial logistic regression models. We believe that the proposed method suits the context mentioned previously and that the lack of available guidelines regarding the best subset of each q-th class-specific with respect to the baseline q^* prioritize the objective Bayesian methodology based on g-priors and its mixtures coupled with Jeffrey's approach. Moreover, the current approach still proves useful in handling sparsity issues of polychotomous response variable when standard Bayesian variable selection methods such as Laplace approximation fail due to the small sample size. However, an obvious disadvantage for the implementation of the underlying method is that the intractability of the regression coefficients lead unavoidably to a sophisticated Metropolis-Hastings step (analogous to a standard GLM) that accounts for the correlations and cross-correlations between the same or different classes. Even this drawback, the desire to create automatic Bayesian variable selection procedures, stems from the appealing property of handling complex structures like multinomial logistic regression and is fulfilled in the present work. To begin with, SSVS usually assumes for the observed values $\boldsymbol{y}$ a fixed sampling density in multinomial logistic regression that captures the linear dependence of the polychotomous response $\boldsymbol{Y}$ and the covariates as the following

$$f(\boldsymbol{y}|\boldsymbol{a},\boldsymbol{\beta}) = \frac{\exp\left(\sum_{q=1}^{Q-1} \boldsymbol{y}_q^T \left(a_q \mathbf{1}_n + \boldsymbol{X}\boldsymbol{\beta}_q\right)\right) \exp\left(\sum_{q=1}^{Q-1} \binom{n}{\boldsymbol{y}_q}\right)}{\mathbf{1}_n^T \left(\mathbf{1}_n + \sum_{q=1}^{Q-1} \exp\left(a_q \mathbf{1}_n + \boldsymbol{X}\boldsymbol{\beta}_q\right)\right)}, \tag{4.9}$$

where notice that the sampling density is suppressed from any dependence on binary vector $\boldsymbol{\gamma}$. The main intuition behind the binary latent vector $\boldsymbol{\gamma}$ lies on monitoring the importance of covariates in such a way, that the important covariates of each q-th class-specific and hence the respective effects are entering the model, whereas the non significant are omitted by shrinking them to an area near zero. In other words, the included and non included variables of each class-specific are separated by adopting a hierarchical mixture representation conditional on features of the model space, equivalently on the q-th class-specific subsets, which allows to decide whether they have to enter or not in the respective class-specific subset based on the information of the data $\boldsymbol{y}$. Often, this hierarchical prior specification includes a two-component spike-slab representation which is controlled by tuning parameters τ and c that adjust the level of shrinkage tracking down the important variables that really affect the target variable. However, their input values need care in order to ensure good mixing of the chain and separation. In the case of multinomial logistic regression, the hierarchical prior specification for the regression coefficients conditional on the latent vector $\boldsymbol{\gamma}$ is defined

$$\boldsymbol{\beta}|Q, g, \boldsymbol{\gamma} \sim N_{p(q-1)}\left(\mathbf{0}_{p_q(Q-1)}, \boldsymbol{D}\boldsymbol{R}^{(BH)}\boldsymbol{D}\right), \tag{4.10}$$

where $\boldsymbol{D}$ is a diagonal matrix of dimension $(Q-1)p_q \times (Q-1)p_q$ matrix with j-th entry equal to $\gamma_j c_j \tau_j + (1-\gamma_j)\tau_j$, for $j = 1, \ldots, (Q-1)p_q$ and which can be alternatively seen for each q-th class-specific $\boldsymbol{\gamma}_q$ as subsets of $\boldsymbol{\gamma}$ as

$$\boldsymbol{D} = \begin{pmatrix} \boldsymbol{D}_1 & \mathbf{0}_{p_q} & \mathbf{0}_{p_q} \\ \mathbf{0}_{p_q} & \ddots & \mathbf{0}_{p_q} \\ \mathbf{0}_{p_q} & \mathbf{0}_{p_q} & \boldsymbol{D}_{Q-1} \end{pmatrix},$$

where each $\boldsymbol{D}_q$ denote partitioned matrices of dimension $p_q \times p_q$ with diagonal elements $\gamma_{j,q} c_{j,q} \tau_{j,q} + (1-\gamma_{j,q})\tau_{j,q}$ for $j = 1, \ldots, p_q$ and $q = 1, \ldots, Q-1$ referring to each q-th class-specific $\boldsymbol{\gamma}_q$ and this notation will be useful in the next sections of augmented multinomial logistic regression, $\boldsymbol{R}^{(BH)} = gQ^2 \mathcal{I}^{(BH)}(\mathbf{0}_{(Q-1)p_q})^{-1}$ is the prior correlation matrix under Bové and Held (2011), where $\mathcal{I}^{(BH)}(\mathbf{0}_{(Q-1)p_q})$ denotes the expected Fisher information matrix which encapsulates the variance-covariance and covariance between the same $q = q'$ or two different class specifics $q \neq q'$

$$\mathcal{I}^{(BH)}(\mathbf{0}_{p_q}, \mathbf{0}_{p_q}) = \begin{cases} \mathcal{I}^{(BH)}(\mathbf{0}_{p_q}) = (Q-1)\boldsymbol{X}^T\boldsymbol{X} & , q = q' \\ -\boldsymbol{X}^T\boldsymbol{X} & , q \neq q' \end{cases}, \tag{4.11}$$

where notice that the above expected Fisher information matrix is of fixed dimension suppressed from the dependence of $\boldsymbol{\gamma}$. To deal with the issue of the prior specification, we adopt the joint hierarchical prior specification for parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, g and $\boldsymbol{\gamma}$ based on Bové and Held (2011) with directions over the objective Bayesian ideas of Jeffreys (1961) and mixtures of g-priors Zellner (1986) and Liang et al. (2008) for multinomial logistic regression as

$$\pi^{SSVS}(\boldsymbol{a}, \beta, g, \boldsymbol{\gamma}) = \prod_{q=1}^{Q-1} \pi^{(BH)}(a_q)\pi^{SSVS}(\boldsymbol{\beta}|Q, g, \boldsymbol{\gamma})\pi(g)\pi(\boldsymbol{\gamma}), \tag{4.12}$$

where in the above equation we define

$$\pi^{(BH)}(a_q) \propto 1, \ \text{ for } \ q = 1, \ldots, Q-1,$$

where notice the model indicator $\boldsymbol{\gamma}$ is dropped for simplicity from the joint improper prior and $\pi^{SSVS}(.|Q, g, \boldsymbol{\gamma})$ denotes the hierarchical mixture prior of SSVS for $\boldsymbol{\beta}$ under Bové and Held (2011) generalized g-prior (4.8). Choices of priors on model space include an independent Bernoulli in case of indifference among respective subsets in each baseline logit as

$$\pi(\boldsymbol{\gamma}) = \prod_{j=1}^{p} w_j^{\gamma_j}(1-w_j)^{1-\gamma_j} = \prod_{j=1}^{p}\prod_{q=1}^{Q-1} w_{j,q}^{\gamma_{j,q}}(1-w_{j,q})^{1-\gamma_{j,q}},$$

with prior probabilities weights of inclusion w_i or $w_{j,q}$. The latter can be also extended using the hierarchical prior of Scott and Berger (2010) for w_i or $w_{j,q}$. Regarding the prior $\pi(g)$, mixtures of g-priors based on Zellner and Siow (1980) and hyper-g-prior Liang et al. (2008) allow to account for the additional uncertainty of g parameter.
Apart from the prior specification, SSVS has been distinguished among other Bayesian variable selection methods mainly for its fixed dimension of the model space over the sampling density which is updated successfully with the hierarchical prior (4.12) leading to the joint posterior of parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, g and $\boldsymbol{\gamma}$

$$\pi^{SSVS}(\boldsymbol{a},\boldsymbol{\beta},g,\boldsymbol{\gamma}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{a},\boldsymbol{\beta})\prod_{q=1}^{Q-1}\pi^{(BH)}(a_q)\pi^{SSVS}(\boldsymbol{\beta}|Q,g,\boldsymbol{\gamma})\pi(g)\pi(\boldsymbol{\gamma}), \tag{4.13}$$

where the joint posterior (4.13) remains in an intractable and form which favors the employment of MCMC procedures surpassing the common obstacles of formal Bayesian variable selection methods. In particular, an MCMC method is constructed on the joint parameter and model space that allows to deliver a simulated sample from the unknown joint posterior (4.13) by iterating over the full conditionals of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, g and $\boldsymbol{\gamma}$. This is a natural consequence of the constant dimension of sampling density that accelerates the convergence to the target posterior since the full conditional of $\boldsymbol{\gamma}$, central to variable selection, does not depend on the likelihood and hence the retrieved sample contains relevant information for Bayesian variable selection regarding the possible subsets of all 2^{p_q} subsets of each q-th class.
However, likewise GLMs in multinomial logistic regression, serious issues arise within the implementation of SSVS. Since only the full conditionals of g for Zellner-Siow prior and $\boldsymbol{\gamma}$ are amenable to Gibbs sampling, suggesting additional Metropolis-Hastings steps for the rest of parameters, especially for each class-specific regression coefficients $\boldsymbol{\beta}$ and intercepts $\boldsymbol{\alpha}$ respectively, one has to select carefully proposals that mimic or reproduce the underlying genuine structure in order to provide an ideal representative of the target posterior. In this way, SSVS is based on successive Metropolis-Hastings stages within Gibbs sampler and outlined as follows updating jointly the parameter $\boldsymbol{\beta}$ including all class-specific regression coefficients $\boldsymbol{\beta}_q$ given baseline class and separately the parameter $\boldsymbol{a}$ based on each class-specific intercepts $\boldsymbol{a}_q$

A. Set initial values $\boldsymbol{\gamma}^{(0)}$, $\boldsymbol{\beta}^{(0)}$, $a_1^{(0)}, \dots, a_{Q-1}^{(0)}$ and $g^{(0)}$. For fixed $g = n$, delete **Step 5**.

B. For iterations $s = 1, \dots, S$:

Step 1: Set current values equal to $\boldsymbol{\gamma} = \boldsymbol{\gamma}^{(s-1)}$, $\boldsymbol{\beta} = \boldsymbol{\beta}^{(s-1)}$, $a_1 = a_1^{(s-1)}, \dots, a_{Q-1} = a_{Q-1}^{(s-1)}$ and $g = g^{(s-1)}$.

Step 2: Sample $\gamma_j^{(s)} \sim Bern\left(\pi_j^{SSVS}\right)$, for $j = 1, \dots, (Q-1)p_q$, given the current states of $\boldsymbol{\gamma}_{-j}^{(s-1)}$ $\boldsymbol{\beta}^{(s-1)}$, $\sigma^{2(s-1)}$, $a^{(s-1)}$ and $g^{(s-1)}$, where $\boldsymbol{\gamma}_{-j}$ are the components of $\boldsymbol{\gamma}$ except element γ_j

(a) with probability inclusion of j-th covariate $\pi_j^{SSVS} = O_j^{SSVS}/(1 + O_j^{SSVS})$,

(b) with posterior odds O_j^{SSVS}

$$O_j^{SSVS} = \frac{\pi^{SSVS}(\boldsymbol{\beta}|Q, g, \boldsymbol{\delta}, \gamma_j = 1, \boldsymbol{\gamma}_{-j})\pi(\gamma_j = 1, \boldsymbol{\gamma}_{-j})}{\pi^{SSVS}(\boldsymbol{\beta}|Q, g, \boldsymbol{\delta}, \gamma_j = 0, \boldsymbol{\gamma}_{-j})\pi(\gamma_j = 0, \boldsymbol{\gamma}_{-j})},$$

and set $\boldsymbol{\gamma}^{(s)} = \boldsymbol{\gamma}^{(s-1)}$.

Step 3: Sample $\boldsymbol{\beta}^{(s)}$ given the respective updated and current states $\boldsymbol{\gamma}^{(s)}$, $a_1^{(s-1)}, \dots, a_{Q-1}^{(s-1)}$ and $g^{(s-1)}$ from full conditional $\frac{\exp\left(\sum_{q=1}^{Q-1} \boldsymbol{y}_q^T \boldsymbol{X}\boldsymbol{\beta}_q\right)}{\mathbf{1}_n^T\left(\mathbf{1}_n + \sum_{q=1}^{Q-1} \exp\left(a_q \mathbf{1}_n + \boldsymbol{X}\boldsymbol{\beta}_q\right)\right)}$ $\exp\left(-\frac{\boldsymbol{\beta}^T \boldsymbol{D}^{-1}\mathcal{I}^{(BH)}(\mathbf{0}_{(Q-1)p_q})\boldsymbol{D}^{-1}\boldsymbol{\beta}}{2Q^2 g}\right)$ based on a Metropolis-Hastings random walk candidate density generator with properties

(a) a candidate value $\boldsymbol{\beta}^{(can)}$ is generated as $\boldsymbol{\beta}^{(can)} \sim N_{(Q-1)p_q}(\boldsymbol{\beta}, t\boldsymbol{D}^*)$, where $\boldsymbol{D}^* = \left(\boldsymbol{D}^{-1}\mathcal{I}^{(BH)}(\mathbf{0}_{(Q-1)p_q})\boldsymbol{D}^{-1}/Q^2 g + \mathcal{I}(\widehat{\boldsymbol{\beta}})\right)^{-1}$ denotes the proposal precision matrix and t the tuning of MCMC procedure determining the jumps of posterior exploration. Furthermore, the precision matrix $\boldsymbol{D}^*$ is a combination of two precision matrices, $\boldsymbol{D}^{-1}\mathcal{I}^{(BH)}(\mathbf{0}_{(Q-1)p_q})\boldsymbol{D}^{-1}/Q^2 g$ and $\mathcal{I}(\widehat{\boldsymbol{\beta}})$, that encapsulates both the information from prior and likelihood, hence the genuine structure of the unknown full conditional for $\boldsymbol{\beta}$. The matrix $\mathcal{I}(\widehat{\boldsymbol{\beta}})$ originates from maximum likelihood as the following

$$\mathcal{I}(\widehat{\boldsymbol{\beta}}_q, \widehat{\boldsymbol{\beta}}_q) = \begin{cases} \mathcal{I}(\widehat{\boldsymbol{\beta}}_q) = \boldsymbol{X}^T \boldsymbol{p}_q(\widehat{a}_q, \widehat{\boldsymbol{\beta}}_q)(\mathbf{1}_n - \boldsymbol{p}_q(\widehat{a}_q, \widehat{\boldsymbol{\beta}}_q))^T \boldsymbol{X} & , q = q' \\ -\boldsymbol{X}^T \boldsymbol{p}_q(\widehat{a}_q, \widehat{\boldsymbol{\beta}}_q)\boldsymbol{p}_{q'}(\widehat{a}_{q'}, \widehat{\boldsymbol{\beta}}_{q'})\boldsymbol{X} & , q \neq q' \end{cases},$$

where $\widehat{a}$ and $\widehat{\boldsymbol{\beta}}$ are the maximum likelihood estimators of full multinomial logistic model (4.9) and preserve the constant dimension since they don't depend on each q-th $\boldsymbol{\gamma}_q$ specific class.

(b) an acceptance-rate $A_{\boldsymbol{\beta}}^{(SSVS)}$ of the proposed move in the log-scale

$$\log(A_{\boldsymbol{\beta}}^{(SSVS)}) = \log\left(\frac{\pi^{SSVS}(\boldsymbol{\beta}^{(can)}|\boldsymbol{a},Q,\boldsymbol{\delta},g,\boldsymbol{\gamma},\boldsymbol{y})}{\pi^{SSVS}(\boldsymbol{\beta}|\boldsymbol{a},Q,g,\boldsymbol{\delta},\boldsymbol{\gamma},\boldsymbol{y})}\frac{q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(can)},Q,\boldsymbol{\delta},g,\boldsymbol{\gamma},\boldsymbol{y})}{q(\boldsymbol{\beta}^{(can)}|\boldsymbol{\beta},Q,\boldsymbol{\delta},g,\boldsymbol{\gamma},\boldsymbol{y})}\right)$$
$$\propto \sum_{q=1}^{Q-1}\boldsymbol{y}_q^T\boldsymbol{X}\boldsymbol{\beta}_q^{(can)} - \log\left\{\boldsymbol{1}_n^T\left(\boldsymbol{1}_n + \sum_{q=1}^{Q-1}\exp\left(a_q\boldsymbol{1}_n + \boldsymbol{X}\boldsymbol{\beta}_q^{(can)}\right)\right)\right\}$$
$$-\sum_{q=1}^{Q-1}\boldsymbol{y}_q^T\boldsymbol{X}\boldsymbol{\beta}_q + \log\left\{\boldsymbol{1}_n^T\left(\boldsymbol{1}_n + \sum_{q=1}^{Q-1}\exp\left(a_q\boldsymbol{1}_n + \boldsymbol{X}\boldsymbol{\beta}_q\right)\right)\right\}$$
$$-\frac{\boldsymbol{\beta}^{T(can)}\boldsymbol{D}^{-1}\mathcal{I}^{(BH)}(\boldsymbol{0}_{(Q-1)p_q})\boldsymbol{D}^{-1}\boldsymbol{\beta}^{(can)}}{2Q^2g}$$
$$+\frac{\boldsymbol{\beta}^{T}\boldsymbol{D}^{-1}\mathcal{I}^{(BH)}(\boldsymbol{0}_{(Q-1)p_q})\boldsymbol{D}^{-1}\boldsymbol{\beta}}{2Q^2g},$$

(c) Set $\boldsymbol{\beta}^{(s)} = \begin{cases}\boldsymbol{\beta}^{(can)} & \text{, accept with probability } A_{\boldsymbol{\beta}}^{(SSVS)},\\ \boldsymbol{\beta} & \text{, reject with probability } 1 - A_{\boldsymbol{\beta}}^{(SSVS)},\end{cases}$

where $q(.)$ denotes the candidate density generator and the respective log-acceptance rate $\log(A_{\boldsymbol{\beta}}^{(SSVS)})$ is reduced due to the symmetry of ratio $q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(can)},Q,g,\boldsymbol{\delta},\boldsymbol{\gamma},\boldsymbol{y})/q(\boldsymbol{\beta}^{(can)}|\boldsymbol{\beta},Q,g,\boldsymbol{\delta},\boldsymbol{\gamma},\boldsymbol{y})$ of random walk.

Step 4: Sample $a_q^{(s)}$, for $q = 1,\ldots,Q-1$, given the respective updated and current states $\boldsymbol{\gamma}^{(s)}$, $\boldsymbol{\beta}^{(s)}$, $\boldsymbol{a}_{-q}^{(s-1)}$ and $g^{(s-1)}$, where $\boldsymbol{a}_{-q}$ is the vector of each class-specific except q-th element a_q, from full conditional $\frac{\exp\left(\boldsymbol{y}_q^T a_q \boldsymbol{1}_n\right)}{\boldsymbol{1}_n^T\left(\boldsymbol{1}_n + \sum_{q=1}^{Q-1}\exp\left(a_q\boldsymbol{1}_n + \boldsymbol{X}\boldsymbol{\beta}_q\right)\right)}$ based on a Metropolis-Hastings random walk candidate density generator with properties

(a) a candidate value $a_q^{(can)}$ is generated as $a_q^{(can)} \sim N(a_q, v_{a_q})$, where v_a denotes the proposed variance of the random walk.

(b) an acceptance-rate $A_{a_q}^{(SSVS)}$ of the proposed move in the log-scale

$$\log(A_{a_q}^{(SSVS)}) = \log\left(\frac{\pi^{SSVS}(a_q^{(can)}|\boldsymbol{a}_{-q},\boldsymbol{\beta},\boldsymbol{y})}{\pi^{SSVS}(a_q|\boldsymbol{a}_{-q},\boldsymbol{\beta},\boldsymbol{y})}\frac{q(a_q|a_q^{(can)},v_{a_q})}{q(a_q^{(can)}|a_q,v_{a_q})}\right)$$
$$\propto \boldsymbol{y}^T a_q^{(can)}\boldsymbol{1}_n - \log\left\{\boldsymbol{1}_n^T\left(\boldsymbol{1}_n + \sum_{q=1}^{Q-1}\exp\left(a_q^{(can)}\boldsymbol{1}_n + \boldsymbol{X}\boldsymbol{\beta}_q\right)\right)\right\}$$
$$-\boldsymbol{y}^T a_q^{(cur)}\boldsymbol{1}_n + \log\left\{\boldsymbol{1}_n^T\left(\boldsymbol{1}_n + \sum_{q=1}^{Q-1}\exp\left(a_q^{(cur)}\boldsymbol{1}_n + \boldsymbol{X}\boldsymbol{\beta}_q\right)\right)\right\},$$

(c) Set $a_q^{(s)} = \begin{cases} a_q^{(can)} & \text{, accept with probability } A_{a_q}^{(SSVS)}, \\ a_q & \text{, reject with probability } 1 - A_{a_q}^{(SSVS)}, \end{cases}$

where the ratio $q(a_q|a_q^{(can)}, v_{a_q})/q(a^{(can)}|a, v_{a_q})$ is cancelled due to symmetry.

Step 5: given the updated states $\boldsymbol{\gamma}^{(s)}$, $\boldsymbol{\beta}^{(s)}$ and $a_1^{(s)}, \ldots, a_{Q-1}^{(s)}$

(**A**) if $g \sim IG\left(\frac{1}{2}, \frac{n}{2}\right)$,
sample $g^{(s)} \sim IG\left(\widehat{\lambda}_{0,g}^{(SSVS)}, \widehat{\lambda}_{1,g}^{(SSVS)}\right)$, where $\widehat{\lambda}_{0,g}^{(SSVS)}$ and $\widehat{\lambda}_{1,g}^{(SSVS)}$ denote respectively the posterior shape and scale of g respectively as

(a) $\widehat{\lambda}_{0,g}^{(SSVS)} = ((Q-1)p_q + 1)/2$,

(b) $\widehat{\lambda}_{1,g}^{(SSVS)} = \frac{1}{2}\left[\boldsymbol{\beta}^T \boldsymbol{D}^{-1} \mathcal{I}^{(BH)}(\mathbf{0}_{(Q-1)p_q})^{-1} \boldsymbol{D}^{-1} \boldsymbol{\beta}/Q^2 + n\right]$,

and set $g^{(s)} = g^{(s-1)}$.

(**B**) if $\pi(g) \propto (1+g)^{-\frac{\alpha}{2}}$, sample $g^{(s)}$ from full conditional $(1+g)^{-\frac{\alpha}{2}} g^{-\frac{(Q-1)p_q}{2}}$ $\exp\left(-\frac{\boldsymbol{\beta}^{\boldsymbol{T}} \boldsymbol{D}^{-1} \mathcal{I}^{(BH)}(\mathbf{0}_{(Q-1)p_q})^{-1} \boldsymbol{D}^{-1} \boldsymbol{\beta}}{2Q^2 g}\right)$ after translating the parameter space of g on log-scale based on a Metropolis-Hastings random walk candidate density generator with properties

(a) a candidate value $g^{(can)}$ is generated as $\log(g^{(can)}) \sim N(\log(g), v_g)$ $\Rightarrow g^{(can)} = \exp(\log(g^{(can)}))$, where v_g denotes the tuning variance which determines the amount of jumps or the acceptance rate.

(b) an acceptance-rate $A_g^{(SSVS)}$ of the proposed move in log-scale

$$\begin{aligned}
\log(A_g^{(SSVS)}) &= \log\left(\frac{\pi^{SSVS}(g^{(can)}|\boldsymbol{\beta}, Q, \boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{y})}{\pi^{SSVS}(g|\boldsymbol{\beta}, Q, \boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{y})} \frac{q(g|g^{(can)}, v_g)}{q(g^{(can)}|g, v_g)} \frac{J}{J^{(can)}}\right) \\
&\propto -\frac{\alpha}{2}\log(1+g^{(can)}) - \frac{(Q-1)p_q}{2}\log(g^{(can)}) \\
&+ \frac{\alpha}{2}\log(1+g) + \frac{(Q-1)p_q}{2}\log(g) \\
&- \frac{\boldsymbol{\beta}^{\boldsymbol{T}} \boldsymbol{D}^{-1} \mathcal{I}^{(BH)}(\mathbf{0}_{(Q-1)p_q})^{-1} \boldsymbol{D}^{-1} \boldsymbol{\beta}}{2Q^2 g^{(can)}} \\
&+ \frac{\boldsymbol{\beta}^{\boldsymbol{T}} \boldsymbol{D}^{-1} \mathcal{I}^{(BH)}(\mathbf{0}_{(Q-1)p_q})^{-1} \boldsymbol{D}^{-1} \boldsymbol{\beta}}{2Q^2 g} \\
&+ \log\left(\frac{1}{g}\right) - \log\left(\frac{1}{g^{(can)}}\right) \qquad (4.14)
\end{aligned}$$

where J denotes the associated Jacobian from transformation on the original scale of g. Notice that the corresponding ratio $q(g|v_g)/q(g^{can}|v_g)$ cancel due to random walk.

$$\text{(c) Set } g^{(s)} = \begin{cases} g^{(can)} & \text{, with probability } A_g^{(SSVS)}, \\ g & \text{, with probability } 1 - A_g^{(SSVS)}, \end{cases}$$

C. Repeat all the steps untill convergence.

4.4 GVS in Typical Multinomial Logistic Setup

Gibbs variable selection Ntzoufras (1999) and Dellaportas et al. (2002) has been a flexible alternative to SSVS popularized most notably for the introduction of latent binary vector in the main model body and the notion of pseudo-priors that allow to jump from one model to another of different dimensions. It also provides consistent model selection that summarizes the variable selection uncertainty through the use of MCMC surpassing the drawbacks of formal Bayesian methods. Although it was initially developed to handle the problem of variable selection in linear regression, there are still challenging motivations to apply the same ideas to generalized linear models and more precisely to multinomial logistic regression relating then to the computation of posterior probabilities, posterior intractability and the multiplicative increase of the model space in terms of all class-specific subsets 2^{pq} given baseline q^*. For these reasons, we highlight and extend in detail a Bayesian variable selection procedure based on GVS of Dellaportas et al. (2002) sharing the objective principles of mixtures of g-priors for Bayesian variable selection in multinomial logistic regression models. We think the current approach suits the problem since the lack of information regarding each best subset of each q-th logit favors the objective Bayesian approach combined with Jeffreys conventionality ideas regarding the model space and the g-priors mixtures. Furthermore, it is automatically treated as an objective Bayesian method as there is no requirement to tune the prior inputs likewise SSVS and is more robust than any usual Bayesian variable selection method based on Laplace approximation that collapse in the case of small sample size. However, similar issues arise within GVS implementation with respect to the intractability of conditionals and especially in the case of $\boldsymbol{\beta}$, again a highly complex Metropolis-Hastings is added to deal with all the correlations and cross-correlations of respective coefficients belonging to the same or different class-specific. Even in that case, we are motivated to create a default Bayesian variable selection procedure that rests firmly on the appealing property of constructing complex structures like multinomial logistic regression, which is verified in the present work. The GVS usually starts assuming for the observed values $\boldsymbol{y}$, a model varying sampling density in multinomial logistic regression for the linear relationship of the polychotomous response $\boldsymbol{Y}$ and the covariates, as the following

$$f(\boldsymbol{y}|\boldsymbol{a},\boldsymbol{\beta},\boldsymbol{\gamma}) = \frac{\exp\left(\sum_{q=1}^{Q-1} \boldsymbol{y}_q^T \left(a_q \mathbf{1}_n + \boldsymbol{X}\boldsymbol{\Gamma}_q \boldsymbol{\beta}_q\right)\right) \exp\left(\sum_{q=1}^{Q-1} \binom{n}{\boldsymbol{y}_q}\right)}{\mathbf{1}_n^T \left(\mathbf{1}_n + \sum_{q=1}^{Q-1} \exp\left(a_q \mathbf{1}_n + \boldsymbol{X}\boldsymbol{\Gamma}_q \boldsymbol{\beta}_q\right)\right)}, \tag{4.15}$$

where $\boldsymbol{\Gamma} = \text{diag}(\boldsymbol{\gamma}) = \text{diag}(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{Q-1})$ is of dimension $(Q-1)p_q$ and the sampling density is affected by binary vector $\boldsymbol{\gamma}$ in difference with SSVS. Notice that the matrix $\boldsymbol{\Gamma}$ incorporates each q-th class-specific $\boldsymbol{\gamma}_q$ which allows to update jointly the model uncertainty across different $Q-1$ baseline logits. The clever incorporation of binary vector $\boldsymbol{\gamma}$ in the sampling density (4.15) in conjunction with a special hierarchical prior specification for $\boldsymbol{\beta}$ allows the corresponding MCMC method to travel among model of different size given that the model dimension is balanced through the pseudo-priors. The concepts of pseudo-priors will not be presented again in this section since they were covered in the previous chapters showing their utility. As it was discussed previously, we adopt again the approach of Paroli and Spezia (2006) in order to update the parameter $\boldsymbol{\beta}$ jointly rather than based on the respective subsets $\boldsymbol{\beta}_\gamma$ and $\boldsymbol{\beta}_{-\gamma}$ for the current configuration of $\boldsymbol{\gamma}$ likewise Ntzoufras et al. (2003). In addition, the hierarchical prior specification for $\boldsymbol{\beta}$ in the framework of g-prior mixtures can be extended for multinomial logistic regression as the following

$$\boldsymbol{\beta}|Q, g, \boldsymbol{\gamma} \sim N_p\left(\boldsymbol{\mu}, \widetilde{\boldsymbol{D}}^{-1}\right), \tag{4.16}$$

with prior precision matrix $\widetilde{\boldsymbol{D}}$

$$\widetilde{\boldsymbol{D}} = \left(\boldsymbol{\Gamma}(\boldsymbol{R}^{(BH)})^{-1}\boldsymbol{\Gamma} + \text{diag}(1-\boldsymbol{\gamma})\frac{1}{\bar{\boldsymbol{s}}^2}\right)^{-1}, \tag{4.17}$$

where $\boldsymbol{\mu} = (1-\boldsymbol{\gamma})\bar{\boldsymbol{\mu}}$ and $\bar{\boldsymbol{\mu}}$, $\bar{\boldsymbol{s}}^{\mathbf{2}}$ are the prior mean and variance inputs respectively obtained from a pilot run for pseudo-priors and $\boldsymbol{R}^{(BH)}$ is again the prior precision matrix based on Bové and Held (2011) generalized g-prior approach. This hierarchical prior specification is better illustrated if we write the joint prior of the partitioned vectors $\boldsymbol{\beta}_\gamma$, $\boldsymbol{\beta}_{-\gamma}$ as the following

$$\pi(\boldsymbol{\beta}_\gamma, \boldsymbol{\beta}_{-\gamma}|Q, g) \propto \begin{cases} N_{p_\gamma}\left(\boldsymbol{\beta}_\gamma \middle| \mathbf{0}_{p_\gamma}, gQ^2\mathcal{I}^{(BH)}(0_{p_\gamma})^{-1}\right), & \boldsymbol{\gamma} = \mathbf{1}_{p_\gamma}, \\ N_{p_{-\gamma}}\left(\boldsymbol{\beta}_{-\gamma} \middle| \boldsymbol{\mu}_{-\gamma}, \widetilde{\boldsymbol{D}}_{-\gamma}\right), & \boldsymbol{\gamma} = \mathbf{0}_{p_{-\gamma}} \end{cases} \tag{4.18}$$

where $\mathcal{I}^{(BH)}(0_{p_\gamma})$ is defined by (4.5) (4.6) and (4.8) and from the above the actual prior of included effects $\boldsymbol{\beta}_\gamma$ is generated from the generalized g-prior independently of pseudopriors which suggests that if we decide to update jointly $\boldsymbol{\beta}_\gamma$, $\boldsymbol{\beta}_{-\gamma}$ into the vector $\boldsymbol{\beta}$, the MCMC procedure will not be altered. This is justified by the fact that the non active parameter vector $\boldsymbol{\beta}_{-\gamma}$ is not involved in the likelihood for the respective values of $\boldsymbol{\gamma}$ and their update is only based exclusively on pseudo-priors.
Usually, before initializing any MCMC procedure, one has to set up carefully a prior specification preferably under the objective Bayesian guidelines of Jeffreys (1961) and Zellner (1986) for Bayesian variable selection in multinomial logistic regression

$$\pi^{GVS}(\boldsymbol{a}, \beta, g, \boldsymbol{\gamma}) = \prod_{q=1}^{Q-1} \pi^{(BH)}(a_q)\pi^{GVS}(\boldsymbol{\beta}|Q, g, \boldsymbol{\gamma})\pi(g)\pi(\boldsymbol{\gamma}), \tag{4.19}$$

where $\pi^{(BH)}(\boldsymbol{a})$ remains the same as in SSVS as (4.12), $\pi^{GVS}(.|Q, g, \boldsymbol{\gamma})$ denotes the hierarchical mixture prior of GVS (4.16) for $\boldsymbol{\beta}$ based on generalized g-prior approach of Bové and Held (2011) , $\pi(\boldsymbol{\gamma})$ is defined by (**??**) and $\pi(g)$ allows to extend the prior formulation in the framework of mixtures of g-priors such as Zellner and Siow (1980) or hyper-g Liang et al. (2008). Alternative hierarchical priors regarding the model space, such as Scott and Berger (2010), are possibly adopted likewise SSVS and we will not enter in details again.
In addition, the GVS algorithm is created based on an MCMC procedure that applies to the joint posterior of model specific parameters and model respectively as the following

$$\pi^{GVS}(\boldsymbol{a}, \boldsymbol{\beta}, g, \boldsymbol{\gamma}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \prod_{q=1}^{Q-1} \pi^{(BH)}(a_q)\pi^{GVS}(\boldsymbol{\beta}|Q, g, \boldsymbol{\gamma})\pi(g)\pi(\boldsymbol{\gamma}), \tag{4.20}$$

where the above product of the sampling density (4.15) and the joint hierarchical prior specification (4.12) maintains the difference among models of different types due to the incorporation of pseudo-priors. It also brings to light important information related to variable selection. Moreover, the involved joint posterior (4.20) remains in an intractable form and it becomes impossible to obtain valuable information regarding the variable selection problem and hence MCMC procedures are priority for such situations. In addition, the GVS allows to create a Markov chain that moves over the intractable joint posterior discovering important regions of high posterior model probabilities, which are of high interest for the variable selection, excluding those with negligible posterior model probability. In this way, the implied MCMC procedure based on GVS recovers a simulated sample indirectly originating from the unknown target joint posterior by updating sequentially over all the full conditionals of all parameters, such as $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, g and $\boldsymbol{\gamma}$. In particular, GVS treats similar problems like the SSVS and the MCMC procedure with additional Metropolis-Hastings steps within Gibbs sampling are outlined because the full conditionals of $\boldsymbol{\beta}$, $\boldsymbol{a}$ are only known up to proportionality constant and the same holds for g if hyper-g is adopted, otherwise a Gibbs step substitutes the previous Metropolis-Hastings step. The MCMC method of GVS algorithm is outlined as the following based on the successive simulations of full conditionals over each parameters and more precisely updating jointly $\boldsymbol{\beta}$ and separately each a_q likewise SSVS

A. Same as in SSVS.

B. For iterations $s = 1, \dots, S$:

Step 1: Same as in SSVS.

Step 2: Sample $\gamma_j^{(s)} \sim Bern\left(\pi_j^{GVS}\right)$, for $j = 1, \dots, (Q-1)p_q$, given the current states of $\boldsymbol{\gamma}_{-j}^{(s-1)}$ $\boldsymbol{\beta}^{(s-1)}$, $\sigma^{2(s-1)}$, $a^{(s-1)}$ and $g^{(s-1)}$

(a) with probability inclusion of j-th covariate $\pi_j^{GVS} = O_j^{GVS}/(1 + O_j^{GVS})$,

(b) with posterior odds

$$O_j^{GVS} = \frac{f(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{\beta}, \gamma_j = 1, \boldsymbol{\gamma}_{-j})\pi^{GVS}(\boldsymbol{\beta}|Q, \boldsymbol{\delta}, g, \gamma_j = 1, \boldsymbol{\gamma}_{-j})\pi(\gamma_j = 1, \boldsymbol{\gamma}_{-j})}{f(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{\beta}, \gamma_j = 1, \boldsymbol{\gamma}_{-j})\pi^{GVS}(\boldsymbol{\beta}|Q, \boldsymbol{\delta}, g, \gamma_j = 0, \boldsymbol{\gamma}_{-j})\pi(\gamma_j = 0, \boldsymbol{\gamma}_{-j})},$$

notice the above expression GVS differs substantially from SSVS due to the presence of sampling density.

and set $\boldsymbol{\gamma}^{(s)} = \boldsymbol{\gamma}^{(s-1)}$.

Step 3: Sample $\boldsymbol{\beta}^{(s)}$ given the respective updated and current states $\boldsymbol{\gamma}^{(s)}$, $a_1^{(s-1)}, \dots, a_{Q-1}^{(s-1)}$ and $g^{(s-1)}$ from full conditional $\frac{\exp\left(\sum_{q=1}^{Q-1} \boldsymbol{y}_q^T \boldsymbol{X}\boldsymbol{\Gamma}_q\boldsymbol{\beta}_q\right)}{\mathbf{1}_n^T\left(\mathbf{1}_n + \sum_{q=1}^{Q-1} \exp\left(a_q\mathbf{1}_n + \boldsymbol{X}\boldsymbol{\Gamma}_q\boldsymbol{\beta}_q\right)\right)}$ $\exp\left(-\frac{(\boldsymbol{\beta}-\boldsymbol{\mu})^T\left(\frac{\boldsymbol{\Gamma}\mathcal{I}^{(BH)}(\mathbf{0}_{(Q-1)p_q})^{-1}\boldsymbol{\Gamma}}{Q^2 g} + \text{diag}(1-\gamma)\frac{1}{\bar{\boldsymbol{s}}^2}\right)(\boldsymbol{\beta}-\boldsymbol{\mu})}{2}\right)$ based on a Metropolis-Hastings random walk with properties

(a) a candidate value $\boldsymbol{\beta}^{(can)}$ is generated as $\boldsymbol{\beta}^{(can)} \sim N_{(Q-1)p_q}(\boldsymbol{\beta}, t\widetilde{\boldsymbol{D}}^*)$, where $\widetilde{\boldsymbol{D}}^* = \left(\boldsymbol{\Gamma}\mathcal{I}^{(BH)}(\mathbf{0}_{(Q-1)p})\boldsymbol{\Gamma}/Q^2 g + \text{diag}(1-\boldsymbol{\gamma})1/\bar{\boldsymbol{s}}^2 + \boldsymbol{\Gamma}\widetilde{\mathcal{I}}(\widehat{\boldsymbol{\beta}}_{\gamma})\boldsymbol{\Gamma}\right)^{-1}$ denotes the proposal precision matrix. Furthermore, the precision matrix $\widetilde{\boldsymbol{D}}^*$ can be thought as the composition of two precision matrices,

$\boldsymbol{\Gamma}\mathcal{I}^{(BH)}(\mathbf{0}_{(Q-1)p})\boldsymbol{\Gamma}/Q^2g + \text{diag}(1-\boldsymbol{\gamma})1/\bar{\boldsymbol{s}}^2$ and $\boldsymbol{\Gamma}\widetilde{\mathcal{I}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}})\boldsymbol{\Gamma}$, that envelopes the prior and likelihood features in the resulting authentic posterior structure of the unknown full conditional for $\boldsymbol{\beta}$. The precision matrix $\widetilde{\mathcal{I}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}})$ is computed using maximum likelihood estimation as the following

$$\widetilde{\mathcal{I}}(\widehat{\boldsymbol{\beta}}_{q|\boldsymbol{\gamma}_q}, \widehat{\boldsymbol{\beta}}_{q|\boldsymbol{\gamma}_q}) = \begin{cases} \widetilde{\mathcal{I}}(\widehat{\boldsymbol{\beta}}_q) = \boldsymbol{X}^T_{\boldsymbol{\gamma}_q}\boldsymbol{p}_q(\widehat{a}_q, \widehat{\boldsymbol{\beta}}_{q|\boldsymbol{\gamma}_q})(\mathbf{1}_n - \boldsymbol{p}_q(\widehat{a}_q, \widehat{\boldsymbol{\beta}}_{q|\boldsymbol{\gamma}_q}))^T\boldsymbol{X}_{\boldsymbol{\gamma}_q} & , q = q' \\ -\boldsymbol{X}^T_{\boldsymbol{\gamma}_q}\boldsymbol{p}_q(\widehat{a}_q, \widehat{\boldsymbol{\beta}}_{q|\boldsymbol{\gamma}_q})\boldsymbol{p}_{q'}(\widehat{a}_{q'}, \widehat{\boldsymbol{\beta}}_{q'|\boldsymbol{\gamma}_{q'}})\boldsymbol{X}_{\boldsymbol{\gamma}_{q'}} & , q \neq q' \end{cases}$$

where $\widehat{a}$ and $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ and are the maximum likelihood estimators of multinomial logistic regression density (4.15) for the respective values of $\boldsymbol{\gamma}$, hence containing all the included class-specific regression coefficients. Such a prior choice is simply based on the assumption that only the included effects of each class-specific regression coefficient must contribute to the posterior.

(b) an acceptance-rate $A^{(GVS)}_{\boldsymbol{\beta}}$ of the proposed move in the log-scale

$$\begin{aligned} &\log(A^{(GVS)}_{\boldsymbol{\beta}}) = \log\left(\frac{\pi^{GVS}(\boldsymbol{\beta}^{(can)}|\boldsymbol{a}, Q, \boldsymbol{\delta}, g, \boldsymbol{\gamma}, \boldsymbol{y})}{\pi^{GVS}(\boldsymbol{\beta}|\boldsymbol{a}, Q, \boldsymbol{\delta}, g, \boldsymbol{\gamma}, \boldsymbol{y})}\frac{q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(can)}, Q, \boldsymbol{\delta}, g, \boldsymbol{\gamma}, \boldsymbol{y})}{q(\boldsymbol{\beta}^{(can)}|\boldsymbol{\beta}, Q, \boldsymbol{\delta}, g, \boldsymbol{\gamma}, \boldsymbol{y})}\right) \\ &\propto \sum_{q=1}^{Q-1}\boldsymbol{y}_q^T\boldsymbol{X}\boldsymbol{\Gamma}_q\boldsymbol{\beta}_q^{(can)} - \log\left\{\mathbf{1}_n^T\left(\mathbf{1}_n + \sum_{q=1}^{Q-1}\exp(a_q\mathbf{1}_n + \boldsymbol{X}\boldsymbol{\Gamma}_q\boldsymbol{\beta}_q^{(can)})\right)\right\} \\ &- \sum_{q=1}^{Q-1}\boldsymbol{y}_q^T\boldsymbol{X}\boldsymbol{\Gamma}_q\boldsymbol{\beta}_q + \log\left\{\mathbf{1}_n^T\left(\mathbf{1}_n + \sum_{q=1}^{Q-1}\exp(a_q\mathbf{1}_n + \boldsymbol{X}\boldsymbol{\Gamma}_q\boldsymbol{\beta}_q)\right)\right\} \\ &- \frac{\left(\boldsymbol{\beta}^{(can)} - \boldsymbol{\mu}\right)^T\widetilde{\boldsymbol{D}}^{*(-1)}\left(\boldsymbol{\beta}^{(can)} - \boldsymbol{\mu}\right)}{2} + \frac{(\boldsymbol{\beta} - \boldsymbol{\mu})^T\widetilde{\boldsymbol{D}}^{*(-1)}(\boldsymbol{\beta} - \boldsymbol{\mu})}{2}, \end{aligned}$$

(c) Set $\boldsymbol{\beta}^{(s)} = \begin{cases} \boldsymbol{\beta}^{(can)} & , \text{ accept with probability } A^{(GVS)}_{\boldsymbol{\beta}}, \\ \boldsymbol{\beta} & , \text{reject with probability } 1 - A^{(GVS)}_{\boldsymbol{\beta}}, \end{cases}$

where $q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(can)}, Q, g, \boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{y})/q(\boldsymbol{\beta}^{(can)}|\boldsymbol{\beta}, Q, g, \boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{y})$ is reduced due to symmetry.

Step 4: Sample $a_q^{(s)}$, for $q = 1, \ldots, Q-1$, given the respective updated and current states $\boldsymbol{\gamma}^{(s)}$, $\boldsymbol{\beta}^{(s)}$ and $g^{(s-1)}$, from full conditional $\frac{\exp\left(\boldsymbol{y}_q^T a_q\mathbf{1}_n\right)}{\mathbf{1}_n^T\left(\mathbf{1}_n + \sum_{q=1}^{Q-1}\exp(a_q\mathbf{1}_n + \boldsymbol{X}\boldsymbol{\Gamma}_q\boldsymbol{\beta}_q)\right)}$ based on a Metropolis-Hastings random walk with properties

(a) a candidate value $a_q^{(can)}$ is generated as $a_q^{(can)} \sim N(a_q, v_{a_q})$, where v_a denotes the proposed variance of the random walk.

(b) an acceptance-rate $A_{a_q}^{(GVS)}$ of the proposed move in the log-scale

$$\log(A_{a_q}^{(GVS)}) = \log\left(\frac{\pi^{GVS}(a_q^{(can)}|\boldsymbol{a}_{-q},\boldsymbol{\beta},\boldsymbol{y},\boldsymbol{\gamma})}{\pi^{GVS}(a_q|\boldsymbol{a}_{-q},\boldsymbol{\beta},\boldsymbol{y},\boldsymbol{\gamma})}\frac{q(a_q^{(cur)}|a_q^{(can)},v_{a_q})}{q(a_q^{(can)}|a_q^{(cur)},v_{a_q})}\right)$$
$$\propto \boldsymbol{y}^T a_q^{(can)}\mathbf{1}_n - \log\left\{\mathbf{1}_n^T\left(\mathbf{1}_n + \sum_{q=1}^{Q-1}\exp\left(a_q^{(can)}\mathbf{1}_n + \boldsymbol{X}\boldsymbol{\Gamma}_q\boldsymbol{\beta}_q\right)\right)\right\}$$
$$- \boldsymbol{y}^T a_q^{(cur)}\mathbf{1}_n + \log\left\{\mathbf{1}_n^T\left(\mathbf{1}_n + \sum_{q=1}^{Q-1}\exp\left(a_q^{(cur)}\mathbf{1}_n + \boldsymbol{X}\boldsymbol{\Gamma}_q\boldsymbol{\beta}_q\right)\right)\right\}.$$

(c) Set $a_q^{(s)} = \begin{cases} a_q^{(can)} & \text{, accept with probability } A_{a_q}^{(GVS)}, \\ a_q & \text{,reject with probability } 1 - A_{a_q}^{(GVS)}, \end{cases}$

where the ratio $q(a_q|a_q^{(can)},v_{a_q})/q(a^{(can)}|a,v_{a_q})$ is cancelled due to symmetry.

Step 5: given the updated states $\boldsymbol{\gamma}^{(s)}$, $\boldsymbol{\beta}^{(s)}$ and $a_1^{(s)},\ldots,a_{Q-1}^{(s)}$

(**A**) if $g \sim IG\left(\frac{1}{2},\frac{n}{2}\right)$,
sample $g^{(s)} \sim IG\left(\widehat{\lambda}_{0,g}^{(GVS)},\widehat{\lambda}_{1,g}^{(GVS)}\right)$, where $\widehat{\lambda}_{0,g}^{(GVS)}$ and $\widehat{\lambda}_{1,g}^{(GVS)}$ denote respectively the posterior shape and scale of g respectively as

(a) $\widehat{\lambda}_{0,g}^{(GVS)} = (p_\gamma + 1)/2$,

(b) $\widehat{\lambda}_{1,g}^{(GVS)} = \frac{1}{2}\left[(\boldsymbol{\beta}-\boldsymbol{\mu})^T\boldsymbol{\Gamma}\mathcal{I}^{(BH)}(\mathbf{0}_{(Q-1)p})\boldsymbol{\Gamma}\left(\boldsymbol{\beta}-\boldsymbol{\mu}\right)/Q^2 + n\right]$,

and set $g^{(s)} = g^{(s-1)}$.

(**B**) if $\pi(g) \propto (1+g)^{-\frac{a}{2}}$, sample $g^{(s)}$ from full conditional $(1+g)^{-\frac{\alpha}{2}}g^{-\frac{p_\gamma}{2}}$ $\exp\left(-\frac{(\boldsymbol{\beta}-\boldsymbol{\mu})^T\boldsymbol{\Gamma}\mathcal{I}^{(BH)}(\mathbf{0}_{(Q-1)p})\boldsymbol{\Gamma}(\boldsymbol{\beta}-\boldsymbol{\mu})}{2Q^2g}\right)$ based on a Metropolis-Hastings random walk with properties

(a) The same as in SSVS.

(b) an acceptance-rate $A_g^{(GVS)}$ of the proposed move in log-scale

$$\log(A_g^{(GVS)}) = \log\left(\frac{\pi^{GVS}(g^{(can)}|\boldsymbol{\beta},Q,\boldsymbol{\delta},\boldsymbol{\gamma},\boldsymbol{y})}{\pi^{GVS}(g|\boldsymbol{\beta},Q,\boldsymbol{\delta},\boldsymbol{\gamma},\boldsymbol{y})}\frac{q(g|g^{(can)},v_g)}{q(g^{(can)}|g,v_g)}\frac{J}{J^{(can)}}\right)$$
$$\propto -\frac{\alpha}{2}\log(1+g^{(can)}) - \frac{p_\gamma}{2}\log(g^{(can)}) + \frac{\alpha}{2}\log(1+g) + \frac{p_\gamma}{2}\log(g)$$
$$- \frac{(\boldsymbol{\beta}-\boldsymbol{\mu})^T\boldsymbol{\Gamma}\mathcal{I}^{(BH)}(\mathbf{0}_{(Q-1)p})\boldsymbol{\Gamma}\left(\boldsymbol{\beta}-\boldsymbol{\mu}\right)}{2Q^2g^{(can)}}$$
$$+ \frac{(\boldsymbol{\beta}-\boldsymbol{\mu})^T\boldsymbol{\Gamma}\mathcal{I}^{(BH)}(\mathbf{0}_{(Q-1)p})\boldsymbol{\Gamma}\left(\boldsymbol{\beta}-\boldsymbol{\mu}\right)}{2Q^2g}$$
$$+ \log\left(\frac{1}{g}\right) - \log\left(\frac{1}{g^{(can)}}\right).$$

where notice that the corresponding ratio $q(g|v_g)/q(g^{can}|v_g)$ cancels also due to random walk.

(c) Set $g^{(s)} = \begin{cases} g^{(can)} & \text{, with probability } A_g^{(GVS)}, \\ g & \text{, with probability } 1 - A_g^{(GVS)}, \end{cases}$

C. Repeat all the steps untill convergence.

4.5 SSVS vs GVS Within Typical Logistic Multinomial Setup

The model selection algorithms SSVS and GVS are appropriate whenever the need takes place for variable selection. However, their construction requires a different hierarchical model and prior structure which should be addressed carefully. Moreover, their implementation must be gauged apriori with caution in some instances. The following summarizes the main parts and features of each algorithm in order to familiarize their use to the interesting reader. These are also found in Ntzoufras (1999) and Dellaportas et al. (2002). To begin with, SSVS assumes

(a) a constant model over the model space, such as likelihood (4.9).

(b) a hierarchical prior construction (4.12) conditional on $\boldsymbol{\gamma}$ and hence on $\boldsymbol{\gamma}_q$ of included and excluded components via small and large prior variances of spike-slab.

(c) The included and excluded components of $\boldsymbol{\gamma}$, through hierarchical prior (4.12) contribute respectively to the update of joint posterior and hence to the rest of full conditionals of parameters.

whereas, GVS

(a) a varying model over the model space, such as likelihood (4.15), such that depending on $\boldsymbol{\gamma}$ and hence on each class-specific $\boldsymbol{\gamma}_q$.

(b) a hierarchical prior construction (4.16) given $\boldsymbol{\gamma}$, and hence $\boldsymbol{\gamma}_q$, configuration decomposed as (4.18) into the main prior part for the included effects $\boldsymbol{\beta}_\gamma$ and the pseudo-prior part for the excluded effects $\boldsymbol{\beta}_{-\gamma}$.

(c) Only the included components of $\boldsymbol{\gamma}$ through the main prior part will contribute respectively to the update of joint posterior and hence to the full conditionals, the excluded components based on pseudo-priors vanish.

Moreover the major differences of SSVS and GVS are the following with regard to their initialized steps

- Posterior odds $O_j^{(SSVS)}$ in SSVS doesn't depend from the model dependence $\boldsymbol{\gamma}$, whereas $O_j^{(GVS)}$ in GVS case is present due to the incorporation in the likelihood.
- The matrix $\mathcal{I}(\widehat{\boldsymbol{\beta}})$ originating from maximum likelihood and involved in the proposal generation for $\boldsymbol{\beta}$ is of constant dimension in constrast with that of GVS.
- The full conditional of each class-specific intercept a_q in SSVS doesn't depend on $\boldsymbol{\gamma}$ in contrast with GVS.
- For the generation of g whether Zellner-Siow or hyper-g is adopted, in SSVS are affected by fixed dimension $(Q-1)p$, whether in GVS are affected only from $p_{\boldsymbol{\gamma}}$ and pseudo-priors disappear, that's why it appears only the first prior component devoted to the main prior specificationin the respective posterior measures or log-acceptance rate $\log(A_g^{(GVS)})$.

4.6 Polya-Gamma Data Augmentation

The posterior of model probabilities and regression coefficients has been recognised as the main computational requirement that has spurred the development of many MCMC methods owing to the intractable form in the framework of Bayesian variable selection in generalized linear models and hence in multinomial logistic models, including that with data augmentation Tanner and Wong (1987).
One of the most popular approach is Albert and Chib (1993) data augmentation initially developed for probit regression that for the first time introduced the notion of latent variables in Bayesian inference. Under this approach, a layer of latent variables are incorporated to convert the intractable likelihood into standard known linear model results, and then a sequentially updating of parameters is permitted through Gibbs sampling.
Two decades later, Polson et al. (2013) proposed an analogous method for logistic regression, where they introduced a new class of distributions, namely, Polya-Gamma, which are expressed as an infinite convolution of independent Gamma distributions. The main core of this approach is to parametrize binomial likelihoods in terms of log-odds written as mixture of normals with Polya-Gamma distributions, while their approach is enriched with two novel satisfactory results. The first concerns the computational efficiency of the method which accelerates the convergence owing to the simply closed

form expressions of Laplace transform and the second entails the construction of a general class for Polya-Gamma distribution after translating a sub-class of Polya-Gamma distribution through an exponential tilting. Direct consequences of the previous results, were the tractability of Polya-Gamma moments via the Laplace transform with applications in missing data such as expectation maximization Durante et al. (2018) and their connection to Rieman-Zeta distributions. The versatility of the underlying method rests firmly on the binomial likelihood which is involved also to other statistical models such as the negative binomial regression Polson et al. (2013) and Choi and Román (2017), non linear mixed-effects and spatial for count data Linderman et al. (2015) and hence to multinomial logistic regression models Polson et al. (2013), which is the main topic of this thesis. Moreover, theorical results on uniform ergodicity of Polya-Gamma data augmentation are derived by Choi et al. (2013)
Since the seminal paper of Albert and Chib (1993), the Bayesian community was on the demand for a similar algorithm in logistic regression despite the excessive efforts in scientific research. Although these attempts tried to reproduce the work of Albert and Chib (1993) in logistic regression framework, they haven't accomplished it yet, at least in depth. In particular, the methods of the respective approaches seem more complex than that of Albert and Chib and some simple versions of them are usually inaccurate or misleading; see for example Holmes and Held (2006) and Frühwirth-Schnatter (2016). Using a somewhat different approach, Polson et al. (2013) developed a direct approach of the algorithm of Albert and Chib in logistic regression models. Even though, there seems to be a sort of correspondence between the two methods, the two differ substantially due to the mixture construction of latent variables. For example, the approach of Albert and Chib is a location mixture, while the former approach is a scale mixture. The approach of Polson et al. (2013) introduces a layer of latent variables that are mixtures of normals with independent Polya-Gamma precision terms. In this approach, a random variable ω is said to follow a Polya-Gamma distribution with parameters $b > 0$ and $\eta \in \mathbb{R}$, denoted $\omega \sim PG(b, \eta)$, if

$$\omega \overset{d}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{G_k}{(k - \frac{1}{2})^2 + \frac{\eta^2}{4\pi^2}},$$

where $G_k \sim G(b, 1)$ are independent Gamma random variables distributed with parameters b and 1. In addition, Polson et al. (2013) established two important definitions regarding $PG(b, \eta)$ probability density function. The first circumvents the main core of this approach as they developed the fundamental identity between a typical and a

Polya-Gamma model, for $\zeta \in \mathbb{R}$ and $\eta \in \mathbb{R}$

$$\frac{(\exp(\eta))^{\zeta}}{(1+\exp(\eta))^{b}} = 2^{-b}\exp(k\eta) \int_{0}^{\infty} \exp\left(-\frac{\eta^2\omega}{2}\right)\pi(\omega|b,0)d\omega, \tag{4.21}$$

where $k = \zeta - \frac{b}{2}$ and $\omega|b,0 \sim PG(b,0)$ denotes the Polya-Gamma density with parameters b and 0. The second concerns the conditional distribution $\pi(\omega|b,\eta) \sim PG(b,\eta)$ which arises from an exponential tilting of the $PG(b,0)$

$$\pi(\omega|b,\eta) = \frac{\exp\left(-\frac{\eta^2\omega}{2}\right)\pi(\omega|b,0)}{\mathbb{E}_{\omega}\left\{\exp\left(-\frac{\eta^2\omega}{2}\right)\right\}} = \frac{\exp\left(-\frac{\eta^2\omega}{2}\right)\pi(\omega|b,0)}{\int_0^{\infty}\exp\left(-\frac{\eta^2\omega}{2}\right)\pi(\omega|b,0)d\omega}, \tag{4.22}$$

where the expectation $\mathbb{E}_{\omega}(.)$ is taken with respect to the Polya-Gamma density $PG(b,0)$. In particular, they demonstrate that the conditional distribution (4.22) can be expressed in terms of infinite convolution of Gamma's through the Laplace transform of $PG(b,\eta)$ which coincides with the definition (**??**). A key feature that solves the subsequent calculations in the previous definition, was the Laplace transform of $PG(b,0)$ which results in a special case of the general class of $PG(b,\eta)$ equal to $\cosh^{-b}\left(\sqrt{\frac{t}{2}}\right)$ in conjunction with the Weierstrass factorization theorem. In this way, they took advantage of $PG(b,0)$ Laplace's transform and conditional distribution (4.22) in order to result to the main identity (4.21)

$$\begin{aligned}\frac{(\exp(\eta))^{\zeta}}{(1+\exp(\eta))^{b}} &= \frac{(\exp(\eta))^{\zeta}}{\left(2\cosh\left(\frac{\eta}{2}\right)\exp\left(\frac{\eta}{2}\right)\right)^{b}} = 2^{-b}\frac{(\exp(\eta))^{\zeta-\frac{b}{2}}}{\left(\cosh\left(\frac{\eta}{2}\right)\right)^{b}} \\ &= 2^{-b}\mathbb{E}_{\omega}\left\{\exp\left(-\frac{\eta^2\omega}{2}\right)\right\} = 2^{-b}\exp(k\eta) \int_{0}^{\infty}\exp\left(-\frac{\eta^2\omega}{2}\right)\pi(\omega|b,0)d\omega.\end{aligned}$$

Furthermore, the Laplace transform of the general family $PG(b,c)$ can be computed in closed form

$$\mathbb{E}_{\omega}\left\{\exp(-\omega t)\right\} = \frac{\cosh^{b}\left(\frac{\eta}{2}\right)}{\cosh^{b}\left(\sqrt{\frac{\frac{\eta^2}{2}+t}{2}}\right)},$$

$$\mathbb{E}_{\omega}(\omega) = \frac{b}{2\eta}\left(\frac{\exp(b)-1}{\exp(b)+1}\right),$$

$$\mathbb{E}_{\omega}(\omega^2) = b\left((b+1)\left(\frac{\exp(b)-1}{\exp(b)+1}\right)^2 + \frac{1}{4\eta^2} - \frac{1}{8}\left(\frac{\eta^2}{4}\right)^{-\frac{3}{2}}\right).$$

Furthermore, the variance of a Polya-Gamma class can still be calculated based on the above quantities as the following

$$\begin{aligned} Var(\omega) &= \mathbb{E}_\omega(\omega^2) - \mathbb{E}_\omega(\omega)^2 \\ &= b\left((b+1)\left(\frac{\exp(b)-1}{\exp(b)+1}\right)^2 + \frac{1}{4\eta^2} - \frac{1}{8}\left(\frac{\eta^2}{4}\right)^{-\frac{3}{2}}\right) - \left(\frac{b}{2\eta}\left(\frac{\exp(b)-1}{\exp(b)+1}\right)\right)^2. \end{aligned}$$

On the contrary, the main intuition behind the identity (4.21), is the equivalence of writing binomial likelihoods up to a mixture of normals over Polya-Gamma distributions from which a different model results each time according to the specified pair of ζ, b. The logistic regression likelihood is a special case of this identity for $\eta = \eta_i(a, \boldsymbol{\beta})$, $\zeta = y_i$, $b = 1$. In this way, the Polya-Gamma data augmentation scheme is interpreted as a data contribution of one data point y_i, which is equivalent to an augmented data pair ω_i, y_i as the following does

$$\begin{aligned} f(y_i|a,\boldsymbol{\beta}) = 2^{-b}\exp\left(k_i\eta_i(a,\boldsymbol{\beta})\right) & \int_0^\infty \exp\left(-\frac{\eta_i(a,\boldsymbol{\beta})^2\omega_i}{2}\right)\pi(\omega_i|b,0)d\omega_i \\ = & \int_0^\infty \exp\left\{-\frac{\omega_i}{2}(z_i - \eta_i(a,\boldsymbol{\beta}))^2\right\}\pi(\omega_i|b,0)d\omega_i \\ = & \int_0^\infty f(z_i|a,\boldsymbol{\beta},\omega_i)\pi(\omega_i|b,0)d\omega_i, \end{aligned}$$

where $f(.|a, \boldsymbol{\beta}, \omega_i)$ denotes the Gaussian density with observed z_i with unknown precision terms ω_i mixed with $PG(.|b, 0)$ prior density for ω_i, $z_i = \frac{k_i}{\omega_i}$, $k_i = y_i - \frac{1}{2}$.
Next, all the available information regarding the sample can be obtained in an equivalent way from the augmented likelihood factorization of n data pairs ω_i, y_i, for $i = 1, \ldots, n$, based on

$$f(\boldsymbol{y}|a,\boldsymbol{\beta}) = \prod_{i=1}^{n} \int_0^\infty f(z_i,|a,\boldsymbol{\beta},\omega_i)\pi(\omega_i|b,0)d\omega_i,$$

which means that the information will give the same result if the typical likelihood has been taken into consideration. Based on (4.23), Polson and Scott described a flexible data augmentation scheme applied via Gibbs sampling towards the same approach. This would be obtained for linear models if a multivariate normal prior is adopted for $\boldsymbol{\beta}$ and an improper prior is adopted for a, but this time the Gaussian likelihood depends on the observables $\boldsymbol{z}$, and consequently on $\boldsymbol{y}$, $\boldsymbol{\omega}$ which allows first to sample the parameters of interest $\boldsymbol{\beta}$, α from known full conditionals, and then the n layers of Polya-Gamma using the result of the conditional distribution (4.22). Finally, we point out that the full description of Polya-Gamma random variates generation are highlighed in Polson et al. (2013), and is beyond the scopes of this thesis. The Polya-Gamma random variates are generated through Bayes Logit package in R programming language. To conclude, in the next sections we discuss how the idea of data augmentation incorporates into the problem of Bayesian variable selection and we highlight the benefits of using the Bayesian variable selection algorithms, such as SSVS and GVS for augmented multinomial logistic regressions, combined with latent variables.

4.7 SSVS in Augmented Multinomial Logistic Setup

SSVS has been recognised one of the greatest tools in the history of Bayesian modelling as it was the first procedure that introduced the concepts of Gibbs sampling in order to avoid the computation of posterior probabilities and the exhaustive enumeration of model space for the problem of variable selection, promising a vast number of research publications in this domain within the advent of MCMC methods Gilks et al. (1996). Despite its direct implementation in the linear regression framework, similar thoughts in generalized linear models are based on the idea that the full conditionals are not available in closed forms forcing the sampler into cumbersome Metropolis-Hastings sampling. Especially in multinomial logistic regression the same issue becomes more painful due to the extreme model complexity, hence researchers pursue flexible alternatives through the device of data augmentation. Recently, Polson et al. (2013) introduced a clever data augmentation scheme which approximated binomial likelihoods parametrized in log-odds, by introducing Polya-Gamma latent variates amenable to Gibbs sampling and familiar linear model results, providing similar extensions also to negative binomial and multinomial logistic regression. More precisely, in the case of multinomial logistic regression, they took advantage basically of three main facts: i) the mixture of normal densities with Polya-Gamma random variables, ii) the marginal distributions of the observed response belonging to q-th class-specific $\boldsymbol{y}_q$, follow a binomial distribution and iii) the conditional probabilities of each class-specific were modified in such a way that were obtained for a typical logistic regression, in order to express the typical multinomial logistic model as many individual augmented logistic regression models as the existing classes given the baseline. In particular, they isolated the information contained in the authentic sampling density considering the conditional sampling density of each specific-class regression coefficients for the respective observed values of response $\boldsymbol{y}_q$ with respect to the rest. This allowed the prior specification for each class-specific regression coefficients, to take part in each successive step by extracting successfully an equivalent sample from the unknown joint posterior of the typical multinomial logistic regression if the unknown target posterior was available and this strategy is commonly referred to as a nested Gibbs sampler.

In this section, we present in details a mixed Bayesian variable selection procedure that couples the ideas of George and McCulloch (1993) and Polson et al. (2013), extended for multinomial logistic regression framework. More precisely, we introduce a new SSVS method that incorporates the Polya-Gamma latent variables in order to approximate the Bayesian variable selection uncertainty for multinomial logistic regression, by converting the respective intractable sampling density into a likelihood of convenience as those of linear regression models, via a $Q-1$ nested Gibbs sampler. By this way, we believe that the proposed method is compliant to the objective Bayesian rules since it frees the procedure from any drawbacks of the MCMC including those regarding tuning and proposals. The non available information across the different $Q-1$ baseline logits regarding which subset is the most appropriate for each respective baseline logit, supports the use of default objective Bayesian specification based on g-prior Zellner (1986). Moreover, our method is completed by a detailed prior specification based on Zellner's g prior design by adopting for each class-specific regression coefficients, a reduced Zellner g-prior representative as resulting from the joint g-prior structure that encapsulates the variance-covariance and covariance of the same $q = q'$ or different $q \neq q'$ multinomial logistic regression counterparts. As a consequence, we extend also the Bayesian variable selection problem with mixtures of g-priors such as Zellner-Siow and hyper-g for multinomial logistic regression, but under the linear model perspective. Before presenting the main exposition of the proposed SSVS, it is essential to introduce the reader who is interesting to some basic notions that will help him understand better the structure of the augmented SSVS. To begin with, consider the typical multinomial logistic regression likelihood as proportional to each binomial likelihood of the observed response $\boldsymbol{y}_q$ belonging to q-th class-specific with respective probabilities as the following

$$
\begin{aligned}
& f(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{q=1}^{Q-1} \binom{n}{y_{i,q}} p_{i,q}(a_q, \boldsymbol{\beta}_q)^{y_{i,q}} \propto \prod_{i=1}^{n} p_{i,q}(a_q, \boldsymbol{\beta}_q)^{y_{i,q}} (1 - p_q(a_q, \boldsymbol{\beta}_q))^{1-y_{i,q}} \\
& = \left(\boldsymbol{p}_q(a_q, \boldsymbol{\beta}_q)^T\right)^{\boldsymbol{y}_q} (\mathbf{1}_n - \boldsymbol{p}_q(a_q, \boldsymbol{\beta}_q))^{\mathbf{1}_n - \boldsymbol{y}_q} = f(a_q, \boldsymbol{\beta}_q|\boldsymbol{y}_q, \boldsymbol{\beta}_{-q}, \boldsymbol{a}_{-q}),
\end{aligned}
$$

where $f(.|\boldsymbol{y}_q, \boldsymbol{\beta}_{-q}, \boldsymbol{a}_{-q})$ denotes the conditional likelihood of q-th class-specific coefficients a_q and $\boldsymbol{\beta}_q$ given the rest of regression coefficients and observed values $\boldsymbol{y}_q$, then the probabilities belonging to q-th class-specific given baseline q^* are modified as

$$p_{i,q}(a_q, \boldsymbol{\beta}_q) = \frac{\exp\left(a_q + \boldsymbol{x}_i\boldsymbol{\beta}_q\right)}{1 + \sum_{q=1}^{Q-1} \exp\left(a_q + \boldsymbol{x}_i\boldsymbol{\beta}_q\right)} = \frac{\exp\left(a_q + \boldsymbol{x}_i\boldsymbol{\beta}_q\right)}{\exp\left(a_q + \boldsymbol{x}_i\boldsymbol{\beta}_q\right) + 1 + \sum_{q\neq q'} \exp\left(a_{q'} + \boldsymbol{x}_i\boldsymbol{\beta}_{q'}\right)}$$
$$= \frac{\exp\left(\widetilde{\eta}_{i,q}(a_q, \boldsymbol{\beta}_q)\right)}{1 + \exp\left(\widetilde{\eta}_{i,q}(a_q, \boldsymbol{\beta}_q)\right)}$$

where $\widetilde{\eta}_{i,q}(a_q, \boldsymbol{\beta}_q) = a_q + \boldsymbol{x}_i\boldsymbol{\beta}_q - C_{i,q}$ and $C_{i,q} = \log(1 + \sum_{q\neq q'} \exp\left(a_{q'} + \boldsymbol{x}_i\boldsymbol{\beta}_{q'}\right))$; see for more details Holmes and Held (2006) and Polson et al. (2013). In this way, the basic insight behind each conditional likelihood is to express it as a q-th individual logistic model given baseline class with the above conditional probabilities, which will allow the contribution between one data point $y_{i,q}$ given q-th class and augmented data pair $\omega_{i,q}$, $y_{i,q}$ via Polya-Gamma data augmentation identity (4.21) as the following shows

$$\begin{aligned}
f(a_q, \boldsymbol{\beta}_q | y_{i,q}, \boldsymbol{\beta}_{-q}, \boldsymbol{a}_{-q}) &= 2^{-b} \exp\left(k_{i,q}\widetilde{\eta}_{i,q}(a_q, \boldsymbol{\beta}_q)\right) \\
&\quad \int_0^\infty \exp\left(-\frac{\widetilde{\eta}_{i,q}(a_q, \boldsymbol{\beta}_q))^2\omega_{i,q}}{2}\right)\pi(\omega_{i,q}|b, 0)d\omega_{i,q} \\
&\propto \int_0^\infty \exp\left\{-\frac{\omega_{i,q}}{2}(z_{i,q} - \widetilde{\eta}_{i,q}(a_q, \boldsymbol{\beta}_q))^2\right\}\pi(\omega_{i,q}|b, 0)d\omega_{i,q} \\
&= \int_0^\infty f(z_{i,q}|a_q, \boldsymbol{\beta}_q, \omega_{i,q})\pi(\omega_{i,q}|b, 0)d\omega_{i,q},
\end{aligned}$$

where $f(.|a_q, \boldsymbol{\beta}_q, \omega_{i,q})$ denotes the Gaussian density with observed $z_{i,q}$ with unknown precision terms $\omega_{i,q}$ with respect to $PG(.|b, 0)$ prior density for $\omega_{i,q}$, $z_{i,q} = \frac{k_{i,q}}{\omega_{i,q}}$, $k_{i,q} = y_{i,q} - \frac{1}{2}$.

Next, all information included in each respective conditional likelihood regarding the $\boldsymbol{y}_q$ observables for fixed class-specific q, can be summarized from the augmented likelihood factorization of n data pairs $\omega_{i,q}, y_{i,q}$, for $i = 1, \ldots, n$, based on Polya-Gamma mixture identity

$$f(a_q, \boldsymbol{\beta}_q | \boldsymbol{y}_q, \boldsymbol{\beta}_{-q}, \boldsymbol{a}_{-q}) = \prod_{i=1}^{n} \int_0^\infty f(z_{i,q}|a_q, \boldsymbol{\beta}_q, \omega_{i,q})\pi(\omega_{i,q}|b, 0)d\omega_{i,q}.$$

In addition, all the available information of the sample regarding the typical multinomial logistic sampling density can be obtained by the factorization of each conditional likelihood regarding the $\boldsymbol{y}_q$ observables for varying class-specific q, of n data sampling points $\omega_{i,q}, y_{i,q}$, for $i = 1, \ldots, n$ and $q = 1, \ldots, Q-1$ based on Polya-Gamma latent representation

$$
\begin{aligned}
f(\boldsymbol{y}|\boldsymbol{a},\boldsymbol{\beta}) &= \prod_{q=1}^{Q-1} f(a_q,\boldsymbol{\beta}_q|\boldsymbol{y}_q,\boldsymbol{\beta}_{-q},\boldsymbol{a}_{-q}) \\
&= \prod_{q=1}^{Q-1}\prod_{i=1}^{n} \int\limits_{0}^{\infty} f(z_{i,q}|a_q,\boldsymbol{\beta}_q,\omega_{i,q})\pi(\omega_{i,q}|b,0)d\omega_{i,q},
\end{aligned}
$$

which means that the information will result the same if we had in our disposal the typical multinomial logistic likelihood. In this way, we may write the augmented multinomial logistic regression model as the following

$$
f(\boldsymbol{y},\boldsymbol{\omega}|a,\boldsymbol{\beta}) = \prod_{q=1}^{Q-1}\prod_{i=1}^{n} f(z_{i,q}|a_q,\boldsymbol{\beta}_q,\omega_{i,q})\pi(\omega_{i,q}|b,0). \tag{4.23}
$$

Moreover, any Bayesian method with MCMC like SSVS needs a suitable prior specification for implementing it carefully and this cannot be other from the joint hierarchical prior specification (4.12) based on objective approach of Jeffreys (1961) and Bové and Held (2011) for parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, g and $\boldsymbol{\gamma}$. We illustrated this prior specification in the previous section of SSVS for typical multinomial logistic regression and hence it will not be mentioned again in details. Regarding its implementation, it seems that SSVS inherits the fixed dimensionality like authentic SSVS as the $\boldsymbol{\gamma}$ is not present in the augmented sampling density (4.23), accelerating the convergence and identifiability of high posterior model probabilities when the sampling (4.23) is multiplied with the hierarchical prior specification (4.12). Thus, the joint variable selection and parameter uncertainty can be described in the resulting joint posterior as the following

$$
\begin{aligned}
\widetilde{\pi}^{SSVS}(\boldsymbol{a},\boldsymbol{\beta},g,\boldsymbol{\gamma}|\boldsymbol{y},\boldsymbol{\omega}) \propto & \prod_{q=1}^{Q-1}\prod_{i=1}^{n} f(z_{i,q}|a_q,\boldsymbol{\beta}_q,\omega_{i,q})\pi(\omega_{i,q}|b,0) \\
& \pi^{(BH)}(a_q)\pi^{SSVS}(\boldsymbol{\beta}|Q,g,\boldsymbol{\delta},\boldsymbol{\gamma})\pi(g)\pi(\boldsymbol{\gamma}),
\end{aligned} \tag{4.24}
$$

where in the above it is evident that the latent variables are unseen and that is the main reason we did not consider them explicitly in the main prior specification, whereas they were considered known and contributed to the joint posterior only after seen the complete data $\boldsymbol{y}$, $\boldsymbol{\omega}$.

By this way, the joint model and parameter uncertainty is updated given the complete data $\boldsymbol{y}$, $\boldsymbol{\omega}$ and important aspects of variable selection are summarized, even though the joint posterior (4.24) remains intractable employing the use of MCMC and more precisely the Gibbs sampler. In particular, the respective MCMC based on Polya-Gamma data augmentation offers the computational advantage of recovering the full conditionals of $\boldsymbol{\beta}$ and $\boldsymbol{a}$ in closed forms by considering each time the respective class-specific regression coefficients $\boldsymbol{\beta}_q$ and intercepts a_q for a fixed q-th class-specific in comparison with the typical multinomial logistic regression, which rests only on Metropolis-Hastings sampling within Gibbs sampler. The current SSVS is an immediate extension of George and McCulloch (1993) in multinomial logistic regression and of SSVS with Polya-Gamma data augmentation presented in the previous chapter for Bayesian variable selection in logistic regression with the only difference that each a_q and $\boldsymbol{\beta}_q$ are updated through $N_n\left(\boldsymbol{z}_q|\tilde{\boldsymbol{\eta}}_q(a_q, \boldsymbol{\beta}_q), \boldsymbol{\Omega}_q^{-1}\right)$ for $\boldsymbol{z}_q$, where $\boldsymbol{\Omega}_q = \text{diag}(\boldsymbol{\omega}_q)$, hence the full conditionals look similarly to those of a linear model. The full conditional of $\boldsymbol{\gamma}$, can be obtained respectively by considering each respective full conditional of $\boldsymbol{\gamma}_q$ since the augmented likelihood does not take part in the update.

Finally, regarding the mixtures of g-priors, if Zellner-Siow or hyper-g are adopted they lead to Gibbs sampler or Metropolis-Hastings step respectively.

In addition, in order to illustrate better this exposition of ideas, consider the joint posterior density (4.24) for fixed q class-specific as the following

$$\tilde{\pi}^{SSVS}(a_q, \boldsymbol{\beta}_q, \boldsymbol{\gamma}_q, \boldsymbol{a}_{-q}, \boldsymbol{\beta}_{-q}, g, \boldsymbol{\gamma}_{-q}|\boldsymbol{y}_q, \boldsymbol{\omega}_q) \propto \prod_{i=1}^{n} f(z_{i,q}|a_q, \boldsymbol{\beta}_q, \omega_{i,q})\pi(\omega_{i,q}|b, 0)\pi^{(BH)}(a_q)$$
$$\pi^{SSVS}(\boldsymbol{\beta}|Q, g, \boldsymbol{\delta}, \boldsymbol{\gamma})\pi(g)\pi(\boldsymbol{\gamma}), \qquad (4.25)$$

in that case, a nested Gibbs sampler based on the full conditionals of each class-specific model parameters a_q, $\boldsymbol{\beta}_q$, $\boldsymbol{\gamma}_q$ and g is described as follows

A. Set initial values $\boldsymbol{\gamma}_1^{(0)}, \ldots, \boldsymbol{\gamma}_{Q-1}^{(0)}$, $\boldsymbol{\beta}_1^{(0)}, \ldots, \boldsymbol{\beta}_{Q-1}^{(0)}$, $a_1^{(0)}, \ldots, a_{Q-1}^{(0)}$, $\omega_1^{(s-1)}, \ldots, \omega_{Q-1}^{(s-1)}$ and $g^{(0)}$. For fixed $g = n$, delete **Step 6**.

B. For iterations $s = 1, \ldots, S$:

C. For specific-class $q = 1, \ldots, Q-1$:

Step 1: Set current values equal to $\boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_1^{(s-1)}, \ldots, \boldsymbol{\gamma}_{Q-1} = \boldsymbol{\gamma}_{Q-1}^{(s-1)}$, $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^{(s-1)}, \ldots,$ $\boldsymbol{\beta}_{Q-1} = \boldsymbol{\beta}_{Q-1}^{(s-1)}$, $a_1 = a_1^{(s-1)}, \ldots, a_{Q-1} = a_{Q-1}^{(s-1)}$ and $g = g^{(s-1)}$.

Step 2: Sample $\gamma_{j,q} \sim Bern\left(\tilde{\pi}_{j,q}^{SSVS}\right)$, for $j = 1, \ldots, p_q$, given the current states of $\boldsymbol{\gamma}_{-j,q}^{(s-1)}$, $\boldsymbol{\gamma}_{-q}^{(s-1)}$ $\boldsymbol{\beta}_q^{(s-1)}$ and $g^{(s-1)}$, where $\boldsymbol{\gamma}_{-j,q}$ are the components of $\boldsymbol{\gamma}_q$ except element $\boldsymbol{\gamma}_{j,q}$, $\boldsymbol{\gamma}_{-q}$ are the components of $\boldsymbol{\gamma}$ excluding element $\boldsymbol{\gamma}_q$ and $\boldsymbol{\beta}_{-q}$ are the components of $\boldsymbol{\beta}$ excluding element $\boldsymbol{\beta}_q$

(a) with probability inclusion of j-th covariate given q-th class-specific $\widetilde{\pi}_{j,q}^{SSVS} = \widetilde{O}_{j,q}^{SSVS} / \left(1 + \widetilde{O}_{j,q}^{SSVS}\right)$,

(b) with posterior odds $\widetilde{O}_{j,q}^{SSVS}$

$$\widetilde{O}_{j,q}^{SSVS} = \frac{\widetilde{\pi}^{SSVS}(\boldsymbol{\beta}_q | \boldsymbol{\beta}_{-q}, Q, \boldsymbol{\delta}, g, \gamma_{j,q} = 1, \boldsymbol{\gamma}_{-j,q}, \boldsymbol{\gamma}_{-q}) \pi(\gamma_{j,q} = 1, \boldsymbol{\gamma}_{-j,q})}{\widetilde{\pi}^{SSVS}(\boldsymbol{\beta}_q | \boldsymbol{\beta}_{-q}, Q, \boldsymbol{\delta}, g, \gamma_{j,q} = 0, \boldsymbol{\gamma}_{-j,q}, \boldsymbol{\gamma}_{-q}) \pi(\gamma_{j,q} = 0, \boldsymbol{\gamma}_{-j,q})},$$

where $\boldsymbol{\beta}_q | \boldsymbol{\beta}_{-q}, Q, \boldsymbol{\delta}, g, \boldsymbol{\gamma_q}, \boldsymbol{\gamma}_{-q}$ is a multivariate normal distribution with prior mean $\boldsymbol{\mu}_{\beta_q}^{(SSVS)}$ and variance-covariance matrix $\boldsymbol{V}_{\beta_q}^{(SSVS)}$ defined

$$\boldsymbol{\beta}_q | \boldsymbol{\beta}_{-q}, g, Q, \boldsymbol{\gamma_q}, \boldsymbol{\gamma}_{-q} \sim N_{p_q}\left(\boldsymbol{\mu}_{\beta_q}^{(SSVS)}, g\boldsymbol{V}_{\beta_q}^{(SSVS)}\right) \tag{4.26}$$

(a) $\boldsymbol{\mu}_{\beta_q}^{(SSVS)} = \boldsymbol{V}_{\beta_q}\left(\sum_{q \neq q'} \boldsymbol{D}_q^{-1} \mathcal{I}^{(BH)}(\mathbf{0}_{p_q}, \mathbf{0}_{p_{q'}}) \boldsymbol{D}_{q'}^{-1} \boldsymbol{\beta}_{q'} / Q^2\right)$,

(b) $\boldsymbol{V}_{\beta_q}^{(SSVS)} = \left(\boldsymbol{D}_q^{-1} \mathcal{I}^{(BH)}(\mathbf{0}_{p_q}, \mathbf{0}_{p_q}) \boldsymbol{D}_q^{-1} / Q^2\right)^{-1}$

and set $\boldsymbol{\gamma}_q^{(s)} = \boldsymbol{\gamma}_q^{(s-1)}$.

Step 3: Sample $\boldsymbol{\beta}_q^{(s)} \sim N_{p_q}\left(\widehat{\boldsymbol{\mu}}_{\beta_q}^{(SSVS)}, \widehat{\boldsymbol{V}}_{\beta_q}^{(SSVS)}\right)$, given the respective updated and current states $\boldsymbol{\gamma}_q^{(s)}$, $\boldsymbol{\gamma}_{-q}^{(s-1)}$, $a_q^{(s-1)}$, $a_{-q}^{(s-1)}$, $\boldsymbol{\omega}_q^{(s-1)}$ and $g^{(s-1)}$, where $\widehat{\boldsymbol{\mu}}_{\beta_q}^{(SSVS)}$ and $\widehat{\boldsymbol{V}}_{\beta_q}^{(SSVS)}$ denote the posterior mean and variance-covariance matrix of $\boldsymbol{\beta}$ defined respectively as

(a) $\widehat{\boldsymbol{\mu}}_{\beta_q}^{(SSVS)} = \widehat{\boldsymbol{V}}_{\beta_q}\left(g\boldsymbol{L}_q^{(SSVS)} + \sum_{q \neq q'} \boldsymbol{D}_q^{-1} \mathcal{I}^{(BH)}(\mathbf{0}_{p_q}, \mathbf{0}_{p_{q'}}) \boldsymbol{D}_{q'}^{-1} \boldsymbol{\beta}_{q'} / Q^2\right)$,
where $\boldsymbol{L}_q^{(SSVS)} = \boldsymbol{X}^T \boldsymbol{\Omega}_q \boldsymbol{C}_q + \boldsymbol{X}^T \boldsymbol{\Omega}_q \boldsymbol{z}_q - a_q \boldsymbol{X}^T \boldsymbol{\Omega}_q \mathbf{1}_n$

(b) $\widehat{\boldsymbol{V}}_{\beta_q}^{(SSVS)} = \left(g\boldsymbol{X}^T \boldsymbol{\Omega}_q \boldsymbol{X} + \boldsymbol{D}_q^{-1} \mathcal{I}^{(BH)}(\mathbf{0}_{p_q}, \mathbf{0}_{p_q}) \boldsymbol{D}_q^{-1} / Q^2\right)^{-1}$,

and set $\boldsymbol{\beta}_q^{(s)} = \boldsymbol{\beta}_q^{(s-1)}$.

Step 4: Sample $a_q^{(s)} \sim N\left(\widehat{\mu}_{a_q}^{(SSVS)}, \widehat{\sigma}_{a_q}^{2(SSVS)}\right)$, given the updated and current states respectively $\boldsymbol{a}_{-q}^{(s-1)}$, $\boldsymbol{\beta}_q^{(s-1)}$, $\boldsymbol{\beta}_{-q}^{(s-1)}$ and $\boldsymbol{\omega}_q^{(s-1)}$ where the $\widehat{\mu}_{a_q}^{(SSVS)}$ and $\widehat{\sigma}_{a_q}^{2(SSVS)}$ denote the posterior mean and variance of a_q respectively as

(a) $\widehat{\mu}_{a_q}^{(SSVS)} = \widehat{\sigma}_{\mu_{a_q}}^{2(SSVS)}\left[\mathbf{1}_n^T\left(\boldsymbol{y}_q - \frac{1}{2}\mathbf{1}_n\right) + \mathbf{1}_n^T \boldsymbol{\Omega}_q \boldsymbol{C}_q - \boldsymbol{\beta}_q^T \boldsymbol{X}^T \boldsymbol{\Omega}_q \mathbf{1}_n\right]$,

(b) $\widehat{\sigma}_{a_q}^{2(SSVS)} = \left(\sum_{i=1}^n \omega_{i,q}\right)^{-1}$,

and set $a^{(s)} = a^{(s-1)}$.

Step 5: Sample $\omega_{i,q}^{(s)} \sim PG(b, \widetilde{\eta}_{i,q}(a_q, \boldsymbol{\beta}_q))$, for $i = 1, \ldots, n$ given updated states $a_q^{(s)}$, $\boldsymbol{a}_{-q}^{(s-1)}$, $\boldsymbol{\beta}_q^{(s)}$, $\boldsymbol{\beta}_{-q}^{(s-1)}$, where $\omega_{i,q}$ is the i-element of $\boldsymbol{\omega}_q$
and set $\boldsymbol{\omega}_q^{(s)} = \boldsymbol{\omega}_q^{(s-1)}$

Step 6: End of step **C**..

Step 7: for a fixed q-th class-specific, given the updated states $\boldsymbol{\gamma}_1^{(s)}, \dots, \boldsymbol{\gamma}_{Q-1}^{(s)}$, $\boldsymbol{\beta}_1^{(s)}, \dots, \boldsymbol{\beta}_{Q-1}^{(s)}$,

(**A**) if $g \sim IG\left(\frac{1}{2}, \frac{n}{2}\right)$,
sample $g^{(s)} \sim IG\left(\widetilde{\lambda}_{0,g}^{(SSVS)}, \widetilde{\lambda}_{1,g}^{(SSVS)}\right)$, where $\widetilde{\lambda}_{0,g}^{(SSVS)}$ and $\widetilde{\lambda}_{1,g}^{(SSVS)}$ denote respectively the posterior shape and scale of g respectively as

(a) $\widetilde{\lambda}_{0,g}^{(SSVS)} = ((Q-1)p_q + 1)/2 + \frac{1}{2}$,

(b) $\widetilde{\lambda}_{1,g}^{(SSVS)} \frac{1}{2}\left[\widetilde{\lambda}_q - 2\sum_{q\neq q'} \boldsymbol{\beta}_q^T \boldsymbol{D}_q^{-1} \mathcal{I}^{(BH)}(\mathbf{0}_{p_q}, \mathbf{0}_{p_{q'}}) \boldsymbol{D}_{q'}^{-1} \boldsymbol{\beta}_{q'}/Q^2 + n\right]$,
where $\widetilde{\lambda}_q = \boldsymbol{\beta}_q^T \boldsymbol{D}_q^{-1} \mathcal{I}^{(BH)}(\mathbf{0}_{p_q}, \mathbf{0}_{p_q}) \boldsymbol{D}_q^{-1} \boldsymbol{\beta}_q/Q^2$,

and set $g^{(s)} = g^{(s-1)}$.

(**B**) if $\pi(g) \propto (1+g)^{-\frac{a}{2}}$, sample $g^{(s)}$ from full conditional $(1+g)^{-\frac{\alpha}{2}} g^{-\frac{(Q-1)p_q}{2}}$ $\exp\left(-\frac{\boldsymbol{\beta}_q^T \boldsymbol{D}_q^{-1} \mathcal{I}^{(BH)}(\mathbf{0}_{p_q}, \mathbf{0}_{p_q}) \boldsymbol{D}_q^{-1} \boldsymbol{\beta}_q}{2gQ^2}\right) \exp\left(\frac{\sum_{q\neq q'} \boldsymbol{\beta}_q^T \boldsymbol{D}_q^{-1} \mathcal{I}^{(BH)}(\mathbf{0}_{p_q}, \mathbf{0}_{p_{q'}}) \boldsymbol{D}_{q'}^{-1} \boldsymbol{\beta}_{q'}}{2gQ^2}\right)$ after translating g on log-scale based on a Metropolis-Hastings with properties

(a) The same as in typical SSVS.

(b) an acceptance-rate $\widetilde{A}_g^{(SSVS)}$ of the proposed move in the log-scale

$$
\begin{aligned}
&log(\widetilde{A}_g^{(SSVS)}) = \\
&\log\left(\frac{\widetilde{\pi}^{SSVS}(g^{(can)}|\boldsymbol{\beta}_q, \boldsymbol{\beta}_{-q}, \boldsymbol{\gamma}_q, \boldsymbol{\gamma}_{-q}, \boldsymbol{y})}{\widetilde{\pi}^{SSVS}(g|\boldsymbol{\beta}_q, \boldsymbol{\beta}_{-q}, \boldsymbol{\gamma}_q, \boldsymbol{\gamma}_{-q}, \boldsymbol{y})} \frac{q(g|g^{(can)}, v_g)}{q(g^{(can)}|g, v_g)} \frac{J}{J^{(can)}}\right) \\
&\propto -\frac{\alpha}{2}\log(1+g^{(can)}) - \frac{(Q-1)p_q}{2}\log(g^{(can)}) \\
&+ \frac{\alpha}{2}\log(1+g) + \frac{(Q-1)p_q}{2}\log(g) \\
&- \frac{\boldsymbol{\beta}_q^T \boldsymbol{D}_q \mathcal{I}^{(BH)}(\mathbf{0}_{p_q}, \mathbf{0}_{p_q}) \boldsymbol{D}_q^{-1} \boldsymbol{\beta}_q}{2g^{(can)}Q^2} + \frac{\sum_{q\neq q'} \boldsymbol{\beta}_{q'}^T \boldsymbol{D}_{q'}^{-1} \mathcal{I}^{(BH)}(\mathbf{0}_{p_{q'}}, \mathbf{0}_{p_{q'}}) \boldsymbol{D}_{q'}^{-1} \boldsymbol{\beta}_{q'}}{g^{(can)}Q^2} \\
&+ \frac{\boldsymbol{\beta}_q^T \boldsymbol{D}_q^{-1} \mathcal{I}^{(BH)}(\mathbf{0}_{p_q}, \mathbf{0}_{p_q}) \boldsymbol{D}_q^{-1} \boldsymbol{\beta}_q}{2gQ^2} - \frac{\sum_{q\neq q'} \boldsymbol{\beta}_{q'}^T \boldsymbol{D}_{q'}^{-1} \mathcal{I}^{(BH)}(\mathbf{0}_{p_{q'}}, \mathbf{0}_{p_{q'}}) \boldsymbol{D}_{q'}^{-1} \boldsymbol{\beta}_{q'}}{gQ^2} \\
&+ \log\left(\frac{1}{g}\right) - \log\left(\frac{1}{g^{(can)}}\right),
\end{aligned}
$$

where $q(.)$ denotes candidate density generator and J the associated jacobian which results from transformation on the original scale of g. Notice that the corresponding ratio $q(g^{cur}|v_g)/q(g^{can}|v_g)$ vanishes due to symmetry feature of the normal random walk.

(c) Set $g^{(s)} = \begin{cases} g^{(can)} & \text{, accept with probability } \widetilde{A}_g^{(SSVS)}, \\ g & \text{, reject with probability } 1 - \widetilde{A}_g^{(SSVS)}, \end{cases}$

D. Repeat all the steps untill convergence,

in this way a complete summary of variable selection problem among different baseline-logits for each class-specific is obtained with respect to baseline. A detailed proof of all this Bayesian variable selection procedure is found on Appendix C section C.1.

4.8 GVS in Augmented Multinomial Logistic Setup

GVS of Ntzoufras (1999) and Dellaportas et al. (2002) has been considered a clever alternative to SSVS. It introduced new aspects in the Bayesian variable selection procedure such as the switching sampling density conditional on model dimensionality and the pseudo-priors that allow an ingenious Gibbs sampler to resolve the problems of formal Bayesian variable selection procedures. Characterized by its flexibility to deal with the enumeration of model space and computation of posterior model probabilities, mainly in the linear regression problem, the authors provide equivalent developments also in the GLM by approaching the intractability of the posterior and hence the full conditionals of regression coefficients through the adaptive rejection sampling. However, similar developments in the current thesis are avoided due to the fact that we prefer to approach the problem by adopting MCMC methods based on Metropolis-Hastings to deal with Bayesian variable selection in multinomial logistic regression. Moreover, even in that case, MCMC methods based on Metropolis-Hastings lose their potential as the model dimensionality increases interfering with the decrease of acceptance rates and the ill conditioned variance covariance matrices which reveal the use of data augmentation. In particular, the Polya-Gamma data augmentation of Polson et al. (2013) has been the principal spark for the present thesis that spurred our interest to illuminate the variable selection problem in multinomial logistic regression, as the authors left undefined statements regarding the variable selection uncertainty. As data augmentation Polson et al. (2013) was introduced in the previous section and chapter, it needs no further clarification also because the logic of the GVS is the same with just minor differentiations. In the present section, we propose a detailed hybrid Bayesian variable selection method based on research works of Dellaportas et al. (2002) and Polson et al. (2013) in multinomial logistic regression settings emphasized for Bayesian variable selection uncertainty. In particular, we introduce a new GVS method that takes advantage of the Polya-Gamma data augmentation scheme in order to reduce the problem of Bayesian variable selection in multinomial logistic regression equivalently with that of linear regression by the convienient factorization of latent variables in the resulting sampling density of each q-th class-specific baseline logit given the baseline class. This proves useful in applying the objective Bayesian principles of Jeffreys (1961) and Zellner and Siow (1980) as no guidelines on best subsets according to each baseline logit are available and problems such as any hard trace of MCMC such as tuning and candidate distribution generators are avoided. Furthermore, the versatility of the implied method rests on the pseudo-priors that constitute it as default Bayesian method since there is no need to tune prior inputs unlike SSVS. Moreover, our approach is based again in a prior specification of Zellner's g prior similar to SSVS, which aims to encapsulate the features of each regression coefficient belonging to a class-specific, outlined in a subclass Zellner g-prior as originating from the authentic joint g-prior. This accounts for the possible correlations and cross-correlations within and among specific classes. An immediate consequence includes the Bayesian variable selection problem extension within the g-priors mixtures framework, such as Zellner-Siow and hyper-g in multinomial logistic regression, but under the aspect of linear model formulation. Again, it proves useful to fix some notations and definitions in order to facilate the exposition of the GVS. Assume that the typical multinomial logistic regression sampling density for GVS is up to the proportionality constant for the binomial likelihood of the observed response $\boldsymbol{y}_q$ of q-th class-specific with respective probabilities as the following

$$\begin{aligned} f(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{\beta}, \gamma) &= \prod_{i=1}^{n} \prod_{q=1}^{Q-1} \binom{n}{y_{i,q}} p_{i,q}(a_q, \boldsymbol{\beta}_q|\boldsymbol{\gamma}_q)^{y_{i,q}} \\ &= f(a_q, \boldsymbol{\beta}_q|\boldsymbol{y}_q, \boldsymbol{\beta}_{-q}, \boldsymbol{a}_{-q}, \boldsymbol{\gamma}_q, \boldsymbol{\gamma}_{-q}), \end{aligned}$$

where $f(.|\boldsymbol{y}_q, \boldsymbol{\beta}_{-q}, \boldsymbol{a}_{-q}, \boldsymbol{\gamma}_q, \boldsymbol{\gamma}_{-q})$ denotes the conditional likelihood of q-th class-specific coefficients a_q and $\boldsymbol{\beta}_q$ given the rest, observed values $\boldsymbol{y}_q$, binary latent indicators $\boldsymbol{\gamma}_q$ and $\boldsymbol{\gamma}_{-q}$, then the probabilities of q-th class given baseline q^* are modified as

$$p_{i,q}(a_q, \boldsymbol{\beta}_q|\boldsymbol{\gamma}_q) = \frac{\exp\left(\widetilde{\eta}_{i,q}(a_q, \boldsymbol{\beta}_q|\boldsymbol{\gamma}_q)\right)}{\mathbf{1}_n + \exp\left(\widetilde{\eta}_{i,q}(a_q, \boldsymbol{\beta}_q|\boldsymbol{\gamma}_q)\right)},$$

where $\widetilde{\eta}_{i,q}(a_q, \boldsymbol{\beta}_q|\boldsymbol{\gamma}_q) = a_q + \boldsymbol{x}_i \boldsymbol{\Gamma}_q \boldsymbol{\beta}_q - \widetilde{C}_{i,q}$ and $\widetilde{C}_{i,q} = \log(1 + \sum_{q \neq q'} \exp(a_{q'} + \boldsymbol{x}_i \boldsymbol{\Gamma}_{q'} \boldsymbol{\beta}_{q'}))$; see Holmes and Held (2006) and Polson et al. (2013). The main purpose of each conditional likelihood (**??**) is to express the following identity

$$
\begin{aligned}
f(a_q, \boldsymbol{\beta}_q | y_{i,q}, \boldsymbol{\beta}_{-q}, \boldsymbol{a}_{-q}, \boldsymbol{\gamma}_q, \boldsymbol{\gamma}_{-q}) &= 2^{-b} \exp\left(k_{i,q}\widetilde{\eta}_{i,q}(a_q, \boldsymbol{\beta}_q|\boldsymbol{\gamma}_q)\right) \\
&\quad \int\limits_0^\infty \exp\left(-\frac{\widetilde{\eta}_{i,q}(a_q, \boldsymbol{\beta}_q|\boldsymbol{\gamma}_q))^2 \omega_{i,q}}{2}\right) \pi(\omega_{i,q}|b,0) d\omega_{i,q} \\
&= \int\limits_0^\infty \exp\left\{-\frac{\omega_{i,q}}{2}(z_{i,q} - \widetilde{\eta}_{i,q}(a_q, \boldsymbol{\beta}_q|\boldsymbol{\gamma}_q))^2\right\} \pi(\omega_{i,q}|b,0) d\omega_{i,q} \\
&= \int\limits_0^\infty f(z_{i,q}|a_q, \boldsymbol{\beta}_q, \omega_{i,q}, \boldsymbol{\gamma}_q) \pi(\omega_{i,q}|b,0) d\omega_{i,q},
\end{aligned}
$$

where $f(.|a_q, \boldsymbol{\beta}_q, \omega_{i,q}, \boldsymbol{\gamma})$ denotes the Gaussian density with observed $z_{i,q}$ and unknown precision terms $\omega_{i,q}$ with respect to $PG(.|b,0)$ prior density for $\omega_{i,q}$.
All information regarding each conditional likelihood with respect to the $\boldsymbol{y}_q$ observables for fixed specific class q, can be summarized from the augmented likelihood factorization of n data pairs $\omega_{i,q}, y_{i,q}$, for $i = 1, \ldots, n$, based on Polya-Gamma mixture identity

$$
f(a_q, \boldsymbol{\beta}_q | \boldsymbol{y}_q, \boldsymbol{\beta}_{-q}, \boldsymbol{a}_{-q}, \boldsymbol{\gamma}_q, \boldsymbol{\gamma}_{-q}) = \prod_{i=1}^{n} \int\limits_0^\infty f(z_{i,q}|a_q, \boldsymbol{\beta}_q, \omega_{i,q}, \boldsymbol{\gamma}_q) \pi(\omega_{i,q}|b,0) d\omega_{i,q}.
$$

In addition, all the incorporated information from the sample for the typical multinomial logistic likelihood can be obtained from the joint product of each conditional likelihood regarding the $\boldsymbol{y}_q$ observables for each specific class q, as n data sampling points $\omega_{i,q}, y_{i,q}$, for $i = 1, \ldots, n$ and $q = 1, \ldots, Q-1$ based on Polya-Gamma latent representation

$$
\begin{aligned}
f(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{\beta}, \boldsymbol{\gamma}) &= \prod_{q=1}^{Q-1} f(a_q, \boldsymbol{\beta}_q | \boldsymbol{y}_q, \boldsymbol{\beta}_{-q}, \boldsymbol{a}_{-q}, \boldsymbol{\gamma}_q, \boldsymbol{\gamma}_{-q}) \\
&= \prod_{q=1}^{Q-1} \prod_{i=1}^{n} \int\limits_0^\infty f(z_{i,q}|a_q, \boldsymbol{\beta}_q, \omega_{i,q}, \boldsymbol{\gamma}_q) \pi(\omega_{i,q}|b,0) d\omega_{i,q}.
\end{aligned}
\tag{4.27}
$$

We may write the augmented multinomial logistic regression model as the following

$$
f(\boldsymbol{y}, \boldsymbol{\omega}|a, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_{q=1}^{Q-1} \prod_{i=1}^{n} f(z_{i,q}|a_q, \boldsymbol{\beta}_q, \omega_{i,q}, \boldsymbol{\gamma}_q) \pi(\omega_{i,q}|b,0). \tag{4.28}
$$

In addition, for the implementation of GVS based on Polya-Gamma data augmentation we adopt again the default joint prior (4.19) based on the objective approach of Jeffreys (1961) and Zellner (1986) for parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, g and $\boldsymbol{\gamma}$ and hence will not be described

again.
On the contrary, regarding its implementation, GVS maintains $\boldsymbol{\gamma}$ the balance among different sizes of models, when the sampling (4.28) is combined with the hierachical prior specification (4.19). Thus, the joint variable selection and parameter uncertainty can be updated respectively based on the following joint posterior

$$\widetilde{\pi}^{GVS}(\boldsymbol{a}, \boldsymbol{\beta}, g, \boldsymbol{\gamma}|\boldsymbol{y}, \boldsymbol{\omega}) \propto \prod_{q=1}^{Q-1}\prod_{i=1}^{n} f(z_{i,q}|a_q, \boldsymbol{\beta}_q, \omega_{i,q}, \boldsymbol{\gamma}_q)\pi(\omega_{i,q}|b, 0)$$
$$\pi^{(BH)}(a_q)\pi^{GVS}(\boldsymbol{\beta}|Q, g, \boldsymbol{\gamma})\pi(g)\pi(\boldsymbol{\gamma}), \tag{4.29}$$

where in the above the latent variables did not take part in the main prior specification because they are included in the data augmentation. Then, the joint model and parameter uncertainty is updated respectively based on the complete data $\boldsymbol{y}$, $\boldsymbol{\omega}$ and important information of variable selection is obtained, despite the joint's posterior (4.29) intractable form. The information regarding the Bayesian variable selection problem is provided by unlocking the inctractability of the joint posterior based on Polya-Gamma data augmentation of full conditionals of $\boldsymbol{\beta}$, $\boldsymbol{a}$ that are found in closed forms for the respective regression coefficients $\boldsymbol{\beta}_q$ and a_q for fixed q-th specific class in comparison with the typical multinomial logistic regression. The respective GVS represents a straighforward expansion of Dellaportas et al. (2002) and of GVS with Polya-Gamma data augmentation presented Bayesian variable selection in logistic regression with the only difference that each a_q and $\boldsymbol{\beta}_q$ are updated through $N_n\left(\boldsymbol{z}_q|\widetilde{\boldsymbol{\eta}}_q(a_q, \boldsymbol{\beta}_q|\boldsymbol{\gamma}_q), \boldsymbol{\Omega}_q^{-1}\right)$ for $\boldsymbol{z}_q$, where $\boldsymbol{\Omega}_q = \text{diag}(\boldsymbol{\omega}_q)$, hence the full conditionals resemble those of a linear model. The full conditional of $\boldsymbol{\gamma}$, can be obtained like SSVS by considering each respective full conditional of $\boldsymbol{\gamma}_q$. However, the sampling density based on the latent data structure is included this time. The same things hold for GVS in typical multinomial logistic model.
In order to simplify the description of the underlying MCMC method, assume the joint posterior density (4.29) for fixed q specific class as expressed in the following

$$\widetilde{\pi}^{GVS}(a_q, \boldsymbol{\beta}_q, \boldsymbol{\gamma}_q, \boldsymbol{a}_{-q}, \boldsymbol{\beta}_{-q}, g, \boldsymbol{\gamma}_{-q}|\boldsymbol{y}_q, \boldsymbol{\omega}_q) \propto \prod_{i=1}^{n} f(z_{i,q}|a_q, \boldsymbol{\beta}_q, \omega_{i,q}, \boldsymbol{\gamma}_q)\pi(\omega_{i,q}|b, 0)\pi^{(BH)}(a_q)$$
$$\pi^{GVS}(\boldsymbol{\beta}|Q, g, \boldsymbol{\gamma})\pi(g)\pi(\boldsymbol{\gamma}), \tag{4.30}$$

in this case, the GVS based on the full conditionals of each q-th specific-class model parameters a_q, $\boldsymbol{\beta}_q$, $\boldsymbol{\gamma}_q$, $\boldsymbol{\omega}_q$ and g allow a nested Gibbs sampler based on the full conditionals of each class-specific model parameters a_q, $\boldsymbol{\beta}_q$, $\boldsymbol{\gamma}_q$ and g is described as follows

A. The same as in augmented SSVS.

B. For iterations $s = 1, \dots, S$:

C. For specific-class $q = 1, \dots, Q-1$:

Step 1: The same as in augmented SSVS.

Step 2: Sample $\gamma_{j,q} \sim Bern\left(\widetilde{\pi}^{GVS}_{j,q}\right)$, for $j = 1, \dots, p_q$, given the current states of $\boldsymbol{\gamma}^{(s-1)}_{-j,q}$, $\boldsymbol{\gamma}^{(s-1)}_{-q}$, $\boldsymbol{\beta}^{(s-1)}_q$, $\boldsymbol{\beta}^{(s-1)}_{-q}$, $a^{(s-1)}_q$, $\boldsymbol{a}^{(s-1)}_{-q}$, $\boldsymbol{\omega}^{(s-1)}_q$ and $g^{(s-1)}$

(a) with probability inclusion of j-th covariate given q-th class-specific $\widetilde{\pi}^{GVS}_{j,q} = \widetilde{O}^{GVS}_{j,q} / \left(1 + \widetilde{O}^{GVS}_{j,q}\right)$,

(b) with posterior odds $\widetilde{O}^{GVS}_{j,q}$

$$\widetilde{O}^{GVS}_{j,q} = \frac{f(\boldsymbol{z}_q | a_q, \boldsymbol{\beta}_q, \omega_{i,q}, \gamma_{j,q} = 1, \boldsymbol{\gamma}_{-j,q}, \boldsymbol{\gamma}_{-q})}{f(\boldsymbol{z}_q | a_q, \boldsymbol{\beta}_q, \omega_{i,q}, \gamma_{j,q} = 0, \boldsymbol{\gamma}_{-j,q}, \boldsymbol{\gamma}_{-q})} \frac{\widetilde{\pi}^{GVS}(\boldsymbol{\beta}_q | \boldsymbol{\beta}_{-q}, Q, \boldsymbol{\delta}, g, \gamma_{j,q} = 1, \boldsymbol{\gamma}_{-j,q}, \boldsymbol{\gamma}_{-q}) \pi(\gamma_{j,q} = 1, \boldsymbol{\gamma}_{-j,q})}{\widetilde{\pi}^{GVS}(\boldsymbol{\beta}_q | \boldsymbol{\beta}_{-q}, Q, \boldsymbol{\delta}, g, \gamma_{j,q} = 0, \boldsymbol{\gamma}_{-j,q}, \boldsymbol{\gamma}_{-q}) \pi(\gamma_{j,q} = 0, \boldsymbol{\gamma}_{-j,q})},$$

where $\boldsymbol{\beta}_q | \boldsymbol{\beta}_{-q}, Q, \boldsymbol{\delta}, g, \boldsymbol{\gamma_q}, \boldsymbol{\gamma}_{-q}$ is a multivariate normal distribution with prior mean $\boldsymbol{\mu}^{(GVS)}_{\boldsymbol{\beta}_q}$ and variance-covariance matrix $\boldsymbol{V}^{(GVS)}_{\boldsymbol{\beta}_q}$ defined respectively as

$$\boldsymbol{\beta}_q | \boldsymbol{\beta}_{-q}, Q, \boldsymbol{\delta}, g, \boldsymbol{\gamma_q}, \boldsymbol{\gamma}_{-q} \sim N_{p_q}\left(\boldsymbol{\mu}^{(GVS)}_{\boldsymbol{\beta}_q}, \boldsymbol{C}^{(GVS)}_{\boldsymbol{\beta}_q}\right) \tag{4.31}$$

(a) $\boldsymbol{\mu}^{(GVS)}_{\boldsymbol{\beta}_q} = \boldsymbol{C}^{(GVS)}_{\boldsymbol{\beta}_q} \boldsymbol{F}_q$,
where

$$\boldsymbol{F}_q = \frac{\boldsymbol{\Gamma}_q \mathcal{I}^{(BH)}(\mathbf{0}_{p_q}, \mathbf{0}_{p_q}) \boldsymbol{\mu}_q}{gQ^2} + \frac{\sum_{q \neq q'} \left(\boldsymbol{\Gamma}_q \mathcal{I}^{(BH)}(\mathbf{0}_{p_q}, \mathbf{0}_{p_{q'}}) \boldsymbol{\Gamma}_{q'} \boldsymbol{\beta}_{q'} - \boldsymbol{\Gamma}_q \mathcal{I}^{(BH)}(\mathbf{0}_{p_q}, \mathbf{0}_{p_{q'}}) \boldsymbol{\Gamma}_{q'} \boldsymbol{\mu}_{q'}\right)}{gQ^2}$$

(b) $\boldsymbol{C}^{(GVS)}_{\boldsymbol{\beta}_q} = \widetilde{\boldsymbol{D}}_q^{-1}$,
where $\widetilde{\boldsymbol{D}}_q = \boldsymbol{\Gamma}_q \mathcal{I}^{(BH)}(\mathbf{0}_{p_q}, \mathbf{0}_{p_q}) \boldsymbol{\Gamma}_q / gQ^2 + \widetilde{\boldsymbol{d}}_q$, $\widetilde{\boldsymbol{d}}_q = \text{diag}(1 - \boldsymbol{\gamma}_q) \frac{1}{\bar{\boldsymbol{s}}^2_q}$ and $\boldsymbol{\mu}_q = (1 - \boldsymbol{\gamma}_q) \bar{\boldsymbol{\mu}}_q$. Denoting with $\boldsymbol{F}_q$ components originating from prior of $\boldsymbol{\beta}_q$ and $\widetilde{\boldsymbol{d}}_q$, $\boldsymbol{\mu}_q$ are the counterparts of pseudo-prior for q-th

class-specific regression coefficients, which are regarded as subsets of the prior precision matrix $\widetilde{\boldsymbol{D}}$ in typical multinomial logistic model.

and set $\boldsymbol{\gamma}_q^{(s)} = \boldsymbol{\gamma}_q^{(s-1)}$.

Step 3: Sample $\boldsymbol{\beta}_q^{(s)} \sim N_{p_q}\left(\widehat{\boldsymbol{\mu}}_{\beta_q}^{(GVS)}, \widehat{\boldsymbol{C}}_{\beta_q}^{(GVS)}\right)$, given the respective updated and current states $\boldsymbol{\gamma}_q^{(s)}$, $\boldsymbol{\gamma}_{-q}^{(s-1)}$, $\boldsymbol{\beta}_{-q}^{(s-1)}$, $a_q^{(s-1)}$, $\boldsymbol{a}_{-q}^{(s-1)}$, $\boldsymbol{\omega}_q^{(s-1)}$ and $g^{(s-1)}$, where $\widehat{\boldsymbol{\mu}}_{\beta_q}^{(GVS)}$ and $\widehat{\boldsymbol{C}}_{\beta_q}^{(GVS)}$ denote the posterior mean and variance-covariance matrix of $\boldsymbol{\beta}$ defined respectively as

(a) $\widehat{\boldsymbol{\mu}}_{\beta_q}^{(GVS)} = \widehat{\boldsymbol{V}}_{\beta_q}^{(GVS)}\left(\boldsymbol{L}_q^{(GVS)} + \boldsymbol{F}_q\right)$, where

$$\begin{aligned}
\boldsymbol{L}_q^{(GVS)} &= \left[\boldsymbol{\Gamma}_q\boldsymbol{X}^T\boldsymbol{\Omega}_q\widetilde{\boldsymbol{C}}_q + \boldsymbol{\Gamma}_q\boldsymbol{X}^T\boldsymbol{\Omega}_q\boldsymbol{z}_q - a_q\boldsymbol{\Gamma}_q\boldsymbol{X}^T\boldsymbol{\Omega}_q\boldsymbol{1}_n\right],\\
\boldsymbol{F}_q &= \frac{\boldsymbol{\Gamma}_q\mathcal{I}^{(BH)}(\boldsymbol{0}_{p_q}, \boldsymbol{0}_{p_q})\boldsymbol{\mu}_q}{gQ^2}\\
&+ \frac{\sum_{q\neq q'}\left(\boldsymbol{\Gamma}_q\mathcal{I}^{(BH)}(\boldsymbol{0}_{p_q}, \boldsymbol{0}_{p_{q'}})\boldsymbol{\Gamma}_{q'}\boldsymbol{\beta}_{q'} - \boldsymbol{\Gamma}_q\mathcal{I}^{(BH)}(\boldsymbol{0}_{p_q}, \boldsymbol{0}_{p_{q'}})\boldsymbol{\Gamma}_{q'}\boldsymbol{\mu}_{q'}\right)}{gQ^2}\\
&+ \widetilde{\boldsymbol{d}}_q\boldsymbol{\mu}_q,
\end{aligned}$$

denoting with $\boldsymbol{L}_q^{(GVS)}$ components originating from augmented likelihood of q-th class-specific regression coefficients,

(b) $\widehat{\boldsymbol{V}}_{\beta_q}^{(GVS)} = \left(\widetilde{\boldsymbol{D}}_q + \boldsymbol{\Gamma}_q\boldsymbol{X}\boldsymbol{\Omega}_q\boldsymbol{X}\boldsymbol{\Gamma}_q\right)^{-1}$,

and set $\boldsymbol{\beta}_q^{(s)} = \boldsymbol{\beta}_q^{(s-1)}$.

Step 4: Sample $a_q^{(s)} \sim N\left(\widehat{\mu}_{a_q}^{(GVS)}, \widehat{\sigma}_{a_q}^{2(GVS)}\right)$, given the updated and current states respectively $\boldsymbol{\gamma}_q^{(s)}$, $\boldsymbol{\gamma}_{-q}^{(s-1)}$, $\boldsymbol{a}_{-q}^{(s-1)}$, $\boldsymbol{\beta}_q^{(s-1)}$, $\boldsymbol{\beta}_{-q}^{(s-1)}$ and $\boldsymbol{\omega}_q^{(s-1)}$ where the $\widehat{\mu}_{a_q}^{(GVS)}$ and $\widehat{\sigma}_{a_q}^{2(GVS)}$ denote the posterior mean and variance of a_q respectively as

(a) $\widehat{\mu}_{a_q}^{(GVS)} = \widehat{\sigma}_{\mu_{a_q}}^{2(GVS)}\left[\boldsymbol{1}_n^T\left(\boldsymbol{y} - \frac{1}{2}\boldsymbol{1}_n\right) + \boldsymbol{1}_n^T\boldsymbol{\Omega}_q\widetilde{\boldsymbol{C}}_q - \boldsymbol{\beta}_q^T\boldsymbol{\Gamma}_q\boldsymbol{X}^T\boldsymbol{\Omega}_q\boldsymbol{1}_n\right]$,

(b) $\widehat{\sigma}_{a_q}^{2(GVS)} = \left(\sum_{i=1}^n \omega_{i,q}\right)^{-1}$,

and set $a^{(s)} = a^{(s-1)}$.

Step 5: Sample $\omega_{i,q}^{(s)} \sim PG(b, \widetilde{\eta}_{i,q}(a_q, \boldsymbol{\beta}_q|\boldsymbol{\gamma}_q))$, for $i = 1, \ldots, n$, given updated and current states respectively $\boldsymbol{\gamma}_q^{(s)}$, $\boldsymbol{\gamma}_{-q}^{(s-1)}$ $\boldsymbol{\beta}_q^{(s)}$, $\boldsymbol{\beta}_{-q}^{(s-1)}$, $a_q^{(s)}$, $\boldsymbol{a}_{-q}^{(s-1)}$,
and set $\boldsymbol{\omega}_q^{(s)} = \boldsymbol{\omega}_q^{(s-1)}$

Step 6: End of step **C**..

Step 7: for a fixed q-th class-specific, given the states $\boldsymbol{\gamma}_1^{(s)}, \ldots, \boldsymbol{\gamma}_{Q-1}^{(s)}$, $\boldsymbol{\beta}_1^{(s)}, \ldots, \boldsymbol{\beta}_{Q-1}^{(s)}$,

(**A**) if $g \sim IG\left(\frac{1}{2}, \frac{n}{2}\right)$, sample $g^{(s)} \sim IG\left(\widetilde{\lambda}_{0,g}^{(SSVS)}, \widetilde{\lambda}_{1,g}^{(GVS)}\right)$, where $\widetilde{\lambda}_{0,g}^{(GVS)}$ and $\widetilde{\lambda}_{1,g}^{(SSVS)}$ are denoted as

(a) $\widetilde{\lambda}_{0,g}^{(GVS)} = (p_\gamma + 1)/2 + \frac{1}{2}$,

(b) $\widetilde{\lambda}_{1,g}^{(GVS)} = 0.5\,[\boldsymbol{F}_{1,q} + \boldsymbol{F}_{2,q} + n]$, where

$$\boldsymbol{F}_{1,q} = (\boldsymbol{\beta}_q - \boldsymbol{\mu}_q)^T \boldsymbol{\Gamma}_q \mathcal{I}^{(BH)}(\mathbf{0}_{p_q}, \mathbf{0}_{p_q}) \boldsymbol{\Gamma}_q (\boldsymbol{\beta}_q - \boldsymbol{\mu}_q)/Q^2,$$
$$\boldsymbol{F}_{2,q} = -2 \sum_{q \neq q'} (\boldsymbol{\beta}_q - \boldsymbol{\mu}_q)^T \boldsymbol{\Gamma}_q \mathcal{I}^{(BH)}(\mathbf{0}_{p_q}, \mathbf{0}_{p_{q'}}) \boldsymbol{\Gamma}_{q'} (\boldsymbol{\beta}_{q'} - \boldsymbol{\mu}_{q'})/Q^2,$$

and set $g^{(s)} = g^{(s-1)}$.

(B) if $\pi(g) \propto (1+g)^{-\frac{\alpha}{2}}$, sample $g^{(s)}$ from full conditional $(1+g)^{-\frac{\alpha}{2}} g^{\frac{p_\gamma}{2}}$ $\exp\left(-\frac{\boldsymbol{F}_{1,q}+\boldsymbol{F}_{2,q}}{2gQ^2}\right)$ based on a Metropolis-Hastings with properties

(a) The same as in typical SSVS.

(b) an acceptance-rate $\widetilde{A}_g^{(GVS)}$ of the proposed move in the log-scale

$$\begin{aligned}
&\log(\widetilde{A}_g^{(GVS)}) = \\
&\log\left(\frac{\widetilde{\pi}^{GVS}(g^{(can)}|\boldsymbol{\beta}_q, \boldsymbol{\beta}_{-q}, \boldsymbol{\gamma}_q, \boldsymbol{\gamma}_{-q}, \boldsymbol{y})}{\widetilde{\pi}^{GVS}(g|\boldsymbol{\beta}_q, \boldsymbol{\beta}_{-q}, \boldsymbol{\gamma}_q, \boldsymbol{\gamma}_{-q}, \boldsymbol{y})} \frac{q(g|g^{(can)}, v_g)}{q(g^{(can)}|g, v_g)} \frac{J}{J^{(can)}}\right) \\
&\propto -\frac{\alpha}{2}\log(1+g^{(can)}) - \frac{p_\gamma}{2}\log(g^{(can)}) \\
&+ \frac{\alpha}{2}\log(1+g) + \frac{p_\gamma}{2}\log(g) \\
&- \frac{\boldsymbol{F}_{1,q}+\boldsymbol{F}_{2,q}}{2g^{(can)}Q^2} + \frac{\boldsymbol{F}_{1,q}+\boldsymbol{F}_{2,q}}{2gQ^2} \\
&+ \log\left(\frac{1}{g}\right) - \log\left(\frac{1}{g^{(can)}}\right),
\end{aligned} \tag{4.32}$$

where $q(g^{cur}|v_g)/q(g^{can}|v_g)$ vanishes due to symmetry feature of the normal random walk.

(c) Set $g^{(s)} = \begin{cases} g^{(can)} & , \text{ accept with probability } \widetilde{A}_g^{(GVS)}, \\ g & , \text{ reject with probability } 1 - \widetilde{A}_g^{(GVS)}, \end{cases}$

D. Repeat all the steps untill convergence,

by this way a complete description of variable selection uncertainty is summarized among $Q-1$ different baseline-logits for each class-specific given baseline class, which is obtained indirectly from the sample of augmented posterior (4.29). A detailed proof of this Bayesian variable selection procedure is found on Appendix C section C.2.

4.9 SSVS vs GVS Within Augmented Multinomial setup

Polya-Gamma data augmentation provides flexible alternative within the Bayesian variable selection algorithms SSVS and GVS which suits ideally under the aspect of linear regression. This will allow to surpass the encountered difficulties within the framework of standard MCMC methods, especially in multinomial logistic regression. The implementation of SSVS and GVS stands for its computational ease that allows to recover the full conditionals of each class-specific intercept and regression coefficients respectively into normal distribution in contrast with cumbersome Metropolis-Hastings steps. Moreover, their implementation must be gauged apriori with caution in some instances. Next, we summarizes the main points and parts of each algorithm in order to familiarize their use to the interesting reader. These can be seen as consequent extensions of Ntzoufras (1999) and Dellaportas et al. (2002). To begin with, SSVS includes

(a) an augmented model of fixed dimension (4.23).

(b) the same hierarchical prior construction (4.12) given $\boldsymbol{\gamma}$, is modified in such way to adapt it for each q-th class-specific regression coefficients given the rest, representing a sub-class prior of the original one taking advantage of the variance-covariance structure of expected Fisher information matrix.

whereas, GVS

(a) a likelihood (4.28) changing with model dimension.

(b) the same hierarchical prior construction (4.16) given $\boldsymbol{\gamma}$, which allows to be adapted similarly as SSVS augmented for each q-th class-specific regression coefficients given the rest, encapsulating also respective parts of the main prior specification and pseudo-priors for the same and different class-specific.

The implementations steps of the two approaches differ only by the fact that in GVS is incorporated the the class-specific $\boldsymbol{\gamma}_q$. In the possible disadvantages, of both methods we can add the computational complexity related to the increased computational time. Moreover, GVS augmented version with respect to SSVS, will be more overburden with uncertainty of each q-class specific subset due to the incorporation of the augmented likelihood which depends on the observed of response and configurations of Polya-Gamma latent variables.

4.10 Simulated Experiments

In this section, we used the simulation study of Ghosh et al. (2011) for multinomial regression with fewer predictors and classes regarding the Bayesian variable selection using MCMC methods with mixtures of g-priors. Our simulation study sub-divides the simulation of Ghosh et al. (2011) into two simulated scenarios with $p_q = 10$ covariates to each class and $Q = 3$ classes such that different multinomial logistic regression models are implemented. These covariates were obtained as independent standardized normal vectors $\boldsymbol{X_1}, \ldots, \boldsymbol{X_{10}}$ iid $\sim N_{200}(0, 1)$. Moreover, these two sparse scenarios are examined within each design to describe the true generating mechanism of data as the following implies

Scenario	**Class**	**True regression coefficients**										
no-overlapping	2	a_2	$\beta_{2,1}$	$\beta_{2,2}$	$\beta_{2,3}$	$\beta_{2,4}$	$\beta_{2,5}$	$\beta_{2,6}$	$\beta_{2,7}$	$\beta_{2,8}$	$\beta_{2,9}$	$\beta_{2,10}$
		0.8	1	2	0.9	0	0	0	0	0	0	0
	3	a_3	$\beta_{3,1}$	$\beta_{3,2}$	$\beta_{3,3}$	$\beta_{3,4}$	$\beta_{3,5}$	$\beta_{3,6}$	$\beta_{3,7}$	$\beta_{3,8}$	$\beta_{3,9}$	$\beta_{3,10}$
		0.3	0	0	0	0	0	0	0	-1	1.7	-2
overlapping	2	a_2	$\beta_{2,1}$	$\beta_{2,2}$	$\beta_{2,3}$	$\beta_{2,4}$	$\beta_{2,5}$	$\beta_{2,6}$	$\beta_{2,7}$	$\beta_{2,8}$	$\beta_{2,9}$	$\beta_{2,10}$
		0.8	1	2	0.9	0	0	0	0	0	0	0
	3	a_3	$\beta_{3,1}$	$\beta_{3,2}$	$\beta_{3,3}$	$\beta_{3,4}$	$\beta_{3,5}$	$\beta_{3,6}$	$\beta_{3,7}$	$\beta_{3,8}$	$\beta_{3,9}$	$\beta_{3,10}$
		0.3	1	-2	0.8	0.9	0	0	0	0	0	0

Table 4.1 Multinomial logistic regression sparse scenarios using independent variables

where the coefficients of the Table (4.1) were set to be sparse similarly to Ghosh et al. (2011) and their values were chosen so that the true predictors of classes 2 and 3 differ in the first simulated design, whereas in the second, most of those of classes 2 and 3 are overlapping but with different magnitudes. In both simulated scenarios, in comparison with Ghosh et al. (2011), the third regression coefficient of class 2 is set equal to $\beta_{2,3} = 0.9$ rather than 0.5 because we considered smaller sample size to address class imbalance for the polychotomous respose $\boldsymbol{Y}$. The usefulness of these scenarios rests firmly on describing how the importance of covariates is altered if the same covariates were drawn as important or important and non-important across two different class-specific. Our aim is to assess the performance of Bayesian variable selection methods with MCMC both for a typical and augmented multinomial logistic regression. Furthemore additional analysis is performed to evaluate the computational

efficiency of each method based on effective sample size (ESS) and MC$_e$ Monte Carlo standard error, since we will expect similar results. We generated the class labels from a multinomial distribution (4.1) where the baseline was considered the class 1.
In addition, we illustrate the main results of simulations using again as basic tools the SSVS and GVS computational algorithms in the settings of g-priors and its mixtures adopted each time for the typical and augmented multinomial logistic regression. In particular, among methods using hyper-g we used the usual value $\alpha = 3$ suggested by Liang et al. (2008) and a Metropolis-Hastings random walk step with tuning variance $u_g = 1$ for good mixing of the chain. Regarding the tuning of proposals of intercepts and regression coefficients $\boldsymbol{\beta}$, $\boldsymbol{a}$ of each specific class respectively for both typical methods SSVS and GVS, we used $t = 0.2$, $v_{a_2} = 1$ and $v_{a_2} = 3$, respectively to ensure the good mixing of the chains. With respect to MCMC methods, prior inputs $\tau_j = 0.02$ and $c_j = 50$ for $j = 1, \ldots, (Q-1)p_q$ were set on practical significance for SSVS to achieve similar results with the objective Bayesian methods and $\bar{\boldsymbol{\mu}}$, $\bar{\boldsymbol{s}}^2$ were computed from pilot runs under the full model for GVS of each simulated dataset repetition among the methods for typical multinomial logistic regression, whereas for methods of augmented logistic regression we can easily extract them from the previous prior inputs respectively for SSVS and GVS, if we consider the prior inputs $\tau_{j,q}$, $c_{j,q}$ and $\mu_{j,q}$, $\bar{s}^2_{j,q}$ according to each class-specific covariates for $j = 1, \ldots, p_q$ and $q = 1, \ldots, Q-1$. The option of prior input τ_j and c_j are such that $\tau_j = 0.02 << \tau_j c_j = 1$. A detailed prescription of acronyms and references for all Bayesian variable selection methods are included in Table (4.3).

Prior Inputs-Initial Values	
Parameter	**Value**
τ	0.02
c	50
$\boldsymbol{\gamma}^{(0)}$	$(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)^T$
$\boldsymbol{\beta}^{(0)}$	$\widehat{\boldsymbol{\beta}}$
$\boldsymbol{a}^{(0)}$	$\hat{\boldsymbol{a}}$
$g^{(0)}$	n
$\boldsymbol{\omega}_q^{(0)}$	$(1, \ldots, 1)$

Table 4.2 Prior-inputs and initial values

The no available or little guidelines regarding the choice of subsets of variables for each class-specific covariates suggest the use of objective Bayesian methodology to each specific model parameter and model itself. In particular, we adopted the joint prior specifications of Bové and Held (2011) for both typical and augmented multinomial logistic regresion in SSVS and GVS to depict the possible prior dependences among parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$ and g. All the compared methods under Zellner-Siow prior used prior specification (2.10) for g, whereas those including hyper-g-prior used (2.12). Furthermore, since the number of covariates is equal to 10 of each specific class, which results equal to 20 if we consider the total number of covariates, we will adopt the sparse prior specification of Scott and Berger (2010) regarding the model space in order to avoid the dilution of prior probabilities.

	Acronym	Computational Method	Prior	Model
1	ssvs.hyp.typ	Stochastic Search Variable Selection for $\alpha = 3$, $\tau = 0.02$, $c = 50$, $u_{a_2} = 1$, $u_{a_2} = 1$, $t = 0.2$, $u_g = 1$	Hyper-g	Typical
2	ssvs.hyp.aug	Stochastic Search Variable Selection for $\alpha = 3$, $\tau = 0.02$, $c = 50$, $u_g = 1$	Hyper-g	Augmented
3	gvs.hyp.typ	Gibbs Variable Selection for $\alpha = 3$, $u_{a_2} = 1$, $u_{a_2} = 1$, $t = 0.2$, $u_g = 1$	Hyper-g	Typical
4	gvs.hyp.aug	Gibbs Variable Selection for $u_g = 1$	Hyper-g	Augmented
5	ssvs.ZS.typ	Stochastic Search Variable Selection for $\tau = 0.02$, $c = 50$, $u_{a_2} = 1$, $u_{a_2} = 1$, $t = 0.2$	Zellner-Siow	Typical
6	ssvs.ZS.aug	Stochastic Search Variable Selection for $\tau = 0.02$, $c = 50$	Zellner-Siow	Augmented
7	gvs.ZS.typ	Gibbs Variable Selection for $u_{a_2} = 1$, $u_{a_2} = 1$, $t = 0.2$	Zellner-Siow	Typical
8	gvs.ZS.aug	Gibbs Variable Selection	Zellner-Siow	Augmented
9	ssvs.g.typ	Stochastic Search Variable Selection for $\tau = 0.02$, $c = 50$, $u_{a_2} = 1$, $u_{a_2} = 1$, $t = 0.4$, $g = n$	g-prior	Typical
10	ssvs.g.aug	Stochastic Search Variable Selection for $\tau = 0.02$, $c = 50$, $g = n$	g-prior	Augmented
11	gvs.g.typ	Gibbs Variable Selection for $u_{a_2} = 1$, $u_{a_2} = 1$, $t = 0.4$, $g = n$	g-prior	Typical
12	gvs.g.aug	Gibbs Variable Selection for $g = n$	g-prior	Augmented

Table 4.3 Acronyms of Bayesian variable selection methods with MCMC for multinomial logistic regression

Applying Bayesian variable selection methods to this simulating design, model fitting was performed through a Gibbs sampling with subsequent Metropolis-Hastings steps and pure Gibbs sampler with or no Metropolis-Hastings step depending on the prior choice respectively for typical and augmented Bayesian variable selection methods. In this framework, we generated 40000 valid values of Markov chains to obtain convergence for the Bayesian variable selection methods of typical and augmented multinomial logistic regression. In particular, for typical multinomial logistic regression Bayesian variable selection methods we considered as initial values $\boldsymbol{\alpha}^{(0)}$, $\boldsymbol{\beta}^{(0)}$ the maximum likelihood estimators respectively for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and for the binary indicator $\boldsymbol{\gamma}$ the $\boldsymbol{\gamma}^{(0)}$, whereas for augmented logistic regression Bayesian variable selection methods the initial values are simply resulting as partitions of previous initial values as the following imply $\alpha_q^{(0)}$, $\boldsymbol{\beta}_q^{(0)}$, $\boldsymbol{\gamma}_q^{(0)}$ for $q = 1, \ldots, Q-1$ regarding each specific class given baseline q^* and for $\boldsymbol{\omega}$ the initial value $\boldsymbol{\omega}_q^{(0)}$ is used; see for more information; see for more information Table (4.2).

Results based on the frequency of identifying the true generating mechanism of data through the maximum aposteriori model for the typical and augmented multinomial logistic regression over 100 replicated simulations of each sparse scenario are provided in Tables (4.4) and (4.5) respectively. Paired evaluations of Bayesian variable selection methods with mixtures of g-priors approaches versus the rest of the methods show the following behaviour

i) In general, the Bayesian variable selection methods with mixtures of g-priors perform successfully in 2 out of the 2 scenarios. The best method of identifying the true generating mechanism of the data includes one of the methods with mixtures of g-priors.

ii) In the no-overlapping scenario, all procedures with mixtures of g-priors trace correctly the true generating data model. Both true model rates are very satisfying for the identification of subsets of covariates belonging to class 2 and 3 given baseline class 1.

iii) In the overlapping scenario, all computational methods under the various prior specifications perform poorly with almost tracing the true model. The true model rate of specific class 2 versus baseline class 1 is better with respect to that of class 3 versus the baseline class 2.

iv) In the overlapping scenario, even if the true model identification for class 3 versus baseline class 1 is considerable low, the highest true model rates are those including methods with mixtures of g-priors such gvs.ZS.typ and gvs.ZS.aug. The same holds also for true model rate of class 2 versus baseline class 1 with the only difference that is higher for gvs.hyp.aug and gvs.g.typ

Both mixtures of g-priors and fixed g-priors are perfoming similarly in the two sparse scenarios guaranteeing no particular differences among them. For these reasons, further results comparing mixtures of g-priors versus fixed are depicted in Figures (4.1), (4.2), (4.3) and (4.4) in order to stress out in depth their difference based on the marginal posterior inclusion probabilities and the MPM over the 100 simulated repetitions under the various Bayesian variable selection procedures respectively for specific classes 2 and 3 given baseline class 1. From these results, as it was expected, the non important variables with random g methods exhibit more variability in the resulting marginal posterior inclusion probabilities with respect to fixed g-priors methods. More precisely, are showing higher marginal posterior inclusion probabilities with direction towards 0.5, whereas the Zellner-Siow prior are less affected. This behaviour is actually related to the different model complexity supported by each prior specification of g and the additional variability of random g Bayesian variable selection methods that overweights the non certain covariates, thus hyper-g Bayesian variable selection procedures prefer more complex models with respect to Zellner-Siow methods that exhibit more shrinkage in the non important covariates. Even if the MAP didn't work at all only in the overlapping scenario, we observe that the MPM is identified for both simulated design scenarios separately within each class 2 and 3 given baseline class 1. This is deduced by the fact that the only important covariates that were supposed to enter the model due to simulation construction, coincided with those that exhibited marginal posterior inclusion probabilities over 0.5. Thus, the MPM as shown in this cases maximized the overall predicted profit with directions towards the true conclusions.

On the other hand, additional highlights are underlined based on the comparison of Bayesian variable selection methods with mixtures of g-priors with the respect to the rest of the methods among typical and augmented multinomial logistic regression which are presented as follows

i) Overall, Bayesian variable selection methods of mixtures of g-priors with data augmentation perform satisfactorily as in 2 out of the 2 scenarios the best method of identifying the true generating mechanism of the data was one of the methods with mixtures of g-priors with data augmentation

ii) In general, methods with data augmentation perform satisfactorily as in 2 out of 2 scenarios; the best method of identifying the true generating mechanism of the data included one of the methods with data augmentation.

iii) Generally, there are no striking differences among Bayesian variable selection methods for typical and augmented logistic regression apart only for the current cases as the following (no-overlapping case for class 3: ssvs.hyp.typ-ssvs.hyp.aug, ssvs.ZS.typ-ssvs.hyp.aug and gvs.g.typ-gvs.g.aug, overlapping case for class 2: ssvs.ZS.typ-ssvs.ZS.aug and overlapping case for class 3: ssvs.hyp.typ-ssvs.hyp.aug, ssvs.ZS.typ-ssvs.ZS.aug and ssvs.g.typ-ssvs.g.aug).

iv) All the methods with data augmentation multinomial logistic regression model seem to perform equally well under the different scenarios as their typical analogues apart from the exceptions (in most cases data augmentation outperformed of typical Bayesian variable selection method and vice versa) mentioned above.

In addition, major differentiations are encountered with respect to the properties of mixtures g-priors versus fixed g-priors based on Bayesian variable selection methods of typical and augmented multinomial logistic regression models. In particular, from Figures (4.1), (4.2), (4.3) and (4.4) it is obvious that Bayesian variable selection methods with data augmentation suffer from additional uncertainty accumulated in uncertain covariates for mixtures of g-priors, which are inflated higher towards 0.5 compared to the respective methods of typical multinomial logistic regression, even though the preference of model complexity remains the same among Bayesian variable selection methods of the two types. In particular, in Figures (4.1) and (4.2) there are no highlighted important differences apart from the uncertain covariates as we mentioned previously, whereas in Figures (4.3) and (4.4) Bayesian variable selection methods with data augmentation shrink also the important covariates in difference with the methods of typical multinomial logistic regression. Furthermore, additional information regarding the posterior density of shrinkage factor $\frac{g}{g+1}$ over 100 repeated simulations are found on Figures (4.9), (4.10), (4.11) and (4.12) respectively for both sparse scenarios among typical and augmented multinomial Bayesian variable selection methods with mixtures of g-priors, which don't seem to exhibit strange departures. Since these results based on posterior measures were expected to be close relatively, we examined further the computational efficiency of each method with respect to typical and augmented multinomial regression within SSVS and GVS based on the effective sample size ESS and Monte Carlo standard error MC$_e$ for the respective two scenarios. According to (Holmes and Held, 2006) and (Polson et al., 2013), the effective sample size can be defined for a binary model indicator γ_j as

$$ESS_j = B/(1 + 2\sum_{k=1}^{K} \rho_j(k)),$$

where B is the number of post-burnin-in samples and $\rho_j(k)$ is the k-th autocorrelation of γ_j. The ESS is used to asses how much iterations need to converge and MC$_e$ measures how much error is attributed to sampling of MCMC method. It makes sense to compute each estimate only for the non included covariates and hence the respective binary model indicators. These are computed across the two different scenarios for each class specific 2 and 3 given baseline 1 which are found respectively in Tables (4.6), (4.7), (4.8), (4.9), (4.10), (4.11), (4.12) and (4.13) only for one simulated repeatition out of 100, since the we would expect similar results across the simulated samples. The results suggests that

- All computational methods with data augmentation under the three different prior setups show larger effective sample size ESS with respect to their typical versions.
- All computational methods with data augmentation strategy under the three different prior choices conserve Monte Carlo errors MC$_e$ than their typical setup.

Such behaviour is expected since the incorporation of additional latent variables is proportional to the number of iterations and hence to the model complexity. Regarding the computational efficiency among SSVS and GVS we summarize that

- The typical version of SSVS shows always lower effective sample size and larger sampling error in contrast with the typical of GVS.
- The typical version and augmented version of SSVS are exposed to larger sampling errors in contrast with their respective versions of GVS.
- The typical version and augmented version of GVS are insensitive and behave similarly with respect to sampling errors, despite the large ESS for augmented GVS.

In addition, results based on the frequency of true model identification through the maximum aposteriori model for the typical and augmented multinomial logistic regression over 100 repeated simulated datasets of each sparse scenario are also provided in Tables (4.14) and (4.15) for sample size equal to $n = 500$. The compared results among Bayesian variable selection methods show the following

i) In this case, Bayesian variable selection methods with fixed g-priors are more robust over the random g methods as in 2 out of the 2 scenarios the best method of tracing the true generating mechanism of the data includes one of the methods with fixed g-priors.

ii) In the no-overlapping scenario, all procedures show a very high true model rate as they trace correctly again the true generating mechanism of data. Both true model rates are very high for identifying the subsets of covariates belonging to class 2 and 3 given baseline class 1.

iii) In the overlapping scenario, all Bayesian variable selection methods under the different prior specifications outperform tracing correctly the true model. The true model rates of specific class 2 and 3 versus baseline class 1 are improved and show an increasing tendency towards the true results.

iv) Overall, each Bayesian variable selection method between typical and augmented logistic regression converge in the respective posterior measures as the sample size grows. Even small gaps between them are decreased.

Similar conclusions, come in agreement also with Figures (4.5), (4.6), (4.7) and (4.8) which describe the marginal posterior inclusion probabilities of sample (n=500) over the 100 simulated repeated experiments under the different Bayesian variable selection algorithms for specific class 2 and 3 versus baseline class 1. It is evident that due to the impact of large sample size, there is smaller uncertainty accumulated inside the non certain covariates as shown the respective boxplots. Finally, the model preference of mixtures with g-priors remains the same as in the previous settings. To conclude, we evaluate also the performance of all Bayesian variable selection methods under model selection consistency as long as the sample size grows for values equal to $n = 200, 500, 1000 and 5000$ only for the no-overlapping scenario and one generated data simulation and provided in Figure (4.13). These results are depicted in sub-Figures (**??**) and (4.13b) respectively which show that among typical versus augmented SSVS and GVS separately

- For initial values of sample size posterior model probabilities deviate slightly where some methods prevail of other and viceversa both in SSVS and GVS.

- As long as the sample size increases the difference among typical and augmented respectively for SSVS and GVS diminish as it was expected increasing the probability of the true model towards 1.

- The typical setup methods across the three different priors within SSVS and GVS show a slightly increased posterior model probability in comparison with their respective augmented analogues.

We deduce that the last observation is a consequence of data augmentation schemes which suffer from additional uncertainty accumulated in the generation of binary latent vectors of each q-th class-specific, especially in GVS where the augmented likelihood takes part in the generation of $\boldsymbol{\gamma}_q$. These are reasonable since the priors used in the augmented scheme as part of the original one in the typical setup is just a sub-class prior both in SSVS and GVS. That is why model selection consistency is preserved also for the augmented setup methods.

Scenario	**Class**	**Bayesian variable selection methods**											
		ssvs. hyp. typ	ssvs. hyp. aug	gvs. hyp. typ	gvs. hyp. aug	ssvs. ZS. typ	ssvs. ZS. aug	gvs. ZS. typ	gvs. ZS. aug	ssvs. g. typ	ssvs. g. aug	gvs. g. typ	gvs. g. aug
no-overlapping	2	79	82	**84**	80	83	83	83	81	**84**	83	79	82
	3	76	**83**	76	78	77	82	78	79	80	85	74	80

Table 4.4 Number of 100 simulated samples that the MAP coincides with the true generating model of Table (4.1) for no-overlapping scenario for specific class 2 and 3 versus baseline class 1 (row-wise largest value in bold)

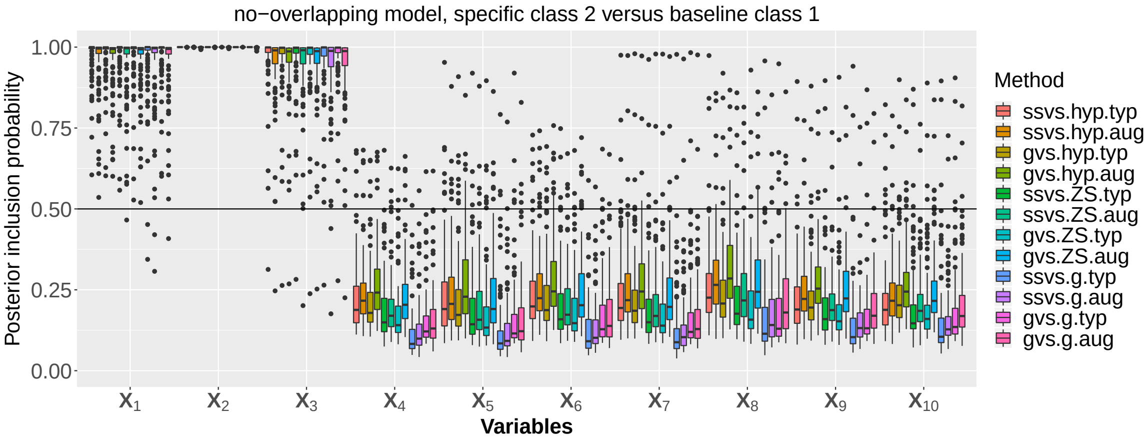


Fig. 4.1 Posterior inclusion probabilities for 100 repetitions of no-overlapping scenario regarding specific class 2 versus baseline class 1

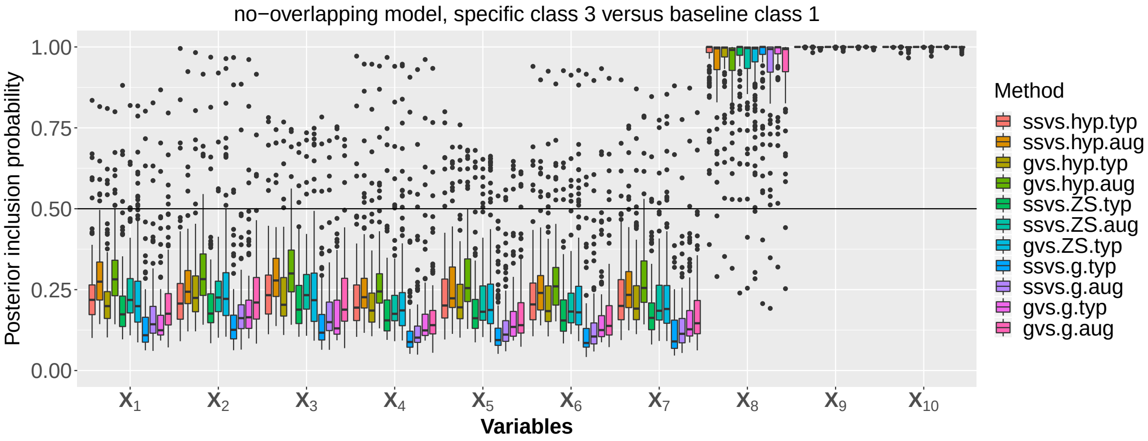


Fig. 4.2 Posterior inclusion probabilities for 100 repetitions of no-overlapping scenario regarding specific class 3 versus baseline class 1

Scenario	Class	Bayesian variable selection methods											
		ssvs. hyp. typ	ssvs. hyp. aug	gvs. hyp. typ	gvs. hyp. aug	ssvs. ZS. typ	ssvs. ZS. aug	gvs. ZS. typ	gvs. ZS. aug	ssvs. g. typ	ssvs. g. aug	gvs. g. typ	gvs. g. aug
overlapping	2	52	52	50	**53**	50	45	51	51	50	48	**53**	52
	3	46	41	45	45	46	36	**49**	**49**	43	35	47	44

Table 4.5 Number of 100 simulated samples that the MAP coincides with the true generating model of Table (4.1) for overlapping scenario for specific class 2 and 3 versus baseline class 1 (row-wise largest value in bold)

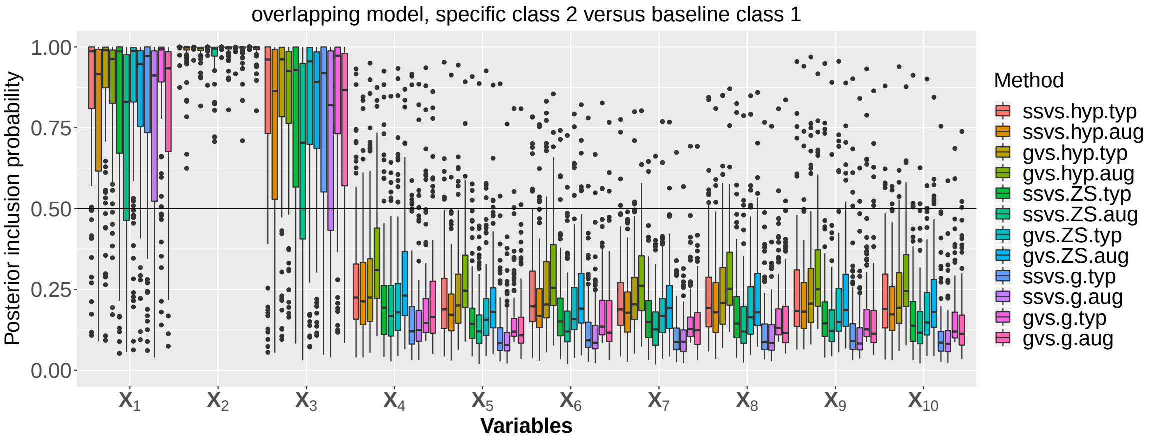


Fig. 4.3 Posterior inclusion probabilities for 100 repetitions of overlapping scenario regarding specific class 2 versus baseline class 1

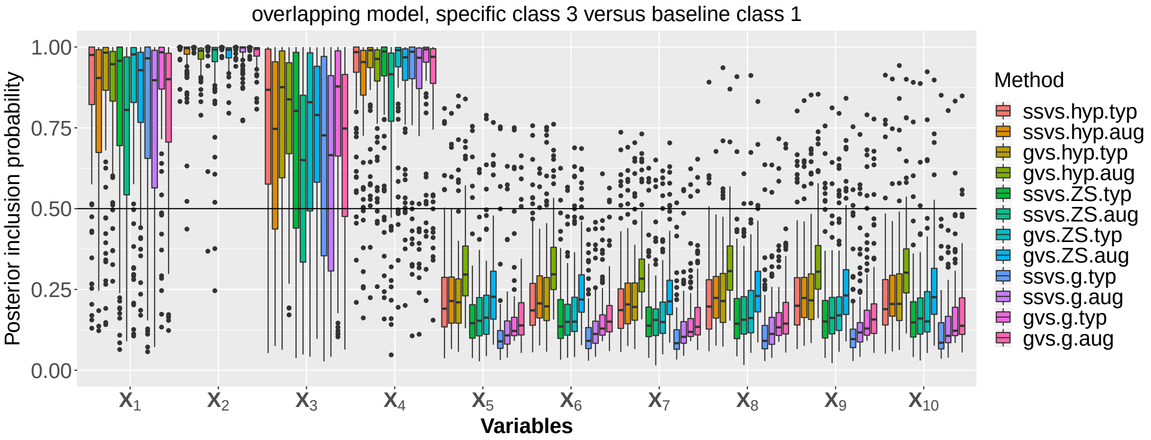


Fig. 4.4 Posterior inclusion probabilities for 100 repetitions of overlapping scenario regarding specific class 3 versus baseline class 1

	ESS, No-overlapping scenario						
	Class 2 versus baseline class 1						
Method	$\gamma_{4,2}$	$\gamma_{5,2}$	$\gamma_{6,2}$	$\gamma_{7,2}$	$\gamma_{8,2}$	$\gamma_{9,2}$	$\gamma_{10,2}$
ssvs.hyp.typ	**374**	**621**	**579**	**299**	**494**	**553**	**415**
ssvs.hyp.aug	1687	2234	2615	917	2290	2737	2157
gvs.hyp.typ	**2078**	**3598**	**4848**	**1224**	**2132**	**3002**	**1496**
gvs.hyp.aug	6650	9723	12652	3726	11042	9323	6975
ssvs.ZS.typ	**421**	**718**	**619**	**283**	**563**	**685**	**599**
ssvs.ZS.aug	2000	2806	3027	1184	2395	3029	2342
gvs.ZS.typ	**2607**	**6432**	**6397**	**1627**	**3314**	**3028**	**2185**
gvs.ZS.aug	6235	10288	11216	3917	9901	11892	7617
ssvs.g.typ	**689**	**1888**	**963**	**459**	**835**	**984**	**1387**
ssvs.g.aug	3630	5634	4334	1652	3165	5471	4633
gvs.g.typ	**6574**	**14151**	**7727**	**2186**	**8093**	**3563**	**5023**
gvs.g.aug	9556	15198	15070	4350	12801	12182	9673

Table 4.6 Effective sample size comparison (in bold lowest value).

	MC$_e$, No-overlapping scenario						
	Class 2 versus baseline class 1						
Method	$\gamma_{4,2}$	$\gamma_{5,2}$	$\gamma_{6,2}$	$\gamma_{7,2}$	$\gamma_{8,2}$	$\gamma_{9,2}$	$\gamma_{10,2}$
ssvs.hyp.typ	0.025	0.017	0.018	0.027	0.021	0.017	0.022
ssvs.hyp.aug	**0.011**	**0.009**	**0.008**	**0.016**	**0.009**	**0.008**	**0.009**
gvs.hyp.typ	0.010	0.007	0.006	0.014	0.009	0.008	0.011
gvs.hyp.aug	**0.006**	**0.004**	**0.004**	**0.008**	**0.004**	**0.004**	**0.005**
ssvs.ZS.typ	0.020	0.014	0.016	0.029	0.018	0.014	0.015
ssvs.ZS.aug	**0.010**	**0.007**	**0.007**	**0.014**	**0.008**	**0.007**	**0.008**
gvs.ZS.typ	0.008	**0.004**	0.005	0.012	0.007	0.006	0.008
gvs.ZS.aug	**0.006**	**0.004**	**0.004**	**0.007**	**0.004**	**0.003**	**0.005**
ssvs.g.typ	0.014	0.006	0.009	0.018	0.011	0.010	0.007
ssvs.g.aug	**0.006**	**0.003**	**0.004**	**0.010**	**0.005**	**0.004**	**0.004**
gvs.g.typ	**0.004**	**0.002**	0.004	0.010	0.003	0.005	0.004
gvs.g.aug	**0.004**	**0.002**	**0.003**	**0.007**	**0.003**	**0.003**	**0.003**

Table 4.7 Monte Carlo error comparison (in bold lowest value).

	ESS, No-overlapping scenario						
	Class 3 versus baseline class 1						
Method	$\gamma_{1,3}$	$\gamma_{2,3}$	$\gamma_{3,3}$	$\gamma_{4,3}$	$\gamma_{5,3}$	$\gamma_{6,3}$	$\gamma_{7,3}$
ssvs.hyp.typ	**380**	**547**	**519**	**477**	**414**	**523**	**268**
ssvs.hyp.aug	1904	2311	1476	1887	1635	2077	824
gvs.hyp.typ	**1546**	**2833**	**3432**	**2116**	**3292**	**4300**	**930**
gvs.hyp.aug	6531	10055	9300	7617	9771	10225	3455
ssvs.ZS.typ	**389**	**603**	**643**	**660**	**380**	**489**	**292**
ssvs.ZS.aug	2021	2831	2347	2429	1879	2768	1106
gvs.ZS.typ	**2104**	**3422**	**4337**	**3585**	**5311**	**8161**	**1273**
gvs.ZS.aug	6727	9662	8518	7339	9182	11668	3156
ssvs.g.typ	**612**	**1527**	**939**	**1548**	**619**	**1251**	**407**
ssvs.g.aug	2396	4471	3359	5641	3045	5431	1468
gvs.g.typ	**3031**	**3127**	**9302**	**5657**	**13451**	**24693**	**1548**
gvs.g.aug	7830	8764	10327	10285	14469	17379	3965

Table 4.8 Effective sample size comparison (in bold lowest value).

	MC$_e$, No-overlapping scenario						
	Class 3 versus baseline class 1						
Method	$\gamma_{1,3}$	$\gamma_{2,3}$	$\gamma_{3,3}$	$\gamma_{4,3}$	$\gamma_{5,3}$	$\gamma_{6,3}$	$\gamma_{7,3}$
ssvs.hyp.typ	0.024	0.017	0.020	0.020	0.020	0.020	0.020
ssvs.hyp.aug	**0.010**	**0.008**	**0.012**	**0.010**	**0.012**	0.010	0.017
gvs.hyp.typ	0.012	0.008	0.007	0.009	0.008	0.006	0.016
gvs.hyp.aug	**0.006**	**0.004**	**0.004**	**0.005**	**0.005**	**0.004**	**0.008**
ssvs.ZS.typ	0.021	0.015	0.016	0.015	0.023	0.019	0.029
ssvs.ZS.aug	**0.009**	**0.007**	**0.009**	**0.008**	**0.010**	**0.008**	**0.015**
gvs.ZS.typ	0.009	0.006	0.006	0.006	0.006	0.004	0.014
gvs.ZS.aug	**0.005**	**0.004**	**0.005**	**0.005**	**0.005**	**0.004**	**0.008**
ssvs.g.typ	0.014	0.007	0.010	0.006	0.014	0.008	0.021
ssvs.g.aug	**0.008**	**0.005**	**0.006**	**0.003**	**0.007**	**0.004**	**0.011**
gvs.g.typ	0.007	0.006	**0.003**	0.004	**0.003**	0.002	0.012
gvs.g.aug	**0.005**	**0.003**	**0.003**	**0.003**	**0.003**	**0.002**	**0.007**

Table 4.9 Monte Carlo error comparison (in bold lowest value).

	ESS, Overlapping scenario						
	Class 2 versus baseline class 1						
Method	$\gamma_{4,2}$	$\gamma_{5,2}$	$\gamma_{6,2}$	$\gamma_{7,2}$	$\gamma_{8,2}$	$\gamma_{9,2}$	$\gamma_{10,2}$
ssvs.hyp.typ	**853**	**843**	**842**	**1166**	**650**	**645**	**470**
ssvs.hyp.aug	3570	5224	5879	4529	3299	3694	2763
gvs.hyp.typ	**7036**	19036	**14946**	**17194**	19144	15126	12196
gvs.hyp.aug	15863	**13728**	15797	17258	**12351**	**16034**	**11564**
ssvs.ZS.typ	**732**	**973**	**1443**	**978**	**657**	**647**	**623**
ssvs.ZS.aug	3993	4939	6063	6550	3395	4929	3059
gvs.ZS.typ	**13661**	22718	**22102**	27400	20176	**18277**	**13654**
gvs.ZS.aug	17942	**21840**	24977	**22387**	**17939**	23731	16339
ssvs.g.typ	**1290**	**1076**	**1191**	**2242**	**629**	**741**	**563**
ssvs.g.aug	5559	7876	9653	8870	6103	8069	4279
gvs.g.typ	**11874**	27965	28369	34863	30140	**18984**	23024
gvs.g.aug	15679	**26503**	**26096**	**32913**	19668	27642	**18665**

Table 4.10 Effective sample size comparison (in bold lowest value).

	MC$_e$, Overlapping scenario						
	Class 2 versus baseline class 1						
Method	$\gamma_{4,2}$	$\gamma_{5,2}$	$\gamma_{6,2}$	$\gamma_{7,2}$	$\gamma_{8,2}$	$\gamma_{9,2}$	$\gamma_{10,2}$
ssvs.hyp.typ	0.010	0.010	0.011	0.008	0.013	0.013	0.017
ssvs.hyp.aug	**0.005**	**0.004**	**0.003**	**0.004**	**0.006**	**0.005**	**0.006**
gvs.hyp.typ	0.003	**0.002**	**0.002**	**0.002**	**0.002**	**0.002**	0.003
gvs.hyp.aug	**0.002**	**0.002**	**0.002**	**0.002**	0.003	0.002	**0.003**
ssvs.ZS.typ	0.011	0.009	0.007	0.009	0.011	0.013	0.014
ssvs.ZS.aug	**0.004**	**0.003**	**0.003**	**0.003**	**0.005**	**0.004**	**0.005**
gvs.ZS.typ	**0.002**	**0.001**	**0.001**	**0.001**	**0.002**	0.002	0.003
gvs.ZS.aug	**0.002**	**0.001**	**0.001**	**0.001**	**0.002**	**0.001**	**0.002**
ssvs.g.typ	0.007	0.007	0.007	0.004	0.012	0.009	0.013
ssvs.g.aug	**0.003**	**0.002**	**0.002**	**0.002**	**0.003**	**0.002**	**0.004**
gvs.g.typ	**0.002**	**0.001**	**0.001**	**0.001**	**0.002**	0.002	**0.002**
gvs.g.aug	**0.002**	**0.001**	**0.001**	**0.001**	**0.002**	0.001	**0.002**

Table 4.11 Monte Carlo error comparison (in bold lowest value).

ESS, Overlapping scenario						
Class 3 versus baseline class 1						
Method	$\gamma_{5,3}$	$\gamma_{6,3}$	$\gamma_{7,3}$	$\gamma_{8,3}$	$\gamma_{9,3}$	$\gamma_{10,3}$
ssvs.hyp.typ	**587**	**838**	**1014**	**866**	**562**	**991**
ssvs.hyp.aug	3110	3492	3533	4550	3044	3896
gvs.hyp.typ	13995	15960	15916	13709	12432	13943
gvs.hyp.aug	**10609**	**12670**	**12717**	**12891**	**10054**	**13297**
ssvs.ZS.typ	**631**	**761**	**931**	**1065**	**662**	**883**
ssvs.ZS.aug	3831	4703	4764	4636	3423	4250
gvs.ZS.typ	21132	22638	**18091**	**16738**	16603	16656
gvs.ZS.aug	**14890**	**16393**	18816	18991	**12218**	**14287**
ssvs.g.typ	**862**	**1273**	**898**	**1106**	**851**	**950**
ssvs.g.aug	5629	6349	7079	7469	4132	5465
gvs.g.typ	29981	32333	25980	22341	30701	20601
gvs.g.aug	**16655**	**23502**	**21549**	**18130**	**15453**	**19446**

Table 4.12 Effective sample size comparison (in bold lowest value).

MC_e, Overlapping scenario						
Class 3 versus baseline class 1						
Method	$\gamma_{5,3}$	$\gamma_{6,3}$	$\gamma_{7,3}$	$\gamma_{8,3}$	$\gamma_{9,3}$	$\gamma_{10,3}$
ssvs.hyp.typ	0.014	0.011	0.009	0.011	0.015	0.009
ssvs.hyp.aug	**0.007**	**0.006**	**0.006**	**0.005**	**0.007**	**0.006**
gvs.hyp.typ	**0.003**	**0.002**	**0.002**	**0.002**	**0.003**	**0.002**
gvs.hyp.aug	0.004	0.003	0.003	0.003	0.004	0.003
ssvs.ZS.typ	0.012	0.010	0.009	0.008	0.014	0.010
ssvs.ZS.aug	**0.006**	**0.005**	**0.005**	**0.005**	**0.006**	**0.005**
gvs.ZS.typ	**0.002**	**0.002**	**0.002**	**0.002**	**0.002**	**0.002**
gvs.ZS.aug	0.003	**0.002**	**0.002**	**0.002**	0.003	0.003
ssvs.g.typ	0.008	0.006	0.009	0.007	0.010	0.008
ssvs.g.aug	**0.004**	**0.003**	**0.003**	**0.003**	**0.005**	**0.004**
gvs.g.typ	**0.002**	**0.001**	**0.001**	**0.001**	**0.002**	**0.002**
gvs.g.aug	**0.002**	0.002	0.002	0.002	**0.002**	**0.002**

Table 4.13 Monte Carlo error comparison (in bold lowest value).

Scenario	**Class**	**Bayesian variable selection methods (n=500)**											
		ssvs. hyp. typ	ssvs. hyp. aug	gvs. hyp. typ	gvs. hyp. aug	ssvs. ZS. typ	ssvs. ZS. aug	gvs. ZS. typ	gvs. ZS. aug	ssvs. g. typ	ssvs. g. aug	gvs. g. typ	gvs. g. aug
no-overlapping	2	87	89	87	87	89	91	87	89	**94**	**94**	85	88
	3	91	96	92	94	94	96	94	96	**97**	**97**	91	96

Table 4.14 Number of 100 simulated samples that the MAP coincides with the true generating model of Table (4.1) for no-overlapping scenario for specific class 2 and 3 versus baseline class 1 (n=500) (row-wise largest value in bold)

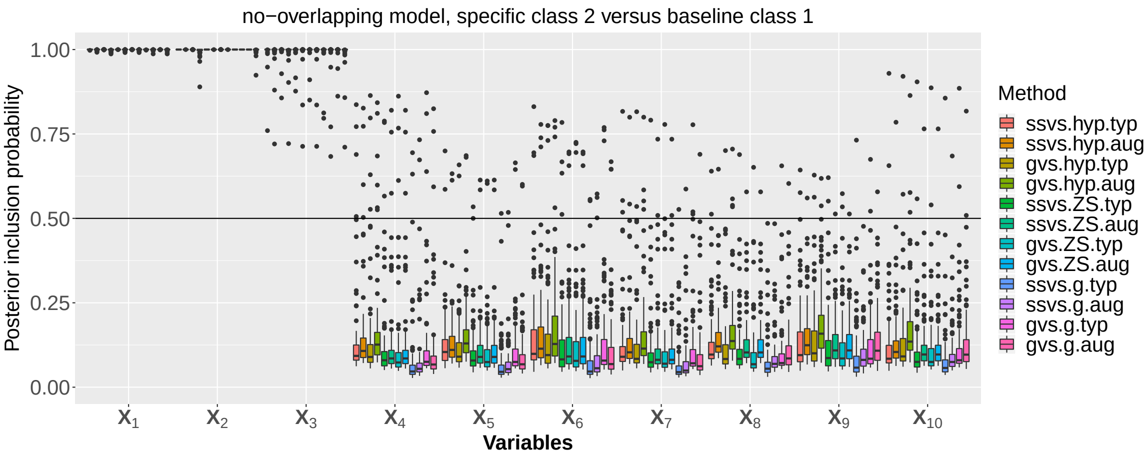


Fig. 4.5 Posterior inclusion probabilities for 100 repetitions of no-overlapping scenario regarding specific class 2 versus baseline class 1 (n=500)

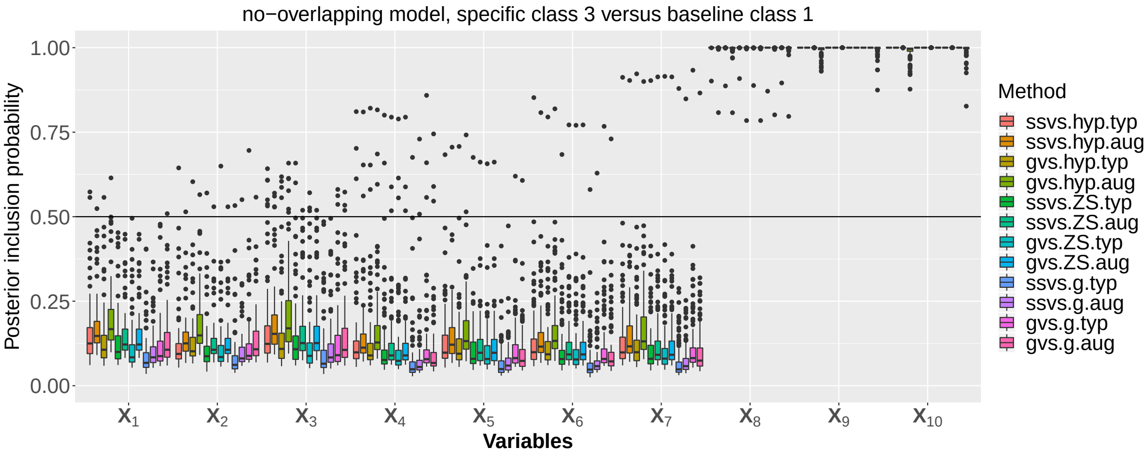


Fig. 4.6 Posterior inclusion probabilities for 100 repetitions of no-overlapping scenario regarding specific class 3 versus baseline class 1 (n=500) (row-wise largest value in bold)

Scenario	**Class**	**Bayesian variable selection methods (n=500)**											
		ssvs. hyp. typ	ssvs. hyp. aug	gvs. hyp. typ	gvs. hyp. aug	ssvs. ZS. typ	ssvs. ZS. aug	gvs. ZS. typ	gvs. ZS. aug	ssvs. g. typ	ssvs. g. aug	gvs. g. typ	gvs. g. aug
no-overlapping	2	82	87	79	82	83	87	81	84	88	**89**	79	86
	3	89	91	83	85	93	91	91	89	**95**	91	93	92

Table 4.15 Number of 100 simulated samples that the MAP coincides with the true generating model of Table (4.1) for overlapping scenario for specific class 2 and 3 versus baseline class 1 (n=500) (row-wise largest value in bold)

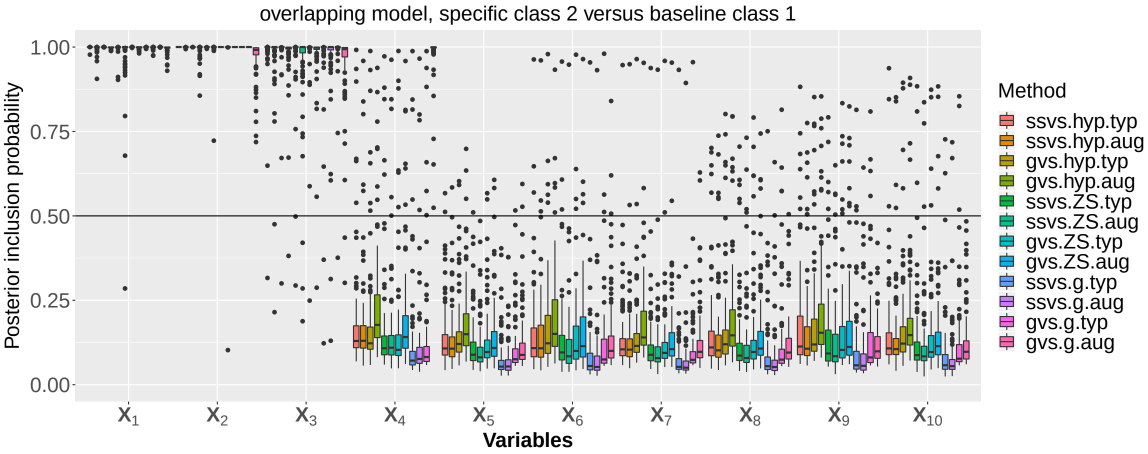


Fig. 4.7 Posterior inclusion probabilities for 100 repetitions of overlapping scenario regarding specific class 2 versus baseline class 1 (n=500)

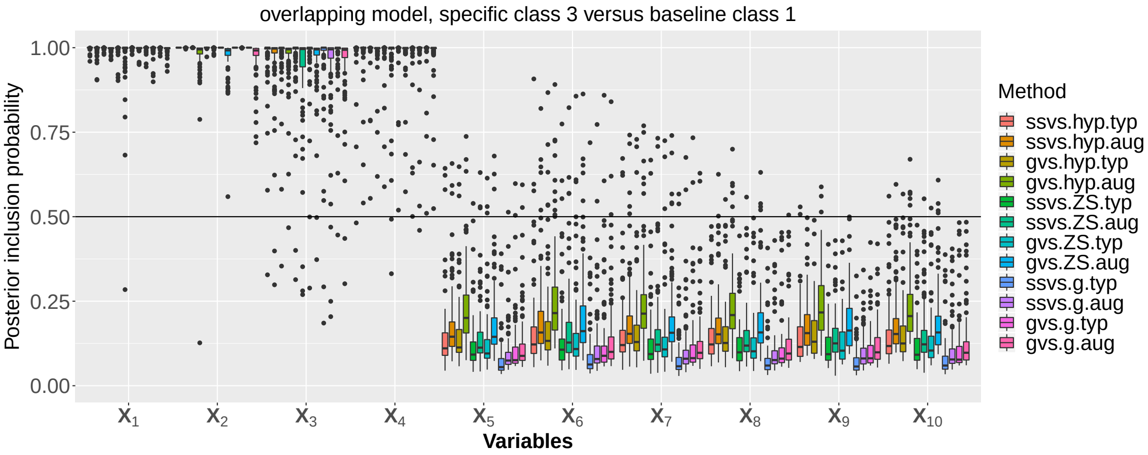
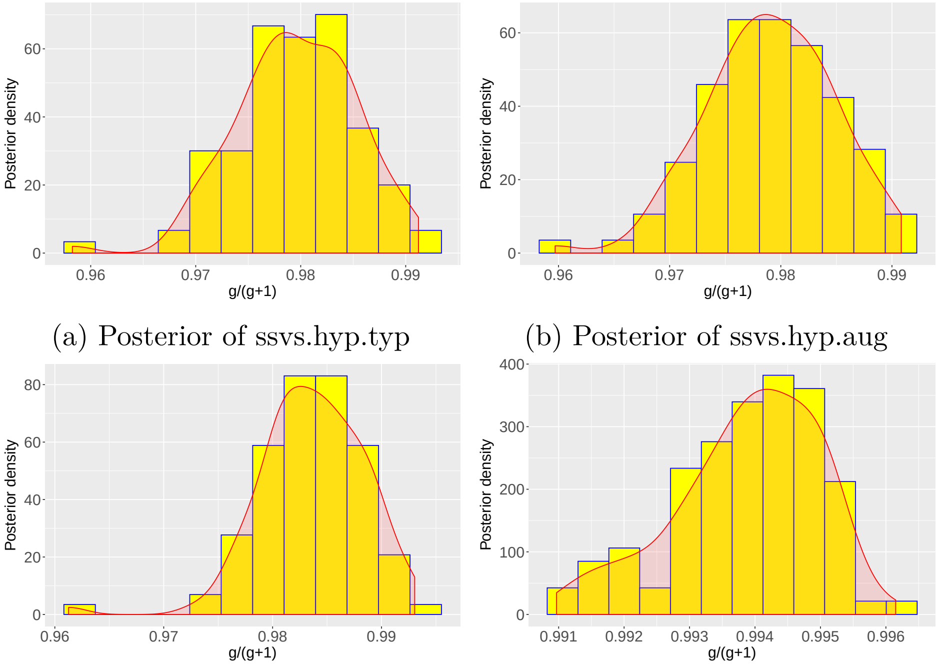


Fig. 4.8 Posterior inclusion probabilities for 100 repetitions of overlapping scenario regarding specific class 2 versus baseline class 1 (n=500)



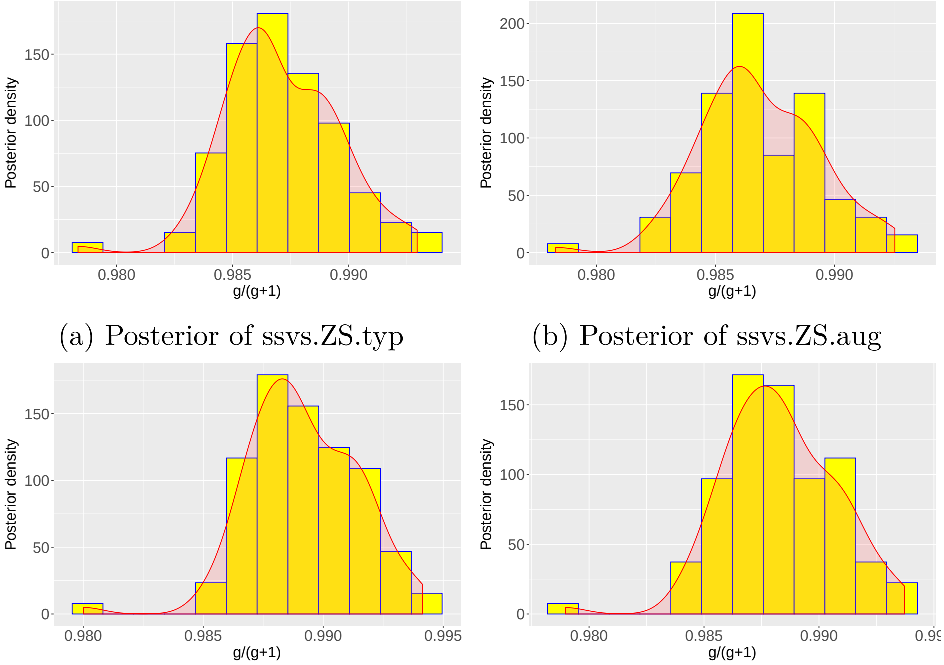
(a) Posterior of ssvs.hyp.typ

(b) Posterior of ssvs.hyp.aug

(c) Posterior of gvs.hyp.typ

(d) Posterior of gvs.hyp.aug

Fig. 4.9 Distributions of $\frac{g}{g+1}$ for 100 repetitions of no-overlapping scenario



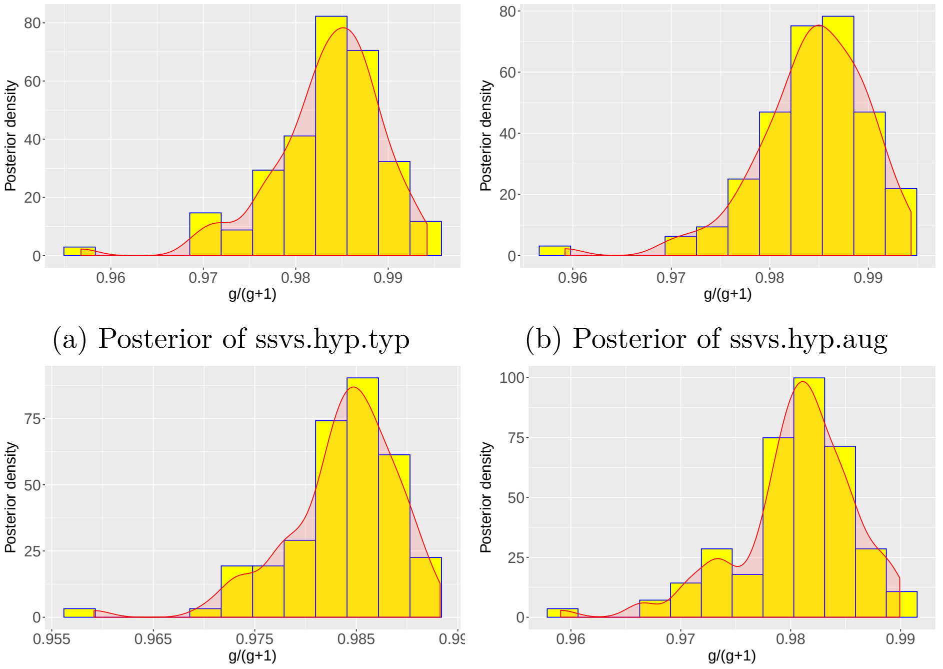
(a) Posterior of ssvs.ZS.typ

(b) Posterior of ssvs.ZS.aug

(c) Posterior of gvs.ZS.typ

(d) Posterior of gvs.ZS.aug

Fig. 4.10 Distributions of $\frac{g}{g+1}$ for 100 repetitions of no-overlapping scenario



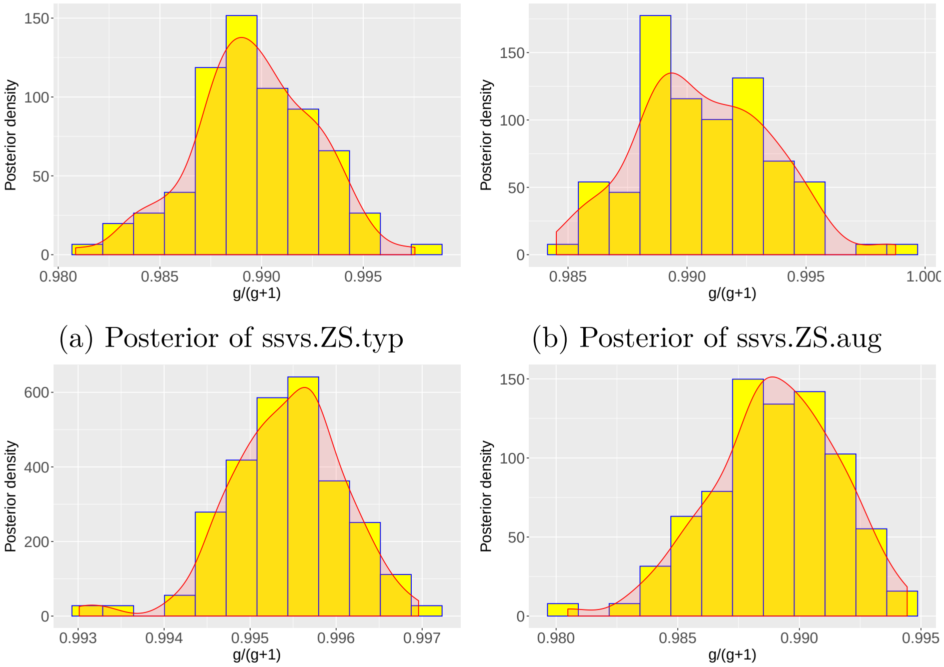
(a) Posterior of ssvs.hyp.typ

(b) Posterior of ssvs.hyp.aug

(c) Posterior of gvs.hyp.typ

(d) Posterior of gvs.hyp.aug

Fig. 4.11 Distributions of $\frac{g}{g+1}$ for 100 repetitions of overlapping scenario



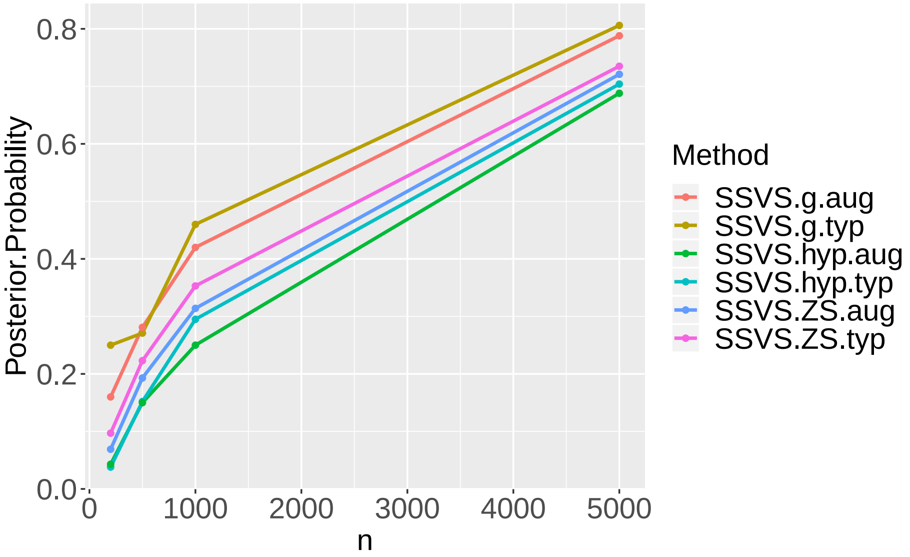
(a) Posterior of ssvs.ZS.typ

(b) Posterior of ssvs.ZS.aug

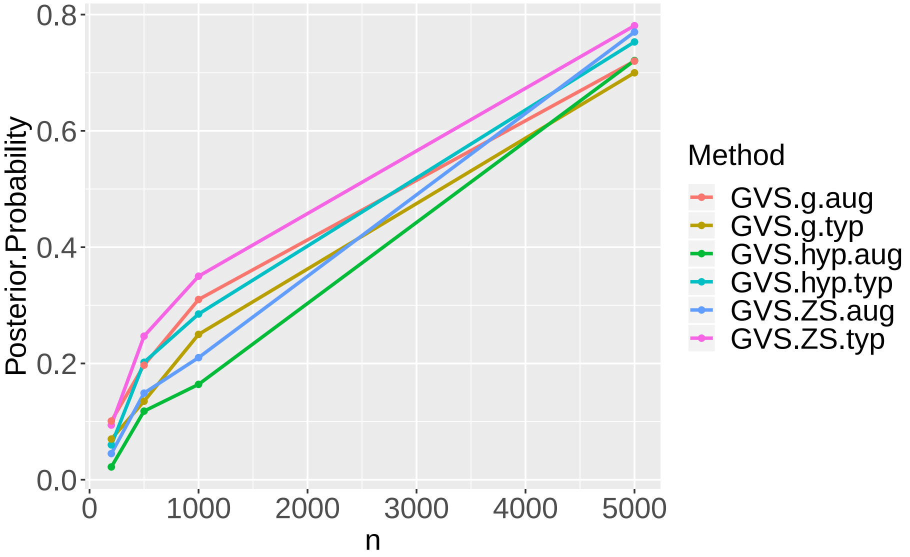
(c) Posterior of gvs.ZS.typ

(d) Posterior of gvs.ZS.typ

Fig. 4.12 Distributions of $\frac{g}{g+1}$ for 100 repetitions of overlapping scenario



(a) Posterior model probability of SSVS typical multinomial logistic vs augmented multinomial logistic



(b) Posterior model probability of SSVS typical multinomial logistic vs augmented multinomial logistic

Fig. 4.13 Model selection consistency of SSVS and GVS in typical vs augmented multinomial logistic setup respectively.

4.11 Real Application

In this section, we illustrate a real application of Bayesian variable selection in multinomial logistic regression with mixtures of g-priors for the Cardiotography dataset Ayres-de Campos et al. (2000), with more emphasis given for hyper-g-prior and fixed g-priors for comparability reasons. In obstetrics, cardiotography is a method of measuring the fetal heartbeat and the uterine contractions during the last trimester of pregnacy and it aids to recognise patterns associated with fetal activity and detect anomalies, thus is prevalent in order to evaluate the well being of the fetus before delivery (Arif, 2015). This dataset was used also by Arif (2015), Kamath and Kamat (2016) and Pereira et al. (2016). This dataset contains 2126 measurements of cardiotocograms collected in the Maternity and Gynecological Clinic (University Hospital of Porto in Portugal) based of fetal heart recordings of prenatal babies and $p_q = 21$ covariates that best describe their profile. The response variable $\boldsymbol{Y}$ is the cardiotography whose outcome was determined by 3 experienced obstetricians with 3 class labels, each one denoting with 1="normal fetal state", 2="suspect fetal state" and 3="pathologic fetal state" and the covariates are the fetal heart baseline value ($\boldsymbol{X_1}$), the accelerations in fetal heart rate ($\boldsymbol{X_2}$), the fetal movement ($\boldsymbol{X_3}$), the uterine contractions ($\boldsymbol{X_4}$), the percentage of time with abnormal short term variability ($\boldsymbol{X_5}$), the mean value of short term variability ($\boldsymbol{X_6}$), the percentage of time with abnormal long term variability ($\boldsymbol{X_7}$), the mean value of long term variability ($\boldsymbol{X_8}$), the light decelerations ($\boldsymbol{X_9}$), the severe decelerations ($\boldsymbol{X_{10}}$), the prolonged decelerations ($\boldsymbol{X_{11}}$), the width of histogram ($\boldsymbol{X_{12}}$), the low frequency of histogram ($\boldsymbol{X_{13}}$), the high frequency of histogram ($\boldsymbol{X_{14}}$), the number of histogram peaks ($\boldsymbol{X_{15}}$), the number of histogram zeros ($\boldsymbol{X_{16}}$), the mode of histogram ($\boldsymbol{X_{17}}$), the mean of histogram ($\boldsymbol{X_{18}}$), the median of histogram ($\boldsymbol{X_{19}}$), the variance of histogram ($\boldsymbol{X_{20}}$) and the histogram density with 3 categories where -1="left asymmetric", 0="symmetric" , 1="right asymmetric" ($\boldsymbol{X_{21}}$). Notice that the last covariate is considered as continuous in this analysis. This application is designed to address issues related to class imbalance of the response and multicollinearity between covariates for which we will try to answer through this application.

Class (Fetal state)	**Number of cases**
Normal	1655
Suspect	295
Pathologic	176

Table 4.16 Class distribution of cardiotography's recordings

More precisely, Tables (4.16) and (4.17) depict respectively the distribution of the response variable based on cardiotography recordings and the highest pair-wise correlations between variables.

Pair-wise variables	**r**
$\boldsymbol{X_1} - \boldsymbol{X_{17}}$	0.708
$\boldsymbol{X_1} - \boldsymbol{X_{18}}$	0.723
$\boldsymbol{X_1} - \boldsymbol{X_{19}}$	0.789
$\boldsymbol{X_{13}} - \boldsymbol{X_{12}}$	-0.898
$\boldsymbol{X_{15}} - \boldsymbol{X_{12}}$	0.747
$\boldsymbol{X_{14}} - \boldsymbol{X_{12}}$	0.690
$\boldsymbol{X_{18}} - \boldsymbol{X_{17}}$	0.893
$\boldsymbol{X_{19}} - \boldsymbol{X_{17}}$	0.933
$\boldsymbol{X_{19}} - \boldsymbol{X_{18}}$	0.948

Table 4.17 Largerst pair-wise correlation among covariates

Our aim in this application is to asses the performance of Bayesian variable selection methods of typical and augmented multinomial logistic regression data with hyper-g-prior and g-prior for out-of-sample values for both SSVS and GVS. We point out to the interesting reader that all Bayesian variable selection methods with hyper-g-prior and fixed g-prior for typical and multinomial logistic regression maintain the same acronyms as those of Table (4.3). Prior to the main analysis we standardized the covariates and we split the data for 70% of training ($n_{tr} = 1489$) and 30% of test ($n_{te} = 637$) sets, in order to evaluate the out-of-sample accuracy. We considered as baseline class 1 in order to have more balance the rest of categories. Then, we evaluate the predictive ability of each MCMC methods based on the maximum aposteriori model MAP, median probability model MPM and Bayesian model averaging (BMA) by calculating the accuracy ($\widehat{ACC}$) and Cohen's kappa statistic ($\widehat{k}_c$). Before the exposition of the main results, we initialized the Bayesian variable selection methods based on prior inputs as following: methods with hyper-g used the usual value $\alpha = 3$ suggested by Liang et al. (2008) and a Metropolis-Hastings random walk step with tuning variance $u_g = 1$ for g and the tuning of proposals of intercepts and regression coefficients $\boldsymbol{\beta}$, $\boldsymbol{a}$ of each specific class respectively for both typical methods SSVS and GVS, we used $t = 0.1$, $v_{a_1} = 1$ and $v_{a_2} = 1$ respectively to ensure the good mixing of the chains proportional to the number of parameters especially for $\boldsymbol{\beta}$, prior inputs $\tau_j = 0.02$ and $c_j = 50$ for $j = 1, \ldots, 42$ were set on practical significance

for SSVS to achieve similar results with the objective Bayesian methods and $\bar{\boldsymbol{\mu}} = (-1.014, -4.628, 0.610, -0.516, 1.731, -0.494, 0.478, 0.205, -0.215, -0.003, 1.612, -0.036, 0.313, 0.437, 0.584, -0.130, -1.289, 2.590, 1.072, 1.042, 0.112, 3.015, -6.011, 1.204, -1.201, 3.674, -0.936, 1.601, 0.434, -0.325, 0.557, 1.863, 0.422, 0.429, 1.625, -1.639, 0.466, -0.756, -1.234, -3.935, 1.911, 0.534)^T$, $\bar{\boldsymbol{s}}^2 = (0.138, 0.457, 0.014, 0.0214, 0.053, 0.101, 0.018, 0.050, 0.125, 735, 0.136, 0.011, 0.011, 0.044, 0.077, 0.042, 0.015, 0.242, 0.922, 0.976, 0.075, 0.037, 0.523, 2.333, 0.047, 0.108, 0.301, 0.268, 0.070, 0.218, 0.169, , 186.2, 0.118, 0.050, 0.117, 0.172, 0.295, 0.095, 0.803, 0.811, 2.034, 0.178, 0.100)^T$ were computed from pilot runs under the frequentist full multinomial logistic regression model for GVS of the real dataset among the methods for typical multinomial logistic regression, whereas for methods of augmented logistic regression, the previous prior inputs respectively for SSVS and GVS result as if we consider the prior inputs $\tau_{j,q}$, $c_{j,q}$ and $\bar{\mu}_{j,q}$, $\bar{s}^2_{j,q}$ according to each class-specific covariates for $j = 1, \ldots, 21$ and $q = 2, 3$. The option of priors input is again τ_j and c_j such that $\tau_j = 0.02 << \tau_j c_j = 1$, whereas $\bar{\boldsymbol{\mu}}_2 = (-1.014, -4.628, 0.610, -0.516, 1.731, -0.494, 0.478, 0.205, -0.215, -0.003, 1.612, -0.036, 0.313, 0.437, 0.584, -0.130, -1.289, 2.590, 1.072, 1.042, 0.112)^T$, $\bar{\boldsymbol{\mu}}_3 = (3.015, -6.011, 1.204, -1.201, 3.674, -0.936, 1.601, 0.434, -0.325, 0.557, 1.863, 0.422, 0.429, 1.625, -1.639, 0.466, -0.756, -1.234, -3.935, 1.911, 0.534)^T$, $\bar{\boldsymbol{s}}^2_2 = (0.138, 0.457, 0.014, 0.0214, 0.053, 0.101, 0.018, 0.050, 0.125, 735, 0.136, 0.011, 0.011, 0.044, 0.077, 0.042, 0.015, 0.242, 0.922, 0.976, 0.075, 0.037)^T$, $\bar{\boldsymbol{s}}^2_3 = (0.523, 2.333, 0.047, 0.108, 0.301, 0.268, 0.070, 0.218, 0.169, 186.2, 0.118, 0.050, 0.117, 0.172, 0.295, 0.095, 0.803, 0.811, 2.034, 0.178, 0.100)^T$.
The little information regarding the choice of variables to enter in each class-specific covariates evidence the use of objective Bayesian methodology to each specific model parameter and model itself. In particular, we adopt again the joint prior specifications of Bové and Held (2011) for both typical and augmented multinomial logistic regresion in SSVS and GVS to express the possible prior dependences among parameters $\boldsymbol{a}, \boldsymbol{\beta}, \boldsymbol{\gamma}$ and g. All the compared methods under hyper-g-prior used prior specification (2.12) for g. Furthermore, since the number of covariates is equal to 21 of each specific class and results equal to 42 totally, if we consider the total number of covariates in each class, we adopt the sparse prior specification of Scott and Berger (2010) regarding the model space in order to avoid the dilution of prior probabilities. Applying Bayesian variable selection methods to this simulating design, model fitting was performed through a Gibbs sampling with additional Metropolis-Hastings steps and pure Gibbs sampler with or no Metropolis-Hastings step depending on the prior choice respectively for typical and augmented Bayesian variable selection methods. In this framework, 40000 valid values were generated from Markov chains to obtain convergence for the Bayesian

variable selection methods of typical and augmented multinomial logistic regression respectively.

In particular, for Bayesian variable selection methods have been considered as initial values $\boldsymbol{\alpha}^{(0)}$, $\boldsymbol{\beta}^{(0)}$ the maximum likelihood estimators respectively for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and for the binary indicator $\boldsymbol{\gamma}$ the $\boldsymbol{\gamma}^{(0)}$ a vector of 42 one's, whereas for augmented logistic regression Bayesian variable selection methods the initial values are simply resulting as partitions of previous initial values as the following imply $\alpha_q^{(0)}$, $\boldsymbol{\beta}_q^{(0)}$, $\boldsymbol{\gamma}_q^{(0)}$ for $q = 2, 3$ regarding each specific class given baseline q^* and for $\boldsymbol{\omega}$ the initial value $\boldsymbol{\omega}_q^{(0)}$ a vector of 1489 one's is used.

The predictive ability of all Bayesian variable selection methods is verified through the predictive distribution of independent and identically distributed random variables $\boldsymbol{Y}^{*(s)} = (\boldsymbol{Y}_{q^*}^{*(s)}, \ldots, \boldsymbol{Y}_{Q-1}^{*(s)})^T$ given baseline q^*, whose observed values $\boldsymbol{y}_i^{*(s)} = (y_{i,q}^{*(s)}, \ldots, y_{i,Q-1}^{*(s)})^T$ are generated for $i = 1, \ldots, n_{te}$ as the following

$$\boldsymbol{Y}_i^*|\widehat{\boldsymbol{a}}^{(s)}, \widehat{\boldsymbol{\beta}}^{(s)} \sim \mathcal{MU}\left(1; p_{i,q^*}^{*(s)}(a_{q^{*(s)}}, \boldsymbol{\beta}_{q^{*(s)}}), p_{i,1}^{*(s)}(\widehat{a}_1^{(s)}, \widehat{\boldsymbol{\beta}}_1^{(s)}), \ldots, p_{i,Q-1}^{*(s)}(\widehat{a}_{Q-1}^{(s)}, \widehat{\boldsymbol{\beta}}_{Q-1}^{(s)})\right),$$

denoting with $p_{i,q}^{*(s)}(\widehat{a}_q^{(s)}, \widehat{\boldsymbol{\beta}}_q^{(s)}) = P^{*(s)}(y_{i,q}^{*(s)} = q|\widehat{a}_q^{(s)}, \widehat{\boldsymbol{\beta}}_q^{(s)})$, the probability the i-th observation belongs to q-th class defined as

$$p_{i,q}^*(\widehat{a}_q, \widehat{\boldsymbol{\beta}}_q) = \begin{cases} \frac{1}{1+\sum_{q=1}^{Q-1} \exp{(\widehat{a}_q^{(s)}+\boldsymbol{x}_i^{te}\widehat{\boldsymbol{\beta}}_q^{(s)})}} & , q = q^* \\ \frac{\exp{(\widehat{a}_q^{(s)}+\boldsymbol{x}_i^{te}\widehat{\boldsymbol{\beta}}_q^{(s)})}}{1+\sum_{q=1}^{Q-1} \exp{(\widehat{a}_q^{(s)}+\boldsymbol{x}_i^{te}\widehat{\boldsymbol{\beta}}_q^{(s)})}} & , q \neq q^* \end{cases},$$

where $\widehat{a}_q^{(s)}$ and $\widehat{\boldsymbol{\beta}}_q^{(s)}$ are the posterior samples of q-th class specific intercepts a_q and regression coefficients $\boldsymbol{\beta}_q^{(s)}$ based either on the MAP, MPM or BMA obtained from the training set of the s-th iteration of the respective MCMC procedure. Then, based on the posterior samples from the predictive distribution, we compute the confusion matrix of the predictive response set $\boldsymbol{y}^{*(s)}$ versus the respective response of test set $\boldsymbol{y}^{te}$

		$\boldsymbol{y}^{te}$		
		Normal	Suspect	Pathologic
	Normal	$Cm_{1,1}^{(s)}$	$Cm_{1,2}^{(s)}$	$Cm_{1,3}$
$\boldsymbol{y}^{*(s)}$	Suspect	$Cm_{2,1}^{(s)}$	$Cm_{2,2}^{(s)}$	$Cm_{2,3}^{(s)}$
	Pathologic	$Cm_{3,1}^{(s)}$	$Cm_{3,2}^{(s)}$	$Cm_{3,3}^{(s)}$

Table 4.18 Confusion matrix at s-th MCMC iteration of the predictive response set $\boldsymbol{y}^{*(s)}$ versus the respective response of test set $\boldsymbol{y}^{te}$

where from the above we can calculate the accuracy as $\widehat{ACC}^{(s)} \approx \frac{\sum_{q=1}^{3} Cm_{q,q}^{(s)}}{\sum_{i=1}^{3}\sum_{q=1}^{3} Cm_{i,q}^{(s)}}$ and Cohen's kappa statistic based on the proportions of observed ($\widehat{oa}^{(s)}$) and expected agreement ($\widehat{ea}^{(s)}$) as the following shows

$$\begin{aligned}
\widehat{oa}^{(s)} &\approx \widehat{ACC}^{(s)}, \\
\widehat{ea}^{(s)} &\approx \frac{\sum_{q=1}^{3} Cm_{1,.}^{(s)} \sum_{i=1}^{3} Cm_{.,1}^{(s)}}{\left(\sum_{i=1}^{3}\sum_{q=1}^{3} Cm_{i,q}^{(s)}\right)^2} \\
&+ \frac{\sum_{q=1}^{3} Cm_{2,.}^{(s)} \sum_{i=1}^{3} Cm_{.,2}^{(s)}}{\left(\sum_{i=1}^{3}\sum_{q=1}^{3} Cm_{i,q}^{(s)}\right)^2} \\
&+ \frac{\sum_{q=1}^{3} Cm_{3,.}^{(s)} \sum_{i=1}^{3} Cm_{.,3}^{(s)}}{\left(\sum_{i=1}^{3}\sum_{q=1}^{3} Cm_{i,q}^{(s)}\right)^2}, \\
\widehat{k}_c^{(s)} &\approx \frac{\widehat{oa}^{(s)} - \widehat{ea}^{(s)}}{1 - \widehat{ea}^{(s)}},
\end{aligned}$$

where from the above measures we would expect that the values $\boldsymbol{y}^{*(s)}$ should resemble as much as with $\boldsymbol{y}^{te}$ in order to achieve good performance rates for trusting the prediction accuracy, thus we accompany the estimated accuracy $\widehat{ACC}$

$$\widehat{ACC} \approx \frac{\sum_{s=1}^{S} \widehat{ACC}^{(s)}}{S},$$

with Cohen's kappa $\widehat{\kappa}_c$ such as

$$\widehat{k}_c \approx \frac{\sum_{s=1}^{S} \widehat{k}_c^{(s)}}{S},$$

since this index is more robust to account for the overall intern observed and expected agreement of each class; the results are presented in Tables (4.20), (4.21) and (4.22) for the MAP, MPM and BMA respectively where only the subscripts of acronyms of important covariates are reported in these Tables.

Generally, we cannot claim that a method is dominant to others in terms of predictive ability as the estimated values of predictions are more or less similar to the three different prior set-ups of typical and augmented multinomial logistic regression models. More precisely, with respect to the compared results of MAP, the highest predictive precision among Bayesian variable selection procedures for typical and augmented multinomial logistic regression models is observed (notice the numbers in bold in column-wise) for ssvs.hyp.aug, gvs.hyp.typ, ssvs.g.aug and gvs.g.typ. Regarding the results based on the MPM findings, the higher predictions are evidenced for ssvs.hyp.aug, gvs.hyp.aug, ssvs.g.aug and gvs.g.aug, whereas in case of BMA, the best methods are ssvs.hyp.aug, gvs.hyp.aug, ssvs.g.aug and gvs.g.typ. Under these settings, Bayesian variable selection methods of augmented setup, tend to select slightly different model complexity with respect to their typical versions, but conserving similar complexity. Notice that the model complexity between Bayesian variable selection methods of typical and augmented setup differ only by 2 or 3 additional covariates mostly. Even if Bayesian variable selection methods among typical and augmented setup seem they didn't support exactly the same model, in the end these models will look very similar and equivalent due to the extreme collinear model space, this is why also the model complexity is quite similar among all Bayesian variable selection methods. Thus, adding or deleting some covariates respectively, may add other variables that will contribute more or less the same as the previous covariates. Additionally, we report the best predictions under the MAP, MPM and BMA that occured for ssvs.hyp.typ resulting with the highest accuracy and Cohen's kappa statistic, including one of the methods with data augmentation. However, we should be aware that SSVS is less trustworthy in contrast to GVS which is exposed to larger in sampling errors due to MCMC method. Finally additional results are illustrated in Figures (4.14) and (4.15) which depict the convergence diagnostics for the shrinkage factor $\frac{g}{g+1}$ over 40000 MCMC values for Bayesian variable selection methods with hyper-g-prior both for augmented and typical multinomial logistic regression which seem reasonable and Table (4.19) shows the satisfying acceptance rates for respective parameters. Furthermore, posterior distributions of $\widehat{ACC}$ and $\widehat{\kappa}_c$ based on BMA from the predictive distributions are depicted in Figures (4.16) and (4.17) for each Bayesian variable selection method. We didn't consider the respective posterior distribution based on the MAP and MPM because the model frequency was negligible in with respect to the number og iterations. To conclude, Bayesian variable selection methods for augmented multinomial logistic regression work better under the MPM and BMA in predicting the state of fetals based on cardiotography recordings, whereas in the case of MAP the truth is lies somewhere in the middle of the two types of Bayesian variable selection methods.

	Acceptance Rate			
Method	a_2	a_3	$\boldsymbol{\beta}$	g
ssvs.hyp.typ	0.135	0.228	0.269	0.264
ssvs.hyp.aug	-	-	-	0.268
gvs.hyp.typ	0.134	0.227	0.269	0.319
gvs.hyp.aug	-	-	-	0.323
ssvs.g.typ	0.134	0.237	0.224	-
ssvs.g.aug		-	-	-
gvs.g.typ	0.136	0.229	0.272	-
gvs.g.aug	-	-	-	-

Table 4.19 Results of acceptance rates for parameters a_2, a_3, $\boldsymbol{\beta}$ and g of each Bayesian variable selection methods for typical and augmented multinomial logistic regression model with mixtures of g-priors over 40000 of MCMC values.

Method	**Class**	**Covariates**	**MAP**	$\widehat{ACC}$	$\widehat{k}_c$
ssvs.hyp.typ	2	9	$1, 2, 3, 4, 5, 7, 11, 15, 18$	0.855	0.585
	3	13	$1, 2, 3, 4, 5, 7, 10, 11, 14, 15, 19, 20, 21$		
ssvs.hyp.aug	2	10	$1, 2, 3, 4, 5, 7, 11, 15, 17, 18, 20$	**0.867**	**0.617**
	3	11	$2, 3, 4, 5, 7, 11, 14, 15, 16, 17, 20$		
gvs.hyp.typ	2	11	$1, 2, 3, 4, 5, 7, 10, 11, 14, 15, 18$	**0.856**	**0.603**
	3	13	$1, 2, 3, 4, 5, 7, 10, 11, 14, 15, 18, 19, 20$		
gvs.hyp.aug	2	9	$1, 2, 3, 4, 5, 7, 11, 15, 18$	0.830	0.567
	3	10	$1, 3, 4, 5, 7, 8, 10, 11, 19, 20$		
ssvs.g.typ	2	12	$1, 2, 3, 4, 5, 7, 10, 11, 15, 18, 19, 20$	0.858	0.554
	3	11	$1, 2, 3, 4, 5, 7, 10, 11, 15, 18, 20$		
ssvs.g.aug	2	11	$1, 2, 3, 4, 5, 7, 10, 11, 15, 18, 20$	**0.862**	**0.567**
	3	12	$1, 2, 3, 4, 5, 7, 10, 11, 14, 15, 19, 20$		
gvs.g.typ	2	11	$1, 2, 3, 4, 5, 7, 11, 15, 17, 18, 20$	**0.854**	**0.597**
	3	12	$1, 2, 3, 4, 5, 7, 10, 11, 12, 15, 19, 20$		
gvs.g.aug	2	9	$1, 2, 3, 4, 5, 7, 11, 15, 18$	0.840	0.560
	3	10	$1, 3, 4, 5, 7, 8, 10, 11, 19, 20$		

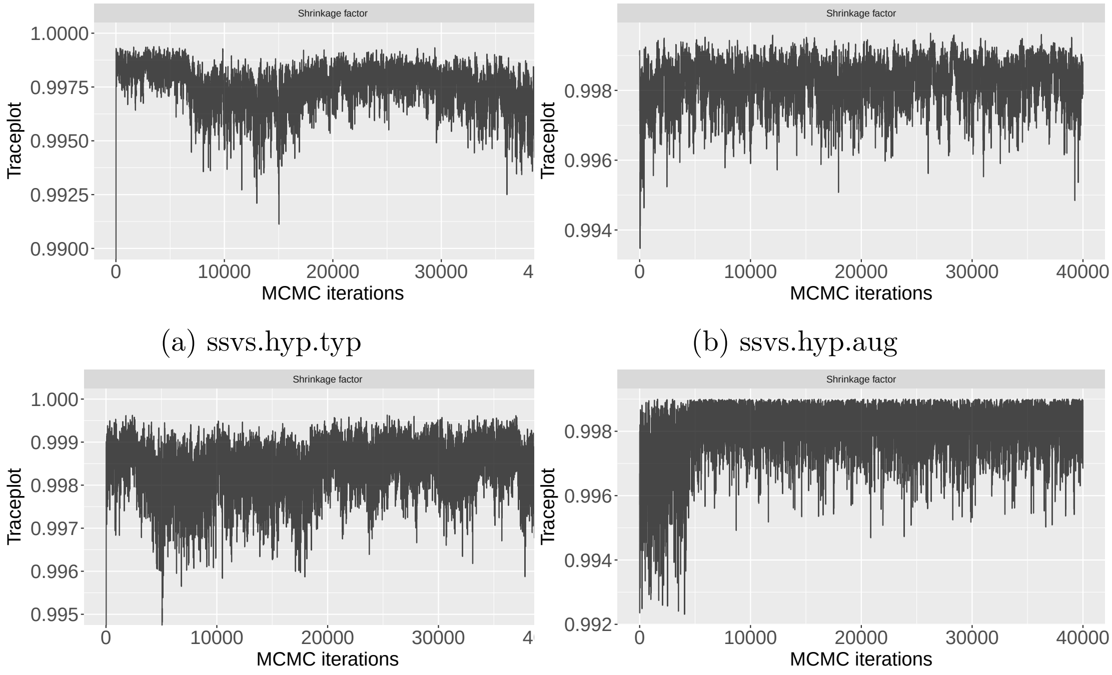
Table 4.20 Results of all Bayesian variable selection methods under the MAP for each specific class 2 and 3 given baseline class 1

Method	**Class**	**Covariates**	**MPM**	$\widehat{ACC}$	$\widehat{\kappa}_c$
ssvs.hyp.typ	2	12	$1, 2, 3, 4, 5, 7, 10, 11, 15, 17, 18, 20$	0.860	0.605
	3	12	$1, 2, 3, 4, 5, 7, 10, 11, 12, 15, 19, 20$		
ssvs.hyp.aug	2	14	$1, 2, 3, 4, 5, 7, 10, 11, 14, 15, 17, 18, 19, 20$	**0.864**	**0.610**
	3	15	$1, 2, 3, 4, 5, 6, 7, 8, 11, 14, 15, 16, 17, 19, 20$		
gvs.hyp.typ	2	11	$1, 2, 3, 4, 5, 7, 11, 15, 17, 18, 20$	0.859	0.604
	3	13	$1, 2, 3, 4, 5, 7, 8, 10, 11, 14, 15, 19, 20$		
gvs.hyp.aug	2	7	$1, 2, 3, 4, 5, 7, 11, 15, 17, 18, 20$	**0.862**	**0.605**
	3	11	$1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 14, 15, 19, 20$		
ssvs.g.typ	2	14	$1, 2, 3, 4, 5, 7, 10, 11, 12, 13, 14, 15, 18, 20$	0.832	0.542
	3	14	$1, 2, 3, 4, 5, 7, 10, 11, 12, 13, 14, 15, 19, 20$		
ssvs.g.aug	2	11	$1, 2, 3, 4, 5, 7, 10, 11, 15, 18, 20$	**0.862**	**0.567**
	3	12	$1, 2, 3, 4, 5, 7, 10, 11, 14, 15, 19, 20$		
gvs.g.typ	2	10	$1, 2, 3, 4, 5, 7, 11, 15, 17, 18, 20$	0.840	0.582
	3	11	$1, 2, 3, 4, 5, 7, 10, 11, 15, 19, 20$		
gvs.g.aug	2	11	$1, 2, 3, 4, 5, 7, 11, 15, 18, 20$	**0.856**	**0.588**
	3	11	$1, 2, 3, 4, 5, 7, 10, 11, 15, 19, 20$		

Table 4.21 Results of all Bayesian variable selection methods under the MPM for each specific class 2 and 3 given baseline class 1

Method	$\widehat{ACC}$	$\widehat{\kappa}_c$
ssvs.hyp.typ	0.860	0.555
ssvs.hyp.aug	**0.863**	**0.565**
gvs.hyp.typ	0.846	0.548
gvs.hyp.aug	**0.849**	**0.554**
ssvs.g.typ	**0.862**	0.566
ssvs.g.aug	**0.862**	**0.606**
gvs.g.typ	**0.861**	**0.562**
gvs.g.aug	0.849	0.554

Table 4.22 Results of all Bayesian variable selection methods under the BMA for each specific class 2 and 3 given baseline class 1



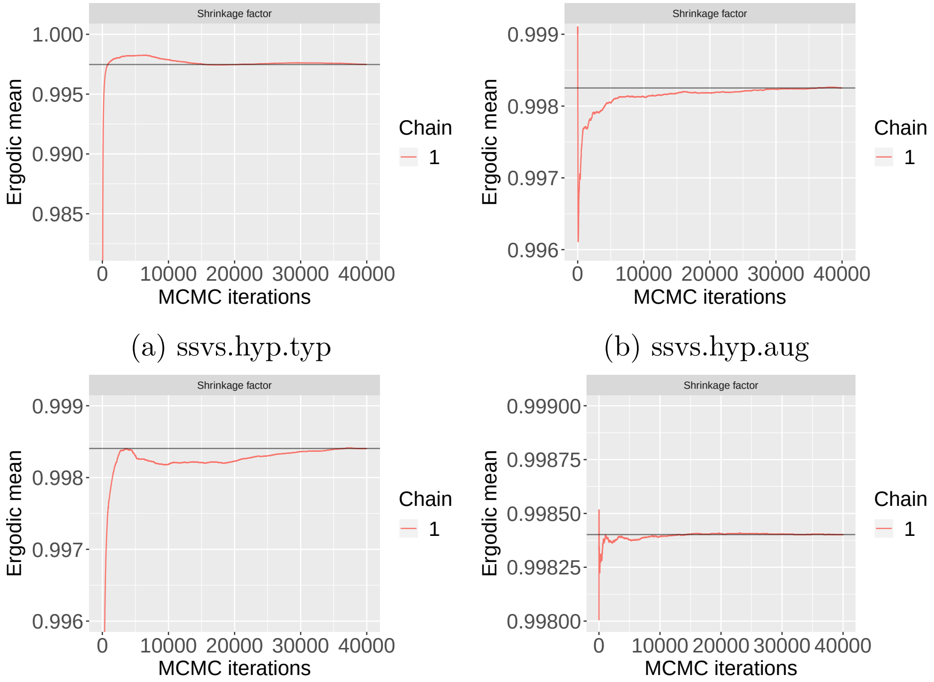
(a) ssvs.hyp.typ

(b) ssvs.hyp.aug

(c) gvs.hyp.typ

(d) gvs.hyp.aug

Fig. 4.14 Traceplots of shrinkage factor $\frac{g}{g+1}$.



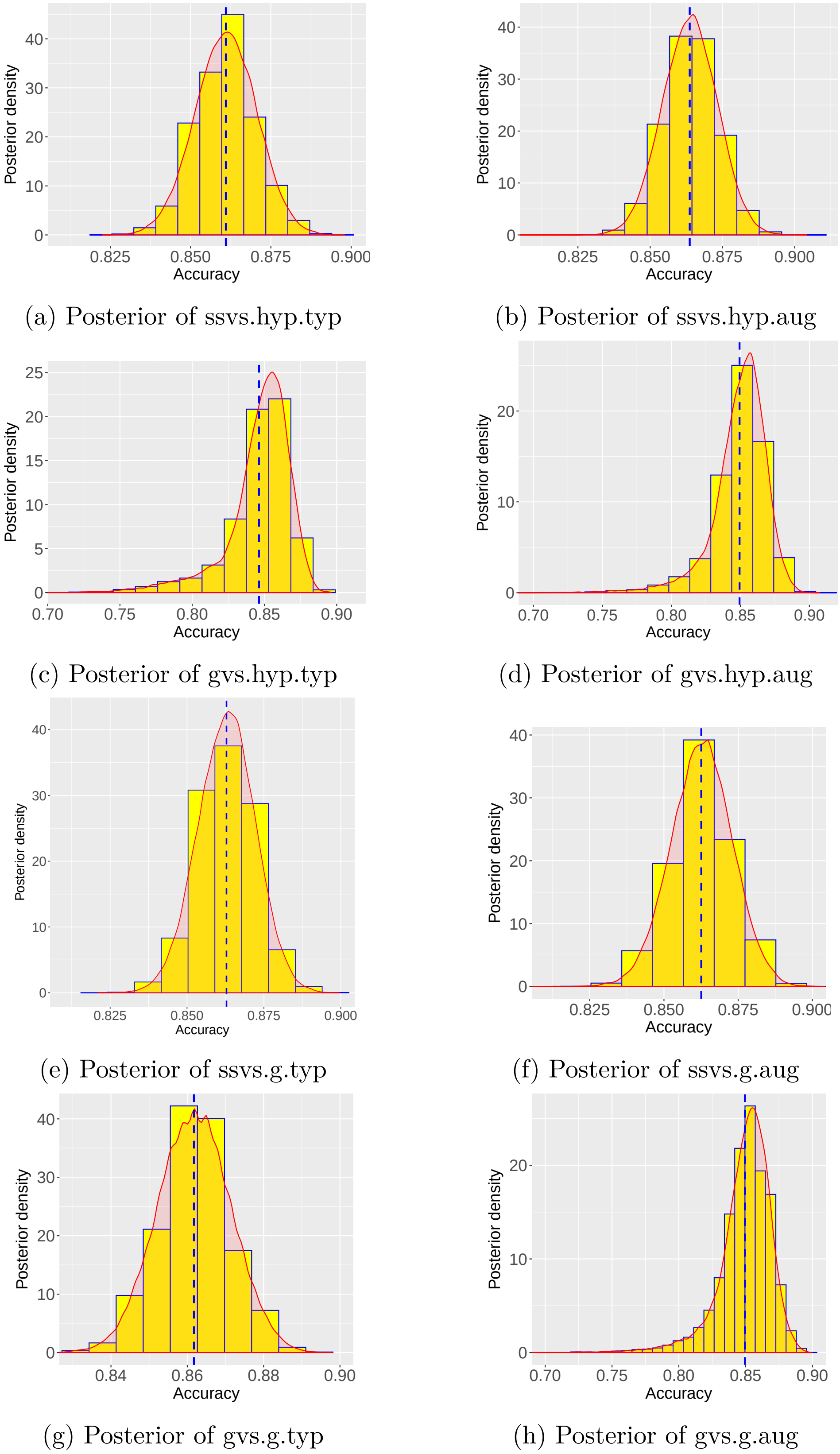
(a) ssvs.hyp.typ

(b) ssvs.hyp.aug

(c) gvs.hyp.typ

(d) gvs.hyp.aug

Fig. 4.15 Ergodic means of shrinkage factor $\frac{g}{g+1}$.



(a) Posterior of ssvs.hyp.typ

(b) Posterior of ssvs.hyp.aug

(c) Posterior of gvs.hyp.typ

(d) Posterior of gvs.hyp.aug

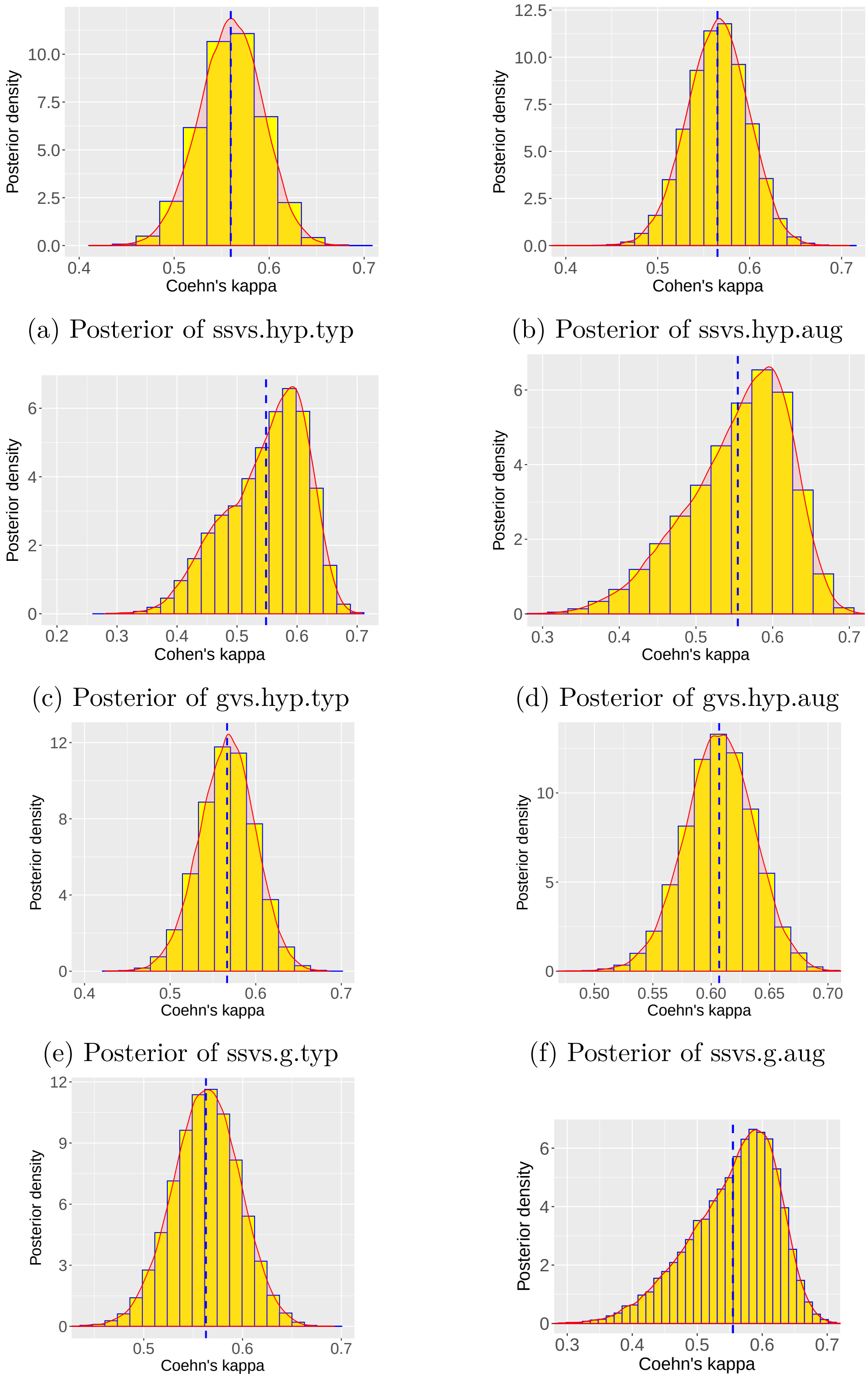
(e) Posterior of ssvs.g.typ

(f) Posterior of ssvs.g.aug

(g) Posterior of gvs.g.typ

(h) Posterior of gvs.g.aug

Fig. 4.16 Posteriors of accuracy $\widehat{ACC}$.



(a) Posterior of ssvs.hyp.typ

(b) Posterior of ssvs.hyp.aug

(c) Posterior of gvs.hyp.typ

(d) Posterior of gvs.hyp.aug

(e) Posterior of ssvs.g.typ

(f) Posterior of ssvs.g.aug

(g) Posterior of gvs.g.typ

(h) Posterior of gvs.g.aug

Fig. 4.17 Posteriors of Coehn's kappa $\widehat{\kappa}_c$.

4.12 Closing Remarks

In this chapter, the problem of variable selection in multinomial logistic regression was enlighted from a pure Bayesian point of view, which is the main research topic of this thesis. We introduced the main aspects concerning objective prior specification and model selection with mixtures of g-priors in the domain of this research area, with emphasis on novel model-based inference via MCMC. Initially, we attempted to provide an extended prior specification as that of Bové and Held (2011) for purposes of multinomial logistic regression, where the centering step facilitated the whole procedure, rendering it priority before initializing any of MCMC methods.

Next, these Bayesian variable selection methods based on the approaches George and McCulloch (1993) and Dellaportas et al. (2000) are presented and described in detail, when additional issues related to class imbalance and different covariate uncertainty except the solid such as posterior intractability and computation of posterior model probabilities, are encountered for the implementation of a typical multinomial logistic regression. We also provided the relative extensions of these Bayesian variable selection methods owing to Polya-Gamma data augmentation, as alternatives in order to cope with appeared problems. Both types of Bayesian variable selection methods were compared and evaluated in the same posterior measures for simulated and real datasets. Overall, with regard to simulated experiments, results showed good performance (depending on the simulated scenario, type of true model identification, the specific class and the sample size). In particular, all Bayesian variable selection methods between typical and augmented multinomial logistic regression worked well for both sparse scenarios under the MPM, whereas in the case of MAP's class-specific, they only worked for no-overlapping scenario. Moreover, with respect to the compared results of Bayesian variable selection methods between typical and augmented multinomial logistic regression, we cannot say that some methods dominated the other, since both methods achieve similar or better performance as the sample sizes increases (model selection consistency) (depending on the simulated scenario, type of true model identification, the specific class and the sample size), whether at initial values of sample size some methods prevail of other and so on. Even, if the between results of these methods were similar on posterior measures identifiability, we compared the computational efficiency of the two types multinomial logistic setups which showed that augmented methods require more iterations to converge proportional to the number of latent variables and are more precise in sampling efficiency. On the other hand, the main analysis of the real dataset was restricted only for Bayesian variable selection methods with hyper-g-prior and fixed g-prior. The aforementioned Bayesian variable selection methods showed good overall performance regarding the real dataset of cardiotocography, in terms of accuracy and Cohen's kappa, despite the high correlation among some variables and class imbalance. The leading results, suggested different model complexity within each class-specific 2 and 3 given baseline class 1 in predicting effectively the fetal state of prenatal babies. These results came to different conclusions towards the selection of final model according to the respective type of model identification, whether it was the MAP or MPM and BMA. In this settings, final conclusions evidenced Bayesian variable selection methods for augmented multinomial logistic regression with higher predictive ability under the MPM and BMA, whereas in the case of BMA both methods performed well in practice. All results even it seems that they traced completely different models across the class-specific, in reality there were considered neighbouring models within the strong collinearity effect which reduced them to similar models among the two types of multinomial logistic setups. To conclude, we contributed by developing and improving the traditional implementation of Bayesian variable selection methods with mixtures of g-priors in multinomial logistic regression via Polya-Gamma data augmentation, which was the principal spark inspired by Polson et al. (2013) to expose and support these ideas. The latter is a direct consequence of a nested Gibbs sampler that operates over $Q-1$ intractable augmented joint posteriors based on Polson et al. (2013) conditional likelihood identity, which result as the product of each specific class-coefficients conditional likelihoods and g-prior, if we might express this original g-prior for each specific class given then rest. In this way, the proposed methodology with the subsequent use of Gibbs sampler amenable to standard linear model results and Bayesian variable selection approaches of Bové and Held (2011) and Dellaportas et al. (2002), substitutes definitely Metropolis-Hastings sampler, surpassing the hard aspects of MCMC. Consequently, we considered also the Bayesian variable selection in logistic models as resulting a special case of multinomial logistic setup in Appendix sections C.3, C.4 and C.5. Both aforementioned Bayesian variable selection methods were compared and evaluated in the same posterior metrics for simulated and real data-sets showing similar results. Regarding the simulated study, results suggest better or similar performance of mixtures of g-priors (depending on the simulated scenario) in terms of the true model rate identification and the posterior marginal inclusion probabilities between typical and augmented logistic regression models for each respective MCMC method. In the contrary, the performance of the posterior coefficient estimators for the real dataset is similar. Practice showed, according to bibliography, that the Bayesian variable selection methods, in relation to each family of g-priors mixtures used, differentiate on terms of model complexity preference. It should be noted that the hyper-g prior family supported the models with the higher complexity. To conclude, we contributed by ameliorating the Bayesian variable selection methods by data augmentation, an idea which was ignited by contemplating on the use of Polya-Gamma data augmentation in the logistic regression context. The subsequent use of the Gibbs Sampler instead of the Metropolis - Hastings gains more stable estimates in terms of precision and autocorrelation.

Chapter 5

Conclusions and Future Research

This thesis investigated the main aspects of variable selection problem concerning objective prior specification and model selection via MCMC model-based methods with mixtures of g-priors in linear regression and generalized linear models, and then emphasized on multinomial logistic regression, which was the main topic of this research project. We developed the traditional implementation of Bayesian variable selection methods George and McCulloch (1993) and Dellaportas et al. (2002) based on prior specification of Bové and Held (2011) under sparse regularity constraints imposed by spike-slab priors, in order to explore sufficiently the model space and to ensure maximized the profit of predictions, when issues of class imbalance, posterior intractability and computation of posterior model probabilities are encountered for a typical multinomial logistic model. In addition, we proposed extensions of the aforementioned methods based on the Polya-Gamma data augmentation Polson et al. (2013) to overcome these difficulties, through a nested Gibbs sampler manageable to familiar results, under the aspect of linear models. We applied the proposed methodologies to sparse simulation settings and to the analysis of cardiotography data Ayres-de Campos et al. (2000), providing insightful comments on the comparison between the two different types of Bayesian variable selection methods. We illustrated through simulations designs on two different sparse scenarios that both methods perform more or less the same, here considered two different metrics of model identifiability respectively, reaching the same conclusions with respect to the true generating mechanism of data. The final conclusions here were the obvious ones, methods with same priors give similar results. Methods with different priors support slightly different models according to the characteristics of these priors. This is why interest was given to the computational efficiency of each method based on effective sample size and Monte Carlo standard errors. On the contrary, with respect to the real data restricting the comparisons only

between hyper-g-prior and fixed g-prior for typical and augmented multinomial logistic regression respectively, the main results of this application suggested that both Bayesian variable selection methods preferred different model complexity under three measures of model identifiability, from which result more or less different models. In particular, all Bayesian variable selection methods for augmented multinomial logistic regression methods achieved better predictive accuracy under the MPM and BMA, whereas the MAP indicated both methods of Bayesian variable selection. Based on this research finding, we shall argue that MPM and MAP are not trustworthy when comparing them with BMA, especially MAP, because they don't take the model uncertainty into account as BMA does.
Furthermore, we clarify that even a magnitude of predictive accuracy in the observed range of 0.83-0.88 in the underlying estimation, can be extremely competitive given the high correlations and class imbalance of this dataset. Moreover, the implementation of all Bayesian variable selection was based on a generalized inverse Moore-Penrose rather than the authentic, in order to surpass the singularity of authentic g-prior. Finally, we didn't consider any further comparison with state of arts methods because we emphasized more on the between comparison of typical and augmented multinomial logistic regression methods.
Finally, we discuss the possible research avenues for future development of this thesis. At first, more synthetic and accurate Bayesian variable selection may be produced assessing the overall significance of covariates in each class-specific subsets possibly, where they will investigate further the irrelevant alternatives domain and problems associated with more classes. Some other important issues in multinomial models are the following

- How the selection of the baseline category influence the results?
- How to extend the methods in order multinomial models?
- How to merge categories of the polychotomous response?
- What is the implication on the resultsand the augmented method if the g-prior is specified conditionallyon the latent variables?.

Next, our proposed Bayesian variable selection methods with or without data augmentation may be extended to, especially for GVS according to (Dellaportas et al. (2002)), to reversible jump of Green (1995) for larger model spaces. In addition, even if we didn't mentioned in GLMs chapter because we focused on moderate model spaces, the approach of Zucknick and Richardson (2014) consists of a complete summary of applications with emphasis on Bayesian variable selection with data augmentation schemes of Holmes and Held (2006) and Albert and Chib (1993) addressing the small n and large p problem Zucknick and Richardson (2014), which can extend the idea of this thesis also in high dimensional settings by integrating the respective class-specific intercepts and regression coefficients since the posteriors are normal.

Appendix A

Bayesian Variable Selection in Linear Regression

A.1 Posterior Interpretation of g-Prior

Let us suppose that we have the linear model with known variance, $\boldsymbol{Y_0}|\boldsymbol{\beta}, \sigma^2, g \sim N\left(\boldsymbol{X_0\beta}, g\sigma^2\boldsymbol{I_n}\right)$ and the corresponding prior is the usual improper one $f(\boldsymbol{\beta}) \propto 1$. The corresponding posterior of $\boldsymbol{\beta}$ can be derived through the Bayes theorem as the following

$$
\begin{aligned}
\pi(\boldsymbol{\beta}|\boldsymbol{y_0}, g, \sigma^2) \propto & f(\boldsymbol{y_0}|\boldsymbol{\beta}, \sigma^2, g)\pi(\boldsymbol{\beta}) \\
& \propto \exp\left\{-\frac{1}{2g\sigma^2}\left[(\boldsymbol{y_0} - \boldsymbol{X_0}\beta)^{\boldsymbol{T}}(\boldsymbol{y_0} - \boldsymbol{X_0}\beta)\right]\right\} \\
& = \exp\left\{-\frac{1}{2g\sigma^2}\left[(\boldsymbol{y_0} - \boldsymbol{X_0}\widehat{\beta})^{\boldsymbol{T}}(\boldsymbol{y_0} - \boldsymbol{X_0}\widehat{\beta}) + 2(\boldsymbol{y_0} - \boldsymbol{X_0}\widehat{\beta})^{\boldsymbol{T}}\boldsymbol{X_0}(\widehat{\beta} - \beta)\right]\right\} \\
& \quad \exp\left\{-\frac{1}{2g\sigma^2}\left[(\widehat{\beta} - \beta^{\boldsymbol{T}})(\boldsymbol{X_0^T X_0})(\hat{\boldsymbol{\beta}} - \beta)\right]\right\} \\
& \propto \exp\left\{-\frac{1}{2g\sigma^2}\left[2(\boldsymbol{y_0} - \boldsymbol{X_0}\widehat{\beta})^{\boldsymbol{T}}\boldsymbol{X_0}(\widehat{\beta} - \beta) + (\beta - \widehat{\beta})^{\boldsymbol{T}}(\boldsymbol{X_0^T X_0})(\beta - \widehat{\beta})\right]\right\},
\end{aligned}
$$

from the last step, we can observe that

$$
\begin{aligned}
(\boldsymbol{y_0} - \boldsymbol{X_0}\hat{\boldsymbol{\beta}})^{\boldsymbol{T}}\boldsymbol{X_0}(\widehat{\beta} - \beta) &= \boldsymbol{y_0^T X_0}\widehat{\beta} - \boldsymbol{y_0^T X_0}\beta - \widehat{\beta}^{\boldsymbol{T}}(\boldsymbol{X_0^T X_0})\widehat{\beta} + \widehat{\beta}^{\boldsymbol{T}}(\boldsymbol{X_0^T X_0})\beta \\
&= \boldsymbol{y_0^T X_0}(\boldsymbol{X_0^T X_0})^{-1}\boldsymbol{y_0} - \boldsymbol{y_0^T X_0}\beta - \boldsymbol{y_0^T X_0}(\boldsymbol{X_0^T X_0})^{-1}\boldsymbol{y_0} \\
&+ \boldsymbol{y_0^T X_0}(\boldsymbol{X_0^T X_0})^{-1}(\boldsymbol{X_0^T X_0})\beta \\
&= 0,
\end{aligned}
$$

where this step leads to the posterior distribution of $\boldsymbol{\beta}$ as the following

$$\pi(\boldsymbol{\beta}|\boldsymbol{y_0}, g, \sigma^2) \propto \exp\left\{-\frac{1}{2g\sigma^2}\left[(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})^{\boldsymbol{T}}(\boldsymbol{X_0^T X_0})(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\right]\right\}, \tag{A.1}$$

whereas (A.1) can be recognized as the kernel of a normal density distribution, $N_p(\widehat{\boldsymbol{\beta}}, g\sigma^2(\boldsymbol{X_0^T X_0})^{-1})$, where $\widehat{\boldsymbol{\beta}}$ reduces to $\mathbf{0}_p$ for a imaginary sample size $\boldsymbol{y}_0 = \mathbf{0}_{n^*}$.

A.2 Deriving the Joint Posterior of a, $\boldsymbol{\beta}_\gamma$ and σ^2

Let's assume that we have the $\boldsymbol{\gamma}$ linear model $\boldsymbol{y}, \boldsymbol{\beta}_\gamma, \sigma^2|\boldsymbol{\gamma} \sim N(\boldsymbol{a}\mathbf{1}_{\boldsymbol{n}} + \boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma, \sigma^2)$ with prior under Zellner's g-prior $\pi(\alpha, \boldsymbol{\beta}_\gamma, \sigma^2|\boldsymbol{\gamma}) = \pi(a, \sigma^2|\boldsymbol{\gamma})\pi(\boldsymbol{\beta}_\gamma|g, \sigma^2, \boldsymbol{\gamma})$, where $\pi(a, \sigma^2)$, $\pi(\boldsymbol{\beta}_\gamma|g, \sigma^2, \boldsymbol{\gamma})$ are denoted through (2.6), (2.7) respectively. The posterior can be calculated again through Bayes theorem as

$$\begin{aligned}
&\pi(\boldsymbol{\beta}_\gamma, a, \sigma^2|\boldsymbol{y}, g) \propto f(\boldsymbol{y}|\boldsymbol{\beta}_\gamma, a, \sigma^2, \boldsymbol{\gamma})\pi(a, \boldsymbol{\beta}_\gamma, \sigma^2|\boldsymbol{\gamma})\\
&= (\sigma^2)^{-1}(\sigma^2)^{-\frac{n}{2}}(\sigma^2)^{-\frac{p\gamma}{2}}\\
&\exp\left\{-\frac{1}{2\sigma^2}\left[(\boldsymbol{y}-\bar{y}\mathbf{1}_{\boldsymbol{n}}-\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma)^{\boldsymbol{T}}(\boldsymbol{y}-\bar{y}\mathbf{1}_{\boldsymbol{n}}-\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma)\right]\right\}\\
&\exp\left\{-\frac{1}{2\sigma^2}\left[2(\boldsymbol{y}-\bar{y}\mathbf{1}_{\boldsymbol{n}}-\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma)^{\boldsymbol{T}}((\bar{y}-a)\mathbf{1}_{\boldsymbol{n}}))\right]\right\}\\
&\exp\left\{-\frac{1}{2\sigma^2}\left[((\bar{y}-a)\mathbf{1}_{\boldsymbol{n}})^{\boldsymbol{T}}((\bar{y}-a)\mathbf{1}_{\boldsymbol{n}})\right]\right\}\exp\left\{-\frac{1}{2g\sigma^2}\left[(\boldsymbol{\beta}_\gamma)(\boldsymbol{X}_\gamma^{\boldsymbol{T}}\boldsymbol{X}_\gamma)(\boldsymbol{\beta}_\gamma)\right]\right\},
\end{aligned}$$

and we notice that the last step can be further reduced as

$$\begin{aligned}
(\boldsymbol{y}-\bar{y}\mathbf{1}_{\boldsymbol{n}}-\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma)^{\boldsymbol{T}}((\bar{y}-a)\mathbf{1}_{\boldsymbol{n}})) &= (\bar{y}-a)(\boldsymbol{y}-\bar{y}\mathbf{1}_{\boldsymbol{n}}-\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma)^{\boldsymbol{T}}\mathbf{1}_{\boldsymbol{n}}\\
&= (\bar{y}-a)(\boldsymbol{y}^{\boldsymbol{T}}\mathbf{1}_{\boldsymbol{n}}-\bar{y}\mathbf{1}_{\boldsymbol{n}}^{\boldsymbol{T}}\mathbf{1}_{\boldsymbol{n}}-\boldsymbol{\beta}_\gamma^{\boldsymbol{T}}\boldsymbol{X}_\gamma^{\boldsymbol{T}}\mathbf{1}_{\boldsymbol{n}})\\
&= n\bar{y}-n\bar{y}\\
&= 0,
\end{aligned}$$

so taking into account the previous steps, we may write the posterior as

$$\begin{aligned}
&\pi(\boldsymbol{\beta}_\gamma, a, \sigma^2|\boldsymbol{y}, g) \propto f(\boldsymbol{y}|\boldsymbol{\beta}_\gamma, a, \sigma^2, \boldsymbol{\gamma})\pi(a, \boldsymbol{\beta}_\gamma, \sigma^2|\boldsymbol{\gamma})\\
&= (\sigma^2)^{-1}(\sigma^2)^{-\frac{n}{2}}(\sigma^2)^{-\frac{p_\gamma}{2}}\\
&\exp\left\{-\frac{1}{2\sigma^2}\left[(\boldsymbol{y} - \bar{y}\boldsymbol{1_n} - \boldsymbol{X}_\gamma\widehat{\boldsymbol{\beta}}_\gamma)^{\boldsymbol{T}}(\boldsymbol{y} - \bar{y}\boldsymbol{1_n} - \boldsymbol{X}_\gamma\widehat{\boldsymbol{\beta}}_\gamma)\right]\right\}\\
&\exp\left\{-\frac{1}{2\sigma^2}\left[2(\boldsymbol{y} - \bar{y}\boldsymbol{1_n} - \boldsymbol{X}_\gamma\widehat{\boldsymbol{\beta}}_\gamma)^{\boldsymbol{T}}\boldsymbol{X}_\gamma(\widehat{\boldsymbol{\beta}}_\gamma - \boldsymbol{\beta}_\gamma)\right]\right\}\\
&\exp\left\{-\frac{1}{2\sigma^2}\left[(\widehat{\boldsymbol{\beta}}_\gamma - \boldsymbol{\beta}_\gamma)^{\boldsymbol{T}}(\boldsymbol{X}_\gamma^{\boldsymbol{T}}\boldsymbol{X}_\gamma)(\widehat{\boldsymbol{\beta}}_\gamma - \boldsymbol{\beta}_\gamma)\right]\right\}\\
&\exp\left\{-\frac{1}{2\sigma^2}\left[n(\bar{y} - a)^2\right]\right\}\exp\left\{-\frac{1}{2g\sigma^2}\left[(\boldsymbol{\beta}_\gamma)^{\boldsymbol{T}}(\boldsymbol{X}_\gamma^{\boldsymbol{T}}\boldsymbol{X}_\gamma)(\boldsymbol{\beta}_\gamma)\right]\right\},
\end{aligned}$$

where the last expression may be simplified as

$$\begin{aligned}
&(\boldsymbol{y} - \bar{y}\boldsymbol{1_n} - \boldsymbol{X}_\gamma\widehat{\boldsymbol{\beta}}_\gamma)^{\boldsymbol{T}}\boldsymbol{X}_\gamma(\widehat{\boldsymbol{\beta}}_\gamma - \boldsymbol{\beta}_\gamma) =\\
&= (\boldsymbol{y} - \bar{y}\boldsymbol{1_n})^{\boldsymbol{T}}\boldsymbol{X}_\gamma\widehat{\boldsymbol{\beta}}_\gamma - (\boldsymbol{y} - \bar{y}\boldsymbol{1_n})^{\boldsymbol{T}}\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma - (\widehat{\boldsymbol{\beta}}_\gamma)^{\boldsymbol{T}}(\boldsymbol{X}_\gamma^{\boldsymbol{T}}\boldsymbol{X}_\gamma)(\widehat{\boldsymbol{\beta}}_\gamma) + (\widehat{\boldsymbol{\beta}}_\gamma)^{\boldsymbol{T}}(\boldsymbol{X}_\gamma^{\boldsymbol{T}}\boldsymbol{X}_\gamma)(\boldsymbol{\beta}_\gamma)\\
&= (\boldsymbol{y} - \bar{y}\boldsymbol{1_n})^{\boldsymbol{T}}\boldsymbol{X}_\gamma(\boldsymbol{X}_\gamma^{\boldsymbol{T}}\boldsymbol{X}_\gamma)^{-1}\boldsymbol{X}_\gamma^{\boldsymbol{T}}(\boldsymbol{y}) - (\boldsymbol{y} - \bar{y}\boldsymbol{1_n})^{\boldsymbol{T}}\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma\\
&- (\boldsymbol{y})^{\boldsymbol{T}}\boldsymbol{X}_\gamma(\boldsymbol{X}_\gamma^{\boldsymbol{T}}\boldsymbol{X}_\gamma)^{-1}\boldsymbol{X}_\gamma^{\boldsymbol{T}}(\boldsymbol{y}) + (\boldsymbol{y})^{\boldsymbol{T}}\boldsymbol{X}_\gamma(\boldsymbol{X}_\gamma^{\boldsymbol{T}}\boldsymbol{X}_\gamma)^{-1}\boldsymbol{X}_\gamma^{\boldsymbol{T}}\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma\\
&= 0.
\end{aligned}$$

Then, we can rexpress the posterior as

$$\begin{aligned}
&\pi(\boldsymbol{\beta}_\gamma, a, \sigma^2|\boldsymbol{y}, g) \propto f(\boldsymbol{y}|\boldsymbol{\beta}_\gamma, a, \sigma^2, \boldsymbol{\gamma})\pi(a, \boldsymbol{\beta}_\gamma, \sigma^2|\boldsymbol{\gamma})\\
&= (\sigma^2)^{-1}(\sigma^2)^{-\frac{n}{2}}(\sigma^2)^{-\frac{p_\gamma}{2}}\\
&\exp\left\{-\frac{1}{2\sigma^2}\left[(\boldsymbol{y} - \bar{y}\boldsymbol{1_n} - \boldsymbol{X}_\gamma\widehat{\boldsymbol{\beta}}_\gamma)^{\boldsymbol{T}}(\boldsymbol{y} - \bar{y}\boldsymbol{1_n} - \boldsymbol{X}_\gamma\widehat{\boldsymbol{\beta}}_\gamma)\right]\right\}\\
&\exp\left\{-\frac{1}{2\sigma^2}\left[(\widehat{\boldsymbol{\beta}}_\gamma - \boldsymbol{\beta}_\gamma)^{\boldsymbol{T}}(\boldsymbol{X}_\gamma^{\boldsymbol{T}}\boldsymbol{X}_\gamma)(\widehat{\boldsymbol{\beta}}_\gamma - \boldsymbol{\beta}_\gamma)\right]\right\}\\
&\exp\left\{-\frac{1}{2\sigma^2}\left[n(\bar{y} - a)^2\right]\right\}\exp\left\{-\frac{1}{2g\sigma^2}\left[(\boldsymbol{\beta}_\gamma)^{\boldsymbol{T}}(\boldsymbol{X}_\gamma^{\boldsymbol{T}}\boldsymbol{X}_\gamma)(\boldsymbol{\beta}_\gamma)\right]\right\},
\end{aligned}$$

and this expression is further reduced as the following

$$\begin{aligned}
&(\widehat{\boldsymbol{\beta}}_\gamma - \boldsymbol{\beta}_\gamma)^{\boldsymbol{T}}(\boldsymbol{X}_\gamma^{\boldsymbol{T}}\boldsymbol{X}_\gamma)(\widehat{\boldsymbol{\beta}}_\gamma - \boldsymbol{\beta}_\gamma) =\\
&= (\boldsymbol{\beta}_\gamma)^{\boldsymbol{T}}(\boldsymbol{X}_\gamma^{\boldsymbol{T}}\boldsymbol{X}_\gamma)\boldsymbol{\beta}_\gamma - 2(\boldsymbol{\beta}_\gamma)^{\boldsymbol{T}}(\boldsymbol{X}_\gamma^{\boldsymbol{T}}\boldsymbol{X}_\gamma)\widehat{\boldsymbol{\beta}}_\gamma + (\widehat{\boldsymbol{\beta}}_\gamma)^{\boldsymbol{T}}(\boldsymbol{X}_\gamma^{\boldsymbol{T}}\boldsymbol{X}_\gamma)(\widehat{\boldsymbol{\beta}}_\gamma),
\end{aligned}$$

where the posterior is modified as follows

$$
\begin{aligned}
&\pi(\boldsymbol{\beta}_\gamma, a, \sigma^2|\boldsymbol{y}, g) \propto f(\boldsymbol{y}|\boldsymbol{\beta}_\gamma, a, \sigma^2, \boldsymbol{\gamma})\pi(a, \boldsymbol{\beta}_\gamma, \sigma^2|\boldsymbol{\gamma})\\
&= (\sigma^2)^{-1}(\sigma^2)^{-\frac{n}{2}}(\sigma^2)^{-\frac{p_\gamma}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[(\boldsymbol{y} - \bar{y}\mathbf{1}_{\boldsymbol{n}} - \boldsymbol{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma)^{\boldsymbol{T}}(\boldsymbol{y} - \bar{y}\mathbf{1}_{\boldsymbol{n}} - \boldsymbol{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma)\right]\right\}\\
&\exp\left\{-\frac{1}{2\sigma^2}\left[(\boldsymbol{\beta}_\gamma)^{\boldsymbol{T}}(\boldsymbol{X}_\gamma^{\boldsymbol{T}}\boldsymbol{X}_\gamma)(\boldsymbol{\beta}_\gamma)\right]\right\}\exp\left\{-\frac{1}{2\sigma^2}\left[-2(\boldsymbol{\beta}_\gamma)^{\boldsymbol{T}}(\boldsymbol{X}_\gamma^{\boldsymbol{T}}\boldsymbol{X}_\gamma)(\widehat{\boldsymbol{\beta}}_\gamma)\right]\right\}\\
&\exp\left\{-\frac{1}{2\sigma^2}\left[(\widehat{\boldsymbol{\beta}}_\gamma)^{\boldsymbol{T}}(\boldsymbol{X}_\gamma^{\boldsymbol{T}}\boldsymbol{X}_\gamma)(\widehat{\boldsymbol{\beta}}_\gamma)\right]\right\}\exp\left\{-\frac{1}{2g\sigma^2}\left[(\boldsymbol{\beta}_\gamma)^{\boldsymbol{T}}(\boldsymbol{X}_\gamma^{\boldsymbol{T}}\boldsymbol{X}_\gamma)(\boldsymbol{\beta}_\gamma)\right]\right\}\exp\left\{-\frac{1}{2\sigma^2}\left[n(\bar{y} - a)^2\right]\right\},
\end{aligned}
$$

and finally we factorize the joint posterior as the following

$$
\begin{aligned}
&\pi(\boldsymbol{\beta}_\gamma, a, \sigma^2|\boldsymbol{y}, g) \propto f(\boldsymbol{y}|\boldsymbol{\beta}_\gamma, a, \sigma^2, \boldsymbol{\gamma})\pi(a, \boldsymbol{\beta}_\gamma, \sigma^2|\boldsymbol{\gamma})\\
&= (\sigma^2)^{-1}(\sigma^2)^{-\frac{n}{2}}(\sigma^2)^{-\frac{p_\gamma}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[(\boldsymbol{y} - \bar{y}\mathbf{1}_{\boldsymbol{n}} - \boldsymbol{X}_\gamma\widehat{\boldsymbol{\beta}}_\gamma)^{\boldsymbol{T}}(\boldsymbol{y} - \bar{y}\mathbf{1}_{\boldsymbol{n}} - \boldsymbol{X}_\gamma\widehat{\boldsymbol{\beta}}_\gamma) - \frac{1}{g}\widehat{\boldsymbol{\mu}}_\gamma^{\boldsymbol{T}}\widehat{\boldsymbol{C}}_\gamma^{-\mathbf{1}}\widehat{\boldsymbol{\mu}}_\gamma\right]\right\}\\
&\exp\left\{-\frac{1}{2\sigma^2}\left[(\widehat{\boldsymbol{\beta}}_\gamma)^{\boldsymbol{T}}(\boldsymbol{X}_\gamma^{\boldsymbol{T}}\boldsymbol{X}_\gamma)(\widehat{\boldsymbol{\beta}}_\gamma)\right]\right\}\exp\left\{-\frac{1}{2g\sigma^2}\left[(\boldsymbol{\beta}_\gamma)^{\boldsymbol{T}}\widehat{\boldsymbol{C}}_\gamma^{-\mathbf{1}}(\boldsymbol{\beta}_\gamma) - 2(\boldsymbol{\beta}_\gamma)^{\boldsymbol{T}}\widehat{\boldsymbol{C}}_\gamma^{-\mathbf{1}}\widehat{\boldsymbol{\mu}}_\gamma + \widehat{\boldsymbol{\mu}}_\gamma^{\boldsymbol{T}}\widehat{\boldsymbol{C}}_\gamma^{-\mathbf{1}}\widehat{\boldsymbol{\mu}}_\gamma\right]\right\}\\
&\exp\left\{-\frac{1}{2\sigma^2}\left[n(a - \bar{y})^2\right]\right\},
\end{aligned}
$$

where $\widehat{\boldsymbol{C}}_\gamma = (g+1)^{-1}(\boldsymbol{X}_\gamma^{\boldsymbol{T}}\boldsymbol{X}_\gamma)^{-1}$ and $\widehat{\boldsymbol{\mu}}_\gamma = g\widehat{\boldsymbol{C}}_\gamma^{-\mathbf{1}}(\boldsymbol{X}_\gamma^{\boldsymbol{T}}\boldsymbol{X}_\gamma)(\widehat{\boldsymbol{\beta}}_\gamma)$. These posterior distributions, for each respective parameter, are denoted as

$$a|\boldsymbol{y}, \sigma^2, \boldsymbol{\gamma} \sim N\left(\hat{a}, \frac{\sigma^2}{n}\right), \tag{A.2}$$

$$\boldsymbol{\beta}_\gamma|g, \boldsymbol{y}, \sigma^2, \boldsymbol{\gamma} \sim N_{p_\gamma}\left(\frac{g}{g+1}\widehat{\boldsymbol{\beta}}_\gamma, \sigma^2\frac{g}{g+1}(\boldsymbol{X}_\gamma^{\boldsymbol{T}}\boldsymbol{X}_\gamma)^{-1}\right), \tag{A.3}$$

$$\sigma^2|g, \boldsymbol{y}, \boldsymbol{\gamma} \sim IG\left(\frac{n-1}{2}, s_\gamma^2 + \frac{1}{g+1}\widehat{\boldsymbol{\beta}}_\gamma^{\boldsymbol{T}}\boldsymbol{X}_\gamma^{\boldsymbol{T}}\boldsymbol{X}_\gamma\widehat{\boldsymbol{\beta}}_\gamma\right), \tag{A.4}$$

where

$$\widehat{\boldsymbol{\beta}}_\gamma = (\boldsymbol{X}_\gamma^{\boldsymbol{T}}\boldsymbol{X}_\gamma)^{-1}\boldsymbol{X}_\gamma^{\boldsymbol{T}}(\boldsymbol{y}), \tag{A.5}$$

$$\hat{a} = \bar{y}, \tag{A.6}$$

denote the maximum likelihood estimators of the parameters $\boldsymbol{\beta}_\gamma$, a , $s_\gamma^2 = ||\boldsymbol{y} - \bar{y}\mathbf{1}_{\boldsymbol{n}} - \boldsymbol{X}_\gamma\widehat{\boldsymbol{\beta}}_\gamma||$ corresponds to classical residual sum of squares for each model $\boldsymbol{\gamma}$ and $||.||$ denotes the $\mathcal{L}_1$ norm

A.3 Laplace Approximation

This approximation consists of expanding a smooth unimodal function twice differentiable $H(g) = \log h(g)$ in a Taylor series expansion of second order around $\hat{g}$, the mode of $H(g)$. The Laplace approximation is precise only if there exists a unique mode $\hat{g}$, which for large samples the behaviour of integrals are depending exclusively from the mode $\hat{g}$; see Tierney et al. (1989). The mode $\hat{g}$ will result as a solution of the equation $\frac{dH(g)}{dg} = 0$. The Laplace approximation can be implemented as following

$$\begin{aligned}
\int_0^\infty \exp\left\{H(g)\right\} dg &\approx \int_0^\infty \exp\left\{ H(\hat{g}) + \left.\frac{dH(g)}{dg}\right|_{g=\hat{g}} (g-\hat{g}) + \frac{1}{2}\left.\frac{d^2H(g)}{dg^2}\right|_{g=\hat{g}} (g-\hat{g})^2 \right\} dg \\
&\approx \int_0^\infty \exp\left\{ H(\hat{g}) + \frac{1}{2}\left.\frac{d^2H(g)}{dg^2}\right|_{g=\hat{g}} (g-\hat{g})^2 \right\} dg \\
&\approx \exp\left\{H(\hat{g})\right\} \int_0^\infty \exp\left\{ \frac{1}{2}\left.\frac{d^2H(g)}{dg^2}\right|_{g=\hat{g}} (g-\hat{g})^2 \right\} dg \\
&\approx \exp\left\{H(\hat{g})\right\} \int_0^\infty \exp\left\{ -\frac{1}{2}\left.\frac{-d^2H(g)}{dg^2}\right|_{g=\hat{g}} (g-\hat{g})^2 \right\} dg,
\end{aligned}$$

in the steps above it is evident $\left.\frac{dH(g)}{dg}\right|_{g=\hat{g}} = 0$ and the term $\int_0^\infty \exp\left\{-\frac{1}{2}\left.\frac{-d^2H(g)}{dg^2}\right|_{g=\hat{g}} (g-\hat{g})^2\right\} dg$, is recognized as the integrated kernel of a normal distribution denoted as $N\left(\hat{g}, \left[\left.\frac{-d^2H(g)}{dg^2}\right|_{g=\hat{g}}\right]^{-1}\right)$, while all steps are leading to a complete Laplace approximation written as

$$\int_0^\infty \exp\left\{H(g)\right\} dg \approx \sqrt{2\pi}\hat{\sigma}_H h(\hat{g}),$$

where $\hat{\sigma}_H = \left[\left.\frac{-d^2H(g)}{dg^2}\right|_{g=\hat{g}}\right]^{-\frac{1}{2}}$. Moreover, it is important to notice that $\hat{g}$ is the mode of $H(g)$, equivalently of $h(g)$, because the logarithm is a bijective function and becomes the mode due to the fact that $\hat{\sigma}_H$ must be a positive quantity and to be positive, this quantity must be negative $\left.\frac{-d^2H(g)}{dg^2}\right|_{g=\hat{g}}$ in order to evaluate the Laplace approximation; see for more information (Liang et al., 2008); (Tierney et al., 1989).

A.4 Laplace Approximation for Zellner-Siow Priors

For the Bayes factor of (2.11) we consider

$$H^{ZS}(g) = \frac{(n-p_\gamma-1)}{2}\log(1+g) - \frac{(n-1)}{2}\log\left[1+(1-R^2_\gamma)g\right] - \frac{3}{2}\log g - \frac{n}{2g}, \quad \text{(A.7)}$$

where the respective derivative is computed as

$$\frac{dH^{ZS}(g)}{dg} = \frac{(n-p_\gamma-1)}{2}\frac{1}{1+g} - \frac{(n-1)}{2}\left[\frac{1-R^2_\gamma}{1+(1-R^2_\gamma)g}\right] - \frac{3}{2}\frac{1}{g} + \frac{n}{2g}, \qquad \text{(A.8)}$$

and in order to obtain the mode $\hat{g}^{ZS}$ of H^{ZS} we have to solve equation $\frac{dH^{ZS}(g)}{dg} = 0$ as the following

$$\begin{aligned}
&2g^2g(1+g)\left[1+(1-R^2_\gamma)g\right]\frac{(n-p_\gamma-1)}{2}\frac{1}{1+g}\\
&-2g^2g(1+g)\left[1+(1-R^2_\gamma)g\right]\frac{(n-1)}{2}\left[\frac{1-R^2_\gamma}{1+(1-R^2_\gamma)g}\right] - \frac{3}{2g}2g^2g(1+g)\left[1+(1-R^2_\gamma)g\right]\\
&2g^2g(1+g)\left[1+(1-R^2_\gamma)g\right]\frac{n}{2g^2} = 0,
\end{aligned}$$

where after some mathematical steps we have

$$\begin{aligned}
&g^2g\left[1+(1-R^2_\gamma)g\right](n-p_\gamma-1) - g^2g(1+g)(n-1)(1-R^2_\gamma)\\
&-3g^2(1+g)\left[1+(1-R^2_\gamma)g\right] + g(1+g)\left[1+(1-R^2_\gamma)g\right]n = 0,
\end{aligned}$$

and going on with calculations we have

$$\begin{aligned}
&g^3(n-p_\gamma-1) + g^4(1-R^2_\gamma)(n-p_\gamma-1) - g^3(n-1)(1-R^2_\gamma) - g^4(n-1)(1-R^2_\gamma)\\
&-3g^3 - 3g^4(1-R^2_\gamma) - 3g^2 - 3g^3(1-R^2_\gamma) + ng + ng^2(1-R^2_\gamma) + ng^2 + ng^3(1-R^2_\gamma) = 0,
\end{aligned}$$

The previous expression is reduced on a cubic equation since $g > 0$ with coefficients a_2, a_1, a_0

$$g^3 + a_2g^2 + a_1g + a_0 = 0,$$

where a_2, a_1, a_0 are expressed as

$$\begin{aligned}
a_2 &= -\left[\frac{n - p_\gamma - 4 - 2(1 - R_\gamma^2)}{(1 - R_\gamma^2)(p_\gamma + 3)}\right],\\
a_1 &= -\left[\frac{n(2 - R_\gamma^2) - 3}{(1 - R_\gamma)(p_\gamma + 3)}\right],\\
a_0 &= -\left[\frac{n}{(1 - R_\gamma^2)(p_\gamma + 3)}\right].
\end{aligned}$$

The previous coefficients are resulted useful, in the calculation of the unique real mode $\hat{g}^{ZS}$ according to (Abramowitz and Stegun, 1970) and Liang et al. (2008) as

$$\hat{g}^{ZS} = \left(\left[r + (q^3 + r^2)^{\frac{1}{2}}\right]^{\frac{1}{2}} + \left[r - (q^3 + r^2)^{\frac{1}{2}}\right]^{\frac{1}{2}}\right) - \frac{a_2}{3}, \tag{A.9}$$

where the quantities r, q are calculated as

$$\begin{aligned}
r &= \frac{1}{6}(a_1 a_2 + 3a_0) - \frac{1}{27}a_2^3,\\
q &= \frac{1}{3}a_1 - \frac{1}{9}a_2^2,
\end{aligned}$$

which coincide with equations. The standard error of the mode $\hat{g}^{ZS}$ is found by plugging in the mode in the second derivative of H_{ZS}. This is done by first calculating the second derivative of H^{ZS} as the following

$$\frac{d^2H^{ZS}(g)}{dg^2} = -\frac{n - p_\gamma - 1}{2}\frac{1}{(g+1)^2} + \frac{n-1}{2}\left[\frac{1 - R_\gamma}{1 + (1 - R_\gamma)g}\right]^2 + \frac{3}{2}\frac{1}{g^2} - \frac{n}{g^3},$$

secondly we compute the standard error $\hat{\sigma}_{H^{ZS}}$ of the mode $\hat{g}^{ZS}$ as

$$\hat{\sigma}_{H^{ZS}} = \left[\left.\frac{-d^2H^{ZS}(g)}{dg^2}\right|_{g=\hat{g}^{ZS}}\right]^{-\frac{1}{2}}. \tag{A.10}$$

A.5 Calculation of Bayes Factor and Posterior Moments of g and $\frac{g}{g+1}$

The Bayes factor (2.14) of hyper-g prior

$$BF^{hy}_{[\gamma:\gamma_0]} = \frac{\alpha-2}{2}\int_0^\infty (1+g)^{\frac{n-1-p_\gamma-\alpha}{2}}[1+g(1-R^2_\gamma)]^{-\frac{n-1}{2}}dg,$$

where in the above integral we use the transformation $u=\frac{g}{g+1}$ in order to deal in a more easier way. The transformed prior can be found by the transform theorem of distributions as the following

$$u=\frac{g}{g+1}\Leftrightarrow g=\frac{u}{1-u}, \tag{A.11}$$

$$\left|\frac{dg}{du}\right| = \left(\frac{1}{1-u}\right)^2, \tag{A.12}$$

$$g>0\Leftrightarrow \frac{u}{1-u}>0\Leftrightarrow u\in(0,1), \tag{A.13}$$

$$\pi(u)=\frac{\alpha-2}{2}u^{\frac{\alpha}{2}-2}, \tag{A.14}$$

where from (A.14) we notice that the transformed variable $u\sim Beta\left(1,\frac{\alpha}{2}-1\right)$ which coincides with the prior distribution of the shrinkage factor $u=\frac{g}{g+1}$. Based on these calculations, the Bayes factor (2.14) of hyper-g prior is modified in terms of u as

$$BF^{hy}_{[\gamma:\gamma_0]} = \frac{\alpha-2}{2}\int_0^1 (1-u)^{\frac{p_\gamma+\alpha}{2}-2}[1-uR^2_\gamma]^{-\frac{n-1}{2}}du,$$

where the above integral consists of a well known form distribution which can be handled by the Gaussian hypergeometric function ${}_2F_1(.)$ as the following

$$\begin{aligned} BF^{hy}_{[\gamma:\gamma_0]} &= \left(\frac{\alpha-2}{2}\right)\frac{\Gamma(1)\Gamma\left(\frac{p_\gamma+\alpha}{2}-1\right)\,{}_2F_1\left(\frac{n-1}{2},1;\frac{p_\gamma+\alpha}{2};R^2_\gamma\right)}{\Gamma\left(\frac{p_\gamma+\alpha}{2}\right)} \\ &= \left(\frac{\alpha-2}{2}\right)\frac{\Gamma\left(\frac{p_\gamma+\alpha}{2}\right)\,{}_2F_1\left(\frac{n-1}{2},1;\frac{p_\gamma+\alpha}{2};R^2_\gamma\right)}{\left(\frac{p_\gamma+\alpha}{2}-1\right)\Gamma\left(\frac{p_\gamma+\alpha}{2}\right)} \\ &= \frac{\alpha-2}{p_\gamma+\alpha-2}\,{}_2F_1\left(\frac{n-1}{2},1;\frac{p_\gamma+\alpha}{2};R^2_\gamma\right), \end{aligned}$$

which corresponds to equation (2.15). On the other hand, the posterior expectation of g is defined as

$$\begin{aligned}
\mathbb{E}(g|\boldsymbol{\gamma},\boldsymbol{y}) &= \frac{\int_0^\infty g\pi(g|\boldsymbol{\gamma},\boldsymbol{y})dg}{\int_0^\infty \pi(g|\boldsymbol{\gamma},\boldsymbol{y})dg} = \\
&= \frac{\int_0^\infty g(1+g)^{\frac{n-1-p_\gamma-\alpha}{2}}[1+g(1-R^2_\gamma)]^{-\frac{n-1}{2}}dg}{\int_0^\infty (1+g)^{\frac{n-1-p_\gamma-\alpha}{2}}[1+g(1-R^2_\gamma)]^{-\frac{n-1}{2}}dg},
\end{aligned}$$

the integral representation in the last step is computed in simpler form using again the transformation $u=\frac{g}{g+1}$ which includes the Gaussian hypergeometric function ${}_2F_1(.)$ as

$$\begin{aligned}
\mathbb{E}(g|\boldsymbol{\gamma},\boldsymbol{y}) &= \frac{\int_0^1 u(1-u)^{\frac{p_\gamma+\alpha}{2}-3}[1-uR^2_\gamma]^{-\frac{n-1}{2}}du}{\int_0^1 (1-u)^{\frac{p_\gamma+\alpha}{2}-2}[1-uR^2_\gamma]^{-\frac{n-1}{2}}du} \\
&= \frac{\Gamma\left(\frac{p_\gamma+\alpha}{2}\right)\Gamma(2)\Gamma\left(\frac{p_\gamma+\alpha}{2}-2\right)\,{}_2F_1\left(\frac{n-1}{2},2;\frac{p_\gamma+\alpha}{2};R^2_\gamma\right)}{\Gamma\left(\frac{p_\gamma+\alpha}{2}\right)\Gamma(1)\Gamma\left(\frac{p_\gamma+\alpha}{2}-1\right)\,{}_2F_1\left(\frac{n-1}{2},1;\frac{p_\gamma+\alpha}{2};R^2_\gamma\right)} \\
&= \frac{\Gamma\left(\frac{p_\gamma+\alpha}{2}-2\right)\,{}_2F_1\left(\frac{n-1}{2},2;\frac{p_\gamma+\alpha}{2};R^2_\gamma\right)}{\left(\frac{p_\gamma+\alpha}{2}-2\right)\Gamma\left(\frac{p_\gamma+\alpha}{2}-2\right)\,{}_2F_1\left(\frac{n-1}{2},1;\frac{p_\gamma+\alpha}{2};R^2_\gamma\right)} \\
&= \frac{2\,{}_2F_1\left(\frac{n-1}{2},2;\frac{p_\gamma+\alpha}{2};R^2_\gamma\right)}{(p_\gamma+\alpha-4)\,{}_2F_1\left(\frac{n-1}{2},1;\frac{p_\gamma+\alpha}{2};R^2_\gamma\right)}. \qquad \text{(A.15)}
\end{aligned}$$

The posterior mean of $\frac{g}{g+1}$ is defined as

$$\begin{aligned}
\mathbb{E}\left(\frac{g}{g+1}\middle|\boldsymbol{\gamma},\boldsymbol{y}\right) &= \frac{\int_0^\infty \left(\frac{g}{g+1}\right)\pi(g|\boldsymbol{\gamma},\boldsymbol{y})dg}{\int_0^\infty f(g|\boldsymbol{\gamma},\boldsymbol{y})dg} \\
&= \frac{\int_0^\infty g(1+g)^{\frac{n-1-p_\gamma-\alpha}{2}-1}[1+g(1-R^2_\gamma)]^{-\frac{n-1}{2}}dg}{\int_0^\infty (1+g)^{\frac{n-1-p_\gamma-\alpha}{2}}[1+g(1-R^2_\gamma)]^{-\frac{n-1}{2}}dg},
\end{aligned}$$

the above integral form in the last step is handled in a more easier way using the transformation $u=\frac{g}{g+1}$ which includes the Gaussian hypergeometric function ${}_2F_1(.)$ as the following

$$\begin{aligned}
\mathbb{E}\left(\frac{g}{g+1}\middle|\boldsymbol{\gamma},\boldsymbol{y}\right) &= \frac{\int_0^1 u(1-u)^{\frac{p_\gamma+\alpha}{2}-2}[1-uR_\gamma^2]^{-\frac{n-1}{2}}du}{\int_0^1 (1-u)^{\frac{p_\gamma+\alpha}{2}-2}[1-uR_\gamma^2]^{-\frac{n-1}{2}}du} \\
&= \frac{\Gamma\left(\frac{p_\gamma+\alpha}{2}\right)\Gamma(2)\Gamma\left(\frac{p_\gamma+\alpha}{2}-1\right)\,{}_2F_1\left(\frac{n-1}{2},2;\frac{p_\gamma+\alpha}{2}+1;R_\gamma^2\right)}{\Gamma\left(\frac{p_\gamma+\alpha}{2}+1\right)\Gamma(1)\Gamma\left(\frac{p_\gamma+\alpha}{2}-1\right)\,{}_2F_1\left(\frac{n-1}{2},1;\frac{p_\gamma+\alpha}{2};R_\gamma^2\right)} \\
&= \frac{\Gamma\left(\frac{p_\gamma+\alpha}{2}\right)\,{}_2F_1\left(\frac{n-1}{2},2;\frac{p_\gamma+\alpha}{2}+1;R_\gamma^2\right)}{\Gamma\left(\frac{p_\gamma+\alpha}{2}+1\right)\,{}_2F_1\left(\frac{n-1}{2},1;\frac{p_\gamma+\alpha}{2};R_\gamma^2\right)} \\
&= \frac{\Gamma\left(\frac{p_\gamma+\alpha}{2}\right)\,{}_2F_1\left(\frac{n-1}{2},2;\frac{p_\gamma+\alpha}{2}+1;R_\gamma^2\right)}{\left(\frac{p_\gamma+\alpha}{2}\right)\Gamma\left(\frac{p_\gamma+\alpha}{2}\right)\,{}_2F_1\left(\frac{n-1}{2},1;\frac{p_\gamma+\alpha}{2};R_\gamma^2\right)} \\
&= \frac{{}_2F_1\left(\frac{n-1}{2},2;\frac{p_\gamma+\alpha}{2}+1;R_\gamma^2\right)}{\left(\frac{p_\gamma+\alpha}{2}\right)\,{}_2F_1\left(\frac{n-1}{2},1;\frac{p_\gamma+\alpha}{2};R_\gamma^2\right)} \\
&= \frac{2\,{}_2F_1\left(\frac{n-1}{2},2;\frac{p_\gamma+\alpha}{2}+1;R_\gamma^2\right)}{(p_\gamma+\alpha)\,{}_2F_1\left(\frac{n-1}{2},1;\frac{p_\gamma+\alpha}{2};R_\gamma^2\right)}. \qquad \text{(A.16)}
\end{aligned}$$

In addition, the posterior mean of g^2 is defined as

$$\mathbb{E}(g^2|\boldsymbol{\gamma},\boldsymbol{y}) = \frac{\int_0^\infty g^2 f(g|\boldsymbol{\gamma},\boldsymbol{y})dg}{\int_0^\infty f(g|\boldsymbol{\gamma},\boldsymbol{y})dg} = \frac{\int_0^\infty g^2(1+g)^{\frac{n-1-p_\gamma-\alpha}{2}}[1+g(1-R_\gamma^2)]^{-\frac{n-1}{2}}dg}{\int_0^\infty (1+g)^{\frac{n-1-p_\gamma-\alpha}{2}}[1+g(1-R_\gamma^2)]^{-\frac{n-1}{2}}dg},$$

where again the integral representation in the last step is computed in simpler form using again the transformation $u=\frac{g}{g+1}$ which includes the Gaussian hypergeometric function ${}_2F_1(.)$ as

$$
\begin{aligned}
\mathbb{E}(g^2|\boldsymbol{\gamma},\boldsymbol{y}) &= \frac{\int_0^1 u^2(1-u)^{\frac{p_\gamma+\alpha}{2}-4}[1-uR^2_\gamma]^{-\frac{n-1}{2}}du}{\int_0^1 (1-u)^{\frac{p_\gamma+\alpha}{2}-2}[1-uR^2_\gamma]^{-\frac{n-1}{2}}du} \\
&= \frac{\Gamma\left(\frac{p_\gamma+\alpha}{2}\right)\Gamma(3)\Gamma\left(\frac{p_\gamma+\alpha}{2}-3\right)\,{}_2F_1\left(\frac{n-1}{2},3;\frac{p_\gamma+\alpha}{2};R^2_\gamma\right)}{\Gamma\left(\frac{p_\gamma+\alpha}{2}\right)\Gamma(1)\Gamma\left(\frac{p_\gamma+\alpha}{2}-1\right)\,{}_2F_1\left(\frac{n-1}{2},1;\frac{p_\gamma+\alpha}{2};R^2_\gamma\right)} \\
&= \frac{2\Gamma\left(\frac{p_\gamma+\alpha}{2}-3\right)\,{}_2F_1\left(\frac{n-1}{2},3;\frac{p_\gamma+\alpha}{2};R^2_\gamma\right)}{\Gamma\left(\frac{p_\gamma+\alpha}{2}-1\right)\,{}_2F_1\left(\frac{n-1}{2},1;\frac{p_\gamma+\alpha}{2};R^2_\gamma\right)} \\
&= \frac{2\Gamma\left(\frac{p_\gamma+\alpha}{2}-2\right)\,{}_2F_1\left(\frac{n-1}{2},3;\frac{p_\gamma+\alpha}{2};R^2_\gamma\right)}{\left(\frac{p_\gamma+\alpha}{2}-3\right)\left(\frac{p_\gamma+\alpha}{2}-2\right)\Gamma\left(\frac{p_\gamma+\alpha}{2}-2\right)\,{}_2F_1\left(\frac{n-1}{2},1;\frac{p_\gamma+\alpha}{2};R^2_\gamma\right)} \\
&= \frac{8\,{}_2F_1\left(\frac{n-1}{2},3;\frac{p_\gamma+\alpha}{2};R^2_\gamma\right)}{(p_\gamma+\alpha-6)\,(p_\gamma+\alpha-4)\,{}_2F_1\left(\frac{n-1}{2},1;\frac{p_\gamma+\alpha}{2};R^2_\gamma\right)}. \qquad \text{(A.17)}
\end{aligned}
$$

The posterior mean of $\left(\frac{g}{g+1}\right)^2$ is defined as

$$
\begin{aligned}
\mathbb{E}\left(\left[\frac{g}{g+1}\right]^2\middle|\boldsymbol{\gamma},\boldsymbol{y}\right) &= \frac{\int_0^\infty \left(\frac{g}{g+1}\right)^2 f(g|\boldsymbol{\gamma},\boldsymbol{y})dg}{\int_0^\infty f(g|\boldsymbol{\gamma},\boldsymbol{y})dg} \\
&= \frac{\int_0^\infty g^2(1+g)^{\frac{n-1-p_\gamma-\alpha}{2}-2}[1+g(1-R^2_\gamma)]^{-\frac{n-1}{2}}dg}{\int_0^\infty (1+g)^{\frac{n-1-p_\gamma-\alpha}{2}}[1+g(1-R^2_\gamma)]^{-\frac{n-1}{2}}dg},
\end{aligned}
$$

the above integral form in the last step is handled in a more easier way using the transformation $u=\frac{g}{g+1}$ which includes the Gaussian hypergeometric function ${}_2F_1(.)$ as the following

$$
\begin{aligned}
\mathbb{E}\left(\left[\frac{g}{g+1}\right]^2\middle|\boldsymbol{\gamma},\boldsymbol{y}\right) &= \frac{\int_0^1 u^2(1-u)^{\frac{p_\gamma+\alpha}{2}-2}[1-uR^2_\gamma]^{-\frac{n-1}{2}}du}{\int_0^1 (1-u)^{\frac{p_\gamma+\alpha}{2}-2}[1-uR^2_\gamma]^{-\frac{n-1}{2}}du} \\
&= \frac{2\Gamma\left(\frac{p_\gamma+\alpha}{2}+1\right)\,{}_2F_1\left(\frac{n-1}{2},3;\frac{p_\gamma+\alpha}{2}+2;R^2_\gamma\right)}{\Gamma\left(\frac{p_\gamma+\alpha}{2}+1\right)\left(\frac{p_\gamma+\alpha}{2}\right)\left(\frac{p_\gamma+\alpha}{2}+1\right)\,{}_2F_1\left(\frac{n-1}{2},1;\frac{p_\gamma+\alpha}{2};R^2_\gamma\right)} \\
&= \frac{8\,{}_2F_1\left(\frac{n-1}{2},3;\frac{p_\gamma+\alpha}{2}+2;R^2_\gamma\right)}{(p_\gamma+\alpha)\,(p_\gamma+\alpha+2)\,{}_2F_1\left(\frac{n-1}{2},1;\frac{p_\gamma+\alpha}{2};R^2_\gamma\right)}. \qquad \text{(A.18)}
\end{aligned}
$$

A.6 Laplace Approximation for Hyper-g Priors

For the Bayes factor of (2.14) after the change of variables $z = \log(g)$ we consider

$$\begin{aligned} H^{hy}(z) = {} & \frac{(n-1-p_\gamma-\alpha)}{2}\log\left[1+\exp(z)\right] \\ & - \frac{(n-1)}{2}\log\left[1+(1-R^2_\gamma)\exp(z)\right] + z, \end{aligned} \tag{A.19}$$

where the respective derivative is written as

$$\begin{aligned} \frac{dH^{hy}(z)}{dz} = {} & \frac{(n-1-p_\gamma-\alpha)}{2}\left[\frac{\exp(z)}{1+\exp(z)}\right] \\ & - \frac{(n-1)}{2}\left[\frac{(1-R^2_\gamma)\exp(z)}{1+(1-R^2_\gamma)\exp(z)}\right] + 1, \end{aligned} \tag{A.20}$$

and the mode $\hat{z}^{hy}$ of H^{hy} is found solving the equation $\frac{dH^{hy}(z)}{dz} = 0$ with respect to z, as the following

$$\begin{aligned} & 2\left[1+\exp(z)\right]\left[1+(1-R^2_\gamma)\exp(z)\right]\frac{(n-1-p_\gamma-\alpha)}{2}\left[\frac{\exp(z)}{1+\exp(z)}\right] \\ & - 2\left[1+\exp(z)\right]\left[1+(1-R^2_\gamma)\exp(z)\right]\frac{(n-1)}{2}\left[\frac{(1-R^2_\gamma)\exp(z)}{1+(1-R^2_\gamma)\exp(z)}\right] \\ & + 2\left[1+\exp(z)\right]\left[1+(1-R^2_\gamma)\exp(z)\right] = 0, \end{aligned}$$

where we obtain after some mathematical steps

$$\begin{aligned} & \exp(z)\left[1+(1-R^2_\gamma)\exp(z)\right](n-1-p_\gamma-\alpha) - \exp(z)\left[1+\exp(z)\right](n-1)(1-R^2_\gamma) \\ & + 2\left[1+\exp(z)\right]\left[1+(1-R^2_\gamma)\exp(z)\right] = 0, \end{aligned}$$

and going on with calculations we have

$$\begin{aligned} & \exp(z)(n-1-p_\gamma-\alpha) + \exp(2z)(n-1-p_\gamma-\alpha)(1-R^2_\gamma) \\ & - \exp(z)(1-R^2_\gamma)(n-1) - \exp(2z)(n-1)(1-R^2_\gamma) = 0, \end{aligned}$$

The previous expression is reduced in a quadratic equation with coefficients κ_2, κ_1, κ_0

$$\kappa_2\exp(2z) + \kappa_1\exp(z) + \kappa_0 = 0,$$

where κ_2, κ_1, κ_0 are calculated as

$$\begin{aligned}
\kappa_2 &= \Big[(2 - p_\gamma - \alpha)(1 - R^2_\gamma)\Big]\,, \\
\kappa_1 &= \Big[4 - p_\gamma - \alpha + R^2_\gamma(n-3)\Big]\,, \\
\kappa_0 &= 2.
\end{aligned}$$

The previous coefficients are resulted useful, in the calculation of the unique real mode $\hat{z}_{hy}$ through the discriminant α

$$\Delta = \sqrt{\kappa_1^2 - 4\kappa_2\kappa_0},$$

According to (Abramowitz and Stegun, 1970) and (Liang et al, 2008); the mode $\hat{z}_{hy}$ is computed as

$$\hat{z}^{hy} = \frac{-\kappa_1 + \sqrt{\Delta}}{2\kappa_2}.$$

and in more precise form as

$$\hat{z}^{hy} = \log\left(\frac{-\Big[4 - p_\gamma - \alpha + R^2_\gamma(n-3)\Big] + \sqrt{\Delta}}{2\Big[(2 - p_\gamma - \alpha)(1 - R^2_\gamma)\Big]}\right). \tag{A.21}$$

The standard error of the mode $\hat{z}_{hy}$ is retrieved by substituting the mode in the second derivative of H^{hy}. Firstly, we start calculating the second derivative of H^{hy} as the following

$$\begin{aligned}
\frac{d^2H^{hy}(z)}{dz^2} = \Bigg\{ & \Bigg[\Big(\frac{n-1-p_\gamma-\alpha}{2}\Big)\Bigg\{\frac{\exp(z)\,[1+\exp(z)] - \exp(2z)}{[1+\exp(z)]^2}\Bigg\}\Bigg] \qquad , \\
& - \frac{(n-1)(1-R^2_\gamma)}{2}\left[\frac{\exp(z)\,\Big[1+(1-R^2_\gamma)\exp(z)\Big] - \exp(2z)}{\Big[1+(1-R^2_\gamma)\exp(z)\Big]^2}\right]\Bigg\}
\end{aligned}$$

and in the next step, we compute the standard error $\hat{\sigma}_{H^{hy}}$ of the mode $\hat{z}^{hy}$ as

$$\hat{\sigma}_{H^{hy}} = \left[\left.\frac{-d^2H^{hy}(z)}{dz^2}\right|_{z=\hat{z}^{hy}}\right]^{-\frac{1}{2}}.$$

which leads to

$$\widehat{\sigma}_{H^{hy}} = \left\{ \left[-\left(\frac{n-1-p_\gamma-\alpha}{2}\right) \left\{ \frac{\exp(\widehat{z}^{hy})\left[1+\exp(\widehat{z}^{hy})\right]-\exp(2\widehat{z}^{hy})}{[1+\exp(\widehat{z}^{hy})]^2}\right\}\right] \right. \\ \left. + \frac{(n-1)(1-R^2_\gamma)}{2}\left[\frac{\exp(\widehat{z}^{hy})\left[1+(1-R^2_\gamma)\exp(\widehat{z}^{hy})\right]-\exp(2\widehat{z}^{hy})}{\left[1+(1-R^2_\gamma)\exp(\widehat{z}^{hy})\right]^2}\right]\right\}^{-\frac{1}{2}}. \tag{A.22}$$

A.7 Implementation of Stochastic Search Variable Selection

The SSVS through Gibbs sampler is outlined as follows

A. Set initial values $\boldsymbol{\gamma}^{(0)}$, $\boldsymbol{\beta}^{(0)}$, $\sigma^{2(0)}$, $a^{(0)}$ and $g^{(0)}$. For fixed $g = n$, delete **Step 6**.

B. For iterations $s = 1, \ldots, S$:

Step 1: Set current values equal to $\boldsymbol{\gamma} = \boldsymbol{\gamma}^{(s-1)}$, $\boldsymbol{\beta} = \boldsymbol{\beta}^{(s-1)}$, $\sigma^2 = \sigma^{2(s-1)}$ $a = a^{(s-1)}$ and $g = g^{(s-1)}$.

Step 2: Sample $\gamma_j^{(s)} \sim Bern\left(\pi_j^{SSVS}\right)$, for $j = 1, \ldots, p$, given the current states of $\boldsymbol{\gamma}_{-j}^{(s-1)}$ $\boldsymbol{\beta}^{(s-1)}$, $\sigma^{2(s-1)}$, $a^{(s-1)}$ and $g^{(s-1)}$, where $\boldsymbol{\gamma}_{-j}$ are the components of $\boldsymbol{\gamma}$ except element γ_j

(a) with probability inclusion of j-th covariate $\pi_j^{SSVS} = O_j^{SSVS}/(1+O_j^{SSVS})$,

(b) with posterior odds O_j^{SSVS}

$$O_j^{SSVS} = \frac{\pi^{SSVS}(\boldsymbol{\beta}|g,\sigma^2,\gamma_j=1,\boldsymbol{\gamma}_{-j})\pi(\gamma_j=1,\boldsymbol{\gamma}_{-j})}{\pi^{SSVS}(\boldsymbol{\beta}|g,\sigma^2,\gamma_j=0,\boldsymbol{\gamma}_{-j})\pi(\gamma_j=0,\boldsymbol{\gamma}_{-j})},$$

and set $\boldsymbol{\gamma}^{(s)} = \boldsymbol{\gamma}^{(s-1)}$.

Step 3: Sample $\boldsymbol{\beta}^{(s)} \sim N_p\left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}^{(SSVS)}, \widehat{\boldsymbol{C}}_{\boldsymbol{\beta}}^{(SSVS)}\right)$, given the respective updated and current states $\boldsymbol{\gamma}^{(s)}$, $\sigma^{2(s-1)}$, $a^{(s-1)}$ and $g^{(s-1)}$, where $\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}^{(SSVS)}$ and $\widehat{\boldsymbol{C}}_{\boldsymbol{\beta}}^{(SSVS)}$ denote the posterior mean and variance-covariance matrix of $\boldsymbol{\beta}$ defined respectively as

(a) $\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}^{(SSVS)} = g\left(g\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} + \boldsymbol{D}^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{D}^{-1}\right)^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}$,

(b) $\widehat{\boldsymbol{C}}_{\boldsymbol{\beta}}^{(SSVS)} = g\sigma^2\left(g\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} + \boldsymbol{D}^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{D}^{-1}\right)^{-1}$, and set $\boldsymbol{\beta}^{(s)} = \boldsymbol{\beta}^{(s-1)}$.

Step 4: Sample $\sigma^{2(s)} \sim IG\left(\widehat{\lambda}_{0,\sigma^2}^{(SSVS)}, \widehat{\lambda}_{1,\sigma^2}^{(SSVS)}\right)$, given the respective updated and current states $\boldsymbol{\gamma}^{(s)}$, $\boldsymbol{\beta}^{(s)}$, $a^{(s-1)}$ and $g^{(s-1)}$, where $\boldsymbol{res}^{(SSVS)} = \Big(\boldsymbol{y} - a\mathbf{1}_{\boldsymbol{n}} - \boldsymbol{X\beta})^T(\boldsymbol{y} - a\mathbf{1}_{\boldsymbol{n}} - \boldsymbol{X\beta}\Big)$, the $\widehat{\lambda}_{0,\sigma^2}^{(SSVS)}$ and $\widehat{\lambda}_{1,\sigma^2}^{(SSVS)}$ denote respectively the posterior shape and scale of σ^2 respectively as

(a) $\widehat{\lambda}_{0,\sigma^2}^{(SSVS)} = (n+p)/2$,

(b) $\widehat{\lambda}_{1,\sigma^2}^{(SSVS)} = \frac{1}{2}\left[\boldsymbol{res}^{(SSVS)} + \boldsymbol{\beta}^T\boldsymbol{D}^{-1}\boldsymbol{X^T X D}^{-1}\boldsymbol{\beta}/g\right]$,

and set $\sigma^{2(s)} = \sigma^{2(s-1)}$.

Step 5: Sample $a^{(s)} \sim N\left(\widehat{\mu}_a, \widehat{\sigma}_a^2\right)$, given the updated state $\sigma^{2(s)}$, where the $\widehat{\mu}_a$ and $\widehat{\sigma}_a^2$ denote the posterior mean and variance of a respectively as

(a) $\widehat{\mu}_a = \bar{y}$,

(b) $\widehat{\sigma}_a^2 = \sigma^2/n$,

and set $a^{(s)} = a^{(s-1)}$

Step 6: given the updated states $\boldsymbol{\gamma}^{(s)}$, $\boldsymbol{\beta}^{(s)}$, $\sigma^{2(s)}$ and $a^{(s)}$

(**A**) if $g \sim IG\left(\frac{1}{2}, \frac{n}{2}\right)$,
sample $g^{(s)} \sim IG\left(\widehat{\lambda}_{0,g}^{(SSVS)}, \widehat{\lambda}_{1,g}^{(SSVS)}\right)$, where $\widehat{\lambda}_{0,g}^{(SSVS)}$ and $\widehat{\lambda}_{1,g}^{(SSVS)}$ denote respectively the posterior shape and scale of g respectively as

(a) $\widehat{\lambda}_{0,g}^{(SSVS)} = (p+1)/2$,

(b) $\widehat{\lambda}_{1,g}^{(SSVS)} = \frac{1}{2}\left[\boldsymbol{\beta}^T\boldsymbol{D}^{-1}\boldsymbol{X^T X D}^{-1}\boldsymbol{\beta}/\sigma^2 + n\right]$,

and set $g^{(s)} = g^{(s-1)}$.

(**B**) if $\pi(g) \propto (1+g)^{-\frac{a}{2}}$, sample $g^{(s)}$ from full conditional $(1+g)^{-\frac{\alpha}{2}}g^{-\frac{p}{2}}$ $\exp\left(-\boldsymbol{\beta^T D}^{-1}\boldsymbol{X^T X D}^{-1}\boldsymbol{\beta}/2g\sigma^2\right)$ based on a Metropolis-Hastings with

(a) a candidate value $g^{(can)}$ is generated as $\log(g^{(can)}) \sim N(\log(g), v_g)$ $\Rightarrow g^{(can)} = \exp(\log(g^{(can)}))$, where v_g denotes the tuning variance.

(b) an acceptance-rate $A_g^{(SSVS)}$ of the proposed move in the log-scale

$$log(A_g^{(SSVS)}) = \log\left(\frac{\pi(g^{(can)}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}, \boldsymbol{y})}{\pi(g|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}, \boldsymbol{y})}\frac{q(g|v_g)}{q(g^{(can)}|v_g)}\frac{J}{J^{(can)}}\right)$$
$$\propto -\frac{\alpha}{2}\log(1+g^{(can)}) - \frac{p}{2}\log(g^{(can)}) - \frac{\boldsymbol{\beta^T D}^{-1}\boldsymbol{X^T X D}^{-1}\boldsymbol{\beta}}{2g^{(can)}\sigma^2}$$
$$+ \frac{\alpha}{2}\log(1+g) + \frac{p}{2}\log(g) + \frac{\boldsymbol{\beta^T D}^{-1}\boldsymbol{X^T X D}^{-1}\boldsymbol{\beta}}{2g\sigma^2}$$
$$- \log\left(\frac{1}{g^{(can)}}\right) + \log\left(\frac{1}{g}\right),$$

where $q(.)$ denotes candidate density generator and J the associated jacobian which results from transformation on the original scale of g. Notice that the corresponding ratio $q(g^{cur}|v_g)/q(g^{can}|v_g)$ vanishes due to symmetry feature of the normal random walk.

(c) Set $g^{(s)} = \begin{cases} g^{(can)} & \text{, accept with probability } A_g^{(SSVS)}, \\ g & \text{, reject with probability } 1 - A_g^{(SSVS)}, \end{cases}$

C. Repeat all the steps untill convergence,

where we clarify that all the detailed steps of full conditionals are ommited for brevity, assuming that the kind reader is familiar with the SSVS George and McCulloch (1993).

A.8 Implementation of Gibbs Variable Selection

The corresponding MCMC procedure is applied through a Gibbs sampler which samples indirectly from the joint posterior distribution as the following

A. Same as in SSVS.

B. Same as in SSVS.

Step 1: Same as in SSVS.

Step 2: Sample $\gamma_j^{(s)} \sim Bern\left(\pi_j^{GVS}\right)$, for $j = 1, \ldots, p$, given the current states of $\boldsymbol{\gamma}_{-j}^{(s-1)}$ $\boldsymbol{\beta}^{(s-1)}$, $\sigma^{2(s-1)}$, $a^{(s-1)}$ and $g^{(s-1)}$, where $\boldsymbol{\gamma}_{-j}$ are the components of $\boldsymbol{\gamma}$ except element γ_j

(a) with probability inclusion of j-th covariate $\pi_j^{GVS} = O_j^{GVS}/(1 + O_j^{GVS})$,

(b) with posterior odds O_j^{GVS}

$$O_j^{GVS} = \frac{f(\boldsymbol{y}|a, \boldsymbol{\beta}, \sigma^2, \gamma_j = 1, \boldsymbol{\gamma}_{-j})\pi^{GVS}(\boldsymbol{\beta}|g, \sigma^2, \gamma_j = 1, \boldsymbol{\gamma}_{-j})\pi(\gamma_j = 1, \boldsymbol{\gamma}_{-j})}{f(\boldsymbol{y}|a, \boldsymbol{\beta}, \sigma^2, \gamma_j = 0, \boldsymbol{\gamma}_{-j})\pi^{GVS}(\boldsymbol{\beta}|g, \sigma^2, \gamma_j = 0, \boldsymbol{\gamma}_{-j})\pi(\gamma_j = 0, \boldsymbol{\gamma}_{-j})},$$

where the O_j^{GVS} is one of the main additional features that differentiates from SSVS method due to appearance of likelihood depending on $\boldsymbol{\gamma}$ configuration and set $\boldsymbol{\gamma}^{(s)} = \boldsymbol{\gamma}^{(s-1)}$.

Step 3: Sample $\boldsymbol{\beta}^{(s)} \sim N_p\left(\widehat{\boldsymbol{\mu}}_\beta^{(GVS)}, \widehat{\boldsymbol{C}}_\beta^{(GVS)}\right)$, given the respective updated and current states $\boldsymbol{\gamma}^{(s)}$, $\sigma^{2(s-1)}$, $a^{(s-1)}$ and $g^{(s-1)}$, where $\widehat{\boldsymbol{\mu}}_\beta^{(GVS)}$ and $\widehat{\boldsymbol{C}}_\beta^{(GVS)}$ denote the posterior mean and variance-covariance matrix of $\boldsymbol{\beta}$ defined respectively as

(a) $\widehat{\boldsymbol{\mu}}_{\beta}^{(GVS)} = \left(\boldsymbol{\Gamma X^T X \Gamma} + \sigma^2 \widetilde{\boldsymbol{D}}^{-1}\right)^{-1} \left(\boldsymbol{\Gamma X^T y} + \sigma^2 \widetilde{\boldsymbol{D}}^{-1} \boldsymbol{\mu}\right)$,

(b) $\widehat{\boldsymbol{C}}_{\beta}^{(GVS)} = \sigma^2 \left(\boldsymbol{\Gamma X^T X \Gamma} + \sigma^2 \widetilde{\boldsymbol{D}}^{-1}\right)^{-1}$,

and set $\boldsymbol{\beta}^{(s)} = \boldsymbol{\beta}^{(s-1)}$.

Step 4: Sample $\sigma^{2(s)} \sim IG\left(\widehat{\lambda}_{0,\sigma^2}^{(SSVS)}, \widehat{\lambda}_{1,\sigma^2}^{(SSVS)}\right)$, given the respective updated and current states $\boldsymbol{\gamma}^{(s)}$, $\boldsymbol{\beta}^{(s)}$, $a^{(s-1)}$ and $g^{(s-1)}$, where $\boldsymbol{res}^{(GVS)} = \left(\boldsymbol{y} - a\boldsymbol{1_n} - \boldsymbol{X\Gamma\beta}\right)^T(\boldsymbol{y} - a\boldsymbol{1_n} - \boldsymbol{X\Gamma\beta}\Big)$, the $\widehat{\lambda}_{0,\sigma^2}^{(GVS)}$ and $\widehat{\lambda}_{1,\sigma^2}^{(GVS)}$ denote respectively the posterior shape and scale of σ^2 respectively as

(a) $\widehat{\lambda}_{0,\sigma^2}^{(GVS)} = (n + p_\gamma)/2$,

(b) $\widehat{\lambda}_{1,\sigma^2}^{(GVS)} = \frac{1}{2}\left[\boldsymbol{res}^{(GVS)} + (\boldsymbol{\beta} - \boldsymbol{\mu})^T \boldsymbol{\Gamma X^T X \Gamma}(\boldsymbol{\beta} - \boldsymbol{\mu})/g\right]$,

and set $\sigma^{2(s)} = \sigma^{2(s-1)}$.

Step 5: Same as in SSVS.

Step 6: given the updated states $\boldsymbol{\gamma}^{(s)}$, $\boldsymbol{\beta}^{(s)}$, $\sigma^{2(s)}$ and $a^{(s)}$

(**A**) if $g \sim IG\left(\frac{1}{2}, \frac{n}{2}\right)$,
sample $g^{(s)} \sim IG\left(\widehat{\lambda}_{0,g}^{(GVS)}, \widehat{\lambda}_{1,g}^{(GVS)}\right)$, where $\widehat{\lambda}_{0,g}^{(GVS)}$ and $\widehat{\lambda}_{1,g}^{(GVS)}$ denote respectively the posterior shape and scale of g respectively as

(a) $\widehat{\lambda}_{0,g}^{(GVS)} = (p_\gamma + 1)/2$,

(b) $\widehat{\lambda}_{1,g}^{(GVS)} = \frac{1}{2}\left[(\boldsymbol{\beta} - \boldsymbol{\mu})^T \boldsymbol{\Gamma X^T X \Gamma}\,(\boldsymbol{\beta} - \boldsymbol{\mu})/\sigma^2 + n\right]$,

and set $g^{(s)} = g^{(s-1)}$.

(**B**) if $\pi(g) \propto (1+g)^{-\frac{a}{2}}$, sample $g^{(s)}$ from full conditional $(1+g)^{-\frac{\alpha}{2}} g^{-p_\gamma}$ $\exp\left(-(\boldsymbol{\beta} - \boldsymbol{\mu})^T \boldsymbol{\Gamma X^T X \Gamma}\,(\boldsymbol{\beta} - \boldsymbol{\mu})/2g\sigma^2\right)$ using a Metropolis-Hastings with properties

(a) Same as in SSVS.

(b) an acceptance-rate $A_g^{(GVS)}$ of the proposed move in log-scale

$$\begin{aligned} log(A_g^{(GVS)}) &= \log\left(\frac{\pi(g^{(can)}|\boldsymbol{\beta}, \sigma^2\boldsymbol{\gamma}, \boldsymbol{y})}{\pi(g|\boldsymbol{\beta}, \sigma^2\boldsymbol{\gamma}, \boldsymbol{y})} \frac{q(g|v_g)}{q(g^{(can)}|v_g)} \frac{J}{J^{(can)}}\right) \\ &\propto -\frac{\alpha}{2}\log(1 + g^{(can)}) - \frac{p_\gamma}{2}\log(g^{(can)}) - \frac{(\boldsymbol{\beta} - \boldsymbol{\mu})^T \boldsymbol{\Gamma X^T X \Gamma}\,(\boldsymbol{\beta} - \boldsymbol{\mu})}{2g^{(can)}\sigma^2} \\ &+ \frac{\alpha}{2}\log(1+g) + \frac{p_\gamma}{2}\log(g) + \frac{(\boldsymbol{\beta} - \boldsymbol{\mu})^T \boldsymbol{\Gamma X^T X \Gamma}\,(\boldsymbol{\beta} - \boldsymbol{\mu})}{2g\sigma^2} \\ &+ \log\left(\frac{1}{g}\right) - \log\left(\frac{1}{g^{(can)}}\right). \end{aligned}$$

Notice again, the corresponding ratio $q(g^{cur}|v_g)/q(g^{can}|v_g)$ vanishes due to symmetry feature of the normal random walk.

(c) Set $g^{(s)} = \begin{cases} g^{(can)} & \text{, accept with probability } A_g^{(SSVS)}, \\ g & \text{, reject with probability } 1 - A_g^{(SSVS)}, \end{cases}$

C. Repeat all the steps untill convergence,

where notice that in the above **Step4** and **Step6**, the posterior of σ^2 and g are affected only by the included components p_γ given the respective configuration of $\boldsymbol{\gamma}$ which is an essential ingredient of pseudo-priors and hence GVS. All the computed steps of full conditionals are avoided for the same reasons likewise SSVS; for additional information see Ntzoufras (1999) and Dellaportas et al. (2002).

A.9 Simulated Experiment

In this section, we used a simulated example of George and McCullogh, (1993) concerning the Bayesian variable selection using full enumeration and MCMC methods with mixtures of g-priors applied for the linear regression. This dataset consists of $p = 5$ covariates of length $n = 60$. The covariates were obtained as independent standardized normal vectors $\boldsymbol{X_1}, \ldots, \boldsymbol{X_5}$ iid $\sim N_{60}(0, 1)$, so that they were uncorrelated. The dependent variable is generated according to the model

$$\boldsymbol{Y} = 2 + \boldsymbol{X_4} + 1.2\boldsymbol{X_5} + \boldsymbol{\epsilon},$$

where the error term $\boldsymbol{\epsilon} \sim N_{60}(0, \sigma^2 \boldsymbol{I}_{60})$ with noise $\sigma = 2.5$. In comparison with the example of George and McCullogh, (1993), the only difference is the intercept included in the generated model. The maximum likelihood estimators for these data were $\hat{a}$, $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}$ and are found respectively in Table (A.1). Each respective covariate $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, $\boldsymbol{X_3}$, $\boldsymbol{X_4}$, $\boldsymbol{X_5}$ was centered for Liang's approach and in order to obtain more comparable results in-sample values. We initially proceed with a multiple scatterplot regarding the relationship of the response variable Y with each respective covariate $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, $\boldsymbol{X_3}$, $\boldsymbol{X_4}$, $\boldsymbol{X_5}$. From Figure (A.1) it is evident that the covariates $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, $\boldsymbol{X_3}$, $\boldsymbol{X_4}$, $\boldsymbol{X_5}$ seem linearly associated with the response variable Y whereas the covariates $\boldsymbol{X_4}$, $\boldsymbol{X_5}$ are positively associated with the response variable Y as it was expected from the model construction. Our primary aim is to evaluate the performance of the MCMC methods for Bayesian variable selection in linear regression with respect to the current simulated dataset.

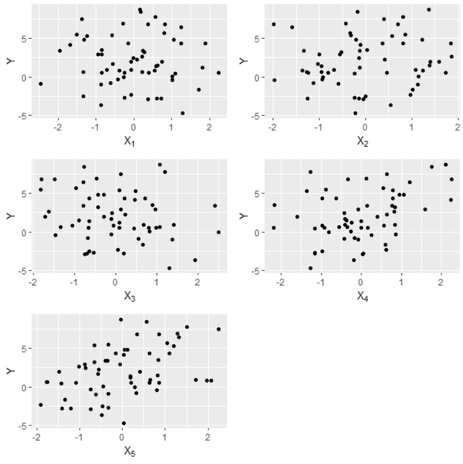


Fig. A.1 Multiple scatterplots of the dependent variable Y versus each covariate $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, $\boldsymbol{X_3}$, $\boldsymbol{X_4}$, $\boldsymbol{X_5}$.

Even if the model space is $2^5 = 32$, which is attractive for Bayesian variable selection with full enumeration, we still prefer to approximate the model space through MCMC methods compared with the formal methods.
We present the results for Bayesian variable selection using SSVS and GVS in the R programming language as the main computational tools, compared with the formal methods of Bayesian variable selection in the context of mixtures of g-priors. Across all methods using hyper-g we considered the suggested value $\alpha = 3$ followed by indications of Liang et al. (2008) and a Metropolis-Hastings random walk scheme with proposed moves $u_g = 1$ which was applied in order to achieve satisfatory acceptance rates a.r for the model search procedures. Additional comparisons are based on the GVS method implemented in WINBUGS Spiegelhalter et al. (2003) and the Bayesian adaptive sampling Clyde et al. (2011) (BAS package) in R in order to verify the accordance of the between results.

Prior Inputs-Initial Values	
Parameter	**Value**
$\widehat{\boldsymbol{\beta}}$	$(-0.236, 0.418, -0.462, 1.220, 1.447)^T$
$\hat{a}$	2.026
$\hat{\sigma}^2$	2.644
τ	0.02
c	50
$\bar{\boldsymbol{\mu}}$	$\widehat{\boldsymbol{\beta}}$
$\bar{\boldsymbol{s}}^2$	$(0.124, 0.124, 0.124, 0.120, 0.119)^T$
$\boldsymbol{\gamma}^{(0)}$	$(1, 1, 1, 1, 1)^T$
$\boldsymbol{\beta}^{(0)}$	$\widehat{\boldsymbol{\beta}}$
$a^{(0)}$	$\hat{a}$
$\sigma^{2(0)}$	$\hat{\sigma}^2$
$g^{(0)}$	n

Table A.1 Prior-inputs and initial values

With regard to MCMC methods, prior inputs τ and c for $j = 1, \ldots, p$ were set on practical significance for SSVS to achieve similar results with the frequentist and formal Bayesian approach and $\bar{\boldsymbol{\mu}}$, $\bar{\boldsymbol{s}}^2$ were computed from a pilot run under the full model for GVS. The option of prior input τ_j and c_j are such that $\tau_j = 0.02 << \tau_j c_j = 1$. All these information are available on Table (A.1).

	Acronym	Computational Method	Prior
1	ssvs^{R}.g	Stochastic Search Variable Selection for $\tau = 0.02$, $c = 50$, $g = n$	g-prior
2	ssvs^{R}.ZS	Stochastic Search Variable Selection for $\tau = 0.02$, $c = 50$	Zellner-Siow
3	ssvs^{R}.hy	Stochastic Search Variable selection for $\tau = 0.02$, $c = 50$, $\alpha = 3$, $u_g = 1$	Hyper-g
4	gvs^{R}.g	Gibbs Variable Selection for $g = n$	g-prior
5	gvs^{R}.ZS	Gibbs Variable Selection Selection	Zellner-Siow
6	gvs^{R}.hyp	Gibbs Variable Selection Selection for $\alpha = 3$, $u_g = 1$	Hyper-g
7	gvs^{W}.g	Gibbs Variable Selection for $g = n$	g-prior
8	gvs^{W}.ZS	Gibbs Variable Selection Selection	Zellner-Siow
9	gvs^{W}.hyp	Gibbs Variable Selection Selection for $\alpha = 3$	Hyper-g
10	fe.g	Full Enumeration for $g = n$	g-prior
11	fe.hyp	Full Enumeration for $\alpha = 3$	Hyper-g
12	bas.g	Bayesian Adaptive Sampling for $g = n$	g-prior
13	bas.ZS	Bayesian Adaptive Sampling	Zeller-Siow
14	bas.hyp	Bayesian Adaptive Sampling for $\alpha = 3$	Hyper-g

Table A.2 Acronyms of Bayesian variable selection methods.

A detailed description of all Bayesian variable selection methods which are used as references in Figures and in Tables, are summarized in Table (A.2). In addition, as there is little information available, Bayesian variable selection turns into an objective approach, assigning objective priors to each respective model specific parameter and to the model. More presicely, we used the joint hierarchical mixture priors (2.16), (2.18) for SSVS and GVS to encapsulate the dependencies among parameters a, σ^2 $\boldsymbol{\beta}$, g, $\boldsymbol{\gamma}$, whereas for standard Bayesian methods based on full enumeration we adopted the combined Liang's g-prior (2.9) in the mixture sense for g.

g-prior	**Highest Posterior Model Probability**				
Model	**Bayesian Variable Selection Methods**				
γ	ssvs$^{\mathrm{R}}$.g	gvs$^{\mathrm{R}}$.g	gvs$^{\mathrm{W}}$.g	fe.g	bas.g
$\boldsymbol{X}_4, \boldsymbol{X}_5$	0.573	0.576	0.577	0.574	0.574
$\boldsymbol{X}_3, \boldsymbol{X}_4, \boldsymbol{X}_5$	0.143	0.139	0.137	0.139	0.139
$\boldsymbol{X}_2, \boldsymbol{X}_4, \boldsymbol{X}_5$	0.112	0.113	0.114	0.114	0.114
$\boldsymbol{X}_1, \boldsymbol{X}_4, \boldsymbol{X}_5$	0.077	0.077	0.078	0.078	0.078
$\boldsymbol{X}_2, \boldsymbol{X}_3, \boldsymbol{X}_4,\ \boldsymbol{X}_5$	0.031	0.033	0.031	0.032	0.032

Table A.3 Posterior model probabilities of top 5 models.

Zellner-Siow	**Highest Posterior Model Probability**			
Model	**Bayesian Variable Selection Methods**			
γ	ssvs$^{\mathrm{R}}$.ZS	gvs$^{\mathrm{R}}$.ZS	gvs$^{\mathrm{W}}$.ZS	bas.ZS
$\boldsymbol{X}_4, \boldsymbol{X}_5$	0.475	0.497	0.495	0.465
$\boldsymbol{X}_3, \boldsymbol{X}_4, \boldsymbol{X}_5$	0.154	0.149	0.151	0.157
$\boldsymbol{X}_2, \boldsymbol{X}_4, \boldsymbol{X}_5$	0.121	0.125	0.125	0.131
$\boldsymbol{X}_1, \boldsymbol{X}_4, \boldsymbol{X}_5$	0.083	0.085	0.086	0.090
$\boldsymbol{X}_2, \boldsymbol{X}_3, \boldsymbol{X}_4,\ \boldsymbol{X}_5$	0.053	0.051	0.051	0.057

Table A.4 Posterior model probabilities of top 5 models.

Hyper-g	**Highest Posterior Model Probability**				
Model	**Bayesian Variable Selection Methods**				
γ	ssvs$^{\mathrm{R}}$.hyp	gvs$^{\mathrm{R}}$.hyp	gvs$^{\mathrm{W}}$.hy	fe.hyp	bas.hyp
$\boldsymbol{X}_4, \boldsymbol{X}_5$	0.323	0.329	0.328	0.327	0.327
$\boldsymbol{X}_2, \boldsymbol{X}_4, \boldsymbol{X}_5$	0.165	0.163	0.162	0.163	0.163
$\boldsymbol{X}_3, \boldsymbol{X}_4, \boldsymbol{X}_5$	0.136	0.140	0.139	0.139	0.139
$\boldsymbol{X}_1, \boldsymbol{X}_4, \boldsymbol{X}_5$	0.096	0.099	0.101	0.101	0.101
$\boldsymbol{X}_2, \boldsymbol{X}_3, \boldsymbol{X}_4,\ \boldsymbol{X}_5$	0.085	0.088	0.089	0.089	0.089

Table A.5 Posterior model probabilities of top 5 models.

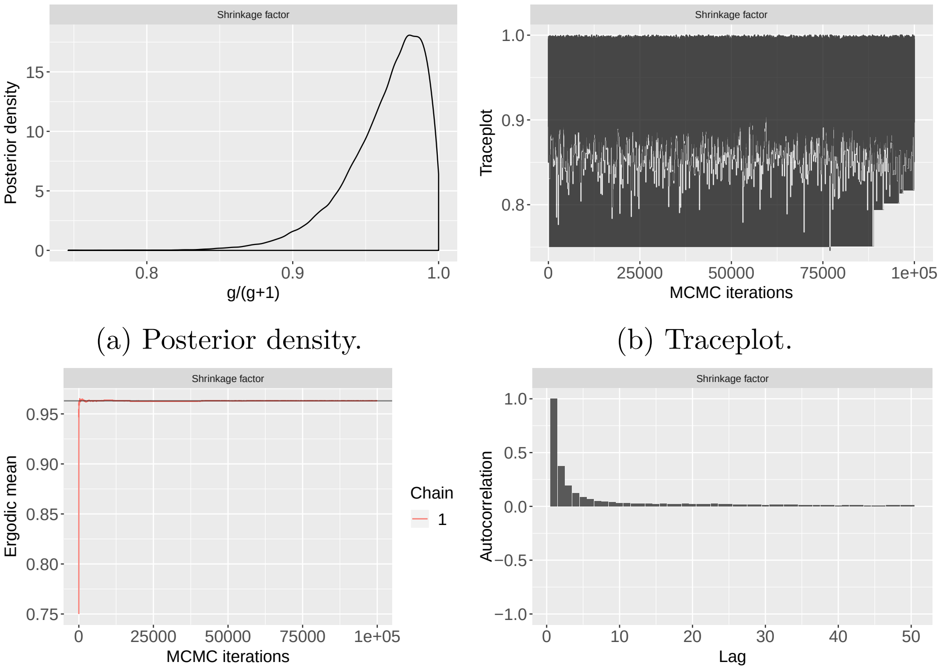
All the compared approaches under Zellner-Siow prior used (2.10), whereas for hyper-g-prior used (2.12). Moreover, regarding the prior of model space, we assigned a uniform prior distribution to each model $\boldsymbol{\gamma}$ reflecting our prior ignorance, regarding the preference in a small model space. In the main analysis, we have used also a frequentist perspective of full linear model for comparison reasons with the main results of Bayesian variable selection methods under prior ignorance, where the variables $\boldsymbol{X_5}$, $\boldsymbol{X_4}$ were statistically significant with p-value lower than 0.05. Applying Bayesian variable selection MCMC methods to these data, the model fitting was performed using a Gibbs algorithm with a Metropolis-Hastings step as implemented in the R programming language and we simulated a Markov chain of 100000 valid values to achieve convergence. More precisely, for parameter vector of effects we have used as initial values $\boldsymbol{\beta}^{(0)}$, for intercept $a^{(0)}$, for variance $\sigma^{2(0)}$, for parameter vector of model indicators $\boldsymbol{\gamma}^{(0)}$ and $g^{(0)}$ for g; the interesting reader may find more details on Table (A.1).

On the other hand, we preferred to illustrate only the best 5 models of each method because they are the only which assign negligible posterior probability mass. All methods seem to perform well in the correct model identification which includes variables $\boldsymbol{X_4}$, $\boldsymbol{X_5}$, but with different posterior model probability for each method. Moreover, the traced submodels of all methods coincide across the three different prior setups. The current model selected by these procedures coincides with the model under the frequentist approach. Predictors $\boldsymbol{X_4}$, $\boldsymbol{X_5}$ are statistically significant with p-value 0.000876, and 0.000106 respectively. With respect to the results of Table (A.3) for methods ssvs.g$^{\mathrm{R}}$, gvs.g$^{\mathrm{R}}$ under g-prior, the posterior model probabilities are larger in magnitude with comparison to those of Zellner-Siow (see Table (A.4), and hyper-g (see Table (A.5) for all respective methods which are known from the bibliography. This is attributed to Jeffrey's-Lindley's paradox which tends to support parsimonius models reflected in the increased posterior model probabilities among all methods. On the contrary, the lower posterior probability is observed in table (A.5) for methods ssvs.hyp$^{\mathrm{R}}$, gvs.hyp$^{\mathrm{R}}$ under hyper-g because of favouring the non significant variables, while methods of Table (A.4) based on Zellner-Siow correspond always to higher posterior model probabilities with respect to hyper-g. Among all models of all methods under each prior setup of Tables (A.3), (A.4) and (A.5) the results are homogeneous with slight differences, apart from the Zellner-Siow approach where the MCMC procedures outperform versus bas.ZS method. In general, the behaviour of posterior model probabilities among GVS, full enumeration and BAS for g-prior and hyper-g are sharing similar behaviour as pointed out by (Perrakis and Ntzoufras, 2015) and (Perrakis and Ntzoufras, 2018) which are convalidated across R and WINBUGS leading to similar posterior model probabilities. In addition, an attractive alternative of maximum posterior probability is provided by the median probability model of Barbieri and Berger (2004). Except the formal methods based on full enumeration, all previously mentioned Bayesian variable selection methods are compared and analyzed through their marginal posterior inclusion probability. From the results of Table (A.6), it is clear that all MCMC methods select covariates $\boldsymbol{X_4}$, $\boldsymbol{X_5}$ as significant, since their marginal posterior inclusion probabilities are greater than 0.5, according to the median posterior probability model by (Barbieri and Berger, 2004). On the contrary, the independent variables $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, $\boldsymbol{X_3}$ have marginal posterior inclusion probabilities lower than 0.5 between all methods and appear to be non significant.

	Median Probability Model				
	Independent Variables				
Method	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$	$\boldsymbol{X_3}$	$\boldsymbol{X_4}$	$\boldsymbol{X_5}$
ssvs$^{\mathrm{R}}$.g	0.121	0.169	0.203	0.982	0.997
gvs$^{\mathrm{R}}$.g	0.121	0.172	0.202	0.984	0.997
gvs$^{\mathrm{W}}$.g	0.122	0.171	0.198	0.983	0.997
bas.g	0.122	0.172	0.201	0.983	0.997
ssvs$^{\mathrm{R}}$.ZS	0.161	0.221	0.260	0.967	0.984
gvs$^{\mathrm{R}}$.ZS	0.162	0.220	0.251	0.984	0.997
gvs$^{\mathrm{W}}$.ZS	0.161	0.221	0.252	0.986	0.997
bas.ZS	0.175	0.238	0.271	0.986	0.997
ssvs$^{\mathrm{R}}$.hyp	0.261	0.334	0.376	0.976	0.992
gvs$^{\mathrm{R}}$.hyp	0.260	0.331	0.363	0.979	0.994
gvs$^{\mathrm{W}}$.hyp	0.262	0.330	0.364	0.980	0.995
bas.hyp	0.262	0.331	0.364	0.979	0.994

Table A.6 Marginal posterior inclusion probabilities for each independent variable $\boldsymbol{X}_j$ regarding Bayesian variable selection methods.

Furthermore, MCMC methods under the g-prior have lower marginal posterior inclusion probabilities for the non significant covariates $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, $\boldsymbol{X_3}$. On the other hand MCMC methods under mixtures of g-priors give results towards 0.5 since the additional randomness related to g increases the uncertainty of the non significant variables $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, $\boldsymbol{X_3}$. The latter behaviour naturally arises as additional randomness of the g parameter affecting the posterior inclusion measures through the adaptive shrinkage of factor $\frac{g}{g+1}$. Moreover, the Bayesian adaptive sampling and GVS in WINBUGS verify the previous results of MCMC procedures in R. Additional analysis has been performed for the shrinkage factor $\frac{g}{g+1}$ in Figures (A.2), (A.3), (A.5) in order to evaluate the efficiency of the MCMC methods. GVS and SSVS under Zellner-Siow posterior distributions of g seem appropriate to describe uncertainty. More precisely, we observe from Table (A.7) that methods ssvs$^{\mathrm{R}}$.hyp, gvs$^{\mathrm{R}}$.hyp are achieving larger posterior means with lower standard errors in comparisons with ssvs$^{\mathrm{R}}$.ZS, gvs$^{\mathrm{R}}$.ZS which seem more confident regarding model uncertainty and share increased marginal inclusion posterior probabilities. The convergence of shrinkage factor $\frac{g}{g+1}$ was monitored through the Figures (A.2), (A.3), (A.4) and (A.5) which do not exhibit strange variation for MCMC methods with random g. To conclude, Bayesian variable selection methods using MCMC perform successfully regarding the approximation of the model space and the tracing of the most probable models producing the same results given by the frequentist approach, namely the methods of full enumeration, Bayesian adaptive sampling and those of GVS implemented in WINBUGS. In accordance with the results of the median posterior probability model and the highest posterior probability model we conclude that $\boldsymbol{X_1}$, $\boldsymbol{X_2}$ are important variables which coincide with the true generating mechanism of the data.



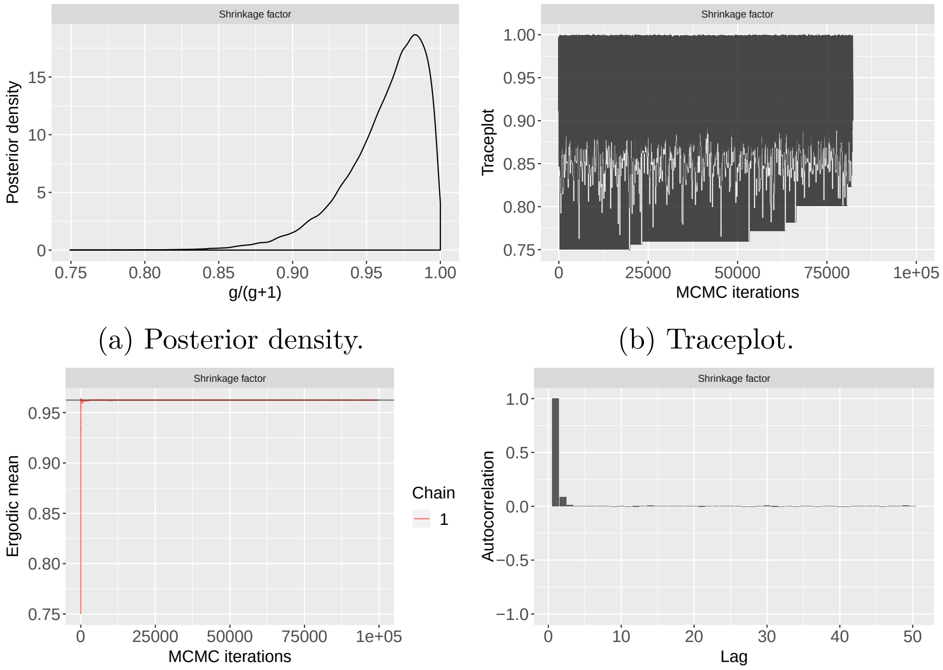
(a) Posterior density.

(b) Traceplot.

(c) Ergodic mean.

(d) Autocorrelation function.

Fig. A.2 Convergence diagnostics of shrinkage factor $\frac{g}{g+1}$ for ssvs$^{\mathrm{R}}$.ZS.



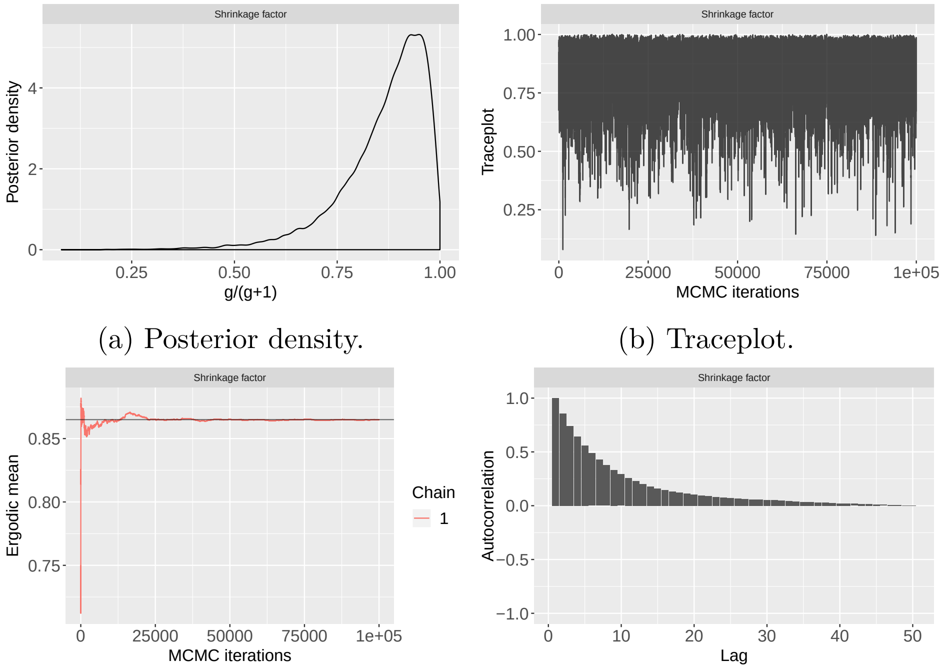
(a) Posterior density.

(b) Traceplot.

(c) Ergodic mean.

(d) Autocorrelation function.

Fig. A.3 Convergence diagnostics of shrinkage factor $\frac{g}{g+1}$ for gvs$^{\mathrm{R}}$.ZS.



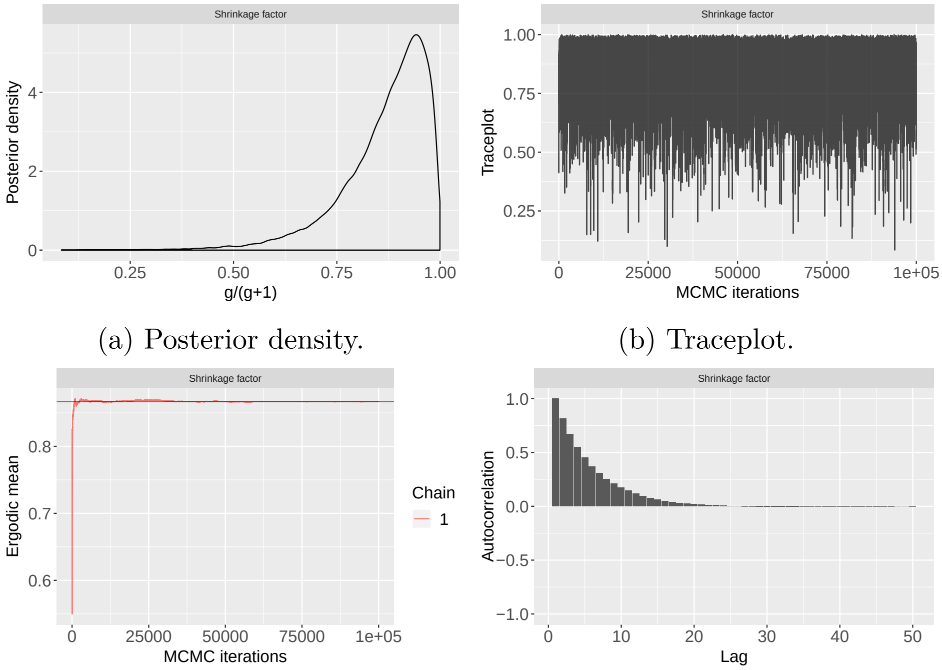
(a) Posterior density.

(b) Traceplot.

(c) Ergodic mean.

(d) Autocorrelation function.

Fig. A.4 Convergence diagnostics of shrinkage factor $\frac{g}{g+1}$ for ssvs$^{\mathrm{R}}$.hyp.



(a) Posterior density.

(b) Traceplot.

(c) Ergodic mean.

(d) Autocorrelation function.

Fig. A.5 Convergence diagnostics of shrinkage factor $\frac{g}{g+1}$ for gvs$^{\text{R}}$.hyp.

	Posterior Summary Statistics		
Method	**Mean**	**Standard Error**	**Acceptance Rate**
ssvs.zs	0.963	0.027	-
gvs.zs	0.962	0.027	-
ssvs.hyp	0.864	0.106	0.566
gvs.hyp	0.866	0.104	0.560

Table A.7 Results of posterior summary statistics regarding each Bayesian variable selection methods for mixtures of g-priors.

A.10 Real Dataset

In this section, we illustrate an application of Bayesian variable selection for the data of prostate cancer Stamey et al. (1989). This dataset consists of $n = 97$ observations and $p = 8$ covariates and was also used by Giron et al. (2006) and Moreno and Girón (2008). The response variable $\boldsymbol{Y}$ is the level of prostate-specific antigen, and the covariates are the logarithm of cancer volume ($\boldsymbol{X_1}$), the logarithm of prostate weight ($\boldsymbol{X_2}$), the age of patient ($\boldsymbol{X_3}$), the amount of benign prostatic hyperplasia ($\boldsymbol{X_4}$) in log-scale, the seminal vesicle invasion ($\boldsymbol{X_5}$), the capsular penetration in log-scale ($\boldsymbol{X_6}$), the gleason score ($\boldsymbol{X_7}$) and the percent of gleason scores 4 and 5 ($\boldsymbol{X_8}$).

Prior Inputs-Initial Values of 1st Analysis	
Parameter	**Value**
$\widehat{\boldsymbol{\beta}}$	$(0.665, 0.266, -0.158, 0.140, 0.315, -0.148, 0.035, 0.125)^T$
$\hat{a}$	2.478
$\hat{\sigma}$	0.699
τ	0.02
c	50
$\bar{\boldsymbol{\mu}}$	$\widehat{\boldsymbol{\beta}}$
$\bar{\boldsymbol{s}}^2$	$(0.103, 0.086, 0.082, 0.084, 0.099, 0.125, 0.112, 0.123)^T$
$\boldsymbol{\gamma}^{(0)}$	$(1, 1, 1, 1, 1, 1, 1, 1)^T$
$\boldsymbol{\beta}^{(0)}$	$\widehat{\boldsymbol{\beta}}$
$a^{(0)}$	$\hat{a}$
$\sigma^{2(0)}$	$\hat{\sigma}^2$
$g^{(0)}$	97

Table A.8 Prior-inputs and initial values of 1st Analysis

Prior Inputs-Initial Values of 2nd Analysis	
Parameter	**Value**
$\widehat{\boldsymbol{\beta}}$	$(0.530, 0.319, -0.201, 0.213, 0.230, 0.041, -0.083, 0.228)^T$
$\hat{a}$	2.626
$\hat{\sigma}$	0.682
τ	0.02
c	50
$\bar{\boldsymbol{\mu}}$	$\widehat{\boldsymbol{\beta}}$
$\bar{\boldsymbol{s}}^2$	$(0.152, 0.112, 0.107, 0.110, 0.158, 0.157, 0.166, 0.177)^T$
$\boldsymbol{\gamma}^{(0)}$	$(1, 1, 1, 1, 1, 1, 1, 1)^T$
$\boldsymbol{\beta}^{(0)}$	$\widehat{\boldsymbol{\beta}}$
$a^{(0)}$	$\hat{a}$
$\sigma^{2(0)}$	$\hat{\sigma}^2$
$g^{(0)}$	50

Table A.9 Prior-inputs and initial values of 2nd Analysis

Furthermore, a maximum likelihood perspective was applied to the full model obtaining estimates $\hat{\boldsymbol{\beta}}$, $\hat{a}$, $\hat{\sigma}$ where $(\boldsymbol{X_1})$, $(\boldsymbol{X_2})$, $(\boldsymbol{X_5})$ were found statistically significant; see for more information Table A.8. Prior to the main analysis, covariates were preprocessed by subtracting their mean in order to adopt Liang's approach for mixtures of g-priors. Our goal is to assess the performance of Bayesian variable selection methods considering both in-sample and out-of-sample values across the three prior setups of this real dataset. At the same time we retain the same acronyms for the methods of Table (A.2) and emphasis is given only to MCMC. For the out-of-sample analysis, we will split the data at half randomly and then compare the performance of MCMC by calculating at each iteration of each method the mean squared error (MSE). More precisely, we begin with a preliminary analysis of Bayesian model selection procedures using MCMC in order to identify important covariates through the highest and median posterior model probability. Then, we proceed with an additional analysis to evaluate the predictive ability of each MCMC method based on the maximum aposteriori model and the median probability model. We perform the analysis for Bayesian variable selection using SSVS and GVS in R programming language in the framework of mixtures of g-priors. Additional comparisons are based on GVS method implemented in WINBUGS Spiegelhalter et al. (2003) and Bayesian adaptive sampling Clyde et al. (2011) BAS package using R in order to verify the intermediate results. The prior inputs of Table (A.2) are considered the same (apart for g-prior where g=97) in order to obtain similar results alongside the frequentist and objective Bayesian approach, while $\bar{\boldsymbol{\mu}}$, $\bar{\boldsymbol{s}}^2$ computed from a pilot run under the full model for GVS for the first analysis and second respectively; see for additional details Tables (A.8) and (A.9).
In addition, when information is not available with respect to the subset of variables, it is preferable to adopt an objective prior elicitation for model parameters and model itself. More presicely, we used again the hierarchical mixture priors (2.16), (2.18) for SSVS and GVS to envelop the prior structures among parameters a, σ^2 $\boldsymbol{\beta}$, g, $\boldsymbol{\gamma}$. When mixtures of g priors are used for SSVS and GVS, are based on (2.10) if Zellner-Siow prior is adopted, whereas for hyper-g-prior is used (2.12). Moreover, regarding the prior of model space, a uniform prior distribution $\boldsymbol{\gamma}$ was adopted reflecting the indifference of models like in the work of Fouskakis and Ntzoufras (2013). Bayesian variable selection model search algorithms were applied to these data, where model fit was implemented using a Gibbs sampler with a Metropolis-Hastings stage in R programming language and a Markov chain of 80000 generated values was simulated to achieve convergence. More precisely, for parameter vector of effects $\boldsymbol{\beta}$ we have used as initial values $\boldsymbol{\beta}^{(0)}$, for variance $\sigma^{2(0)}$, for parameter vector of model indicators initial values $\boldsymbol{\gamma}^{(0)}$ and $g^{(0)}$ for

g; the kind reader may refer to the Tables (A.8) and (A.9) regarding the initial values of first and second analysis.

	Highest Posterior Probability
Method	$\boldsymbol{X_1}+\boldsymbol{X_2}+\boldsymbol{X_5}$
ssvs$^{\mathrm{R}}$.g	0.401
gvs$^{\mathrm{R}}$.g	0.375
gvs$^{\mathrm{W}}$.g	0.370
bas.g	0.372
ssvs$^{\mathrm{R}}$.ZS	0.304
gvs$^{\mathrm{R}}$.ZS	0.320
gvs$^{\mathrm{W}}$.ZS	0.324
bas.ZS	0.304
ssvs$^{\mathrm{R}}$.hyp	0.214
gvs$^{\mathrm{R}}$.hyp	0.254
gvs$^{\mathrm{W}}$.hyp	0.251
bas.hyp	0.255

Table A.10 Results of posterior model probabilities regarding Bayesian variable selection methods.

Moreover, we preferred to illustrate only the best model under consideration of each method because each method traces different submodels. All methods identify the same model which includes as variables $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, $\boldsymbol{X_5}$ but with different posterior model probability for each method across the three prior different setup. The same conclusions are drawn also in the work of Fouskakis and Ntzoufras (2013) The current model selected by these procedures coincides also with the model under the frequentist approach, where predictors $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, $\boldsymbol{X_5}$ are statistically significant. The results of Table (A.10) suggest that under the fixed g-prior, methods ssvs.g$^{\mathrm{R}}$, gvs.g$^{\mathrm{R}}$, show higher posterior model probability versus random g methods including Zellner-Siow and hyper-g. This result is justified again by Jeffrey's-Lindley's paradox which tends to support simpler models due to the influence of the sample $n = 97$. On the other hand, as it was expected regarding the methods ssvs.hyp$^{\mathrm{R}}$, gvs.hyp$^{\mathrm{R}}$ under hyper-g, the posterior probability is smaller compared to the other methods and this is related to the inflation of non important variables towards the "cut-off" of significance, resulting in the decrease of the same posterior measure.

	Median Probability Model							
	Independent Variables							
Method	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$	$\boldsymbol{X_3}$	$\boldsymbol{X_4}$	$\boldsymbol{X_5}$	$\boldsymbol{X_6}$	$\boldsymbol{X_7}$	$\boldsymbol{X_8}$
ssvs$^{\mathrm{R}}$.g	1.000	0.945	0.173	0.271	0.919	0.093	0.106	0.141
gvs$^{\mathrm{R}}$.g	1.000	0.947	0.192	0.250	0.916	0.111	0.124	0.161
gvs$^{\mathrm{W}}$.g	1.000	0.946	0.190	0.251	0.913	0.115	0.123	0.163
bas.g	1.000	0.946	0.192	0.253	0.916	0.110	0.124	0.162
ssvs$^{\mathrm{R}}$.ZS	0.999	0.946	0.240	0.342	0.928	0.089	0.161	0.195
gvs$^{\mathrm{R}}$.ZS	1.000	0.945	0.232	0.289	0.921	0.136	0.151	0.190
gvs$^{\mathrm{W}}$.ZS	1.000	0.948	0.229	0.291	0.921	0.134	0.150	0.186
bas.ZS	1.000	0.949	0.246	0.301	0.924	0.144	0.157	0.200
ssvs$^{\mathrm{R}}$.hyp	1.000	0.946	0.313	0.436	0.953	0.074	0.225	0.282
gvs$^{\mathrm{R}}$.hyp	1.000	0.948	0.284	0.338	0.925	0.172	0.184	0.230
gvs$^{\mathrm{W}}$.hyp	1.000	0.948	0.288	0.338	0.927	0.173	0.184	0.231
bas.hyp	1.000	0.948	0.286	0.337	0.926	0.172	0.184	0.231

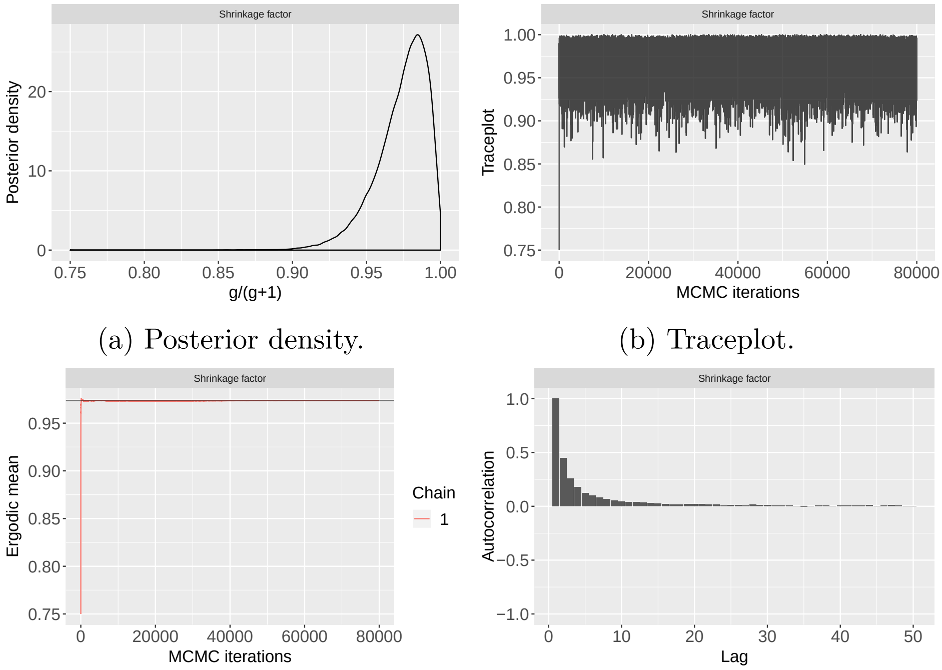
Table A.11 Marginal posterior inclusion probabilities for each independent variable X_j regarding each Bayesian variable selection method.

Among all methods under the three different prior setup, posterior model probabilities show very small differentiations with one exception for the g-prior approach where the SSVS procedure indicates higher posterior measure than the other methods for the g.prior. Posterior model probabilities of GVS implemented in R and WINBUGS are showing similar results with those of BAS for g-prior, hyper-g which are validated in this way. In addition, a median probability model approach was added in the analysis to illustrate better the performance of the MCMC methods.

Regarding the convergence of shrinkage factor $\frac{g}{g+1}$, Figures (A.6), (A.7), (A.8), (A.9) verify the convergence for each Bayesian variable selection method. The computational methods of Table (A.11), are compared in terms of marginal posterior inclusion probability. From the results of Table (A.11) we deduce that MCMC methods included as important covariates $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, $\boldsymbol{X_5}$ since their marginal posterior inclusion probabilities are higher than 0.5. Furthermore, all computed methods for g-prior exhibit a decreased magnitude of marginal posterior inclusion probabilities for the non important covariates $\boldsymbol{X_3}$, $\boldsymbol{X_4}$, $\boldsymbol{X_6}$, $\boldsymbol{X_7}$, $\boldsymbol{X_8}$ whereas in the case of mixtures of g-priors those are directed towards 0.5. The latter behaviour naturally arises as additional randomness of the g parameter which affects the posterior inclusion measures through the adaptive shrinkage of factor $\frac{g}{g+1}$. Moreover, the Bayesian adaptive sampling and GVS in WINBUGS verify the previous results of MCMC procedures in R. On the contrary, the predictive ability of each Bayesian variable selection method is assesed based on the computation of MSE

$$\widehat{MSE} \approx \sqrt{\frac{\sum_{s=1}^{S}\sum_{i=1}^{n_{te}}(y_i^{te} - a^{(s)} - \boldsymbol{x}_i^{te}\boldsymbol{\beta}^{(s)})^2}{Sn_{te}}},$$

where $y_{(\cdot)}^{te}$ are the test values for the response of the tested sample n_{te}, $\boldsymbol{x}_{(\cdot)}^{te}$ are the row-wise of the test design matrix $\boldsymbol{X}^{te}$ and $a^{(s)}$, $\widehat{\boldsymbol{\beta}}^{(s)}$ are the estimated values from each iteration step of each method of intercept and regression coefficients respectively. The maximum aposteriori and median probability model coincided for each computational method identified from the sample of model search algorithms. The predictive ability of SSVS and GVS was similar in terms of MSE with a difference of 2% where SSVS outperformed slightly in Tables (A.12), (A.13), (A.14), (A.15), (A.16), (A.17). Figures (A.10), (A.11), (A.12), (A.13), (A.14), (A.15) depict the posterior distribution of MSE for each MCMC method under the three different prior setups where both subfigures show convergence. To conclude, our analysis seems effective both in model fitting and prediction with including only covariates such as the logarithm of cancer volume ($\boldsymbol{X_1}$), the logarithm of prostate weight ($\boldsymbol{X_2}$) and the seminal vesicle invasion ($\boldsymbol{X_5}$). The same conclusions were found also for maximum aposteriori and median probability model regarding the first and second analysis respectively Fouskakis and Ntzoufras (2013) and Leng et al. (2014).



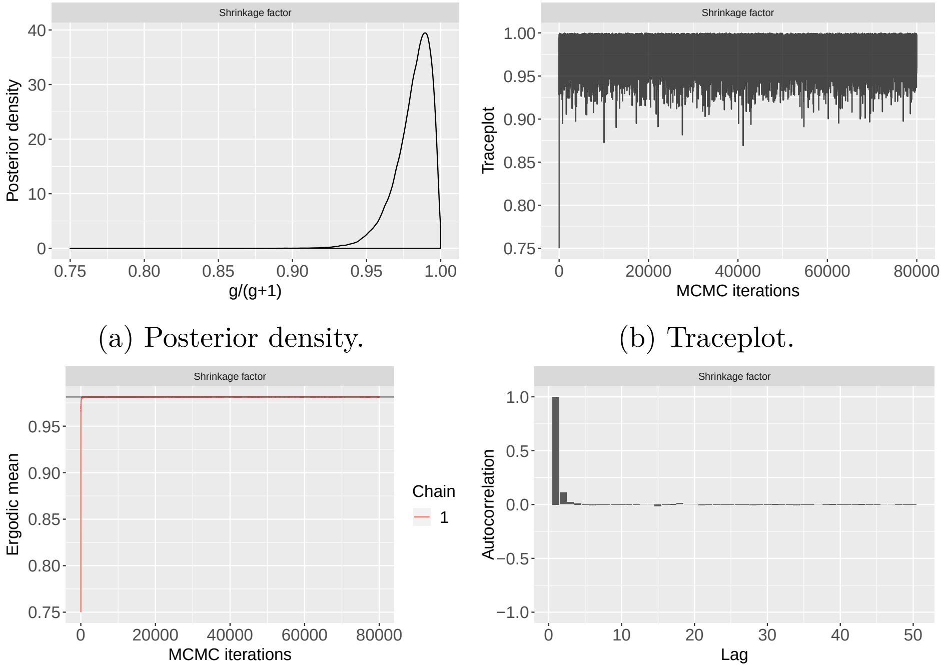
(a) Posterior density.

(b) Traceplot.

(c) Ergodic mean.

(d) Autocorrelation function.

Fig. A.6 Convergence diagnostics of shrinkage factor $\frac{g}{g+1}$ for ssvs$^{\mathrm{R}}$.ZS.



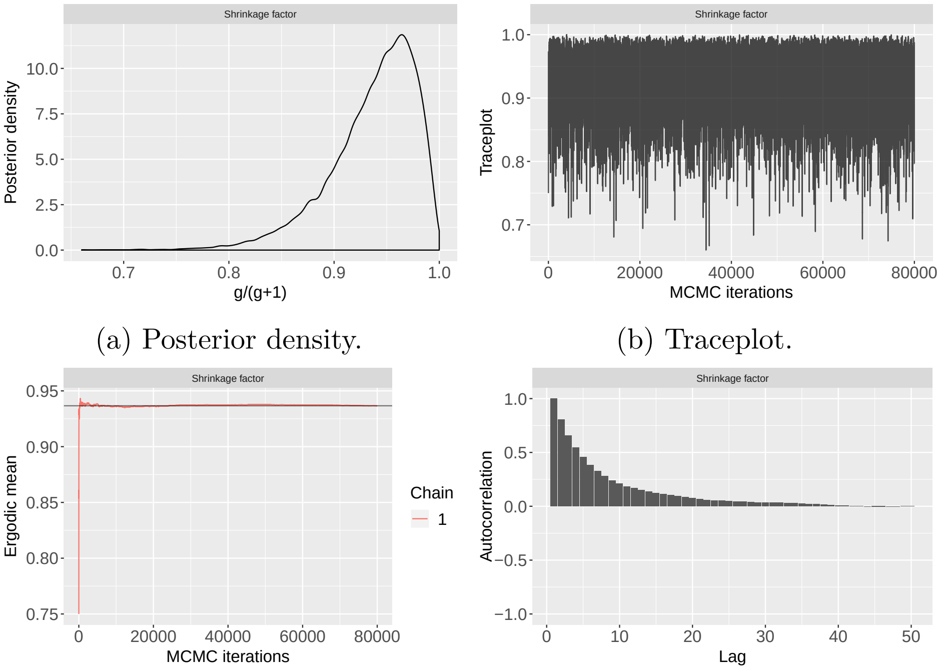
(a) Posterior density.

(b) Traceplot.

(c) Ergodic mean.

(d) Autocorrelation function.

Fig. A.7 Convergence diagnostics of shrinkage factor $\frac{g}{g+1}$ for gvs$^{\mathrm{R}}$.ZS.



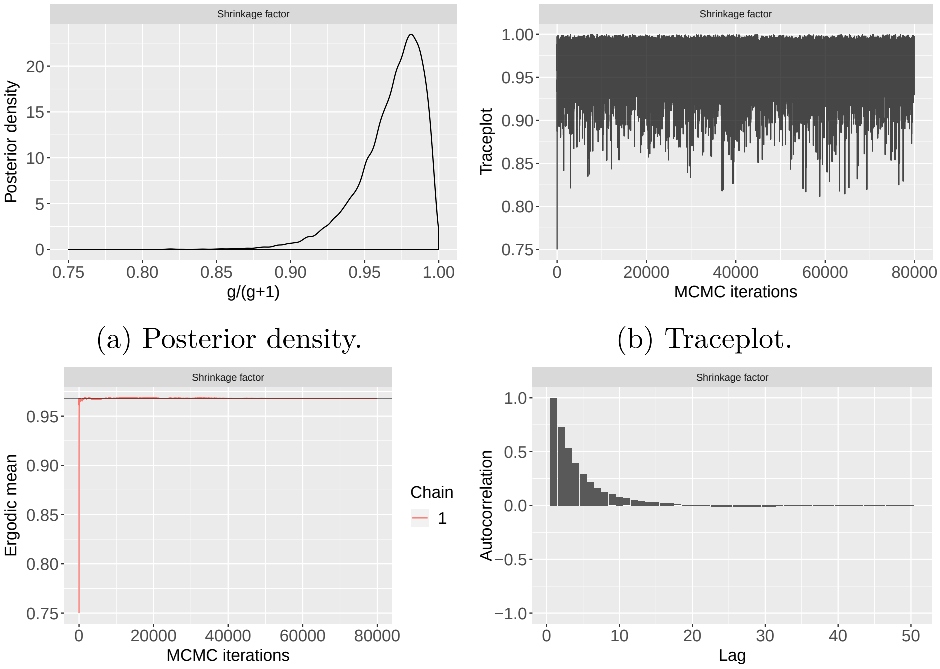
(a) Posterior density.

(b) Traceplot.

(c) Ergodic mean.

(d) Autocorrelation function.

Fig. A.8 Convergence diagnostics of shrinkage factor $\frac{g}{g+1}$ for ssvs$^{\mathrm{R}}$.hyp.



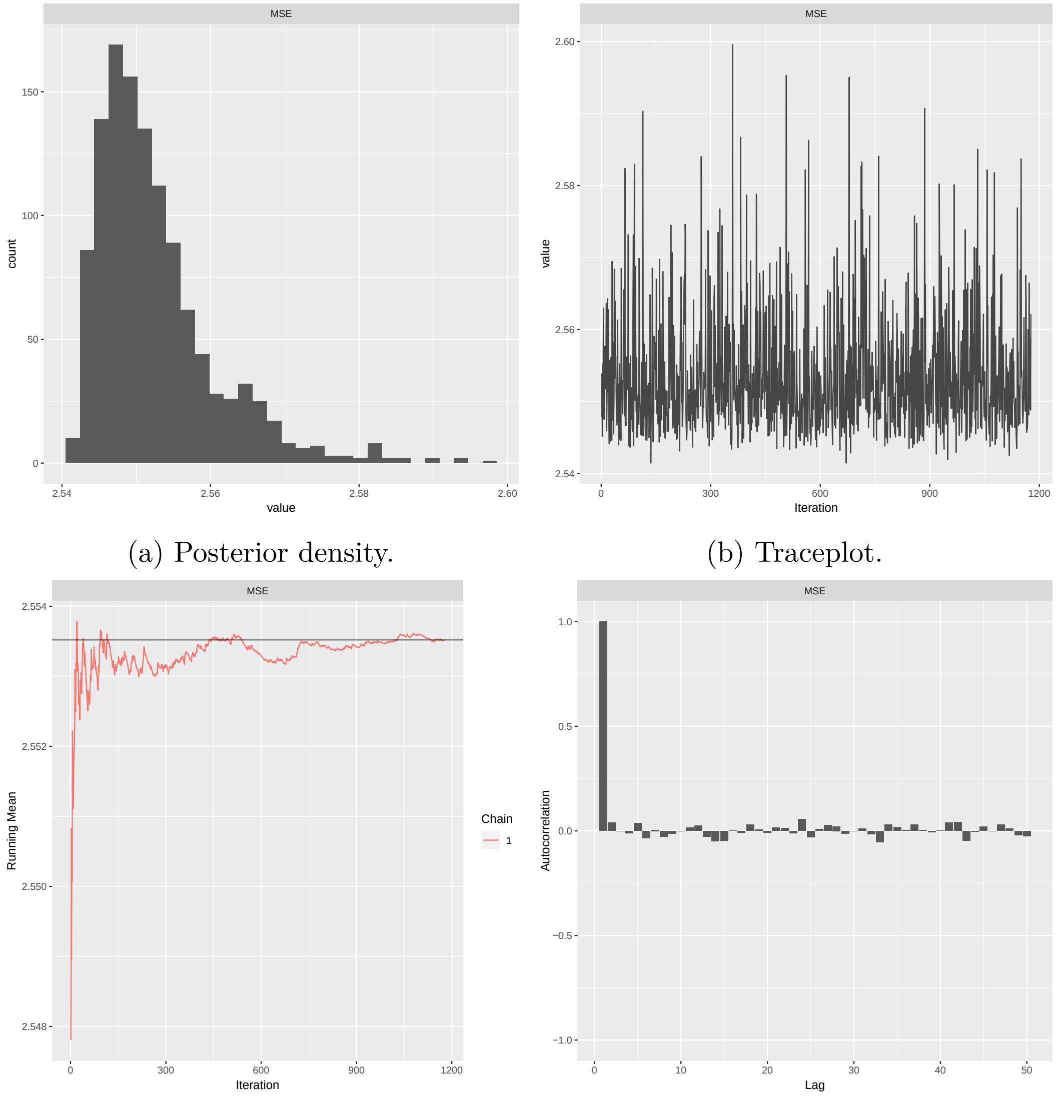
(a) Posterior density.

(b) Traceplot.

(c) Ergodic mean.

(d) Autocorrelation function.

Fig. A.9 Convergence diagnostics of shrinkage factor $\frac{g}{g+1}$ for gvs$^{\mathrm{R}}$.hyp.



(a) Posterior density.

(b) Traceplot.

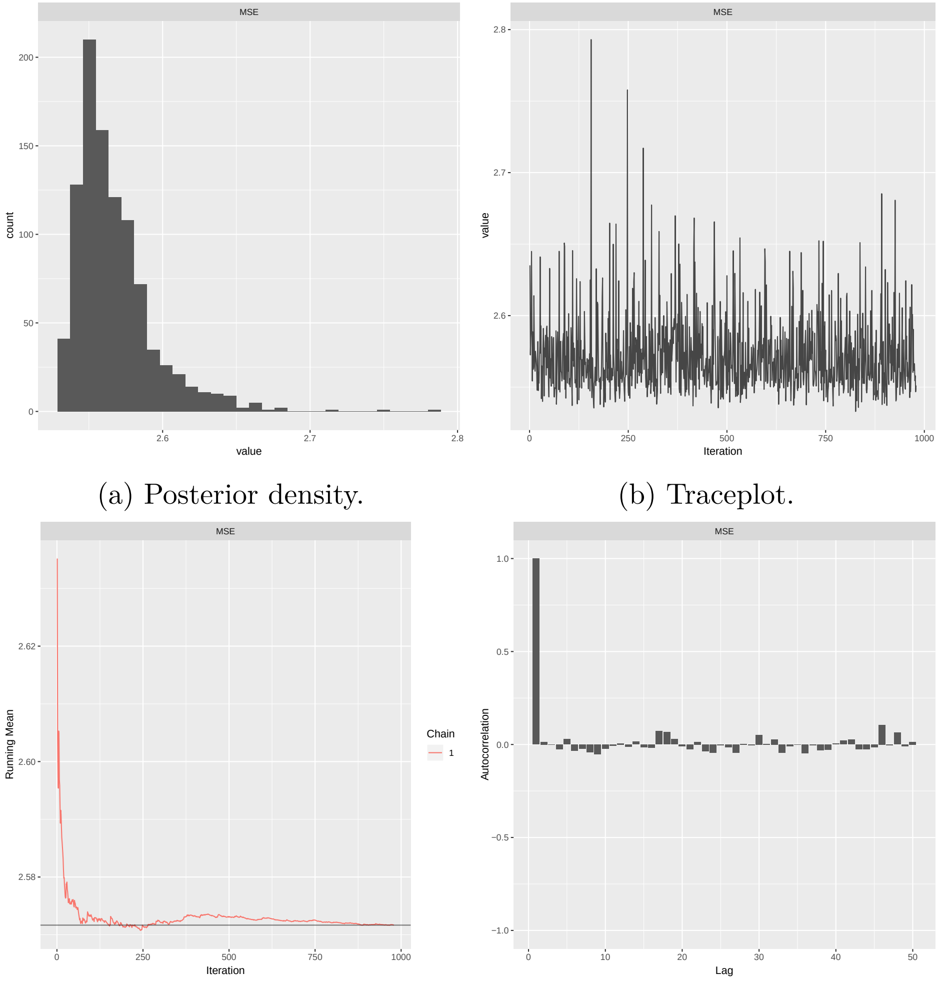
(c) Ergodic mean.

(d) Autocorrelation function.

Fig. A.10 Convergence diagnostics of mean squared error MSE for ssvs$^{\mathrm{R}}$.g.

MSE	**Posterior Summary Statistics**	
Method	**Mean**	**Standard Error**
ssvs.g	2.553	0.008

Table A.12 Posterior summary statistics of mean squared error MSE for ssvs$^{\mathrm{R}}$.g.



(a) Posterior density.

(b) Traceplot.

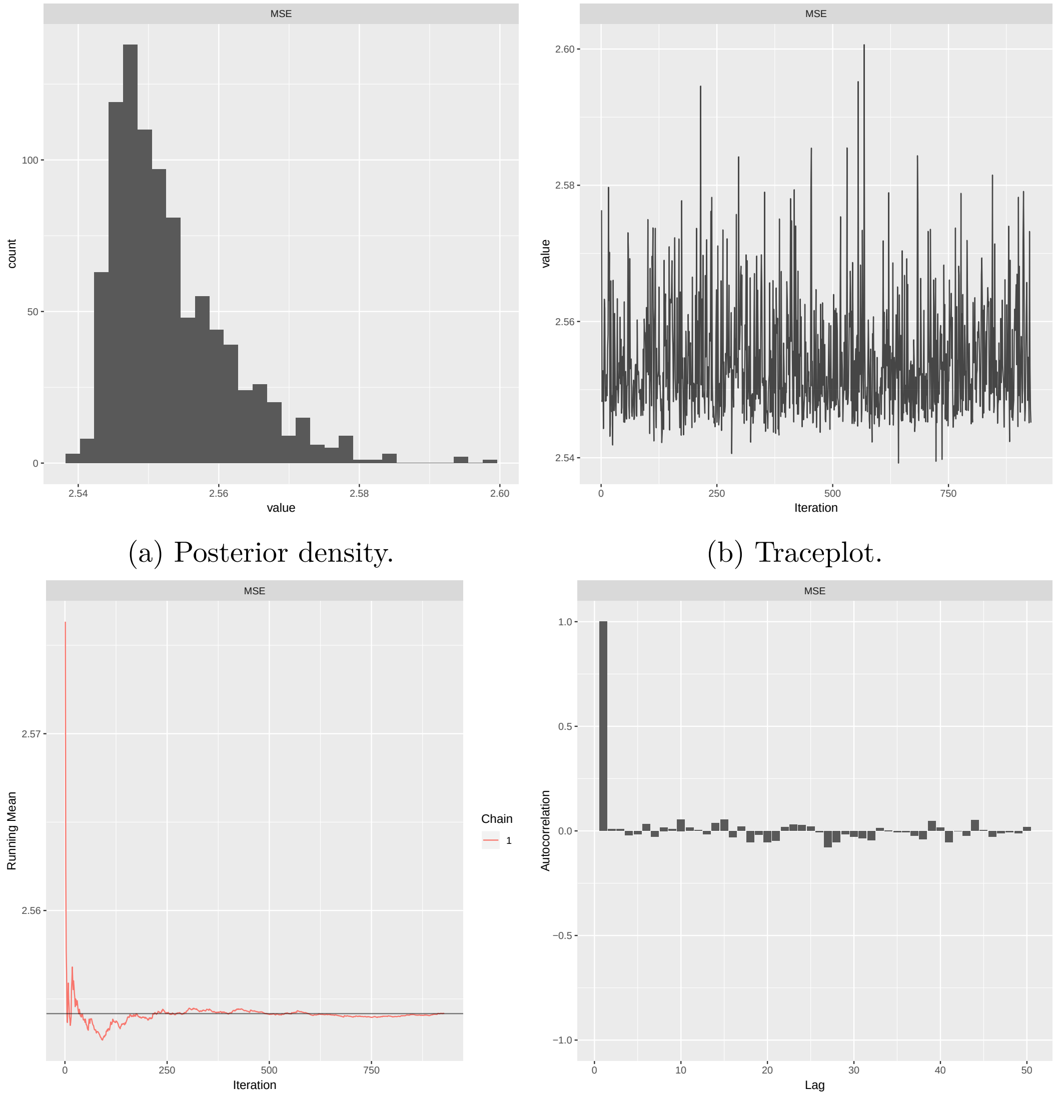
(c) Ergodic mean.

(d) Autocorrelation function.

Fig. A.11 Convergence diagnostics of mean squared error MSE for gvs^{R}.g.

MSE	**Posterior Summary Statistics**	
Method	**Mean**	**Standard Error**
gvs.g	2.571	0.027

Table A.13 Posterior summary statistics of mean squared error MSE for gvs^{R}.g.



(a) Posterior density.

(b) Traceplot.

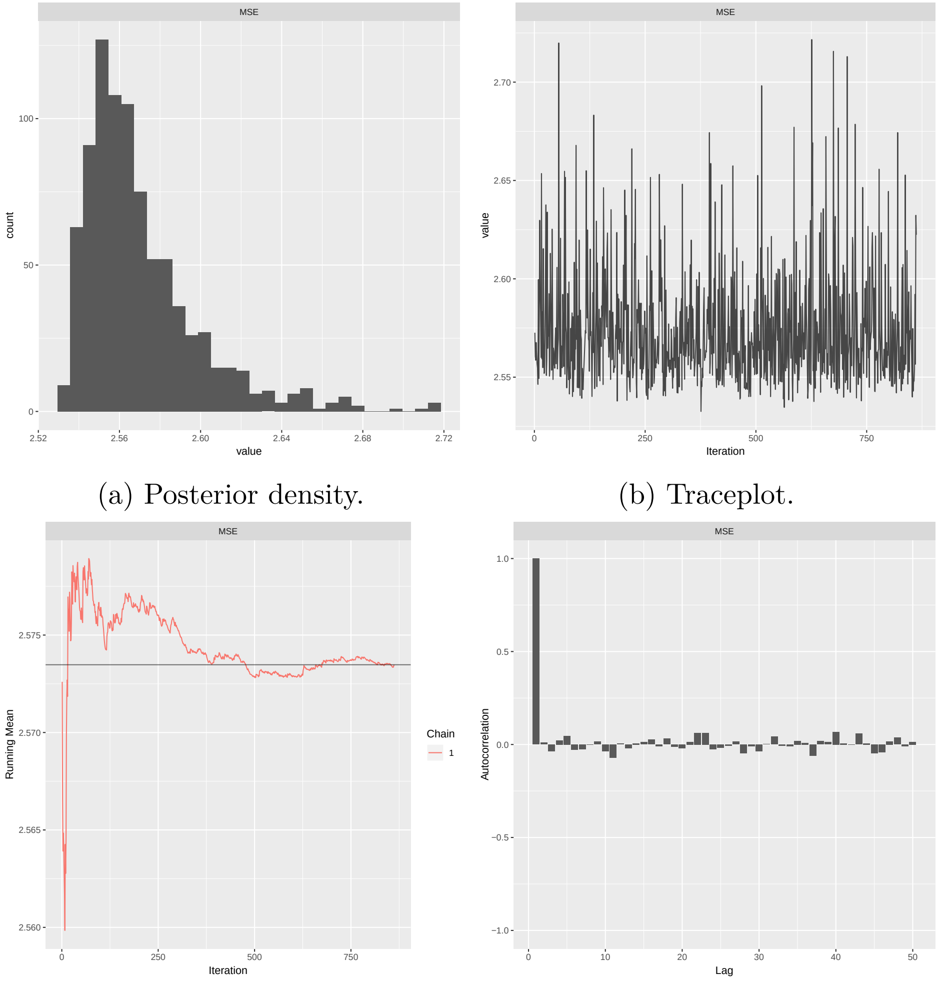
(c) Ergodic mean.

(d) Autocorrelation function.

Fig. A.12 Convergence diagnostics of mean squared error MSE for ssvs^{R}.ZS.

MSE	**Posterior Summary Statistics**	
Method	**Mean**	**Standard Error**
ssvs.ZS	2.554	0.008

Table A.14 Posterior summary statistics of mean squared error MSE for ssvs^{R}.ZS.



(a) Posterior density.

(b) Traceplot.

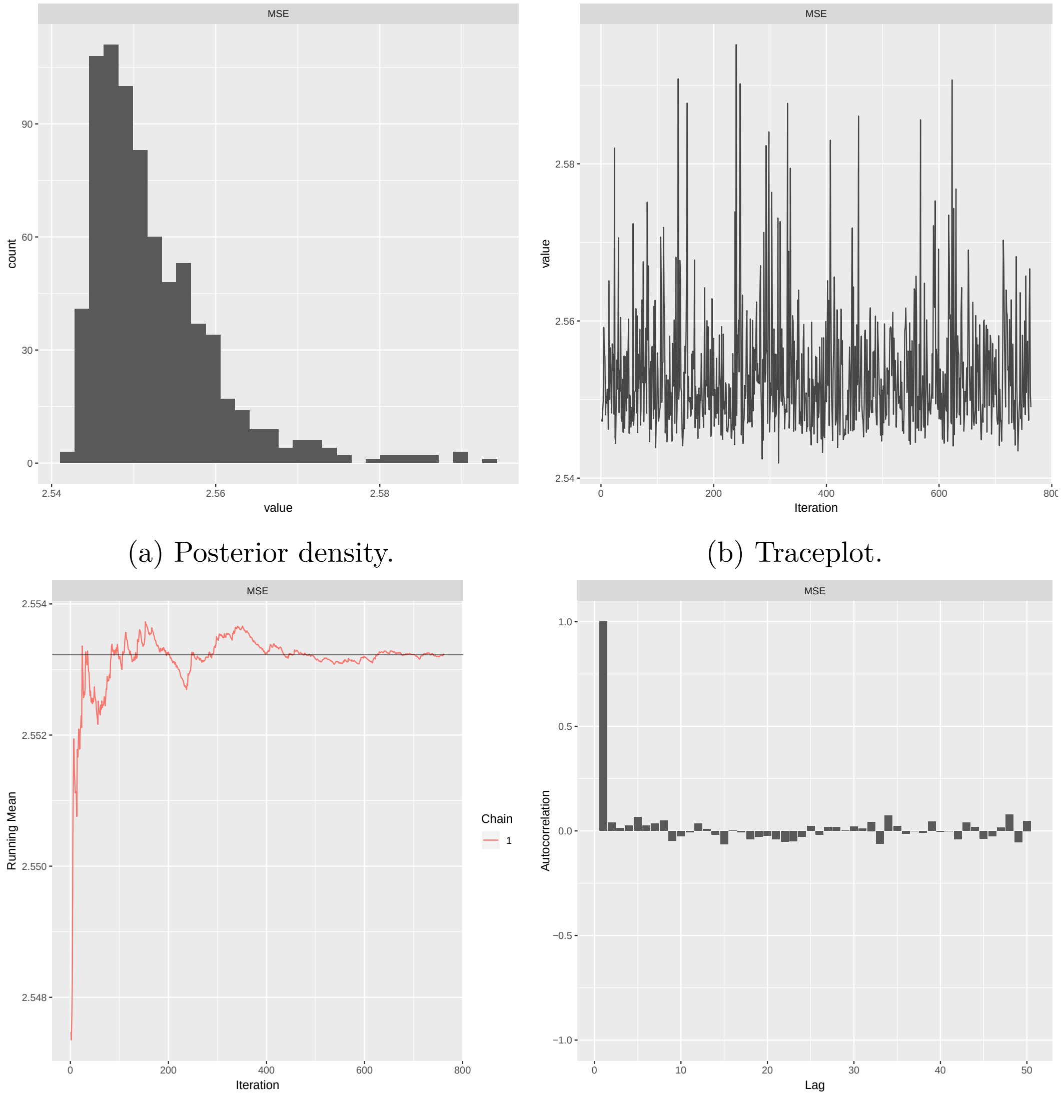
(c) Ergodic mean.

(d) Autocorrelation function.

Fig. A.13 Convergence diagnostics of mean squared error MSE for gvs$^{\mathrm{R}}$.ZS.

MSE	**Posterior Summary Statistics**	
Method	**Mean**	**Standard Error**
gvs.ZS	2.573	0.028

Table A.15 Posterior summary statistics of mean squared error MSE for gvs$^{\mathrm{R}}$.ZS.



(a) Posterior density.

(b) Traceplot.

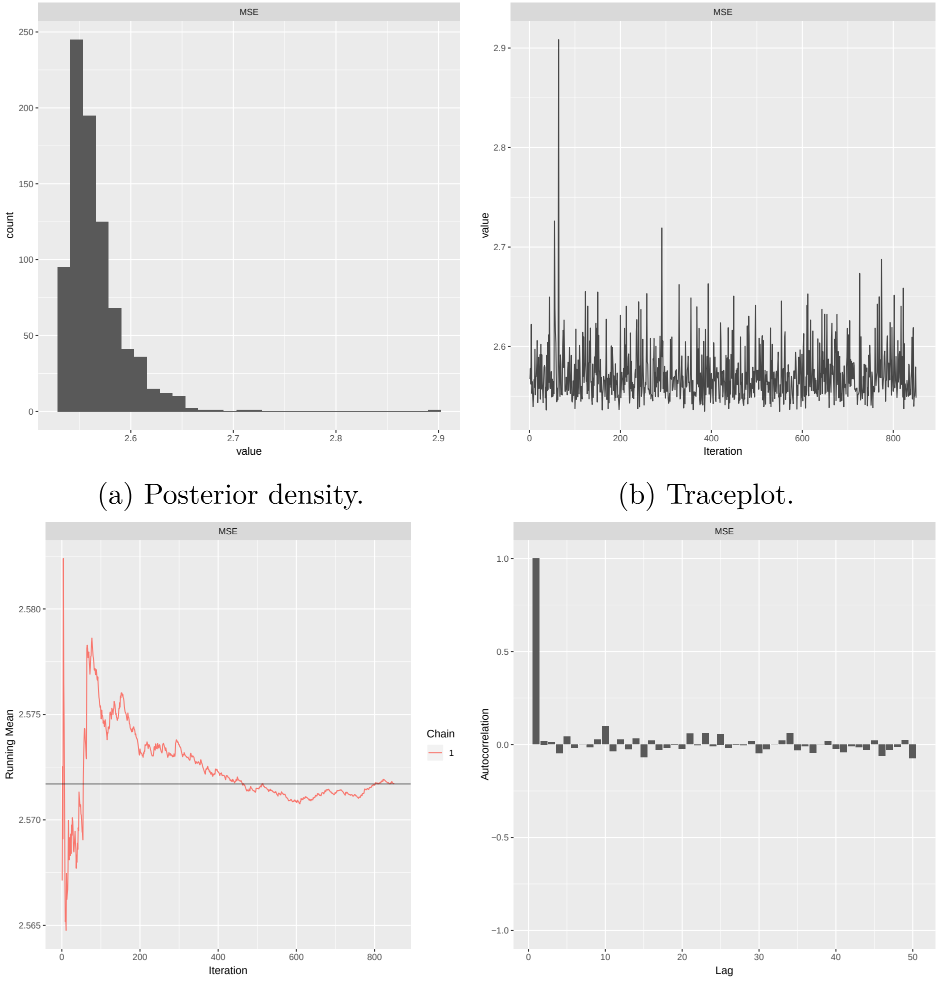
(c) Ergodic mean.

(d) Autocorrelation function.

Fig. A.14 Convergence diagnostics of mean squared error MSE for ssvs$^{\mathrm{R}}$.hyp.

MSE	**Posterior Summary Statistics**	
Method	**Mean**	**Standard Error**
ssvs.hy	2.553	0.007

Table A.16 Posterior summary statistics of mean squared error MSE for ssvs$^{\mathrm{R}}$.hyp.



(a) Posterior density.

(b) Traceplot.

(c) Ergodic mean.

(d) Autocorrelation function.

Fig. A.15 Convergence diagnostics of mean squared error MSE for gvs^{R}.hyp.

MSE	**Posterior Summary Statistics**	
Method	**Mean**	**Standard Error**
gvs.hy	2.571	0.0283

Table A.17 Posterior summary statistics of mean squared error MSE for gvs^{R}.hyp.

Appendix B

Bayesian Variable Selection in Generalized Linear Models

B.1 Laplace Approximation of Bove and Held

The authors provide a Laplace approximation of second order Taylor series expanding the unormalized log-posterior of $\pi^{(BH)}(\boldsymbol{\beta}_{\gamma+1}|\boldsymbol{y}, g, \boldsymbol{\phi})$ around it's posterior mode $\widehat{\boldsymbol{\mu}}_{\gamma+1}$ with precision matrix $\widehat{\boldsymbol{R}}_{\gamma+1}$ evaluated $\widehat{\boldsymbol{\mu}}_{\gamma+1}$ as the following

$$\log\left\{f(\boldsymbol{y}|\boldsymbol{\beta}_{\gamma+1}, \boldsymbol{\phi}, \boldsymbol{\gamma})\pi^{(BH)}(\boldsymbol{\beta}_{\gamma+1}|g, \boldsymbol{\phi}, \delta, \boldsymbol{\gamma})\right\} \approx \log\left\{f(\boldsymbol{y}|\widehat{\boldsymbol{\mu}}_{\gamma+1}, \boldsymbol{\phi}, \boldsymbol{\gamma})\pi^{(BH)}(\widehat{\boldsymbol{\mu}}_{\gamma+1}|g, \boldsymbol{\phi}, \delta, \boldsymbol{\gamma})\right\}$$
$$+\left(\boldsymbol{\beta}_{\gamma+1} - \widehat{\boldsymbol{\mu}}_{\gamma+1}\right)^T \left(\left.\frac{\partial \log\left\{f(\boldsymbol{y}|\boldsymbol{\beta}_{\gamma+1}, \boldsymbol{\phi}, \boldsymbol{\gamma})\pi^{(BH)}(\boldsymbol{\beta}_{\gamma+1}|g, \boldsymbol{\phi}, \delta, \boldsymbol{\gamma})\right\}}{\partial \boldsymbol{\beta}_{\gamma+1}}\right|_{\boldsymbol{\beta}_{\gamma+1}=\widehat{\boldsymbol{\mu}}_{\gamma+1}}\right)$$
$$-\frac{1}{2}\left(\boldsymbol{\beta}_{\gamma+1} - \widehat{\boldsymbol{\mu}}_{\gamma+1}\right)^T \widehat{\boldsymbol{R}}_{\gamma+1}\left(\boldsymbol{\beta}_{\gamma+1} - \widehat{\boldsymbol{\mu}}_{\gamma+1}\right),$$

where $\left(\boldsymbol{\beta}_{\gamma+1} - \widehat{\boldsymbol{\mu}}_{\gamma+1}\right)^T \left(\left.\frac{\partial \log\left\{f(\boldsymbol{y}|\boldsymbol{\beta}_{\gamma+1}, \boldsymbol{\phi}, \boldsymbol{\gamma})\pi^{(BH)}(\boldsymbol{\beta}_{\gamma+1}|g, \boldsymbol{\phi}, \delta, \boldsymbol{\gamma})\right\}}{\partial \boldsymbol{\beta}_{\gamma+1}}\right|_{\boldsymbol{\beta}_{\gamma+1}=\widehat{\boldsymbol{\mu}}_{\gamma+1}}\right) = 0,$

because $\widehat{\boldsymbol{\mu}}_{\gamma+1}$ is the mode with prior precision matrix $\widehat{\boldsymbol{R}}_{\gamma+1}$

$$\widehat{\boldsymbol{R}}_{\gamma+1} = \left(-\left.\frac{\partial^2 \log\left\{f(\boldsymbol{y}|\boldsymbol{\beta}_{\gamma+1}, \boldsymbol{\phi}, \boldsymbol{\gamma})\pi^{(BH)}(\boldsymbol{\beta}_{\gamma+1}|g, \boldsymbol{\phi}, \delta, \boldsymbol{\gamma})\right\}}{\partial \boldsymbol{\beta}^2_{\gamma+1}}\right|_{\boldsymbol{\beta}_{\gamma+1}=\widehat{\boldsymbol{\mu}}_{\gamma+1}}\right)$$
$$= \begin{pmatrix} -\left.\frac{\partial^2 \log\left\{\pi^{(BH)}(\boldsymbol{\beta}_{\gamma+1}|\boldsymbol{y}, g, \boldsymbol{\phi})\right\}}{\partial \boldsymbol{\beta}^2_{\gamma}}\right|_{\boldsymbol{\beta}_{\gamma+1}=\widehat{\boldsymbol{\mu}}_{\gamma+1}} & -\left.\frac{\partial^2 \log\left\{\pi^{(BH)}(\boldsymbol{\beta}_{\gamma+1}|\boldsymbol{y}, g, \boldsymbol{\phi})\right\}}{\partial a \partial \boldsymbol{\beta}_{\gamma}}\right|_{\boldsymbol{\beta}_{\gamma+1}=\widehat{\boldsymbol{\mu}}_{\gamma+1}} \\ -\left.\frac{\partial^2 \log\left\{\pi^{(BH)}(\boldsymbol{\beta}_{\gamma+1}|\boldsymbol{y}, g, \boldsymbol{\phi})\right\}}{\partial \boldsymbol{\beta}_{\gamma} \partial a}\right|_{\boldsymbol{\beta}_{\gamma+1}=\widehat{\boldsymbol{\mu}}_{\gamma+1}} & -\left.\frac{\partial^2 \log\left\{\pi^{(BH)}(\boldsymbol{\beta}_{\gamma+1}|\boldsymbol{y}, g, \boldsymbol{\phi})\right\}}{\partial a^2}\right|_{\boldsymbol{\beta}_{\gamma+1}=\widehat{\boldsymbol{\mu}}_{\gamma+1}} \end{pmatrix},$$

the above Laplace approximation may be reduced in the following

$$\log\left\{f(\boldsymbol{y}|\boldsymbol{\beta}_{\gamma+1},\boldsymbol{\phi},\boldsymbol{\gamma})\pi^{(BH)}(\boldsymbol{\beta}_{\gamma+1}|g,\boldsymbol{\phi},\delta,\boldsymbol{\gamma})\right\} \approx \log\left\{f(\boldsymbol{y}|\widehat{\boldsymbol{\mu}}_{\gamma+1},\boldsymbol{\phi},\boldsymbol{\gamma})\pi^{(BH)}(\widehat{\boldsymbol{\mu}}_{\gamma+1}|g,\boldsymbol{\phi},\delta,\boldsymbol{\gamma})\right\} - \frac{1}{2}\left(\boldsymbol{\beta}_{\gamma+1}-\widehat{\boldsymbol{\mu}}_{\gamma+1}\right)^{T}\widehat{\boldsymbol{R}}_{\gamma+1}\left(\boldsymbol{\beta}_{\gamma+1}-\widehat{\boldsymbol{\mu}}_{\gamma+1}\right),$$

using this expression into the marginal likelihood (3.10) for the joint vector $\boldsymbol{\beta}_{\gamma+1}$ we have

$$m^{(BH)}(\boldsymbol{y}|\boldsymbol{\gamma},g) = \int\limits_{\boldsymbol{\beta}_{\gamma+1}} \exp\left\{\log\left\{f(\boldsymbol{y}|\boldsymbol{\beta}_{\gamma+1},\boldsymbol{\phi},\boldsymbol{\gamma})\pi^{(BH)}(\boldsymbol{\beta}_{\gamma+1}|g,\boldsymbol{\phi},\delta,\boldsymbol{\gamma})\right\}\right\}d\boldsymbol{\beta}_{\gamma+1}$$

$$\approx f(\boldsymbol{y}|\widehat{\boldsymbol{\mu}}_{\gamma+1},\boldsymbol{\phi},\boldsymbol{\gamma})\pi^{(BH)}(\widehat{\boldsymbol{\mu}}_{\gamma+1}|g,\boldsymbol{\phi},\delta,\boldsymbol{\gamma}) \int_{\boldsymbol{\beta}_{\gamma+1}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}_{\gamma+1}-\widehat{\boldsymbol{\mu}}_{\gamma+1}\right)^{T}\widehat{\boldsymbol{R}}_{\gamma+1}\left(\boldsymbol{\beta}_{\gamma+1}-\widehat{\boldsymbol{\mu}}_{\gamma+1}\right)\right\}d\boldsymbol{\beta}_{\gamma+1},$$

the last step of the above expression may be reduced using (3.9), (3.7)

$$m^{(BH)}(\boldsymbol{y}|\boldsymbol{\gamma},g) \approx f(\boldsymbol{y}|\widehat{\boldsymbol{\mu}}_{\gamma+1},\boldsymbol{\phi},\boldsymbol{\gamma})(2\pi g\delta\phi)^{-\frac{p_\gamma}{2}}\det(\boldsymbol{X}_{\gamma}^{\boldsymbol{T}}\boldsymbol{X}_{\gamma})^{\frac{1}{2}}\exp\left\{-\frac{1}{2g\phi\delta}\widehat{\boldsymbol{\mu}}_{\gamma}^{T}\boldsymbol{X}_{\gamma}^{\boldsymbol{T}}\boldsymbol{X}_{\gamma}\widehat{\boldsymbol{\mu}}_{\gamma}^{T}\right\} (2\pi)^{\frac{p_\gamma+1}{2}}\det(\widehat{\boldsymbol{R}}_{\gamma+1})^{-\frac{1}{2}},$$

hence, the last step is equal to (3.11) is verified due to the integrated normal kernel of the multivariate normal distribution $N_{p_\gamma}\left(\widehat{\boldsymbol{\mu}}_{\gamma},\widehat{\boldsymbol{R}}_{\gamma+1}^{-1}\right)$.

B.2 Gaussian Quadrature Approximation with Mixtures Of g-priors

It is known that univariate integrals of the form for a real valued function $f(z)$

$$\int\limits_{-\infty}^{\infty}\exp\left(-z^{2}\right)f(z)dz,$$

may be approximated with a Gaussian-type formula

$$\sum_{j=1}^{N}v_{j}f(z_{j}),$$

where

$$v_j = \frac{2^{N-1}N!\sqrt{\pi}}{N^2\left[He_{n-1}(z_j)\right]}$$

and z_j is the j-th zero Hermite polynomial $He_{n-1}(z)$; for more details see Davis and Rabinowitz (1986). Moreover, the remainder function has the form

$$R_N = \frac{N!\sqrt{\pi}}{2^N(2N)!}f^{(2N)}(q),$$

for some q, if $f(z)$ is indeed a polynomial of degree $2N-1$, the Gaussian quadrature (B.1) will be precise and will result with a remainder function $R_N = 0$. If $F(z)$ is a suitably regular function, then

$$g(z) = F(z)(2\pi r^2)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left(\frac{z-\bar{m}}{s_{\bar{m}}}\right)^2\right\},$$

where $\bar{m}$, $s_{\bar{m}}$ are the mean and standard deviation of z respectively and after the substitution of variables $z = \bar{m} + \sqrt{2}s_{\bar{m}}t$ using

$$\frac{dz}{dt} = \sqrt{2}s_{\bar{m}},$$

considering the transformation step for the integral

$$\int\limits_{-\infty}^{\infty} g(z)dz = \int\limits_{-\infty}^{\infty}\frac{1}{\sqrt{\pi}}F(\bar{m}+\sqrt{2}s_{\bar{m}}t)\exp\left\{-t^2\right\}dt,$$

which has the form of (B.1). Using Gaussian quadrature formula (B.1), the above integral is approximated as

$$\int\limits_{-\infty}^{\infty} g(z)dz \approx \sum_{j=1}^{N} v_j\frac{1}{\sqrt{\pi}}F(\bar{m}+\sqrt{2}s_{\bar{m}}t_j) \tag{B.1}$$

$$\approx \sum_{j=1}^{N} v_j\exp{(t_j^2)}\sqrt{2}s_{\bar{m}}F(\bar{m}+\sqrt{2}s_{\bar{m}}t_j) \tag{B.2}$$

$$\approx \sum_{j=1}^{N} m_jF(z_j),$$

where we denote the actual weights $m_j = v_j\exp{(t_j^2)}\sqrt{2}s_{\bar{m}}$ and nodes $z_j = \bar{m}+\sqrt{2}s_{\bar{m}}t_j$. Tables of t_j,v_j and $v_j\exp{(t_j^2)}$ are available for N and the error term decreases if $F(z)$ is approximately a polynomial. Since, in Bayesian inference the integrals involve usually posterior densities, $\bar{m}$, $s_{\bar{m}}$ will be substituted by the posterior mean and standard deviation $\widehat{z}$, $\widehat{\sigma}_z$. The Gaussian quadrature will have satisfactory results if the posterior density is approximated by the product of a normal density and a polynomial degree at most of order $2N-3$.

B.3 Laplace Approximation of Li and Clyde

The authors provide a Laplace approximation (integrated Laplace approximation) based on a Taylor series expansion of second order for the log-likelihood function $f(\boldsymbol{y}|a, \boldsymbol{\beta}_\gamma, \boldsymbol{\phi}, \boldsymbol{\gamma})$ around the maximum likelihood estimator $(\widehat{a}, \widehat{\boldsymbol{\beta}}_\gamma)$ of model $\boldsymbol{\gamma}$

$$\begin{aligned}\log\left\{f(\boldsymbol{y}|a, \boldsymbol{\beta}_\gamma, \boldsymbol{\phi}, \boldsymbol{\gamma})\right\} \approx{}& \log\left\{f(\boldsymbol{y}|\widehat{a}, \widehat{\boldsymbol{\beta}}_\gamma, \boldsymbol{\phi}, \boldsymbol{\gamma})\right\} \\ &+ \begin{pmatrix} a-\widehat{a} \\ \boldsymbol{\beta}_\gamma - \widehat{\boldsymbol{\beta}}_\gamma \end{pmatrix}^T \left(\left.\frac{\partial \log\left\{f(\boldsymbol{y}|\widehat{a}, \widehat{\boldsymbol{\beta}}_\gamma, \boldsymbol{\phi}, \boldsymbol{\gamma})\right\}}{\partial \widehat{\boldsymbol{\psi}}_\gamma}\right|_{(\widehat{a},\widehat{\boldsymbol{\beta}}_\gamma)=(\widehat{a},\widehat{\boldsymbol{\beta}}_\gamma)} \right) \\ &- \frac{1}{2} \begin{pmatrix} a-\widehat{a} \\ \boldsymbol{\beta}_\gamma - \widehat{\boldsymbol{\beta}}_\gamma \end{pmatrix}^T \mathcal{I}^{(LC)}(\widehat{\boldsymbol{\psi}}_\gamma) \begin{pmatrix} a-\widehat{a} \\ \boldsymbol{\beta}_\gamma - \widehat{\boldsymbol{\beta}}_\gamma \end{pmatrix},\end{aligned}$$

where the following expression hold $\left.\frac{\partial \log\left\{f(\boldsymbol{y}|\widehat{a},\widehat{\boldsymbol{\beta}}_\gamma,\boldsymbol{\phi},\gamma)\right\}}{\partial \widehat{\boldsymbol{\psi}}_\gamma}\right|_{(\widehat{a},\widehat{\boldsymbol{\beta}}_\gamma)=(\widehat{a},\widehat{\boldsymbol{\beta}}_\gamma)} = \begin{pmatrix} \frac{\partial \log\left\{(f(\boldsymbol{y}|\widehat{a},\widehat{\boldsymbol{\beta}}_\gamma,\boldsymbol{\phi},\gamma)\right\}}{\partial \widehat{a}} \\ \frac{\partial \log\left(f(\boldsymbol{y}|\widehat{a},\widehat{\boldsymbol{\beta}}_\gamma,\boldsymbol{\phi},\gamma)\right)}{\partial \widehat{\boldsymbol{\beta}}_\gamma} \end{pmatrix}$, $\begin{pmatrix} a-\widehat{a} \\ \boldsymbol{\beta}_\gamma - \widehat{\boldsymbol{\beta}}_\gamma \end{pmatrix}^T \left(\left.\frac{\partial \log\left\{f(\boldsymbol{y}|\widehat{a},\widehat{\boldsymbol{\beta}}_\gamma,\boldsymbol{\phi},\gamma)\right\}}{\partial \widehat{\boldsymbol{\psi}}_\gamma}\right|_{(\widehat{a},\widehat{\boldsymbol{\beta}}_\gamma)=(\widehat{a},\widehat{\boldsymbol{\beta}}_\gamma)} \right) = 0$, because of the maximum likelihood estimator $(\widehat{a}, \widehat{\boldsymbol{\beta}}_\gamma)$ with observed Fisher information matrix $\mathcal{I}^{(LC)}(\widehat{\boldsymbol{\psi}}_\gamma)$ denoted as

$$\begin{aligned}\mathcal{I}^{(LC)}(\widehat{\boldsymbol{\psi}}_\gamma) &= \left(-\left.\frac{\partial^2 \log\left\{f(\boldsymbol{y}|\widehat{a}, \widehat{\boldsymbol{\beta}}_\gamma, \boldsymbol{\phi}, \boldsymbol{\gamma})\right\}}{\partial \widehat{\boldsymbol{\psi}}_\gamma^2}\right|_{(\widehat{a},\widehat{\boldsymbol{\beta}}_\gamma)=(\widehat{a},\widehat{\boldsymbol{\beta}}_\gamma)} \right) \\ &= \begin{pmatrix} \mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma}) & \mathbf{0}_p \\ \mathbf{0}_p^T & \mathcal{I}^{(LC)}(\widehat{\boldsymbol{\beta}}_\gamma) \end{pmatrix},\end{aligned}$$

the above Laplace approximation may be reduced in the following

$$\log\left\{f(\boldsymbol{y}|a, \boldsymbol{\beta}_\gamma, \boldsymbol{\phi}, \boldsymbol{\gamma})\right\} \approx \log\left\{f(\boldsymbol{y}|\widehat{a}, \widehat{\boldsymbol{\beta}}_\gamma, \boldsymbol{\phi}, \boldsymbol{\gamma})\right\} - \frac{1}{2} \begin{pmatrix} a-\widehat{a} \\ \boldsymbol{\beta}_\gamma - \widehat{\boldsymbol{\beta}}_\gamma \end{pmatrix}^T \mathcal{I}^{(LC)}(\widehat{\boldsymbol{\psi}}_\gamma) \begin{pmatrix} a-\widehat{a} \\ \boldsymbol{\beta}_\gamma - \widehat{\boldsymbol{\beta}}_\gamma \end{pmatrix},$$

and after some linear algebra steps the above expression is reduced as

$$\begin{aligned}\log\left\{f(\boldsymbol{y}|a, \boldsymbol{\beta}_\gamma, \boldsymbol{\phi}, \boldsymbol{\gamma})\right\} \approx{}& \log\left\{f(\boldsymbol{y}|\widehat{a}, \widehat{\boldsymbol{\beta}}_\gamma, \boldsymbol{\phi}, \boldsymbol{\gamma})\right\} - \frac{1}{2}\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})\,(a-\widehat{a})^2 \\ &- \frac{1}{2}\left(\boldsymbol{\beta}_\gamma - \widehat{\boldsymbol{\beta}}_\gamma)^T \mathcal{I}^{(LC)}(\widehat{\boldsymbol{\beta}}_\gamma)(\boldsymbol{\beta}_\gamma - \widehat{\boldsymbol{\beta}}_\gamma\right),\end{aligned}$$

substituting this expression into the marginal likelihood (3.18) we have

$$m^{(LC)}(\boldsymbol{y}|\boldsymbol{\gamma},g) \approx f(\boldsymbol{y}|a,\boldsymbol{\beta}_\gamma,\boldsymbol{\phi},\boldsymbol{\gamma}) \quad \int\limits_a \exp\left\{-\frac{1}{2}\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})\left(a-\widehat{a}\right)^2\right\}\pi^{(LC)}(a|\boldsymbol{\gamma})da$$
$$\int\limits_{\boldsymbol{\beta}_\gamma} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}_\gamma-\widehat{\boldsymbol{\beta}}_\gamma)^T\mathcal{I}^{(LC)}(\widehat{\boldsymbol{\beta}}_\gamma)(\boldsymbol{\beta}_\gamma-\widehat{\boldsymbol{\beta}}_\gamma\right)\right\}\pi^{(LC)}(\boldsymbol{\beta}_\gamma|\widehat{a},\widehat{\boldsymbol{\beta}}_\gamma,g,\boldsymbol{\phi},\boldsymbol{\gamma})d\boldsymbol{\beta}_\gamma,$$

the last step of the above expression may be reduced in the following using (3.16), (3.17) as

$$m^{(LC)}(\boldsymbol{y}|\boldsymbol{\gamma},g) \approx (2\pi nv)^{\frac{1}{2}}(2\pi)^{-\frac{p_\gamma}{2}}\det(\mathcal{I}^{(LC)}(\widehat{\boldsymbol{\beta}}_\gamma))^{\frac{1}{2}}g^{-\frac{p_\gamma}{2}}f(\boldsymbol{y}|a,\boldsymbol{\beta}_\gamma,\boldsymbol{\phi},\boldsymbol{\gamma})$$
$$\exp\left\{-\frac{1}{2}\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})\widehat{a}^2\right\} \quad \int\limits_a \exp\left\{-\frac{1}{2}\left[\left(\frac{1+nv\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})}{nv}\right)a^2-2a\widehat{a}\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})\right]\right\}da$$
$$\exp\left\{-\frac{1}{2}\widehat{\boldsymbol{\beta}}_\gamma^T\mathcal{I}^{(LC)}(\widehat{\boldsymbol{\beta}}_\gamma)\widehat{\boldsymbol{\beta}}_\gamma\right\}$$
$$\int\limits_{\boldsymbol{\beta}_\gamma} \exp\left\{-\frac{1}{2g}\left[(g+1)\boldsymbol{\beta}_\gamma^T\mathcal{I}^{(LC)}(\widehat{\boldsymbol{\beta}}_\gamma)\boldsymbol{\beta}_\gamma-2g\boldsymbol{\beta}_\gamma^T\mathcal{I}^{(LC)}(\widehat{\boldsymbol{\beta}}_\gamma)\widehat{\boldsymbol{\beta}}_\gamma\right]\right\}d\boldsymbol{\beta}_\gamma,$$

$$m^{(LC)}(\boldsymbol{y}|\boldsymbol{\gamma},g) \approx (2\pi nv)^{\frac{1}{2}}(2\pi)^{-\frac{p_\gamma}{2}}\det(\mathcal{I}^{(LC)}(\widehat{\boldsymbol{\beta}}_\gamma))^{\frac{1}{2}}g^{-\frac{p_\gamma}{2}}f(\boldsymbol{y}|a,\boldsymbol{\beta}_\gamma,\boldsymbol{\phi},\boldsymbol{\gamma})$$
$$\exp\left\{-\frac{1}{2}\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})\widehat{a}^2\right\}\exp\left\{\frac{1}{2}\left(\frac{1+nc\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})}{nv}\right)\left[a^2\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})^2\left(\frac{nv}{1+nv\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})}\right)^2\right]\right\}$$
$$\int\limits_a \exp\left\{-\frac{1}{2}\left(\frac{1+nc\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})}{nv}\right)\left[a^2-2a\widehat{a}\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})\left(\frac{nv}{1+nv\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})}\right)\right]\right\}da$$
$$\exp\left\{-\frac{1}{2}\widehat{\boldsymbol{\beta}}_\gamma^T\mathcal{I}^{(LC)}(\widehat{\boldsymbol{\beta}}_\gamma)\widehat{\boldsymbol{\beta}}_\gamma\right\}\exp\left\{\frac{1}{2g}\widehat{\boldsymbol{v}}_\gamma^T\widehat{\boldsymbol{V}}_\gamma^{-1}\widehat{\boldsymbol{v}}_\gamma\right\}$$
$$\int\limits_{\boldsymbol{\beta}_\gamma} \exp\left\{-\frac{1}{2g}\left[\boldsymbol{\beta}_\gamma^T\widehat{\boldsymbol{V}}_\gamma^{-1}\boldsymbol{\beta}_\gamma-2\boldsymbol{\beta}_\gamma^T\widehat{\boldsymbol{V}}_\gamma^{-1}\widehat{\boldsymbol{v}}_\gamma+\widehat{\boldsymbol{v}}_\gamma^T\widehat{\boldsymbol{V}}_\gamma^{-1}\widehat{\boldsymbol{v}}_\gamma\right]\right\}d\boldsymbol{\beta}_\gamma,$$

where $\widehat{\boldsymbol{V}}_\gamma = (g+1)^{-1}\mathcal{I}^{(LC)}(\widehat{\boldsymbol{\beta}}_\gamma)^{-1}$, $\widehat{\boldsymbol{v}}_\gamma = g\widehat{\boldsymbol{V}}_\gamma^{-1}\mathcal{I}^{(LC)}(\widehat{\boldsymbol{\beta}}_\gamma)\boldsymbol{\beta}_\gamma$ and after some algebra we recognise that in the final step the integrals are just the integrated kernels of the posterior distributions $\boldsymbol{\beta}_\gamma|\boldsymbol{y},g,\boldsymbol{\gamma} \sim N_{p_\gamma}\left(\widehat{\boldsymbol{v}}_\gamma, g\widehat{\boldsymbol{V}}_\gamma\right)$,
$a|\boldsymbol{y},\boldsymbol{\gamma} \sim N\left(\widehat{a}\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})\left(\frac{nv}{1+nv\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})}\right), \left(\frac{1+\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})}{nv}\right)^{-1}\right)$.
Moreover, after some elementary algebra these posteriors are written equivalently $\boldsymbol{\beta}_\gamma|\boldsymbol{y},g,\boldsymbol{\gamma} \sim N_{p_\gamma}\left(\frac{g}{g+1}\widehat{\boldsymbol{\beta}}_\gamma, \frac{g}{g+1}\mathcal{I}^{(LC)}(\widehat{\boldsymbol{\beta}}_\gamma)^{-1}\right)$,

$a|\boldsymbol{y},\boldsymbol{\gamma} \sim N\left(\widehat{a}\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})\left(\frac{nv}{1+nv\mathcal{I}^{(LC)}(\widehat{a}|\gamma)}\right), \left(\frac{1+\mathcal{I}^{(LC)}(\widehat{a}|\gamma)}{nv}\right)^{-1}\right)$ and the marginal likelihood finally can be written as

$$\widehat{m}^{(LC)}(\boldsymbol{y}|g,\boldsymbol{\gamma}) \approx f(\boldsymbol{y}|\widehat{a},\widehat{\boldsymbol{\beta}}_{\gamma},\boldsymbol{\phi},\boldsymbol{\gamma})[1+nv\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})]^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\frac{\widehat{a}^2\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})}{1+nv\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})}\right)\right\}$$
$$(1+g)^{-\frac{p_\gamma}{2}} \exp\left\{-\frac{Q_\gamma}{2(g+1)}\right\}$$

which is identical to (3.20).

B.4 Hyper-g-prior Laplace Approximation of Li and Clyde

The computation of Bayes factor (3.24) under hyper-g-prior is derived by

$$\widehat{BF}^{(LC)}_{[\gamma:\gamma_0]} \approx \frac{f(\boldsymbol{y}|\widehat{a},\widehat{\boldsymbol{\beta}}_{\gamma},\boldsymbol{\phi},\boldsymbol{\gamma})}{f(\boldsymbol{y}|\widehat{a},\boldsymbol{\phi},\boldsymbol{\gamma}_0)}\left[\frac{\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})}{\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma}_0)}\right]^{-\frac{1}{2}}$$
$$\frac{\alpha-2}{2}\int_0^\infty (1+g)^{-\frac{p_\gamma}{2}} \exp\left\{-\frac{Q_\gamma}{2(g+1)}\right\}(1+g)^{-\frac{\alpha}{2}}dg,$$

where in the above expression we can use the transformation $u=\frac{1}{g+1}$ in order to handle calculations in an easier way. The transformed prior can be found by the transform theorem of distributions as the following

$$\begin{aligned} u = \frac{1}{g+1} &\Leftrightarrow g = \frac{1-u}{u}, \\ \left|\frac{dg}{du}\right| &= \frac{1}{u^2}, \\ g > 0 \Leftrightarrow \frac{1-u}{u} &> 0 \Leftrightarrow u \in (0,1), \\ \pi(u) &= \frac{\alpha-2}{2}u^{\frac{\alpha}{2}-2}, \end{aligned} \tag{B.3}$$

where from (B.3) we notice that the transformed variable $u \sim Beta\left(1,\frac{\alpha}{2}-1\right)$ which coincides with the prior distribution of the shrinkage factor $u=\frac{g}{g+1}$. Based on these calculations, the Bayes factor (3.24) of hyper-g-prior is modified in terms of u as

$$\widehat{BF}^{(LC)}_{[\gamma:\gamma_0]} \approx \frac{f(\boldsymbol{y}|\widehat{a},\widehat{\boldsymbol{\beta}}_\gamma,\boldsymbol{\phi},\boldsymbol{\gamma})}{f(\boldsymbol{y}|\widehat{a},\boldsymbol{\phi},\boldsymbol{\gamma}_0)}\left[\frac{\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})}{\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma}_0)}\right]^{-\frac{1}{2}}\frac{\alpha-2}{2}\int\limits_0^1 u^{\frac{p_\gamma}{2}+\frac{\alpha}{2}-2}\exp\left\{-u\frac{Q_\gamma}{2}\right\}du,$$

the above integral although it appears difficult to manipulate, it belongs to the general family of distributions which are called confluent hypergeometric distributions; see for more Gordy (1998). So, taking in consideration the result of the confluent hypergeometric distribution, the Bayes factor is reformulated as

$$\begin{aligned}\widehat{BF}^{(LC)}_{[\gamma:\gamma_0]} \approx & \frac{\alpha-2}{2}\frac{f(\boldsymbol{y}|\widehat{a},\widehat{\boldsymbol{\beta}}_\gamma,\boldsymbol{\phi},\boldsymbol{\gamma})}{f(\boldsymbol{y}|\widehat{a},\boldsymbol{\phi},\boldsymbol{\gamma}_0)}\left[\frac{\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})}{\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma}_0)}\right]^{-\frac{1}{2}}\\ & B\left(\frac{p_\gamma+\alpha}{2}-1,1\right){}_1F_1\left(\frac{p_\gamma+\alpha}{2}-1,\frac{p_\gamma+\alpha}{2},-\frac{Q_\gamma}{2}\right),\end{aligned}$$

where the last step is identical to (3.25) and verified due to the fact that the term $\int_0^1 u^{\frac{p_\gamma}{2}+\frac{\alpha}{2}-2}\exp\left\{-u\frac{Q_\gamma}{2}\right\}du$ can be recognized as the normalizing constant of the posterior distribution
$CH\left(\frac{p_\gamma+\alpha}{2}-1,\frac{p_\gamma+\alpha}{2},-\frac{Q_\gamma}{2}\right)$.

B.5 Confluent Hypergeometric Laplace Approximation of Li and Clyde

After accounting for the hyper-prior (3.26), we can compute Bayes factor (3.26)

$$\begin{aligned}\widehat{BF}^{(LC)}_{[\gamma:\gamma_0]} \approx & \frac{f(\boldsymbol{y}|\widehat{a},\widehat{\boldsymbol{\beta}}_\gamma,\boldsymbol{\phi},\boldsymbol{\gamma})}{f(\boldsymbol{y}|\widehat{a},\boldsymbol{\phi},\boldsymbol{\gamma}_0)}\left[\frac{\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})}{\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma}_0)}\right]^{-\frac{1}{2}}\\ & \frac{1}{B(\frac{e}{2},\frac{d}{2})\,{}_1F_1(\frac{e}{2},\frac{e+d}{2},\frac{r}{2})}\int\limits_0^\infty g^{\frac{e}{2}-1}(1+g)^{-\frac{p_\gamma+e+d}{2}}\exp\left\{-\frac{1}{2}\left(\frac{Q_\gamma-rg}{g+1}\right)\right\}dg,\end{aligned}$$

where the transformation $u=\frac{1}{g+1}$ accommodates a more appropriate form of the above integral through the confluent hypergeometric distribution. The transformed prior can be found by the transform theorem of distributions as the following

$$
\begin{aligned}
u = \frac{1}{g+1} &\Leftrightarrow g = \frac{1-u}{u},\\
\left|\frac{dg}{du}\right| &= \frac{1}{u^2},\\
g > 0 &\Leftrightarrow \frac{1-u}{u} > 0 \Leftrightarrow u \in (0,1),\\
\pi(u) &= \frac{u^{\frac{d}{2}-1}(1+g)^{\frac{e}{2}-1}\exp\left\{\frac{r}{2}(1-u)\right\}}{B(\frac{e}{2},\frac{d}{2})\,{}_1F_1(\frac{e}{2},\frac{e+d}{2},\frac{r}{2})},
\end{aligned}
\tag{B.4}
$$

where from (B.4) we notice that the transformed variable $u \sim CH\left(\frac{d}{2},\frac{e}{2},\frac{r}{2}\right)$; see Li and Clyde (2013). The Bayes factor can be rewritten in terms of u as

$$
\widehat{BF}^{(LC)}_{[\gamma:\gamma_0]} \approx \frac{f(\boldsymbol{y}|\widehat{a},\widehat{\boldsymbol{\beta}}_{\gamma},\boldsymbol{\phi},\boldsymbol{\gamma})}{f(\boldsymbol{y}|\widehat{a},\boldsymbol{\phi},\boldsymbol{\gamma}_0)}\left[\frac{\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})}{\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma}_0)}\right]^{-\frac{1}{2}}
$$

$$
\frac{1}{B(\frac{e}{2},\frac{d}{2})\,{}_1F_1(\frac{e}{2},\frac{e+d}{2},\frac{r}{2})}\int\limits_0^1 u^{\frac{d+p_\gamma}{2}-1}u^{\frac{e+d+p_\gamma}{2}-1}\exp\left\{-\frac{1}{2}\left[u(Q_\gamma+r)-r\right]\right\}du
$$

where in the above appears the normalizing constant,
$\int_0^1 u^{\frac{d+p_\gamma}{2}-1}u^{\frac{e+d+p_\gamma}{2}-1}\exp\left\{-\frac{1}{2}\left[u(Q_\gamma+r)-r\right]\right\}du$
$\propto \int_0^1 u^{\frac{d+p_\gamma}{2}-1}u^{\frac{e+d+p_\gamma}{2}-1}\exp\left\{-\frac{1}{2}\left[u(Q_\gamma+r)\right]\right\}du$, of $CH\left(\frac{d+p_\gamma}{2},\frac{e+d+p_\gamma}{2},\frac{r+Q_\gamma}{2}\right)$. Therefore taking in consideration the result of confluent hypergeometric distribution, the Bayes factor is modified as the following

$$
\widehat{BF}^{(LC)}_{[\gamma:\gamma_0]} \approx \frac{f(\boldsymbol{y}|\widehat{a},\widehat{\boldsymbol{\beta}}_{\gamma},\boldsymbol{\phi},\boldsymbol{\gamma})}{f(\boldsymbol{y}|\widehat{a},\boldsymbol{\phi},\boldsymbol{\gamma}_0)}\left[\frac{\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma})}{\mathcal{I}^{(LC)}(\widehat{a}|\boldsymbol{\gamma}_0)}\right]^{-\frac{1}{2}}\frac{B\left(\frac{e+d+p_\gamma}{2},\frac{d+p_\gamma}{2}\right)\,{}_1F_1\left(\frac{d+p_\gamma}{2},\frac{e+d+p_\gamma}{2},\frac{r+Q_\gamma}{2}\right)}{B(\frac{e}{2},\frac{d}{2})\,{}_1F_1(\frac{e}{2},\frac{e+d}{2},\frac{r}{2})}.
$$

which is the same as (3.27).

Appendix C

Bayesian Variable Selection in Multinomial Logistic Regression

C.1 Proof of SSVS in Augmented Multinomial Logistic Setup

Consider again the joint posterior (4.25) for fixed q and notice that the prior for $\boldsymbol{\beta}$ may be written as follows in terms of each regression coefficients specific class $\boldsymbol{\beta}_q$ with respect to the rest given the baseline class q^* as follows

$$
\begin{aligned}
&\pi^{SSVS}(\boldsymbol{\beta}|Q,\boldsymbol{\delta},g,\boldsymbol{\gamma}) \propto \exp\left(-\frac{\boldsymbol{\beta}_q^T\boldsymbol{D}_q^{-1}\mathcal{I}^{(BH)}(\mathbf{0}_{p_q},\mathbf{0}_{p_q})\boldsymbol{D}_q^{-1}\boldsymbol{\beta}_q}{2gQ^2}\right)\\
&\exp\left(-\frac{\sum_{q\neq q'}\boldsymbol{\beta}_{q'}^T\boldsymbol{D}_{q'}^{-1}\mathcal{I}^{(BH)}(\mathbf{0}_{p_{q'}},\mathbf{0}_{p_{q'}})\boldsymbol{D}_{q'}^{-1}\boldsymbol{\beta}_{q'}}{2gQ^2}\right)\\
&\exp\left(-\frac{-2\sum_{q\neq q'}\boldsymbol{\beta}_q^T\boldsymbol{D}_q^{-1}\mathcal{I}^{(BH)}(\mathbf{0}_{p_q},\mathbf{0}_{p_{q'}})\boldsymbol{D}_{q'}^{-1}\boldsymbol{\beta}_{q'}}{2gQ^2}\right)\\
&=\exp\left(-\frac{\boldsymbol{\beta}_q^T\boldsymbol{D}_q^{-1}\mathcal{I}^{(BH)}(\mathbf{0}_{p_q},\mathbf{0}_{p_q})\boldsymbol{D}_q^{-1}\boldsymbol{\beta}_q-2\sum_{q\neq q'}\boldsymbol{\beta}_q^T\boldsymbol{D}_q^{-1}\mathcal{I}^{(BH)}(\mathbf{0}_{p_q},\mathbf{0}_{p_{q'}})\boldsymbol{D}_{q'}^{-1}\boldsymbol{\beta}_{q'}}{2gQ^2}\right)\\
&\exp\left(-\frac{\sum_{q\neq q'}\boldsymbol{\beta}_{q'}^T\boldsymbol{D}_{q'}^{-1}\mathcal{I}^{(BH)}(\mathbf{0}_{p_{q'}},\mathbf{0}_{p_{q'}})\boldsymbol{D}_{q'}^{-1}\boldsymbol{\beta}_{q'}}{2gQ^2}\right) \qquad \text{(C.1)}\\
&=\widetilde{\pi}^{SSVS}(\boldsymbol{\beta}_q|\boldsymbol{\beta}_{-q},Q,\boldsymbol{\delta},g,\boldsymbol{\gamma}_{\boldsymbol{q}},\boldsymbol{\gamma}_{-q})\prod_{q\neq q'}\pi(\boldsymbol{\beta}_{q'}|Q,\boldsymbol{\delta},g,\boldsymbol{\gamma}_{q'}), \qquad \text{(C.2)}
\end{aligned}
$$

where $\widetilde{\pi}^{SSVS}(\boldsymbol{\beta}_q|.)$ is defined by (4.26) and
$\boldsymbol{\beta}_{q'}|Q,\boldsymbol{\delta},g,\boldsymbol{\gamma}_{q'}\sim N_{p_{q'}}(\mathbf{0}_{p'_q},Q^2g\boldsymbol{D}_{q'}\mathcal{I}^{(BH)}(\mathbf{0}_{p_{q'}},\mathbf{0}_{p_{q'}})^{-1}\boldsymbol{D}_{q'})$ that allows to write the joint

posterior (4.25) as the following

$$\begin{aligned}\widetilde{\pi}^{SSVS}(a_q,\boldsymbol{\beta}_q,\boldsymbol{\gamma}_q,\boldsymbol{a}_{-q},\boldsymbol{\beta}_{-q},g,\boldsymbol{\gamma}_{-q}|\boldsymbol{y}_j,\boldsymbol{\omega}_q) &\propto \prod_{i=1}^{n} f(z_{i,q}|a_q,\boldsymbol{\beta}_q,\omega_{i,q})\pi(\omega_{i,q}|b,0)\pi^{(BH)}(a_q)\\ &\widetilde{\pi}^{SSVS}(\boldsymbol{\beta}_q|\boldsymbol{\beta}_{-q},Q,\boldsymbol{\delta},g,\boldsymbol{\gamma}_q,\boldsymbol{\gamma}_{-q})\pi(\boldsymbol{\beta}_{q'}|Q,\boldsymbol{\delta},g,\boldsymbol{\gamma}_{q'})\pi(g)\pi(\boldsymbol{\gamma}).\end{aligned} \tag{C.3}$$

By this way, we aim to apply the Gibbs sampler by starting from the full conditional of a_q as the following does

$$\begin{aligned}&\widetilde{\pi}^{SSVS}(a_q,\boldsymbol{\beta}_q,\boldsymbol{\gamma}_q,\boldsymbol{a}_{-q},\boldsymbol{\beta}_{-q},g,\boldsymbol{\gamma}_{-q}|\boldsymbol{y}_q,\boldsymbol{\omega}_q) \propto \prod_{i=1}^{n} f(z_{i,q}|a_q,\boldsymbol{\beta}_q,\omega_{i,q})\pi(\omega_{i,q}|b,0)\pi^{(BH)}(a_q)\\ &\widetilde{\pi}^{SSVS}(\boldsymbol{\beta}_q|\boldsymbol{\beta}_{-q}Q,\boldsymbol{\delta},g,\boldsymbol{\gamma_q},\boldsymbol{\gamma}_{-q})\pi(\boldsymbol{\beta}_{q'}|Q,\boldsymbol{\delta},g,\boldsymbol{\gamma}_{q'})\pi(g)\pi(\boldsymbol{\gamma})\\ \propto&\prod_{i=1}^{n} f(z_{i,q}|a_q,\boldsymbol{\beta}_q,\omega_{i,q})\pi^{(BH)}(a_q)\\ \propto& a_q^2\sum_{i=1}^{n}\omega_{i,q} + 2a_q\mathbf{1}_n^T\boldsymbol{\Omega}_q\boldsymbol{X}\boldsymbol{\beta}_q - 2a_q\mathbf{1}_n^T\boldsymbol{\Omega}_q\boldsymbol{C}_q - 2a_q\mathbf{1}_n^T\boldsymbol{\Omega}_q\boldsymbol{z}_q\\ &= \pi(a_q|\boldsymbol{a}_{-q},\boldsymbol{\beta}_q,\boldsymbol{\beta}_{-q}\boldsymbol{\omega}_q,\boldsymbol{y}_q),\end{aligned} \tag{C.4}$$

where $\pi(a_q|.)$ is found in the implementation **Step 4:** of augmented SSVS. Next, we are interested in retrieving the full conditional of $\boldsymbol{\beta}_q$ based on the joint posterior (C.3) by writting the joint posterior as follows

$$\begin{aligned}&\widetilde{\pi}^{SSVS}(a_q,\boldsymbol{\beta}_q,\boldsymbol{\gamma}_q,\boldsymbol{a}_{-q},\boldsymbol{\beta}_{-q},g,\boldsymbol{\gamma}_{-q}|\boldsymbol{y}_q,\boldsymbol{\omega}_q) \propto \prod_{i=1}^{n} f(z_{i,q}|a_q,\boldsymbol{\beta}_q,\omega_{i,q})\pi(\omega_{i,q}|b,0)\pi^{(BH)}(a_q)\\ &\widetilde{\pi}^{SSVS}(\boldsymbol{\beta}_q|\boldsymbol{\beta}_{-q},Q,\boldsymbol{\delta},g,\boldsymbol{\gamma_q},\boldsymbol{\gamma}_{-q})\pi(\boldsymbol{\beta}_{q'}|Q,\boldsymbol{\delta},g,\boldsymbol{\gamma}_{q'})\pi(g)\pi(\boldsymbol{\gamma})\\ \propto&\prod_{i=1}^{n} f(z_{i,q}|a_q,\boldsymbol{\beta}_q,\omega_{i,q})\widetilde{\pi}^{SSVS}(\boldsymbol{\beta}_q|\boldsymbol{\beta}_{-q},Q,\boldsymbol{\delta},g,\boldsymbol{\gamma_q},\boldsymbol{\gamma}_{-q})\\ \propto&\exp\left(-\frac{1}{2g}\boldsymbol{\beta}_q^T\left(g\boldsymbol{X}^T\boldsymbol{\Omega}_q\boldsymbol{X}+\boldsymbol{D}_q^{-1}\mathcal{I}^{(BH)}(\mathbf{0}_{p_q},\mathbf{0}_{p_q})\boldsymbol{D}_q^{-1}/Q^2\right)\boldsymbol{\beta}_q^T\right)\\ &\exp\left(-\frac{1}{2g}\left(-2\boldsymbol{\beta}_q^T\left(g\left[\boldsymbol{X}^T\boldsymbol{\Omega}_q\boldsymbol{C}_q+\boldsymbol{X}^T\boldsymbol{\Omega}_q\boldsymbol{z}_q-a_q\boldsymbol{X}^T\boldsymbol{\Omega}_q\mathbf{1}_n\right]\right)\right)\right)\\ &\exp\left(-\frac{1}{2}\left(-2\boldsymbol{\beta}_q^T\left(\sum_{q\neq q'}\boldsymbol{D}_q^{-1}\mathcal{I}^{(BH)}(\mathbf{0}_{p_q},\mathbf{0}_{p_{q'}})\boldsymbol{D}_{q'}^{-1}\boldsymbol{\beta}_{q'}/Q^2\right)\right)\right) \qquad (\text{C.5})\\ &= \pi(\boldsymbol{\beta}_q|\boldsymbol{\beta}_{-q},a_q,\boldsymbol{a}_{-q},g,\boldsymbol{\delta},\boldsymbol{\gamma}_q,\boldsymbol{\gamma}_{-q},\boldsymbol{y}_q,\boldsymbol{\omega}_q),\end{aligned} \tag{C.6}$$

where $\pi(\boldsymbol{\beta}_q|.)$ is found in the implementation **Step 3:** of augmented SSVS. The next step it to obtain the full conditional of $\boldsymbol{\gamma}_q$ expressing the joint prior (C.3) as follows

$$\widetilde{\pi}^{SSVS}(a_q, \boldsymbol{\beta}_q, \boldsymbol{\gamma}_q, \boldsymbol{a}_{-q}, \boldsymbol{\beta}_{-q}, g, \boldsymbol{\gamma}_{-q}|\boldsymbol{y}_q, \boldsymbol{\omega}_q) \propto \prod_{i=1}^{n} f(z_{i,q}|a_q, \boldsymbol{\beta}_q, \omega_{i,q})\pi(\omega_{i,q}|b, 0)\pi^{(BH)}(a_q)$$
$$\widetilde{\pi}^{SSVS}(\boldsymbol{\beta}_q|\boldsymbol{\beta}_{-q}, Q, \boldsymbol{\delta}, g, \boldsymbol{\gamma_q}, \boldsymbol{\gamma}_{-q})\pi(\boldsymbol{\beta}_{q'}|Q, \boldsymbol{\delta}, g, \boldsymbol{\gamma}_{q'})\pi(g)\pi(\boldsymbol{\gamma})$$
$$= \prod_{i=1}^{n} f(z_{i,q}|a_q, \boldsymbol{\beta}_q, \omega_{i,q})\pi(\omega_{i,q}|b, 0)\pi^{(BH)}(a_q)$$
$$\widetilde{\pi}^{SSVS}(\boldsymbol{\beta}_q|\boldsymbol{\beta}_{-q}, Q, \boldsymbol{\delta}, g, \boldsymbol{\gamma_q}, \boldsymbol{\gamma}_{-q})\pi(\boldsymbol{\beta}_{q'}|Q, g, \boldsymbol{\gamma}_{q'})\pi(g)\prod_{q=1}^{Q-1}\pi(\boldsymbol{\gamma})$$
$$\propto \widetilde{\pi}^{SSVS}(\boldsymbol{\beta}_q|Q, \boldsymbol{\beta}_{-q}, g, \boldsymbol{\gamma_q}, \boldsymbol{\gamma}_{-q})\pi(\boldsymbol{\gamma}_q) \tag{C.7}$$
$$= \pi(\boldsymbol{\gamma}_q|\boldsymbol{\beta}_q, \boldsymbol{\beta}_{-q}, g, \boldsymbol{\gamma}_{-q}), \tag{C.8}$$

where $\pi(\boldsymbol{\gamma}_q|.)$ is obtained in the implementation **Step 2:** of augmented SSVS. Afterwards, the full conditional for g is obtained based on the joint posterior (C.1)

$$\widetilde{\pi}^{SSVS}(a_q, \boldsymbol{\beta}_q, \boldsymbol{\gamma}_q, \boldsymbol{a}_{-q}, \boldsymbol{\beta}_{-q}, g, \boldsymbol{\gamma}_{-q}|\boldsymbol{y}_q, \boldsymbol{\omega}_q) \propto$$
$$\exp\left(-\frac{\boldsymbol{\beta}_q^T\boldsymbol{D}_q^{-1}\mathcal{I}^{(BH)}(\boldsymbol{0}_{p_q}, \boldsymbol{0}_{p_q})\boldsymbol{D}_q^{-1}\boldsymbol{\beta}_q - 2\sum_{q\neq q'}\boldsymbol{\beta}_q^T\boldsymbol{D}_q^{-1}\mathcal{I}^{(BH)}(\boldsymbol{0}_{p_q}, \boldsymbol{0}_{p_{q'}})\boldsymbol{D}_{q'}^{-1}\boldsymbol{\beta}_{q'}}{2gQ^2}\right)$$
$$\exp\left(-\frac{\sum_{q\neq q'}\boldsymbol{\beta}_{q'}^T\boldsymbol{D}_{q'}^{-1}\mathcal{I}^{(BH)}(\boldsymbol{0}_{p_{q'}}, \boldsymbol{0}_{p_{q'}})\boldsymbol{D}_{q'}^{-1}\boldsymbol{\beta}_{q'}}{2gQ^2}\right)$$
$$\propto g^{-\frac{(Q-1)p_q}{2}}$$
$$\exp\left(-\frac{\boldsymbol{\beta}_q^T\boldsymbol{D}_q^{-1}\mathcal{I}^{(BH)}(\boldsymbol{0}_{p_q}, \boldsymbol{0}_{p_q})\boldsymbol{D}_q^{-1}\boldsymbol{\beta}_q - 2\sum_{q\neq q'}\boldsymbol{\beta}_q^T\boldsymbol{D}_q^{-1}\mathcal{I}^{(BH)}(\boldsymbol{0}_{p_q}, \boldsymbol{0}_{p_{q'}})\boldsymbol{D}_{q'}^{-1}\boldsymbol{\beta}_{q'}}{2gQ^2}\right)\pi(g),$$

where if Zellner-Siow is adopted for g, then the full conditional of g is found as

$$\widetilde{\pi}^{SSVS}(a_q, \boldsymbol{\beta}_q, \boldsymbol{\gamma}_q, \boldsymbol{a}_{-q}, \boldsymbol{\beta}_{-q}, g, \boldsymbol{\gamma}_{-q}|\boldsymbol{y}_q, \boldsymbol{\omega}_q) \propto g^{-\frac{(Q-1)p_q}{2}}$$
$$\exp\left(-\frac{\boldsymbol{\beta}_q^T\boldsymbol{D}_q^{-1}\mathcal{I}^{(BH)}(\boldsymbol{0}_{p_q}, \boldsymbol{0}_{p_q})\boldsymbol{D}_q^{-1}\boldsymbol{\beta}_q - 2\sum_{q\neq q'}\boldsymbol{\beta}_q^T\boldsymbol{D}_q^{-1}\mathcal{I}^{(BH)}(\boldsymbol{0}_{p_q}, \boldsymbol{0}_{p_{q'}})\boldsymbol{D}_{q'}^{-1}\boldsymbol{\beta}_{q'}}{2gQ^2}\right)\pi^{ZS}(g)$$
$$\propto \exp\left(-\frac{\boldsymbol{\beta}_q^T\boldsymbol{D}_q^{-1}\mathcal{I}^{(BH)}(\boldsymbol{0}_{p_q}, \boldsymbol{0}_{p_q})\boldsymbol{D}_q^{-1}\boldsymbol{\beta}_q - 2\sum_{q\neq q'}\boldsymbol{\beta}_q^T\boldsymbol{D}_q^{-1}\mathcal{I}^{(BH)}(\boldsymbol{0}_{p_q}, \boldsymbol{0}_{p_{q'}})\boldsymbol{D}_{q'}^{-1}\boldsymbol{\beta}_{q'}}{2gQ^2}\right)$$
$$g^{(-\frac{(Q-1)p_q}{2}-\frac{3}{2})}\exp\left(-\frac{n}{2}g\right) \tag{C.9}$$
$$= \pi^{ZS}(g|Q, \boldsymbol{\beta}_q, \boldsymbol{\beta}_{-q}, \boldsymbol{\gamma}_q, \boldsymbol{\gamma}_{-q}, \boldsymbol{y}), \tag{C.10}$$

where $\pi^{ZS}(g|.)$ is obtained in the implementation **Step 7: A.** of augmented SSVS, otherwise for hyper-g the full conditional remains in an intractable form suggesting the use of Metropolis-Hastings

$$
\begin{aligned}
&\widetilde{\pi}^{SSVS}(a_q, \boldsymbol{\beta}_q, \boldsymbol{\gamma}_q, \boldsymbol{a}_{-q}, \boldsymbol{\beta}_{-q}, g, \boldsymbol{\gamma}_{-q} | \boldsymbol{y}_q, \boldsymbol{\omega}_q) \propto g^{-\frac{(Q-1)p_q}{2}} \\
&\exp\left(-\frac{\boldsymbol{\beta}_q^T \boldsymbol{D}_q^{-1} \mathcal{I}^{(BH)}(\mathbf{0}_{p_q}, \mathbf{0}_{p_q}) \boldsymbol{D}_q^{-1} \boldsymbol{\beta}_q - 2\sum_{q\neq q'} \boldsymbol{\beta}_q^T \boldsymbol{D}_q^{-1} \mathcal{I}^{(BH)}(\mathbf{0}_{p_q}, \mathbf{0}_{p_{q'}}) \boldsymbol{D}_{q'}^{-1} \boldsymbol{\beta}_{q'}}{2gQ^2}\right) \pi^{hy}(g) \\
\propto &\exp\left(-\frac{\boldsymbol{\beta}_q^T \boldsymbol{D}_q^{-1} \mathcal{I}^{(BH)}(\mathbf{0}_{p_q}, \mathbf{0}_{p_q}) \boldsymbol{D}_q^{-1} \boldsymbol{\beta}_q - 2\sum_{q\neq q'} \boldsymbol{\beta}_q^T \boldsymbol{D}_q^{-1} \mathcal{I}^{(BH)}(\mathbf{0}_{p_q}, \mathbf{0}_{p_{q'}}) \boldsymbol{D}_{q'}^{-1} \boldsymbol{\beta}_{q'}}{2gQ^2}\right) \\
&g^{-\frac{(Q-1)p_q}{2}} (1+g)^{-\frac{\alpha}{2}} \\
&= \pi^{hy}(g | Q, \boldsymbol{\beta}_q, \boldsymbol{\beta}_{-q}, \boldsymbol{\gamma}_q, \boldsymbol{\gamma}_{-q}, \boldsymbol{y}),
\end{aligned}
\tag{C.11}
$$

where $\pi^{hy}(g|.)$ is obtained in the implementation **Step 7: B.** of augmented SSVS.

C.2 Proof of GVS in Augmented Multinomial Logistic Setup

Consider again the joint posterior (4.30) for fixed q and notice that the prior (4.16) for $\boldsymbol{\beta}$ may be written as follows in terms of each regression coefficients specific class $\boldsymbol{\beta}_q$ with respect to the rest given the baseline class q^* as follows

$$
\begin{aligned}
\pi^{GVS}(\boldsymbol{\beta} | Q, \boldsymbol{\delta}, g, \boldsymbol{\gamma}) \propto &\exp\left(-\frac{(\boldsymbol{\beta}_q - \boldsymbol{\mu}_q)^T \left(\frac{\boldsymbol{\Gamma}_q \mathcal{I}^{(BH)}(\mathbf{0}_{p_q}, \mathbf{0}_{p_q}) \boldsymbol{\Gamma}_q}{gQ^2} + \widetilde{\boldsymbol{d}_q}\right)(\boldsymbol{\beta}_q - \boldsymbol{\mu}_q)}{2}\right) \\
&\exp\left(-\frac{-2\sum_{q\neq q'} (\boldsymbol{\beta}_q - \boldsymbol{\mu}_q)^T \boldsymbol{\Gamma}_q \mathcal{I}^{(BH)}(\mathbf{0}_{p_q}, \mathbf{0}_{p_{q'}}) \boldsymbol{\Gamma}_{q'} (\boldsymbol{\beta}_{q'} - \boldsymbol{\mu}_{q'})}{2gQ^2}\right) \\
&\exp\left(-\frac{\sum_{q\neq q'} (\boldsymbol{\beta}_{q'} - \boldsymbol{\mu}_{q'})^T \left(\frac{\boldsymbol{\Gamma}_{q'} \mathcal{I}^{(BH)}(\mathbf{0}_{p_{q'}}, \mathbf{0}_{p_{q'}}) \boldsymbol{\Gamma}_{q'}}{gQ^2} + \tilde{\boldsymbol{d}}_{q'}\right)(\boldsymbol{\beta}_{q'} - \boldsymbol{\mu}_{q'})}{2}\right)
\end{aligned}
\tag{C.12}
$$

$$
\begin{aligned}
&\pi^{GVS}(\boldsymbol{\beta}|Q,\boldsymbol{\delta},g,\boldsymbol{\gamma}) \propto \exp\left(-\frac{\boldsymbol{\beta}_q^T\boldsymbol{\Gamma}_q\mathcal{I}^{(BH)}(\mathbf{0}_{p_q},\mathbf{0}_{p_q}\boldsymbol{\beta}_q)\boldsymbol{\Gamma}_q\boldsymbol{\beta}}{2gQ^2}\right)\exp\left(-\frac{-2\boldsymbol{\beta}_q^T\boldsymbol{\Gamma}_q\mathcal{I}^{(BH)}(\mathbf{0}_{p_q},\mathbf{0}_{p_q})\boldsymbol{\mu}_q}{2gQ^2}\right)\\
&\exp\left(-\frac{-2\sum_{q\neq q'}\boldsymbol{\beta}_q^T\boldsymbol{\Gamma}_q\mathcal{I}^{(BH)}(\mathbf{0}_{p_q},\mathbf{0}_{p_{q'}})\boldsymbol{\Gamma}_{q'}\boldsymbol{\beta}_{q'}}{2gQ^2}\right)\exp\left(-\frac{-2\sum_{q\neq q'}\boldsymbol{\beta}_q^T\boldsymbol{\Gamma}_q\mathcal{I}^{(BH)}(\mathbf{0}_{p_q},\mathbf{0}_{p_{q'}})\boldsymbol{\Gamma}_{q'}\boldsymbol{\mu}_{q'}}{2gQ^2}\right)\\
&\exp\left(-\frac{1}{2}\left[\boldsymbol{\beta}_q^T\widetilde{\boldsymbol{d}}_q\boldsymbol{\beta}_q-2\boldsymbol{\beta}_q^T\widetilde{\boldsymbol{d}}_q\boldsymbol{\mu}_q\right]\right)\exp\left(-\frac{1}{2}\left[\sum_{q\neq q'}\left(\boldsymbol{\beta}_{q'}^T\widetilde{\boldsymbol{d}}_{q'}\boldsymbol{\beta}_{q'}-2\boldsymbol{\beta}_{q'}^T\widetilde{\boldsymbol{d}}_{q'}\boldsymbol{\mu}_{q'}\right)\right]\right)\\
&=\widetilde{\pi}^{GVS}(\boldsymbol{\beta}_q|\boldsymbol{\beta}_{-q},Q,\boldsymbol{\delta},g,\boldsymbol{\gamma}_q,\boldsymbol{\gamma}_{-q})\prod_{q\neq q'}\pi(\boldsymbol{\beta}_{q'}|Q,\boldsymbol{\delta},g,\boldsymbol{\gamma}_{q'}), \qquad \text{(C.13)}
\end{aligned}
$$

where $\widetilde{\pi}^{GVS}(\boldsymbol{\beta}_q|.$ is defined by (4.31) and $\boldsymbol{\beta}_{q'}|Q,\boldsymbol{\delta},g,\boldsymbol{\gamma}_{q'}\sim N_{p_{q'}}(\boldsymbol{\mu}_{q'},\widetilde{\boldsymbol{d}}_{q'}^{-1})$ permits to express the joint posterior (4.30) as the following

$$
\begin{aligned}
&\widetilde{\pi}^{GVS}(a_q,\boldsymbol{\beta}_q,\boldsymbol{\gamma}_q,\boldsymbol{a}_{-q},\boldsymbol{\beta}_{-q},g,\boldsymbol{\gamma}_{-q}|\boldsymbol{y}_q,\boldsymbol{\omega}_q)\propto\prod_{i=1}^{n}f(z_{i,q}|a_q,\boldsymbol{\beta}_q,\omega_{i,q},\boldsymbol{\gamma}_q)\pi(\omega_{i,q}|b,0)\pi^{(BH)}(a_q)\\
&\widetilde{\pi}^{GVS}(\boldsymbol{\beta}_q|\boldsymbol{\beta}_{-q},Q,\boldsymbol{\delta},g,\boldsymbol{\gamma_q},\boldsymbol{\gamma}_{-q})\prod_{q\neq q'}\pi(\boldsymbol{\beta}_{q'}|Q,\boldsymbol{\delta},g,\boldsymbol{\gamma}_{q'})\pi(g)\pi(\boldsymbol{\gamma}). \qquad \text{(C.14)}
\end{aligned}
$$

From the above joint posterior, we outline the Gibbs sampler basically starting from the full conditional of a_q as the following

$$
\begin{aligned}
\widetilde{\pi}^{GVS}(a_q,\boldsymbol{\beta}_q,\boldsymbol{\gamma}_q,\boldsymbol{a}_{-q},\boldsymbol{\beta}_{-q},g,\boldsymbol{\gamma}_{-q}|\boldsymbol{y}_q,\boldsymbol{\omega}_q)&\propto\prod_{i=1}^{n}f(z_{i,q}|a_q,\boldsymbol{\beta}_q,\omega_{i,q},\boldsymbol{\gamma}_q)\pi(\omega_{i,q}|b,0)\pi^{(BH)}(a_q)\\
&\quad\widetilde{\pi}^{GVS}(\boldsymbol{\beta}_q|\boldsymbol{\beta}_{-q},Q,\boldsymbol{\delta},g,\boldsymbol{\gamma_q},\boldsymbol{\gamma}_{-q})\pi(\boldsymbol{\beta}_{q'}|Q,\boldsymbol{\delta},g,\boldsymbol{\gamma}_{q'})\pi(g)\pi(\boldsymbol{\gamma})\\
&\propto\prod_{i=1}^{n}f(z_{i,q}|a_q,\boldsymbol{\beta}_q,\omega_{i,q},\boldsymbol{\gamma}_q)\pi^{(BH)}(a_q)\\
&\propto a_q^2\sum_{i=1}^{n}\omega_{i,q}+2a_q\boldsymbol{\beta}_q^T\boldsymbol{\Gamma}_q\boldsymbol{X}\boldsymbol{\Omega}_q\mathbf{1}_n-2a_q\mathbf{1}_n^T\boldsymbol{\Omega}_q\widetilde{\boldsymbol{C}}_q-2a_q\mathbf{1}_n^T\boldsymbol{\Omega}_q\boldsymbol{z}_q\\
&=\pi(a_q|\boldsymbol{a}_{-q},\boldsymbol{\beta}_q,\boldsymbol{\beta}_{-q}\boldsymbol{\omega}_q,\boldsymbol{\gamma}_q,\boldsymbol{\gamma}_{-q},\boldsymbol{y}_q), \qquad \text{(C.15)}
\end{aligned}
$$

where $\pi(a_q|.)$ is defined in in the implementation **Step 4:** of augmented GVS. The next step involves the full conditional of $\boldsymbol{\beta}_q$ which can be retrieved in closed form by the joint posterior (C.14) as follows

$$
\begin{aligned}
&\widetilde{\pi}^{GVS}(a_q, \boldsymbol{\beta}_q, \boldsymbol{\gamma}_q, \boldsymbol{a}_{-q}, \boldsymbol{\beta}_{-q}, g, \boldsymbol{\gamma}_{-q} | \boldsymbol{y}_q, \boldsymbol{\omega}_q) \propto \prod_{i=1}^{n} f(z_{i,q}|a_q, \boldsymbol{\beta}_q, \omega_{i,q}, \boldsymbol{\gamma}_q)\pi(\omega_{i,q}|b,0)\pi^{(BH)}(a_q) \\
&\widetilde{\pi}^{GVS}(\boldsymbol{\beta}_q|\boldsymbol{\beta}_{-q}, Q, \boldsymbol{\delta}, g, \boldsymbol{\gamma}_q, \boldsymbol{\gamma}_{-q}) \prod_{q \neq q'} \pi(\boldsymbol{\beta}_{q'}|Q, \boldsymbol{\delta}, g, \boldsymbol{\gamma}_{q'})\pi(g)\pi(\boldsymbol{\gamma}) \\
&\propto \prod_{i=1}^{n} f(z_{i,q}|a_q, \boldsymbol{\beta}_q, \omega_{i,q}, \boldsymbol{\gamma}_q)\widetilde{\pi}^{GVS}(\boldsymbol{\beta}_q|\boldsymbol{\beta}_{-q}, Q, \boldsymbol{\delta}, g, \boldsymbol{\gamma_q}, \boldsymbol{\gamma}_{-q}) \\
&\propto \exp\left(-\frac{1}{2}\boldsymbol{\beta}_q^T \left(\frac{\boldsymbol{\Gamma}_q \mathcal{I}^{(BH)}(\mathbf{0}_{p_q}, \mathbf{0}_{p_q})\boldsymbol{\Gamma}_q}{gQ^2} + \boldsymbol{\Gamma}_q \boldsymbol{X}^T \boldsymbol{\Omega}_q \boldsymbol{X} \boldsymbol{\Gamma}_q + \widetilde{\boldsymbol{d}}_{q'}\right) \boldsymbol{\beta}_q^T\right) \\
&\quad \exp\left(-\frac{1}{2}\left(-2\boldsymbol{\beta}_q^T \left(\left[\boldsymbol{\Gamma}_q \boldsymbol{X}^T \boldsymbol{\Omega}_q \widetilde{\boldsymbol{C}}_q + \boldsymbol{\Gamma}_q \boldsymbol{X}^T \boldsymbol{\Omega}_q \boldsymbol{z}_q - a_q \boldsymbol{\Gamma}_q \boldsymbol{X}^T \boldsymbol{\Omega}_q \mathbf{1}_n\right]\right)\right)\right) \\
&\quad \exp\left(-\frac{1}{2}\left(-2\boldsymbol{\beta}_q^T \left(\frac{\boldsymbol{\Gamma}_q \mathcal{I}^{(BH)}(\mathbf{0}_{p_q}, \mathbf{0}_{p_q})\boldsymbol{\Gamma}_q \boldsymbol{\mu}_q}{gQ^2}\right)\right)\right) \\
&\quad \exp\left(-\frac{1}{2}\left(-2\boldsymbol{\beta}_q^T \left(\frac{\sum_{q \neq q'} \left(\boldsymbol{\Gamma}_q \mathcal{I}^{(BH)}(\mathbf{0}_{p_q}, \mathbf{0}_{p_{q'}})\boldsymbol{\Gamma}_{q'} \boldsymbol{\beta}_{q'} - \boldsymbol{\Gamma}_q \mathcal{I}^{(BH)}(\mathbf{0}_{p_q}, \mathbf{0}_{p_{q'}})\boldsymbol{\Gamma}_{q'} \boldsymbol{\mu}_{q'}\right)}{gQ^2}\right)\right)\right) \\
&\quad \exp\left(-\frac{1}{2}\left(-2\boldsymbol{\beta}_q^T \widetilde{\boldsymbol{d}}_q \boldsymbol{\mu}_q\right)\right) \qquad \text{(C.16)} \\
&\quad = \pi(\boldsymbol{\beta}_q|Q, \boldsymbol{\beta}_{-q}, g, \boldsymbol{\gamma}_q, \boldsymbol{\gamma}_{-q}, \boldsymbol{y}_q, \boldsymbol{\omega}_q), \qquad \text{(C.17)}
\end{aligned}
$$

where $\pi(\boldsymbol{\beta}_q|.)$ is defined in the implementation **Step 3:** of augmented GVS.
Then proceeding with the Gibbs sampler, the full conditional of $\boldsymbol{\gamma}_q$ is extracted in closed form from the joint prior (C.14) as follows

$$
\begin{aligned}
&\widetilde{\pi}^{GVS}(a_q, \boldsymbol{\beta}_q, \boldsymbol{\gamma}_q, \boldsymbol{a}_{-q}, \boldsymbol{\beta}_{-q}, g, \boldsymbol{\gamma}_{-q} | \boldsymbol{y}_q, \boldsymbol{\omega}_q) \propto \prod_{i=1}^{n} f(z_{i,q}|a_q, \boldsymbol{\beta}_q, \omega_{i,q}, \boldsymbol{\gamma}_q)\pi(\omega_{i,q}|b,0)\pi^{(BH)}(a_q) \\
&\widetilde{\pi}^{GVS}(\boldsymbol{\beta}_q|\boldsymbol{\beta}_{-q} Q, \boldsymbol{\delta}, g, \boldsymbol{\gamma_q}, \boldsymbol{\gamma}_{-q})\pi(\boldsymbol{\beta}_{q'}|Q, \boldsymbol{\delta}, g, \boldsymbol{\gamma}_{q'})\pi(g)\pi(\boldsymbol{\gamma}) \\
&= \prod_{i=1}^{n} f(z_{i,q}|a_q, \boldsymbol{\beta}_q, \omega_{i,q}, \boldsymbol{\gamma}_q)\pi(\omega_{i,q}|b,0)\pi^{(BH)}(a_q) \\
&\widetilde{\pi}^{GVS}(\boldsymbol{\beta}_q|\boldsymbol{\beta}_{-q}, Q, \boldsymbol{\delta}, g, \boldsymbol{\gamma_q}, \boldsymbol{\gamma}_{-q})\pi(\boldsymbol{\beta}_{q'}|Q, \boldsymbol{\delta}, g, \boldsymbol{\gamma}_{q'})\pi(g) \prod_{q=1}^{Q-1} \pi(\boldsymbol{\gamma}) \\
&\propto \widetilde{\pi}^{GVS}(\boldsymbol{\beta}_q|\boldsymbol{\beta}_{-q}, Q, \boldsymbol{\delta}, g, \boldsymbol{\gamma_q}, \boldsymbol{\gamma}_{-q})\pi(\boldsymbol{\gamma}_q) \qquad \text{(C.18)} \\
&= \pi(\boldsymbol{\gamma}_q|\boldsymbol{\beta}_q, \boldsymbol{\beta}_{-q}, a_q, \boldsymbol{a}_{-q}, g, \boldsymbol{\gamma}_{-q}), \qquad \text{(C.19)}
\end{aligned}
$$

where $\pi(\boldsymbol{\gamma}_q|.)$ is defined in the implementation **Step 2:** of augmented GVS. Afterwards, the full conditional for g is obtained based on the joint posterior (C.14)

$$\begin{aligned}
&\widetilde{\pi}^{GVS}(a_q, \boldsymbol{\beta}_q, \boldsymbol{\gamma}_q, \boldsymbol{a}_{-q}, \boldsymbol{\beta}_{-q}, g, \boldsymbol{\gamma}_{-q}|\boldsymbol{y}_q, \boldsymbol{\omega}_q) \propto \prod_{i=1}^{n} f(z_{i,q}|a_q, \boldsymbol{\beta}_q, \omega_{i,q}, \boldsymbol{\gamma}_q)\pi(\omega_{i,q}|b, 0)\pi^{(BH)}(a_q)\\
&\widetilde{\pi}^{GVS}(\boldsymbol{\beta}_q|\boldsymbol{\beta}_{-q}, Q, \boldsymbol{\delta}, g, \boldsymbol{\gamma_q}, \boldsymbol{\gamma}_{-q})\pi(\boldsymbol{\beta}_{q'}|Q, \boldsymbol{\delta}, g, \boldsymbol{\gamma}_{q'})\pi(g)\pi(\boldsymbol{\gamma})\\
&\ \propto \widetilde{\pi}^{GVS}(\boldsymbol{\beta}_q|\boldsymbol{\beta}_{-q}Q, \boldsymbol{\delta}, g, \boldsymbol{\gamma_q}, \boldsymbol{\gamma}_{-q})\pi(\boldsymbol{\beta}_{q'}|Q, \boldsymbol{\delta}, g, \boldsymbol{\gamma}_{q'})\pi(g), \qquad \text{(C.20)}
\end{aligned}$$

where if a Zellner-Siow is adopted for g, then the following full conditional is retrieved in closed form as following

$$\begin{aligned}
&\widetilde{\pi}^{GVS}(a_q, \boldsymbol{\beta}_q, \boldsymbol{\gamma}_q, \boldsymbol{a}_{-q}, \boldsymbol{\beta}_{-q}, g, \boldsymbol{\gamma}_{-q}|\boldsymbol{y}_q, \boldsymbol{\omega}_q) \propto\\
&\widetilde{\pi}^{GVS}(\boldsymbol{\beta}_q|\boldsymbol{\beta}_{-q}, Q, \boldsymbol{\delta}, g, \boldsymbol{\gamma_q}, \boldsymbol{\gamma}_{-q})\pi(\boldsymbol{\beta}_{q'}|Q, \boldsymbol{\delta}, g, \boldsymbol{\gamma}_{q'})\pi^{ZS}(g)\\
\propto&\exp\left(-\frac{1}{2gQ^2}\left(\boldsymbol{\beta}_q - \boldsymbol{\mu}_q\right)^T \boldsymbol{\Gamma}_q \mathcal{I}^{(BH)}(\mathbf{0}_{p_q}, \mathbf{0}_{p_q})\boldsymbol{\Gamma}_q\left(\boldsymbol{\beta}_q - \boldsymbol{\mu}_q\right)\right)\\
&\exp\left(\frac{1}{gQ^2}\sum_{q\neq q'}\left(\boldsymbol{\beta}_q - \boldsymbol{\mu}_q\right)^T \boldsymbol{\Gamma}_q \mathcal{I}^{(BH)}(\mathbf{0}_{p_q}, \mathbf{0}_{p_{q'}})\boldsymbol{\Gamma}_{q'}\left(\boldsymbol{\beta}_{q'} - \boldsymbol{\mu}_{q'}\right)\right) \qquad \text{(C.21)}\\
&= \pi^{ZS}(g|Q, \boldsymbol{\beta}_q, \boldsymbol{\beta}_{-q}, \boldsymbol{\gamma}_q, \boldsymbol{\gamma}_{-q}, \boldsymbol{y}), \qquad \text{(C.22)}
\end{aligned}$$

where $\pi^{ZS}(g|.)$ is obtained in the implementation **Step 7: A.** of augmented GVS, otherwise for hyper-g

$$\begin{aligned}
&\widetilde{\pi}^{GVS}(a_q, \boldsymbol{\beta}_q, \boldsymbol{\gamma}_q, \boldsymbol{a}_{-q}, \boldsymbol{\beta}_{-q}, g, \boldsymbol{\gamma}_{-q}|\boldsymbol{y}_q, \boldsymbol{\omega}_q) \propto\\
&\widetilde{\pi}^{GVS}(\boldsymbol{\beta}_q|\boldsymbol{\beta}_{-q}, Q, \boldsymbol{\delta}, g, \boldsymbol{\gamma_q}, \boldsymbol{\gamma}_{-q})\pi(\boldsymbol{\beta}_{q'}|Q, \boldsymbol{\delta}, g, \boldsymbol{\gamma}_{q'})\pi^{hy}(g)\\
&\exp\left(-\frac{1}{2gQ^2}\left(\boldsymbol{\beta}_q - \boldsymbol{\mu}_q\right)^T \boldsymbol{\Gamma}_q \mathcal{I}^{(BH)}(\mathbf{0}_{p_q}, \mathbf{0}_{p_q})\boldsymbol{\Gamma}_q\left(\boldsymbol{\beta}_q - \boldsymbol{\mu}_q\right)\right)\\
&\exp\left(\frac{1}{gQ^2}\sum_{q\neq q'}\left(\boldsymbol{\beta}_q - \boldsymbol{\mu}_q\right)^T \boldsymbol{\Gamma}_q \mathcal{I}^{(BH)}(\mathbf{0}_{p_q}, \mathbf{0}_{p_{q'}})\boldsymbol{\Gamma}_{q'}\left(\boldsymbol{\beta}_{q'} - \boldsymbol{\mu}_{q'}\right)\right) \qquad \text{(C.23)}\\
&(1+g)^{-\frac{\alpha}{2}} g^{-\frac{p_{\boldsymbol{\gamma}}}{2}} \qquad \text{(C.24)}\\
&= \pi^{hy}(g|Q, \boldsymbol{\beta}_q, \boldsymbol{\beta}_{-q}, \boldsymbol{\gamma}_q, \boldsymbol{\gamma}_{-q}, \boldsymbol{y}) \qquad \text{(C.25)}
\end{aligned}$$

where $\pi^{hy}(g|Q, \boldsymbol{\beta}_q, \boldsymbol{\beta}_{-q}, \boldsymbol{\gamma}_q, \boldsymbol{\gamma}_{-q}, \boldsymbol{y})$ is obtained in the implementation **Step 7: B.** of augmented GVS.

C.3 Bayesian Variable Selection in Logistic Regression

Logistic regression models can be considered a special family of multinomial logistic regression reducing only to a binary response variable. All MCMC of Bayesian variable selection of typical and augmented logistic regression are special cases of multinomial logistic regression presented in the third chapter of this thesis and for the sake of feasibility we cannot describe in detail, since they are created similarly. In this setup, the variance-covariance and covariance structure of expected Fisher information matrix between the same or different logistic regression will be dropped especially for the covariance structure, since only the involved part of variance-covariance of the same class will be used and for $Q = 2$, we recover the logistic regression problem.

C.4 Simulated Experiments

In this section, we used two simulation examples, with independent and correlated predictors for logistic regression presented also in Hansen and Yu (2003), Li and Clyde (2013) and Chen et al. (2008) regarding the Bayesian variable selection methods using MCMC methods with mixtures of g-priors. These datasets contain $p = 5$ covariates of $n = 100$ values. In case of the independent design ($r = 0$) the covariates were obtained as independent standardized normal vectors $\boldsymbol{X_1}, \ldots, \boldsymbol{X_5}$ iid $\sim N_{100}(0, 1)$, whereas in the correlated design ($r = 0.75$) each one was drawn from standardized normal distribution with pairwise correlations given by

$$\text{corr}(X_i, X_j) = r^{|i-j|}, \;\; 1 \leq i < j \leq p. \tag{C.26}$$

Scenario	**Logistic**					
	a	β_1	β_2	β_3	β_4	β_5
null	0.1	0	0	0	0	0
sparse	0.1	0.7	0	0	0	0
medium	0.1	1.6	0.8	-1.5	0	0
full	0.1	1.75	1.5	-1.1	1.4	0.5

Table C.1 Logistic regression scenarios using independent ($r = 0$) and correlated predictors ($r = 0.75$).

Moreover, four sparse scenarios are examined within each design to describe the true generating models based on Table (C.1), where the coefficients of this Table were smaller than of Hansen and Yu (2003) set such that the odds ratios were 2, 2.5 and 3.5 for sparse, medium and full scenarios respectively based on Fouskakis et al. (2018).

Prior Inputs-Initial Values	
Parameter	**Value**
τ	0.02
c	50
$\boldsymbol{\gamma}^{(0)}$	$(1,1,1,1,1)^T$
$\boldsymbol{\beta}^{(0)}$	$\widehat{\boldsymbol{\beta}}$
$a^{(0)}$	$\hat{a}$
$g^{(0)}$	n
$\boldsymbol{\omega}^{(0)}$	$(1,\ldots,1)$

Table C.2 Prior-inputs and initial values

Our goal is to assess the performance of MCMC for Bayesian variable selection methods. In particular, we present the main body of the results using as basis the SSVS and GVS computational methods in the framework of g-priors and its mixtures adopted each time for the typical and augmented logistic regression. Moreover, we examine further also the computational efficiency based on the effective sample size and Monte Carlo standard error estimates related to the discrepancy of iterations convergence and sampling error attributed to the MCMC method. In particular, among methods using hyper-g we considered the proposed value $\alpha = 3$ by Liang et al., (2008) and a Metropolis-Hastings random walk step with tuning variance $u_g = 1$ in order to obtain convergence. Regarding the tuning of proposals of a, $\boldsymbol{\beta}$ for both SSVS and GVS, we used $t = 0.4$ and $v_a = 1$ respectively to ensure the good mixing of the chains. With respect to MCMC methods, prior inputs τ_j and c_j for $j = 1, \ldots, p$ were set on practical significance for SSVS to achieve similar results with the objective Bayesian methods and $\bar{\boldsymbol{\mu}}$, $\bar{\boldsymbol{s}}^2$ were computed from pilot run under the full model for GVS of each simulated dataset repetition. The option of prior input τ_j and c_j are such that $\tau_j = 0.02 << \tau_j c_j = 1$; see for more Table C.2. A detailed description of all Bayesian variable selection methods which are used as references in Figures and in Tables, are summarized in Table C.3.

	Acronym	Computational Method	Prior	Model
1	ssvs.hyp.typ	Stochastic Search Variable Selection for $\alpha = 3$, $\tau = 0.02$, $c = 50$, $u_a = 1$, $t = 0.4$, $u_g = 1$	Hyper-g	Typical
2	ssvs.hyp.aug	Stochastic Search Variable Selection for $\alpha = 3$, $u_g = 1$	Hyper-g	Augmented
3	gvs.hyp.typ	Gibbs Variable Selection for $\alpha = 3$, $u_a = 1$, $t = 0.4$, $u_g = 1$	Hyper-g	Typical
4	gvs.hyp.aug	Gibbs Variable Selection for $\alpha = 3$, $t = 0.4$	Hyper-g	Augmented
5	ssvs.ZS.typ	Stochastic Search Variable Selection for $\tau = 0.02$, $c = 50$, $u_a = 1$, $t = 0.4$	Zellner-Siow	Typical
6	ssvs.ZS.aug	Stochastic Search Variable Selection for $\tau = 0.02$, $c = 50$	Zellner-Siow	Augmented
7	gvs.ZS.typ	Gibbs Variable Selection for $u_a = 1$, $t = 0.4$	Zellner-Siow	Typical
8	gvs.ZS.aug	Gibbs Variable Selection for $u_a = 1$, $t = 0.4$	Zellner-Siow	Augmented
9	ssvs.g.typ	Stochastic Search Variable Selection for $\tau = 0.02$, $c = 50$, $u_a = 1$, $t = 0.4$, $g = n$	g-prior	Typical
10	ssvs.g.aug	Stochastic Search Variable Selection for $\tau = 0.02$, $c = 50$, $g = n$	g-prior	Augmented
11	gvs.hyp.typ	Gibbs Variable Selection for $u_a = 1$, $t = 0.4$, $g = n$	g-prior	Typical
12	gvs.hyp.aug	Gibbs Variable Selection for $g = n$	g-prior	Augmented

Table C.3 Acronyms of Bayesian variable selection methods with MCMC for logistic regression.

Scarce information regarding which variables to include or not in the model, favours the adoption of objective priors to each respective model, specific parameter and to the model itself. In particular, we used the joint hierarchical mixture priors (4.12), (4.19) of Bové and Held (2011) adopted in the style of logistic regression in SSVS and GVS to account for the joint prior dependences among parameters a, $\boldsymbol{\beta}$, g, $\boldsymbol{\gamma}$. All the compared approaches under Zellner-Siow prior used (2.10), whereas for hyper-g-prior was used (2.12). Moreover, since the number of covariates p is small, we use a uniform prior on model space to reflect our prior ignorance on which models to prefer. Implementing

Bayesian variable selection MCMC methods to these simulated data, model fitting was applied through a Gibbs algorithm with successive Metropolis-Hastings steps using the R programming language. We simulated Markov chains of 40000 valid values to achieve convergence for both typical and augmented logistic regression Bayesian variable selection methods. More precisely, for all methods we used the maximum likelihood estimators for parameters $\boldsymbol{\beta}$ and intercept a as initial values for each repeated dataset $\boldsymbol{\beta}^{(0)}$, $a^{(0)}$, for the vector of model indicators $\boldsymbol{\gamma}$ we used as initial values $\boldsymbol{\gamma}^{(0)}$, for g we used $g^{(0)}$ and for the methods with the latent data we used $\boldsymbol{\omega}^{(0)}$; please refer to Table (C.2).

Results based on the frequency of identifying the true data-generating model through the maximum aposteriori model for the typical and augmented logistic regression over 100 repeated simulations of each scenario and correlation design are provided in Table (C.4). Comparison of Bayesian variable selection methods with mixtures of g-priors approaches versus the rest of the methods shows the following

Scenario	**r**	**Bayesian variable selection methods**											
		ssvs.	ssvs.	gvs.	gvs.	ssvs.	ssvs.	gvs.	gvs.	ssvs.	ssvs.	gvs.	gvs.
		hyp.	hyp.	hyp.	hyp.	ZS.	ZS.	ZS.	ZS.	g.	g.	g.	g.
		typ	aug	typ	aug	typ	aug	typ	aug	typ	aug	typ	aug
null	0.00	45	40	55	46	90	**91**	90	90	87	85	85	84
	0.75	53	58	53	47	92	**93**	85	85	89	90	85	85
sparse	0.00	63	67	64	65	67	66	67	67	**71**	69	68	69
	0.75	63	61	57	57	70	69	70	71	73	**74**	73	73
medium	0.00	**80**	**80**	79	**80**	79	78	78	78	79	79	78	79
	0.75	**36**	34	31	30	32	30	27	27	31	33	26	26
full	0.00	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**
	0.75	21	**23**	17	17	20	21	14	15	18	19	12	12

Table C.4 Number of 100 simulated samples that the MAP coincides with the true generating model of Table (C.1) under various scenarios for independent and correlated covariates (row-wise largest value in bold).

i) Generally, the procedures with mixtures of g-priors perform successfully in 6 out of the 8 scenarios. The best method of identifying the true generating mechanism of the data includes one of the methods with mixtures of g-priors

ii) In the null scenario, all methods with some exceptions (ssvs.hyp.typ, ssvs.hyp.aug, gvs.hyp.aug) of mixtures g-priors trace correctly the true generating mechanism in the independent covariates design, whereas in correlated covariates, the mentioned exceptions along with the rest of the methods show an increasing tendency of the true model rate

iii) In the sparse scenario, all computational methods under the various prior choices perform well with good true model rate both in the independent and correlated covariates design, whereas in the medium scenario they only perform very well in the independent in comparison with the correlated covariates design.

iv) In the full scenario, all Bayesian variable selection methods seem to perform equally poorly in identifying the true data generating model with true model rate 3% in the independent covariates design, whereas in the correlated covariates design their performance is improved with increasing magnitude in the identification of the true model rate.

Regarding the comparison of methods with fixed g versus random, we observe that they are more robust. Moreover, Zellner-Siow priors and g-priors behave similarly in comparison with the hyper-g. This is not a surprise since both methods tend to prefer sparser models resulting in similar results. Additional information based on comparisons made between all the methods are found in Figures (C.1), (C.2) which depict the marginal posterior inclusion probabilities over the 100 simulated repetitions under the various Bayesian variable selection methods. From these results we observe that mixtures of g-priors exhibit larger posterior inclusion probabilities only for the non important covariates with respect to fixed g-priors methods. In particular, hyper-g priors suffer from the inflation of posterior inclusion probabilities towards 0.5, whereas the Zellner-Siow posterior inclusion probabilities are less inflated. This is a direct consequence of the additional stochasticity of the random g accumulated in the non certain covariates. This behaviour was expected since hyper-g priors tend to support saturated models in comparison with the Zellner-Siow priors. Moreover, Zellner-Siow priors are the only mixtures of g-priors that exhibit strong shrinkage in terms of the non important covariates. Similar conclusions hold for the posterior regression coefficients as the marginal posterior inclusion probabilities of all Bayesian variable selection methods based on Figures (C.3) and (C.4). On the other hand, additional remarks based on comparison of Bayesian variable selection methods with mixtures of g-priors approaches versus the rest of the methods between typical and augmented logistic regression models are also summarized as follows

i) Overall, methods with mixtures of g-priors with data augmentation perform satisfactorily as in 6 out of the 8 scenarios the best method of identifying the true generating mechanism of the data includes one of the methods with mixtures of g-priors with data augmentation

ii) In general, methods with data augmentation perform satisfactorily in 7 out of the 8 scenarios; the best method of identifying the true generating mechanism of the data includes one of the methods with data augmentation.

iii) Generally, there are no quite differences among Bayesian variable selection methods for typical and augmented logistic regression apart only in the case (null case for r=0.00: gvs.hyp.typ-gvs.hyp.aug, null case for r=0.75: ssvs.hyp.typ-ssvs.hyp.aug, medium case: r=0.75: gvs.g.typ-gvs.g.aug).

iv) All the methods with data augmentation logistic regression model seem to perform equally well under the various scenarios and design correlations as the typical analogues apart from the exceptions specified above

Scenario	**r**	**Bayesian variable selection methods (n=500)**											
		ssvs. hyp. typ	ssvs. hyp. aug	gvs. hyp. typ	gvs. hyp. aug	ssvs. ZS. typ	ssvs. ZS. aug	gvs. ZS. typ	gvs. ZS. aug	ssvs. g. typ	ssvs. g. aug	gvs. g. typ	gvs. g. aug
null	0.00	49	55	59	54	97	**98**	95	95	95	96	95	94
	0.75	68	69	68	67	**99**	**99**	96	96	98	98	96	96
sparse	0.00	82	82	80	80	**96**	**96**	93	93	**96**	96	93	93
	0.75	78	79	81	82	**95**	**95**	92	92	**95**	**95**	92	92
medium	0.00	91	90	89	89	**95**	**95**	92	92	**95**	**95**	94	93
	0.75	92	93	93	92	**94**	**94**	91	91	93	**94**	91	90
full	0.00	93	**94**	94	94	93	93	**94**	**94**	93	92	**94**	**94**
	0.75	68	**69**	61	60	62	60	56	56	58	58	54	52

Table C.5 Number of 100 simulated samples that the MAP coincides with the true generating model of Table (C.1) under various scenarios for independent and correlated covariates (row-wise largest value in bold).

Regarding the properties of mixtures g-priors versus fixed g-priors based on Bayesian variable selection methods of typical and augmented logistic regression models, the conclusions are the same for all the scenarios under the independent and correlated design. Since conclusions based on posterior measures were expected to be in agreement, we further investigated the computational efficiency of each method with respect to typical and augmented logistic regression within SSVS and GVS based on the effective sample size ESS and Monte Carlo standard error MC_e only for the scenarios that traced the true model. These are computed only for the non important covariates across the five different scenarios which are found respectively in Tables (C.6), (C.7), (C.8), (C.9), (C.10), (C.11), (C.12), (C.13), (C.14) and (C.15) for only one simulated repetition, since we will expect similar results across the rest of 99 samples. In Particular, the results suggests that

- All computational methods within data augmentation under the three different prior setups show larger effective sample size ESS with respect to their typical versions.
- All computational methods with data augmentation scheme for the three different prior choices are more prone to lower Monte Carlo errors MC_e than their typical setup.

This seams reasonable since model complexity in data augmentation overburdens with the incorporation of additional latent variables the ESS and hence increase the number of iterations to converge. Regarding the computational efficiency among SSVS and GVS we summarize that

- The typical version of SSVS shows always lower effective sample size and larger sampling error in contrast with the typical of GVS.
- The typical version and augmented version of SSVS are exposed always to larger sampling errors in contrast with their respective versions of GVS.
- The typical version and augmented version of GVS are more accurate and trustworthy with respect to sampling errors.

In addition, results based on the frequency of identifying the true data-generating model through the maximum aposteriori model for the typical and augmented logistic regression over 100 repeated simulations of each scenario and correlation design are also provided in Table (C.5) for sample size equal to $n = 500$. The compared results show the following situations

i) In general, the Bayesian variable selection methods with mixtures of g-priors perform successfully as in 8 out of the 8 scenarios, the best method of identifying the true generating mechanism of the data still remains one of the methods with mixtures of g-priors

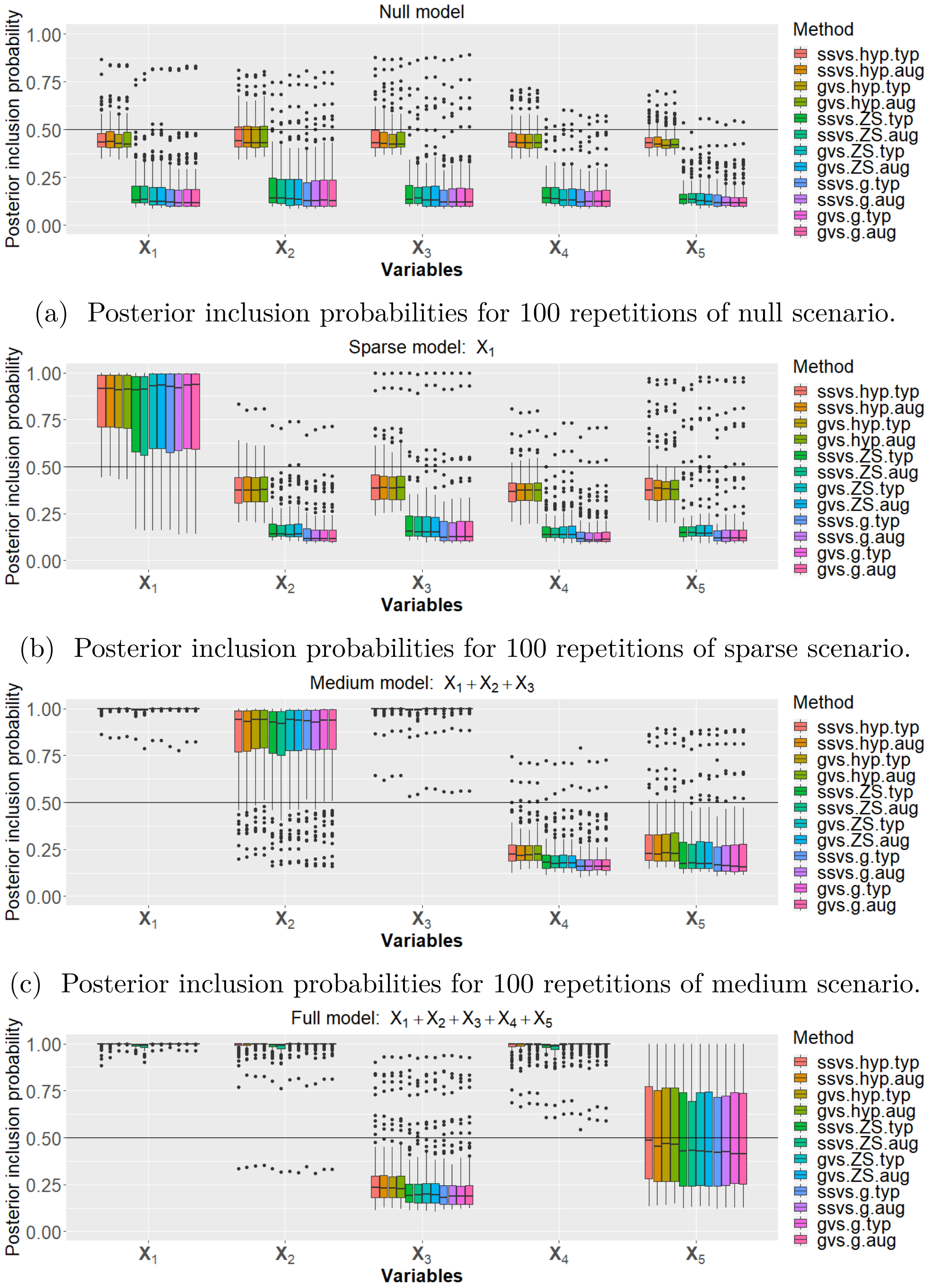
ii) In the null scenario, apart from (ssvs.hyp.typ) methods of mixtures g-priors trace correctly the true generating mechanism in the independent covariates design and in the correlated covariates. For all Bayesian variable selection methods, an improved rate of true model identifiability is observed.

iii) In the medium and sparse scenario, all computational techniques of Bayesian variable selection under the different prior choices, are showing an increased true model identification in both the independent and correlated design.

iv) In the full scenario, all Bayesian variable selection methods surpass the previous 3% of true model rate in the independent covariates design with a very high magnitude, whereas in the correlated design their performance is also improved with a moderate model identification rate.

v) As it was expected, each Bayesian variable selection method among typical and augmented logistic regression converge in the between posterior metrics as the sample size increases. Even minor differences are vanished due to the impact of large sample size $n = 500$.

Similar thoughts, come in agreement also with Figures (C.5) and (C.6) which describe the marginal posterior inclusion probabilities over the 100 simulated repeated experiments under the different Bayesian variable selection algorithms. It is evident that due to the impact of large sample size, there is little uncertainty accumulated in the non-important covariates resulting narrower in the respective boxplots. Finally, the prior choice of mixtures of g-priors affects the specific preference of model resulting with medium or larger complexity.



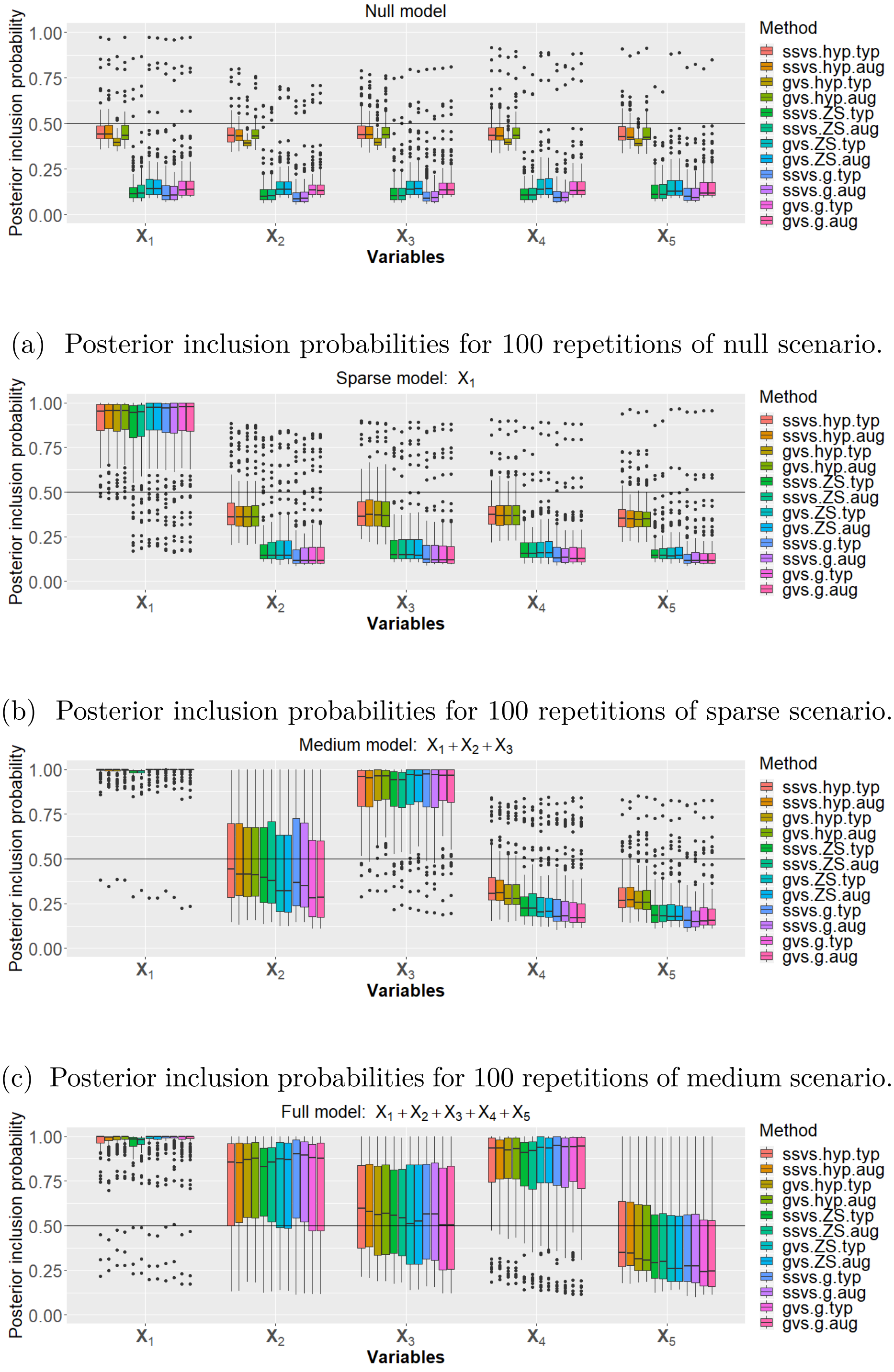
(a) Posterior inclusion probabilities for 100 repetitions of null scenario.

(b) Posterior inclusion probabilities for 100 repetitions of sparse scenario.

(c) Posterior inclusion probabilities for 100 repetitions of medium scenario.

(d) Posterior inclusion probabilities for 100 repetitions of full scenario.

Fig. C.1 Posterior inclusion probabilities under the various methods from 100 repetitions of independent $(r = 0)$ covariates for different scenarios.



(a) Posterior inclusion probabilities for 100 repetitions of null scenario.

(b) Posterior inclusion probabilities for 100 repetitions of sparse scenario.

(c) Posterior inclusion probabilities for 100 repetitions of medium scenario.

(d) Posterior inclusion probabilities for 100 repetitions of full scenario.

Fig. C.2 Posterior inclusion probabilities under the various methods from 100 repetitions of correlated ($r = 0.75$) covariates for different scenarios.

	Ess, Null scenario, r = 0.00				
Method	γ_1	γ_2	γ_3	γ_4	γ_5
ssvs.hyp.typ	**789**	**1053**	**1044**	**994**	**922**
ssvs.hyp.aug	1768	1873	2425	1873	2300
gvs.hyp.typ	**3570**	**7515**	**9712**	**5518**	**8754**
gvs.hyp.aug	12589	15810	17102	13553	15747
ssvs.ZS.typ	**1150**	**2036**	**3417**	**1658**	**2269**
ssvs.ZS.aug	3485	7395	8001	5011	7197
gvs.ZS.typ	**26736**	35758	**37278**	**30838**	**32289**
gvs.ZS.aug	29319	**33519**	40231	34595	36447
ssvs.g.typ	**911**	**2237**	**3795**	**1712**	**2025**
ssvs.g.aug	4901	9791	10281	6808	9560
gvs.g.typ	**28393**	45091	**40724**	**28984**	**40413**
gvs.g.aug	45964	**41355**	42311	35604	43571

Table C.6 Effective sample size comparison (in bold lowest value).

	MC$_e$, Null scenario, r = 0.00				
Method	γ_1	γ_2	γ_3	γ_4	γ_5
ssvs.hyp.typ	0.017	0.015	0.015	0.015	0.016
ssvs.hyp.aug	**0.011**	**0.011**	**0.010**	**0.011**	**0.010**
gvs.hyp.typ	0.008	0.005	0.004	0.006	0.005
gvs.hyp.aug	**0.004**	**0.004**	**0.003**	**0.004**	**0.004**
ssvs.ZS.typ	0.012	0.007	0.005	0.008	0.006
ssvs.ZS.aug	**0.007**	**0.003**	**0.003**	**0.005**	**0.003**
gvs.ZS.typ	**0.002**	**0.001**	**0.001**	**0.001**	**0.001**
gvs.ZS.aug	**0.002**	**0.001**	**0.001**	**0.001**	**0.001**
ssvs.g.typ	0.013	0.006	0.004	0.008	0.007
ssvs.g.aug	**0.005**	**0.003**	**0.003**	**0.004**	**0.003**
gvs.g.typ	0.002	0.001	0.001	0.002	**0.001**
gvs.g.aug	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**

Table C.7 Monte Carlo error comparison (in bold lowest value).

	Ess, Null scenario, r = 0.75				
Method	γ_1	γ_2	γ_3	γ_4	γ_5
ssvs.hyp.typ	**801**	**648**	**567**	**608**	**746**
ssvs.hyp.aug	1443	1249	1390	1081	1544
gvs.hyp.typ	**8463**	**3742**	**4299**	**5423**	**9598**
gvs.hyp.aug	14226	10035	8830	10708	13255
ssvs.ZS.typ	**2556**	**815**	**776**	**742**	**1657**
ssvs.ZS.aug	**5740**	**2941**	**2374**	**3022**	**5055**
gvs.ZS.typ	**25502**	**9588**	**11190**	**16834**	**17952**
gvs.ZS.aug	28172	19567	18879	20463	30373
ssvs.g.typ	**3650**	**1395**	**968**	**733**	**1717**
ssvs.g.aug	8338	5414	3278	3468	6557
gvs.g.typ	**26066**	**12560**	**12688**	**16651**	**13749**
gvs.g.aug	27302	24939	26158	29872	27801

Table C.8 Effective sample size comparison (in bold lowest value).

	MC$_e$, Null scenario, r = 0.75				
Method	γ_1	γ_2	γ_3	γ_4	γ_5
ssvs.hyp.typ	0.017	0.019	0.020	0.020	0.017
ssvs.hyp.aug	**0.012**	**0.014**	**0.013**	**0.015**	**0.012**
gvs.hyp.typ	0.005	0.008	0.007	0.006	0.005
gvs.hyp.aug	**0.004**	**0.004**	**0.005**	**0.004**	**0.004**
ssvs.ZS.typ	0.005	0.011	0.013	0.015	0.007
ssvs.ZS.aug	**0.003**	**0.005**	**0.007**	**0.007**	**0.004**
gvs.ZS.typ	0.002	0.003	0.003	**0.003**	**0.002**
gvs.ZS.aug	**0.001**	**0.002**	**0.002**	**0.003**	**0.002**
ssvs.g.typ	0.004	0.007	0.011	0.014	0.007
ssvs.g.aug	**0.003**	**0.003**	**0.006**	**0.006**	**0.003**
gvs.g.typ	**0.001**	**0.002**	0.003	0.003	**0.002**
gvs.g.aug	**0.001**	**0.002**	**0.002**	**0.002**	**0.002**

Table C.9 Monte Carlo error comparison (in bold lowest value).

	Ess, Sparse scenario, r = 0.00			
Method	γ_2	γ_3	γ_4	γ_5
ssvs.hyp.typ	**978**	**881**	**1145**	**1155**
ssvs.hyp.aug	1836	2426	2265	1968
gvs.hyp.typ	**13548**	**13451**	**13511**	**16099**
gvs.hyp.aug	23634	22254	17206	20729
ssvs.ZS.typ	**1493**	**1496**	**2518**	**1643**
ssvs.ZS.aug	4952	4752	7114	5519
gvs.ZS.typ	32597	**31223**	**33537**	**34404**
gvs.ZS.aug	**30775**	35159	35743	36800
ssvs.g.typ	**1103**	**1652**	**2709**	**1685**
ssvs.g.aug	7305	6714	8493	7575
gvs.g.typ	38390	**31729**	**33630**	**29118**
gvs.g.aug	**35017**	33216	35315	34858

Table C.10 Effective sample size comparison (in bold lowest value).

	MC$_e$**, Sparse scenario, r** = 0.00			
Method	γ_2	γ_3	γ_4	γ_5
ssvs.hyp.typ	0.015	0.015	0.013	0.014
ssvs.hyp.aug	**0.011**	**0.009**	**0.009**	**0.010**
gvs.hyp.typ	0.004	0.004	0.004	**0.003**
gvs.hyp.aug	**0.003**	**0.003**	**0.003**	**0.003**
ssvs.ZS.typ	0.010	0.009	0.006	0.009
ssvs.ZS.aug	**0.005**	**0.005**	**0.004**	**0.004**
gvs.ZS.typ	**0.002**	**0.002**	**0.001**	**0.001**
gvs.ZS.aug	**0.002**	**0.002**	**0.001**	0.002
ssvs.g.typ	0.011	0.008	0.005	0.008
ssvs.g.aug	**0.004**	**0.004**	**0.003**	**0.004**
gvs.g.typ	**0.001**	**0.002**	**0.001**	0.002
gvs.g.aug	**0.001**	**0.002**	**0.001**	**0.001**

Table C.11 Monte Carlo error comparison (in bold lowest value).

	Ess, Sparse scenario, r $= 0.75$			
Method	γ_2	γ_3	γ_4	γ_5
ssvs.hyp.typ	**577**	**669**	**720**	**855**
ssvs.hyp.aug	1204	1210	1394	1524
gvs.hyp.typ	**3096**	**3517**	**6643**	**10321**
gvs.hyp.aug	12066	9101	12174	19034
ssvs.ZS.typ	**1225**	**995**	**1849**	**2197**
ssvs.ZS.aug	4350	3423	4280	5521
gvs.ZS.typ	**6375**	**8248**	**16521**	**25019**
gvs.ZS.aug	17666	16975	27469	26365
ssvs.g.typ	**4454**	**1011**	**2049**	**2175**
ssvs.g.aug	4672	3307	7155	6484
gvs.g.typ	**7460**	**9865**	**19790**	**24134**
gvs.g.aug	16259	18289	28040	32510

Table C.12 Effective sample size comparison (in bold lowest value).

	$\mathbf{MC}_e$**, Sparse scenario, r** $= 0.75$			
Method	γ_2	γ_3	γ_4	γ_5
ssvs.hyp.typ	0.019	0.019	0.018	0.016
ssvs.hyp.aug	**0.013**	**0.014**	**0.012**	**0.012**
gvs.hyp.typ	0.008	0.008	0.005	0.004
gvs.hyp.aug	**0.004**	**0.005**	**0.004**	**0.003**
ssvs.ZS.typ	0.010	0.011	0.006	0.006
ssvs.ZS.aug	**0.005**	**0.005**	**0.004**	**0.004**
gvs.ZS.typ	0.004	0.004	0.002	0.002
gvs.ZS.aug	**0.002**	**0.002**	**0.002**	**0.002**
ssvs.g.typ	0.008	0.010	0.006	0.006
ssvs.g.aug	**0.005**	**0.005**	**0.003**	**0.003**
gvs.g.typ	0.004	0.003	0.002	0.002
gvs.g.aug	**0.002**	**0.002**	**0.001**	**0.001**

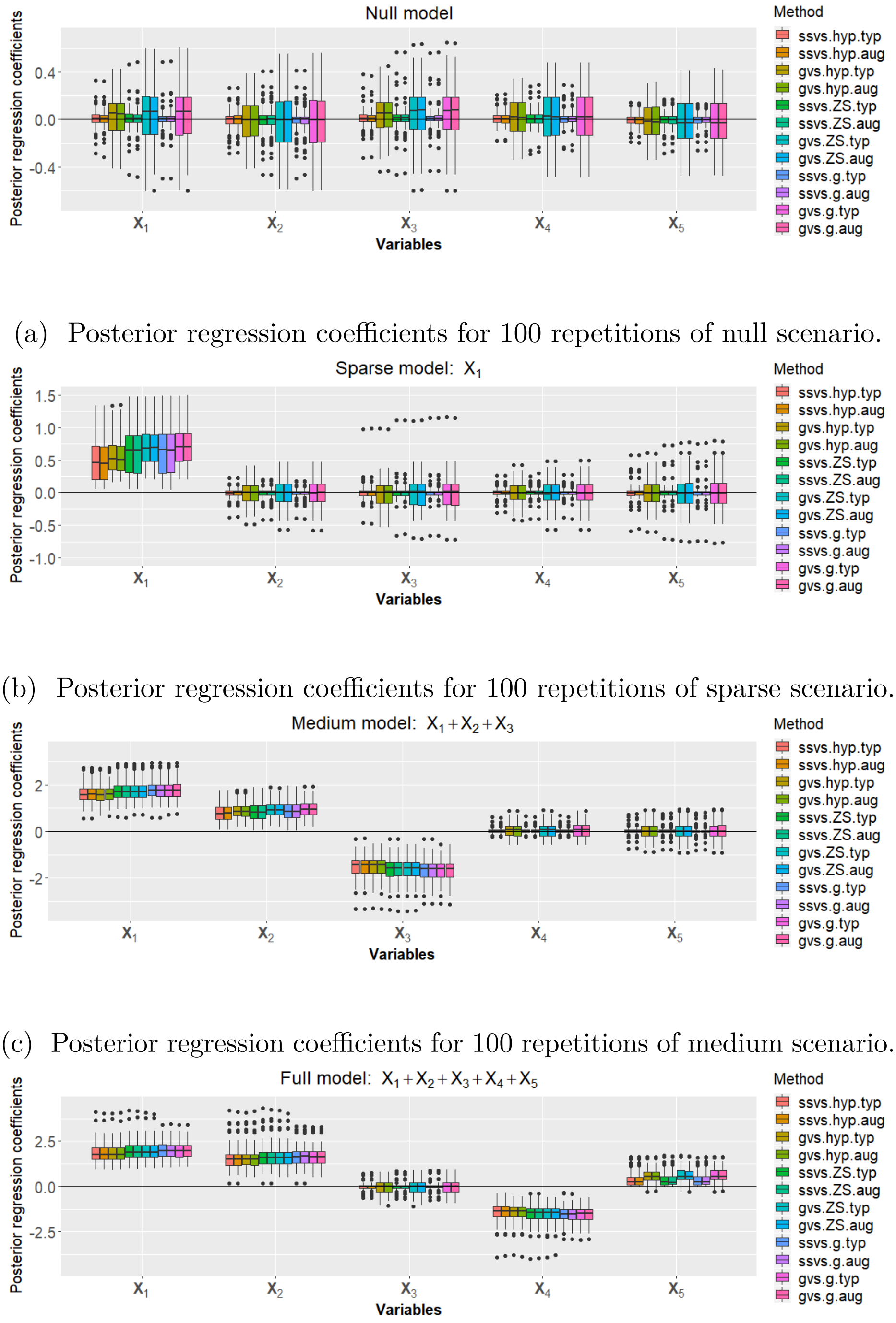
Table C.13 Monte Carlo error comparison (in bold lowest value).

Ess, Medium scenario, r = 0.00		
Method	γ_4	γ_5
ssvs.hyp.typ	**1083**	**1135**
ssvs.hyp.aug	3154	3044
gvs.hyp.typ	**25306**	**20654**
gvs.hyp.aug	25538	24708
ssvs.ZS.typ	**1231**	**1481**
ssvs.ZS.aug	4105	4294
gvs.ZS.typ	**32679**	**28052**
gvs.ZS.aug	33326	30629
ssvs.g.typ	**1766**	**1474**
ssvs.g.aug	6332	7312
gvs.g.typ	34154	31301
gvs.g.aug	**21537**	**22334**

Table C.14 Effective sample size comparison (in bold lowest value).

MC$_e$, Medium scenario, r = 0.00		
Method	γ_4	γ_5
ssvs.hyp.typ	0.013	0.013
ssvs.hyp.aug	**0.007**	**0.008**
gvs.hyp.typ	**0.002**	0.003
gvs.hyp.aug	**0.002**	**0.002**
ssvs.ZS.typ	0.011	0.010
ssvs.ZS.aug	**0.006**	**0.005**
gvs.ZS.typ	**0.002**	**0.002**
gvs.ZS.aug	**0.002**	**0.002**
ssvs.g.typ	0.007	0.009
ssvs.g.aug	**0.004**	**0.004**
gvs.g.typ	**0.001**	**0.002**
gvs.g.aug	0.003	0.003

Table C.15 Monte Carlo error comparison (in bold lowest value).



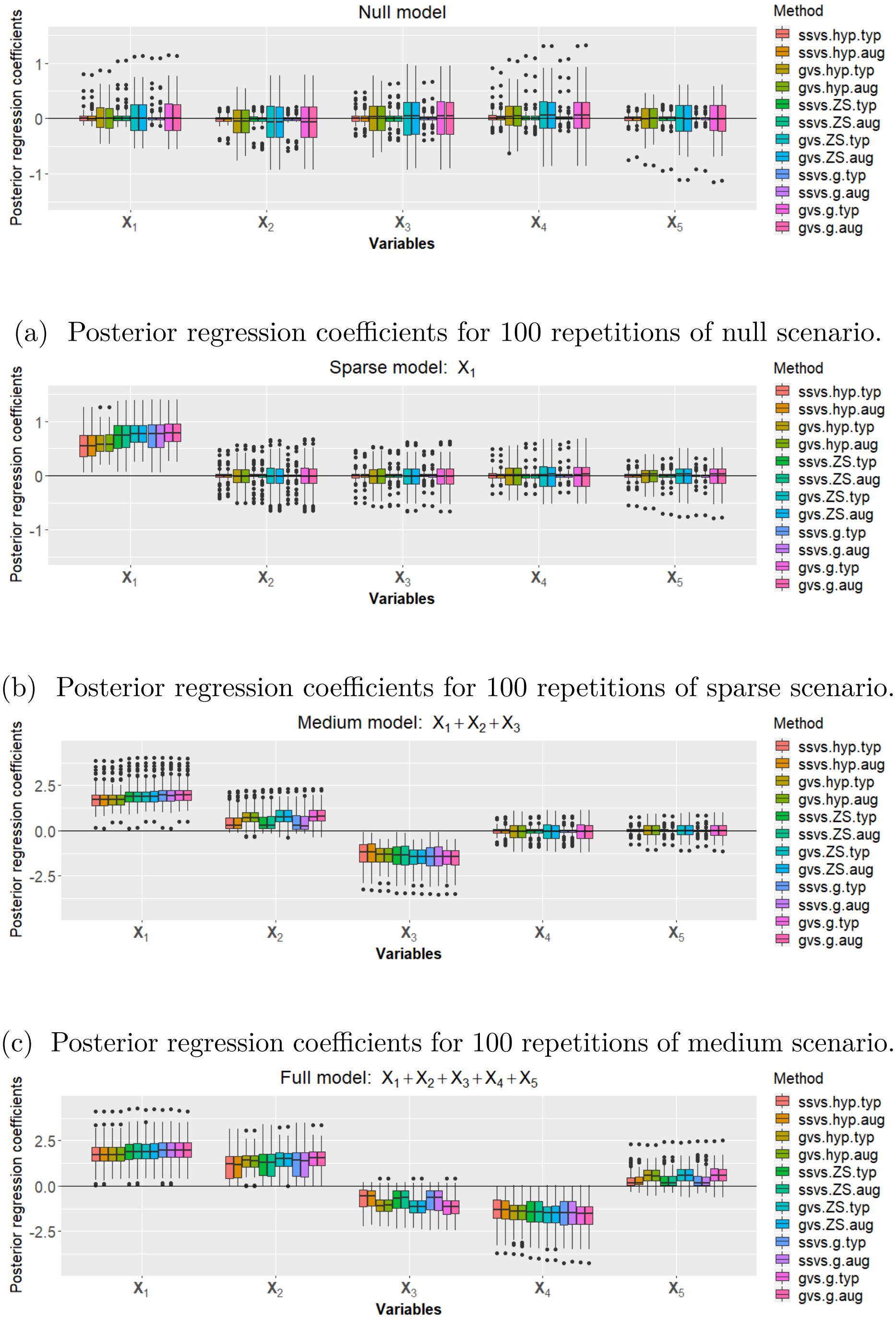
(a) Posterior regression coefficients for 100 repetitions of null scenario.

(b) Posterior regression coefficients for 100 repetitions of sparse scenario.

(c) Posterior regression coefficients for 100 repetitions of medium scenario.

(d) Posterior regression coefficients for 100 repetitions of full scenario.

Fig. C.3 Posterior regression coefficients under the various methods from 100 repetitions of independent ($r = 0$) covariates for different scenarios.



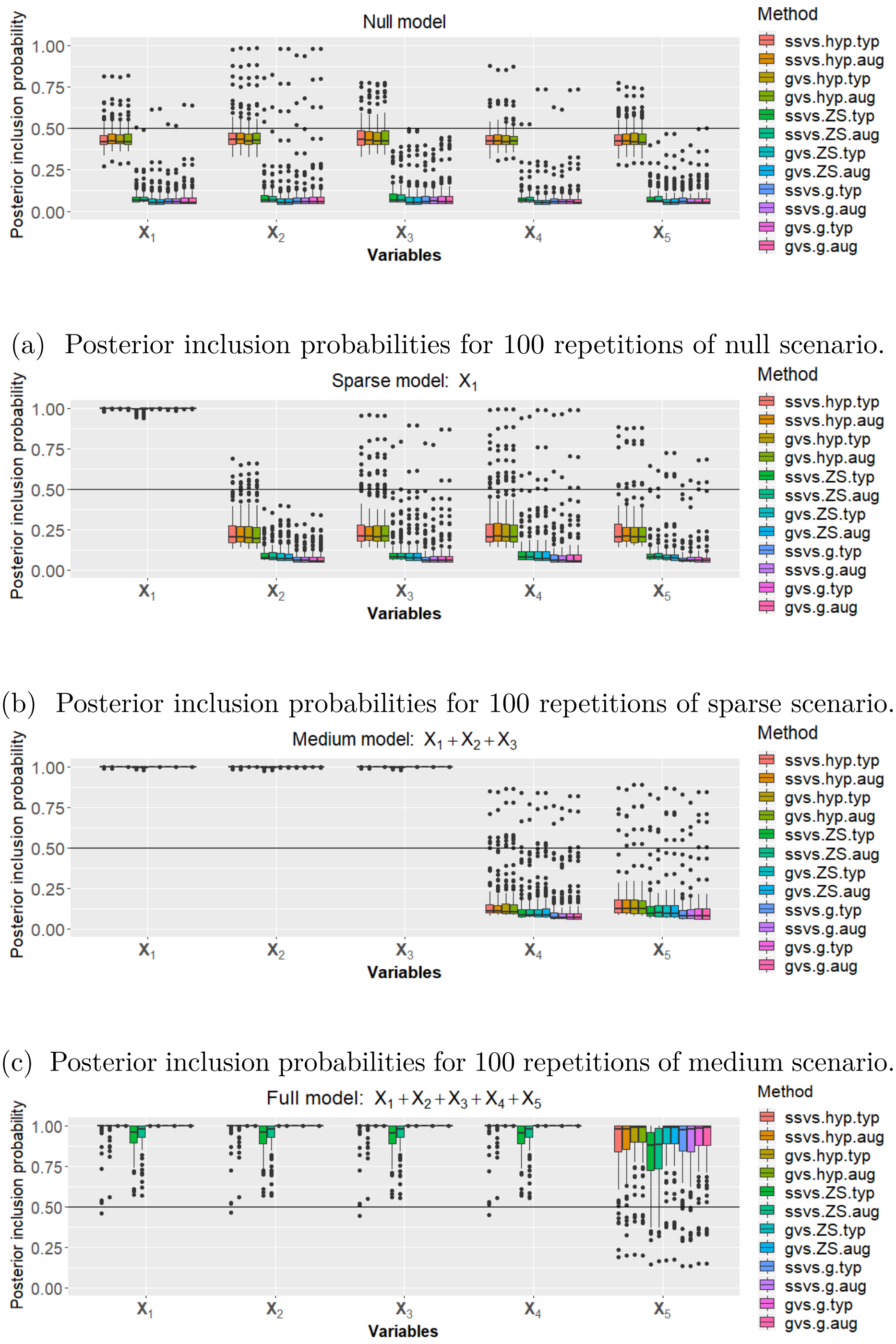
(a) Posterior regression coefficients for 100 repetitions of null scenario.

(b) Posterior regression coefficients for 100 repetitions of sparse scenario.

(c) Posterior regression coefficients for 100 repetitions of medium scenario.

(d) Posterior regression coefficients for 100 repetitions of full scenario.

Fig. C.4 Posterior regression coefficients under the various methods from 100 repetitions of correlated ($r = 0.75$) covariates for different scenarios.



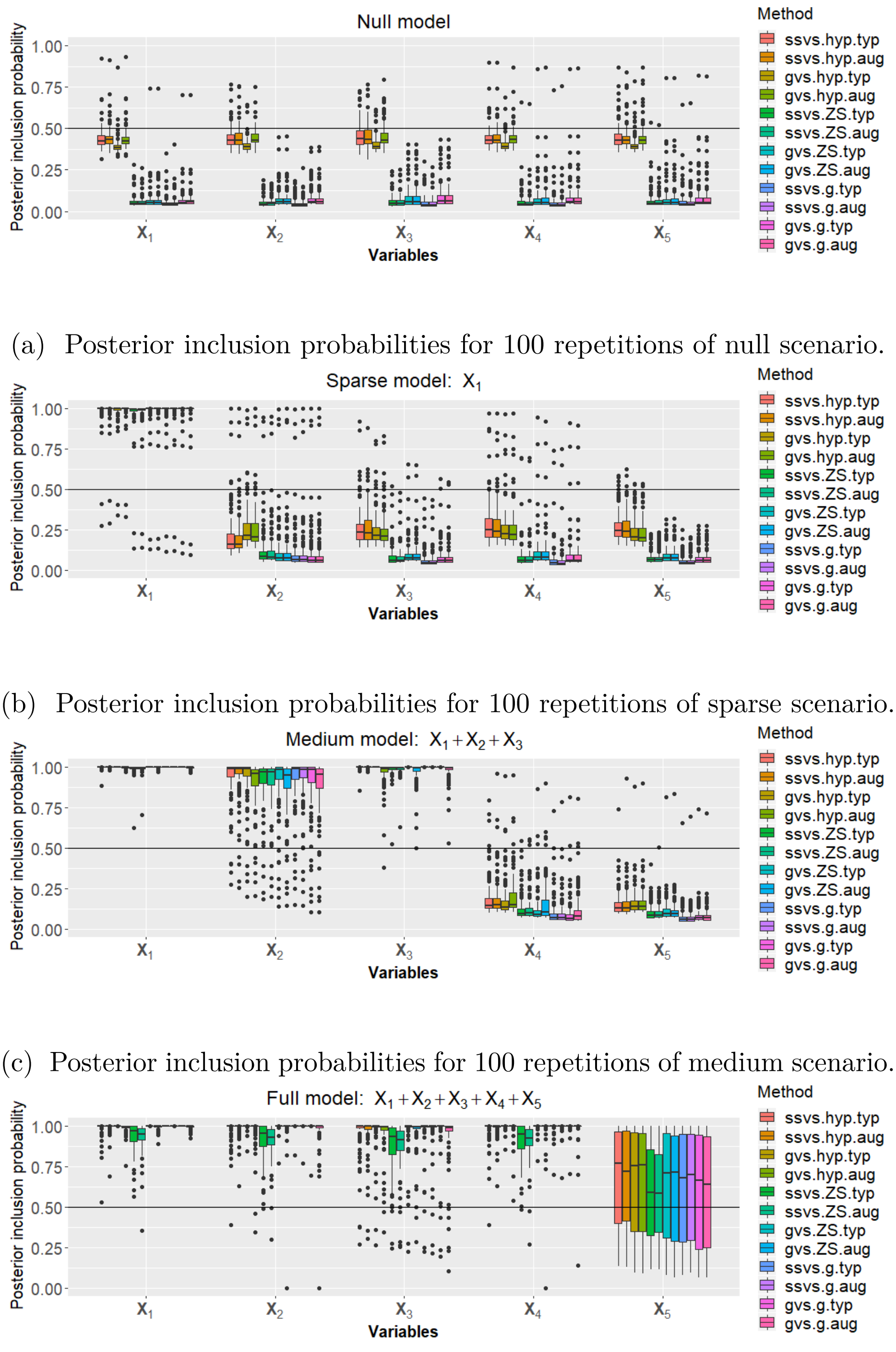
(a) Posterior inclusion probabilities for 100 repetitions of null scenario.

(b) Posterior inclusion probabilities for 100 repetitions of sparse scenario.

(c) Posterior inclusion probabilities for 100 repetitions of medium scenario.

(d) Posterior inclusion probabilities for 100 repetitions of full scenario.

Fig. C.5 Posterior inclusion probabilities under the various methods from 100 repetitions of independent ($r = 0$) covariates for different scenarios ($n = 500$).



(a) Posterior inclusion probabilities for 100 repetitions of null scenario.

(b) Posterior inclusion probabilities for 100 repetitions of sparse scenario.

(c) Posterior inclusion probabilities for 100 repetitions of medium scenario.

(d) Posterior inclusion probabilities for 100 repetitions of full scenario.

Fig. C.6 Posterior inclusion probabilities under the various methods from 100 repetitions of correlated ($r = 0.75$) covariates for different scenarios ($n = 500$).

C.5 Real Dataset

Finally, we illustrate an application of Bayesian variable selection in logistic regression with mixtures of g-priors for the Pima Indians dataset of diabetes Ripley (1996). This dataset consists of $n = 532$ measurements of women patients with diabetes and $p = 8$ covariates that describe their profile which was also analysed by Holmes and Held (2006), Bové and Held (2011) and by Fouskakis et al. (2018) where they applied the power expected posterior prior methodology. The response variable $\boldsymbol{Y}$ is binary indicating the presence or not of diabetes and the covariates are the number of pregnancies ($\boldsymbol{X_1}$), the plasma glucose concentration ($\boldsymbol{X_2}$), the diastolic blood pressure ($\boldsymbol{X_3}$), the triceps skin fold thickness ($\boldsymbol{X_4}$), the body mass index ($\boldsymbol{X_5}$), the diabetes pedigree function ($\boldsymbol{X_6}$) and age ($\boldsymbol{X_7}$). Furthermore, with regard to the Bayesian variable selection analysis, a maximum likelihood perspective was also applied to the full model obtaining estimates $\hat{\boldsymbol{\beta}}$, where only ($\boldsymbol{X_1}$), ($\boldsymbol{X_2}$), ($\boldsymbol{X_5}$), ($\boldsymbol{X_6}$) were found as statistically significant at 5%, whereas ($\boldsymbol{X_7}$) was at the border of insignificance with p-value $= 0.059$. The main task of this application is to assess again the performance of Bayesian variable selection methods of typical and augmented logistic regression based on mixtures of g-priors both in-sample and out-of-sample values across the three different prior set-ups for the computational methods of SSVS and GVS. It should be stated that we change the acronyms for the methods based on Table (C.16) with an added R as superscript to denote the methods implemented in the R programming language in order to compare them and differentiate from GVS for typical logistic regression method implemented in WINBUGS, with acronyms gvs$^{\text{W}}$.hyp.typ, gvs$^{\text{W}}$.ZS.typ, gvs$^{\text{W}}$.g.typ under the three different prior set-ups respectively. For the out-of-sample analysis, we will split the data at half randomly and then compare the performance of MCMC by calculating at each iteration of each method the false negative ($\widehat{FN}$), the false positive ($\widehat{FP}$), the accuracy ($\widehat{ACC}$) and the missclassification error ($\widehat{ERR}$). In other words,, we begin with a preliminary analysis of Bayesian variable selection procedures using MCMC in order to identify important covariates through the median posterior model probability. Then, we proceed with an additional analysis to evaluate the predictive ability of each MCMC method based on the maximum a posteriori model MAP and the median probability model MPM.

Among all methods with Hyper-g we considered again the value of $\alpha = 3$ proposed by Liang et al., (2008) and a Metropolis-Hastings random walk was added with tuning variance $u_g = 1$ in order to obtain convergence in both GVS and SSVS. Regarding the scaling of proposals of a, $\boldsymbol{\beta}$ for both SSVS and GVS, we used $t = 0.8$ suggested by Roberts and Rosenthal (2001) and $u_a = \widehat{\sigma}_a^2$ respectively to ensure the good mixing of the chains with high respective acceptance rates, where $\widehat{\sigma}_a^2 = 0.015$ and $\widehat{\sigma}_a^2 = 0.030$ is the variance of the intercept under the full logistic regression model in the first and second analysis respectively.

	Acronym	Computational Method	Prior	Model
1	ssvs$^{\mathrm{R}}$.hyp.typ	Stochastic Search Variable Selection for $\alpha = 3$, $\tau = 0.02$, $c = 50$, $u_g = 1$, $u_a = 0.015$, $t = 0.8$	Hyper-g	Typical
2	ssvs$^{\mathrm{R}}$.hyp.aug	Stochastic Search Variable Selection for $\alpha = 3$, $\tau = 0.02$, $c = 50$, $u_g = 1$	Hyper-g	Augmented
3	gvs$^{\mathrm{R}}$.hyp.typ	Gibbs Variable Selection for $\alpha = 3$, $u_g = 1$, $u_a = 0.015$, $t = 0.8$	Hyper-g	Typical
4	gvs$^{\mathrm{R}}$.hyp.aug	Gibbs Variable Selection for $\alpha = 3$, $u_g = 1$, $u_a = 0.015$, $t = 0.8$	Hyper-g	Augmented
5	gvs$^{\mathrm{W}}$.hyp.typ	Gibbs Variable Selection for $\alpha = 3$	Hyper-g	Typical
6	ssvs$^{\mathrm{R}}$.ZS.typ	Stochastic Search Variable Selection for $\tau = 0.02$, $c = 50$,$u_a = 0.015$, $t = 0.8$	Zellner-Siow	Typical
7	ssvs$^{\mathrm{R}}$.ZS.aug	Stochastic Search Variable Selection for $\tau = 0.02$, $c = 50$	Zellner-Siow	Augmented
8	gvs$^{\mathrm{R}}$.ZS.typ	Gibbs Variable Selection for $u_a = 0.015$, $t = 0.8$	Zellner-Siow	Typical
9	gvs$^{\mathrm{R}}$.ZS.aug	Gibbs Variable Selection for $u_a = 0.015$, $t = 0.8$	Zellner-Siow	Augmented
10	gvs$^{\mathrm{W}}$.ZS.typ	Gibbs Variable Selection	Zellner-Siow	Typical
11	ssvs$^{\mathrm{R}}$.g.typ	Stochastic Search Variable Selection for $\tau = 0.02$, $c = 50$, $u_a = 0.015$, $t = 0.8$, $g = n$	g-prior	Typical
12	ssvs$^{\mathrm{R}}$.g.aug	Stochastic Search Variable Selection for $\tau = 0.02$, $c = 50$, $g = n$	g-prior	Augmented
13	gvs$^{\mathrm{R}}$.g.typ	Gibbs Variable Selection for $g = n$	g-prior	Typical
14	gvs$^{\mathrm{R}}$.g.aug	Gibbs Variable Selection for $g = n$	g-prior	Augmented
15	gvs$^{\mathrm{W}}$.g.typ	Gibbs Variable Selection for $g = n$	g-prior	Typical

Table C.16 Acronyms of Bayesian variable selection methods with MCMC for logistic regression.

With respect to MCMC methods, prior inputs $\tau_j = 0.02$ and $c_j = 50$ for $j = 1, \ldots, p$ were set on practical significance for SSVS to achieve similar results with the objective

Bayesian methods and $\bar{\boldsymbol{\mu}}$, $\bar{\boldsymbol{s}}^2$ were computed from pilot runs under the full model for GVS of each simulated dataset repetition. The option of prior input τ_j and c_j are as indicated in Table (C.16). A detailed description of all Bayesian variable selection methods which are used as references in Figures and in Tables, are found in Table (C.16). The prior inputs of Table (C.16) were considered the same (apart for g-prior where g=532) only for the first analysis, whereas in the second we used $g = 266$. The prior inputs for GVS were set as the following $\bar{\boldsymbol{\mu}}$, $\bar{\boldsymbol{s}}^2$ for the first analysis adn second analysis respectively.

	Median Probability Model						
	Independent Variables						
Method	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$	$\boldsymbol{X_3}$	$\boldsymbol{X_4}$	$\boldsymbol{X_5}$	$\boldsymbol{X_6}$	$\boldsymbol{X_7}$
ssvs$^{\mathrm{R}}$.hyp.typ	0.980	1.000	0.347	0.352	0.997	0.997	0.680
ssvs$^{\mathrm{R}}$.hyp.aug	0.977	1.000	0.353	0.339	0.998	0.996	0.669
gvs$^{\mathrm{R}}$.hyp.typ	0.969	1.000	0.381	0.372	0.997	0.996	0.654
gvs$^{\mathrm{R}}$.hyp.aug	0.970	0.999	0.384	0.376	0.997	0.996	0.657
gvs$^{\mathrm{w}}$.hyp.typ	0.970	1.000	0.383	0.376	0.998	0.995	0.655
ssvs$^{\mathrm{R}}$.ZS.typ	0.891	0.936	0.250	0.270	0.917	0.908	0.543
ssvs$^{\mathrm{R}}$.ZS.aug	0.887	0.932	0.241	0.252	0.917	0.910	0.555
gvs$^{\mathrm{R}}$.ZS.typ	0.954	1.000	0.248	0.253	0.998	0.994	0.526
gvs$^{\mathrm{R}}$.ZS.aug	0.962	0.999	0.248	0.243	0.998	0.994	0.526
gvs$^{\mathrm{W}}$.ZS.typ	0.961	1.000	0.258	0.251	0.998	0.994	0.535
ssvs$^{\mathrm{R}}$.g.typ	0.942	1.000	0.160	0.179	0.992	0.989	0.465
ssvs$^{\mathrm{R}}$.g.aug	0.954	1.000	0.152	0.184	0.995	0.985	0.459
gvs$^{\mathrm{R}}$.g.typ	0.950	1.000	0.135	0.136	0.997	0.991	0.391
gvs$^{\mathrm{R}}$.g.aug	0.951	0.999	0.133	0.136	0.998	0.990	0.385
gvs$^{\mathrm{W}}$.g.typ	0.951	0.999	0.133	0.136	0.998	0.990	0.385

Table C.17 Marginal posterior inclusion probabilities for each independent variable X_j regarding each Bayesian variable selection method for 40000 MCMC iterations.

When information is not available with respect to the subset of variables, it is preferable to adopt an objective prior elicitation for the model parameters and the model itself. Again, we considered (4.12), (4.19) of Bové and Held (2011) adopted in the style of logistic regression in SSVS and GVS to account for the joint prior dependences among parameters a, $\boldsymbol{\beta}$, g, $\boldsymbol{\gamma}$.

Prior Inputs-Initial Values of 1st Analysis	
Parameter	**Value**
$\widehat{\boldsymbol{\beta}}$	$(0.405, 1.094, -0.094, 0.071, 0.568, 0.450, 0.283)^T$
$\hat{a}$	-0.990
τ	0.02
c	50
$\bar{\boldsymbol{\mu}}$	$\widehat{\boldsymbol{\beta}}$
$\bar{\boldsymbol{s}}^2$	$(0.020, 0.017, 0.016, 0.024, 0.025, 0.015, 0.022)^T$
$\boldsymbol{\gamma}^{(0)}$	$(1, 1, 1, 1, 1, 1, 1)^T$
$\boldsymbol{\beta}^{(0)}$	$\widehat{\boldsymbol{\beta}}$
$a^{(0)}$	$\hat{a}$
$g^{(0)}$	532
$\boldsymbol{\omega}^0$	$(1, \dots, 1)$

Table C.18 Prior-inputs and initial values of 1st Analysis

Prior Inputs-Initial Values of 2nd Analysis	
Parameter	**Value**
$\widehat{\boldsymbol{\beta}}$	$(0.483, 1.145, -0.040, -0.143, 0.544, 0.600, 0.418)^T$
$\hat{a}$	-0.944
τ	0.02
c	50
$\bar{\boldsymbol{\mu}}$	$\widehat{\boldsymbol{\beta}}$
$\bar{\boldsymbol{s}}^2$	$(0.044, 0.034, 0.028, 0.040, 0.044, 0.034, 0.051)^T$
$\boldsymbol{\gamma}^{(0)}$	$(1, 1, 1, 1, 1, 1, 1)^T$
$\boldsymbol{\beta}^{(0)}$	$\widehat{\boldsymbol{\beta}}$
$a^{(0)}$	$\hat{a}$
$g^{(0)}$	n_{te}
$\boldsymbol{\omega}^0$	$(1, \dots, 1)$

Table C.19 Prior-inputs and initial values of 2nd Analysis

All the compared approaches under Zellner-Siow prior used (2.10), whereas for hyper-g-prior was used (2.12). Regarding the prior of model space, we assigned a beta-binomial with both hyper-parameters equal to one to preserve sparsity based on Scott and Berger (2010).

Implementing Bayesian variable selection MCMC methods to these real data, model fitting was applied through a Gibbs algorithm with successive Metropolis-Hastings steps as implemented in the R programming language simulating Markov chains of 40000 valid values to achieve convergence for both typical and augmented logistic regression Bayesian variable selection methods. The same was implemented for the GVS method in WINBUGS under the three different prior set-ups but only with regard to the first analysis. More precisely, in the first analysis for all methods we used the maximum likelihood estimators for parameters $\boldsymbol{\beta}$ and intercept a as initial values for each MCMC method

$\boldsymbol{\beta}^{(0)}$, $a^{(0)}$, for the vector of model indicators $\boldsymbol{\gamma}$ we have used as initial values $\boldsymbol{\gamma}^{\mathbf{(0)}}$, for g we used $g^{(0)}$ and for the methods with the latent data we used $\boldsymbol{\omega}^{(0)}$, whereas for the second analysis we have used the maximum likelihood estimates

$\boldsymbol{\beta}^{(0)}$, $a^{(0)}$ and $\boldsymbol{\gamma}^{\mathbf{(0)}}, g^{(0)}, \boldsymbol{\omega}^{(0)}$ are set the same as in the first analysis; please refer to Tables (C.18) and (C.19) respectively. Table (C.17) shows the marginal posterior inclusion probabilities of each covariate under the various Bayesian variable selection methods for the three different prior set-ups according to the MPM. In particular, we observe that the independent variables $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_5$, $\boldsymbol{X}_6$ are relevant since their marginal posterior inclusion probabilities are larger than 0.5 showing no important differences among all the Bayesian variable selection methods. On the contrary, the marginal posterior inclusion probabilities for the non important covariates $\boldsymbol{X}_3$, $\boldsymbol{X}_4$, $\boldsymbol{X}_7$, change substantially under the three different prior set-ups. More precisely, we observe that for the methods with mixtures of g-priors their marginal posterior inclusion probabilities of these covariates are higher in comparison with fixed g methods, a deduction which comes in agreement with the behaviour of mixtures of g-priors. For instance, the posterior marginal inclusion probabilities are inflated towards 0.5 for Bayesian variable selection methods with hyper-g priors, whereas for Zellner-Siow prior methods they are slightly lower than the hyper-g. This behaviour is related to the additional variability of the random g overburdening the uncertainty of including $\boldsymbol{X}_3$, $\boldsymbol{X}_4$ and more especially of $\boldsymbol{X}_7$ where only under the mixtures of g-priors becomes significant in contradiction with the fixed g-prior approaches. This is justified due to the stronger shrinkage under the fixed methods g-prior where the value $g = 532$ tends to conserve sparser models due to the activation of Jeffreys-Lindleys paradox. Regarding the shrinkage of g-priors mixtures, additional analysis was provided based on the computational methods of SSVS and GVS for typical and augmented logistic regression respectively as seen in Figures (C.7), (C.8), (C.9) and (C.10) where the posterior distribution is indicative of its behaviour. From these Figures it is evident that the posterior distribution of the shrinkage factor $\frac{g}{g+1}$ for the Bayesian variable selection methods with Zellner-Siow priors is concentrated close to one, resulting in a stronger shrinkage than those with hyper-g priors. Moreover, in the same Figures the posterior densities, the ergodic means, the autocorrelations and the traceplots are also depicted.

	Acceptance Rate		
Method	a	$\boldsymbol{\beta}$	g
ssvs$^{\mathrm{R}}$.hyp.typ	0.687	0.268	0.511
ssvs$^{\mathrm{R}}$.hyp.aug	-	-	0.511
gvs$^{\mathrm{R}}$.hyp.typ	0.689	0.273	0.545
gvs$^{\mathrm{R}}$.hyp.aug	-	-	0.552
ssvs$^{\mathrm{R}}$.ZS.typ	0.692	0.275	-
ssvs$^{\mathrm{R}}$.ZS.aug	-	-	-
gvs$^{\mathrm{R}}$.ZS.typ	0.698	0.274	-
gvs$^{\mathrm{R}}$.ZS.aug	-	-	-
ssvs$^{\mathrm{R}}$.g.typ	0.686	0.279	-
ssvs$^{\mathrm{R}}$.g.aug	-	-	-
gvs$^{\mathrm{R}}$.g.typ	0.689	0.275	-
gvs$^{\mathrm{R}}$.g.aug	-	-	-

Table C.20 Results of acceptance rates for parameters a, $\boldsymbol{\beta}$ and g of each Bayesian variable selection methods for typical and augmented logistic regression model with mixtures of g-priors.

They show convergence through the MCMC iterations among all Bayesian variable selection methods with mixtures of g-priors. Additional information is found on Table (C.20) for the acceptance rates of parameters a, $\boldsymbol{\beta}$, g which shows that all Bayesian variable selection methods perform efficiently with high acceptance rates. Regarding the comparison of all typical and augmented logistic regression Bayesian variable selection methods for each different prior set-ups, we observe similar results in terms of marginal posterior inclusion probabilities. Generally, regarding the comparison of GVS and SSVS based on the above results, it seems that for the Bayesian variable selection methods with GVS under the Zellner-Siow and fixed g the marginal posterior inclusion probabilities are more shrunk for the non important covariates $\boldsymbol{X}_3$, $\boldsymbol{X}_4$, $\boldsymbol{X}_7$ in comparison with the analogues of SSVS, whereas for the Bayesian variable selection methods with SSVS under hyper-g are more shrunk with respect to GVS. Furthermore, SSVS under Zellner-Siow priors both for typical and augmented logistic regression model exhibit strong shrinkage even for the important covariates $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ in contrast to the covariate $\boldsymbol{X}_7$. Finally, our analysis seems effective as it is verified based on the results obtained across all GVS methods under the three different prior set-ups implemented in the R programming language compared with those obtained in WINBUGS leading to the same results both for typical and augmented logistic regression. Similar conclusions, based in terms of posterior marginal inclusion probabilities regarding the importance of covariates are found in the works of Holmes and Held (2006) and Fouskakis et al. (2018).

We end the first part of this application with real data by examining the out-of-sample predictive accuracy of the above Bayesian variable selection methods for the typical and augmented logistic regression model. The predictive ability of all Bayesian variable selection methods is assessed by the predictive distribution of independent and identically distributed random variables $\boldsymbol{Y}^* = (\boldsymbol{Y}^*_1, \ldots, \boldsymbol{Y}^*_{n_{te}})^T$ which generate observed values $\boldsymbol{y}^{*(s)} = (y^{*(s)}_1, \ldots, y^{*(s)}_{n_{te}})^T$ as the following

$$\begin{aligned}\boldsymbol{Y}^*_i | a^{(s)}, \boldsymbol{\beta}^{(s)} &\sim Bern(1, p^{*(s)}_i),\\ \pi^{*(s)}_i &\approx \frac{\exp\left(a^{(s)} + \boldsymbol{x}^{te}_i \boldsymbol{\beta}^{(s)}\right)}{1 + \exp\left(a^{(s)} + \boldsymbol{x}^{te}_i \boldsymbol{\beta}^{(s)}\right)},\end{aligned}$$

where $p^{*(s)}_i = P(y^*_i = 1 | a^{(s)}, \boldsymbol{\beta}^{(s)})$, $a^{(s)}$ and $\boldsymbol{\beta}^{(s)}$ are the posterior samples of the intercept a and the regression coefficients $\boldsymbol{\beta}$ of MAP and MPM obtained from the training set based on the s-th iteration of the MCMC respective procedure and $\boldsymbol{x}^{te}_{(.)}$ are the row-wise of the test design matrix $\boldsymbol{X}^{te}$. Then, based on the posterior samples from the predictive distribution of s-th MCMC iteration, we construct the confusion matrix of the predictive response set $\boldsymbol{y}^{*(s)}$ versus the respective response test set $\boldsymbol{y}^{te}$

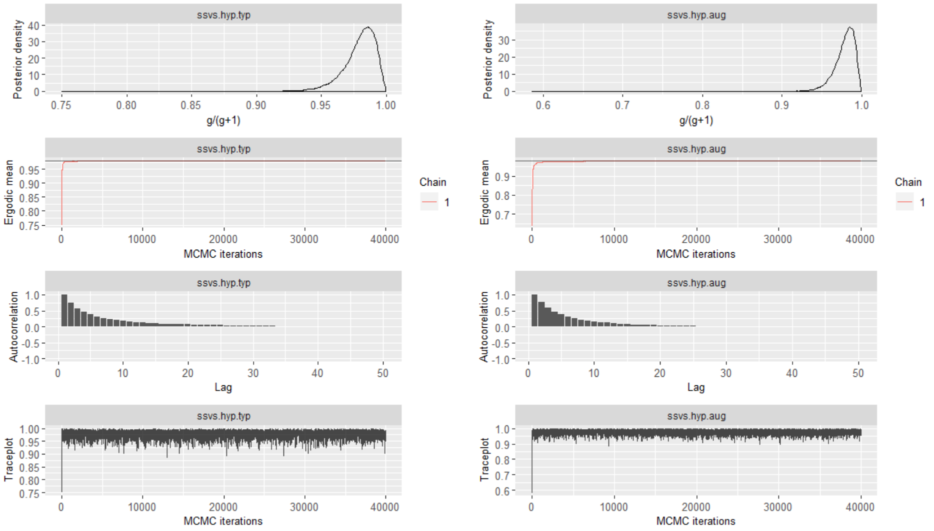
		$\boldsymbol{y}^{te}$	
		Negative	Positive
$\boldsymbol{y}^{*(s)}$	Negative	$TN^{(s)}$	$FN^{(s)}$
	Positive	$FP^{(s)}$	$TP^{(s)}$

Table C.21 Confusion matrix at s-th MCMC iteration of the predictive response set $\boldsymbol{y}^{*(s)}$ versus the respective response test set $\boldsymbol{y}^{te}$.

where we denote as $FN^{(s)} =$ P(Negative|Positive), $FP^{(s)} =$ P(Positive|Negative) and calculate the accuracy $ACC^{(s)} \approx \frac{TN^{(s)}+TP^{(s)}}{TN^{(s)}+TP^{(s)}+FN^{(s)}+FP^{(s)}}$ and the missclassification error $ERR^{(s)} \approx \frac{FN^{(s)}+FP^{(s)}}{TN^{(s)}+TP^{(s)}+FN^{(s)}+FP^{(s)}}$. In this way, we compute the averages of the posterior distribution for these quantities based on the respective MCMC iteration for the MAP and MPM under each Bayesian variable selection method as the following

$$\widehat{FN} \approx \sum_{s=1}^{S} \frac{FN^{(s)}}{S}$$
$$\widehat{FP} \approx \sum_{s=1}^{S} \frac{FP^{(s)}}{S}$$
$$\widehat{ACC} \approx \sum_{s=1}^{S} \frac{ACC^{(s)}}{S}$$
$$\widehat{ERR} \approx \sum_{s=1}^{S} \frac{ERR^{(s)}}{S}$$

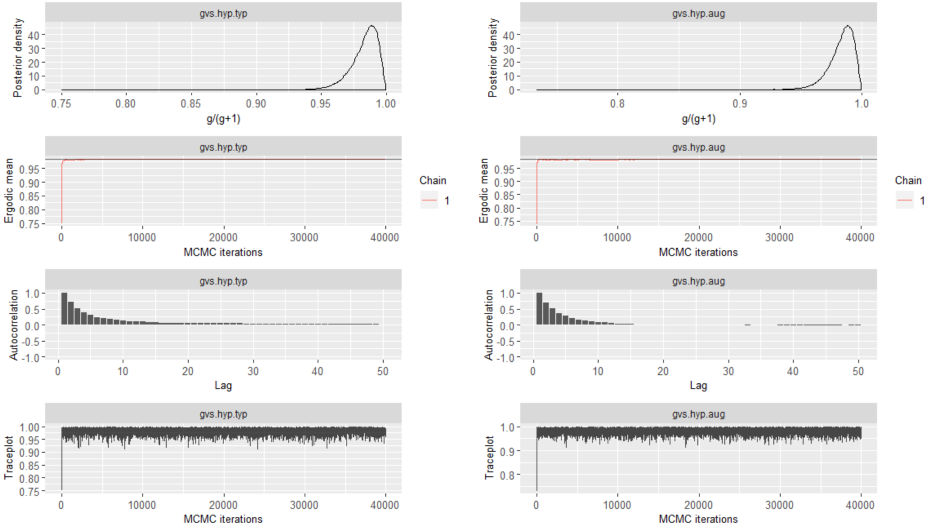
where from the above indices we would expect that the values $\boldsymbol{y}^{*(s)}$ should match as much possible with $\boldsymbol{y}^{te}$ in order to obtain good predictive performance rates; the results are presented in Tables (C.22) and (C.23). Overall, we cannot say that a method prevails to others in terms of predictive performance as the predictions are more or less the same across the three different prior set-ups of typical and augmented logistic regression models with higher accuracy and lower false negative detections in comparison with those of false positive. In particular, regarding the comparisons under the MAP, the highest predictive performance across the pairwise comparisons of typical and augmented logistic regression is observed in the augmented logistic regression models (notice the numbers in bold in column-wise) for GVS under hyper-g, Zellner-Siow, g-prior and only for SSVS under Zellner-Siow, g-prior, whereas for the MPM for GVS under hyper-g, Zellner-Siow and for SSVS under Zellner-Siow, g-prior. Based on this finding, it seems that Bayesian variable selection methods with data augmentation under mixtures of g-priors work well in practice in predicting the women's outcome of being diabetic or not. However, it is not of surprise that the worse Bayesian variable selection methods both for MAP and MPM in terms of predictive performance are observed for those with hyper-g priors due to the complexity of the models. Based on this finding, the MAP under SSVS method with hyper-g priors is sparser with higher predictive ability than GVS due to the additional shrinkage imposed by the shrinkage parameters, whereas for the MPM both methods are sparser with the latter being lower in predictive performance. On the other hand, the MAP and MPM for Bayesian variable selection methods with SSVS under Zellner-Siow, g-prior approaches coincide, whereas for GVS they coincide only for Zellner-Siow. In case of g-prior approach for GVS, the MPM differs from the MAP only by the variable $\boldsymbol{X}_7$ which causes lower predictive accuracy since this covariate is uncertain. Generally, the latter behaviour is expected since SSVS and GVS tend to prefer sparser models under Zellner-Siow and fixed g-prior approach in contradiction with Bayesian variable selection methods with hyper-g priors. Even in that case, SSVS for the three different prior set-ups both for typical and augmented logistic regression show larger predictive performance with respect to GVS prior set-up analogues. Generally, all Bayesian variable selection methods prefer in majority different models of complexity according to each prior set-up for MAP, whereas in the case of MPM all methods trace the same model but with different predictive performance. To conclude, the best predictive performance for the MAP and MPM is observed for ssvs^{R}.g.aug which compromises a technique with data augmentation, but we should be aware that the SSVS in contrast with GVS is more sensitive to large sample error deviating.



(a) ssvs$^{\mathrm{R}}$.hyp.typ.

(b) ssvs$^{\mathrm{R}}$.hyp.aug.

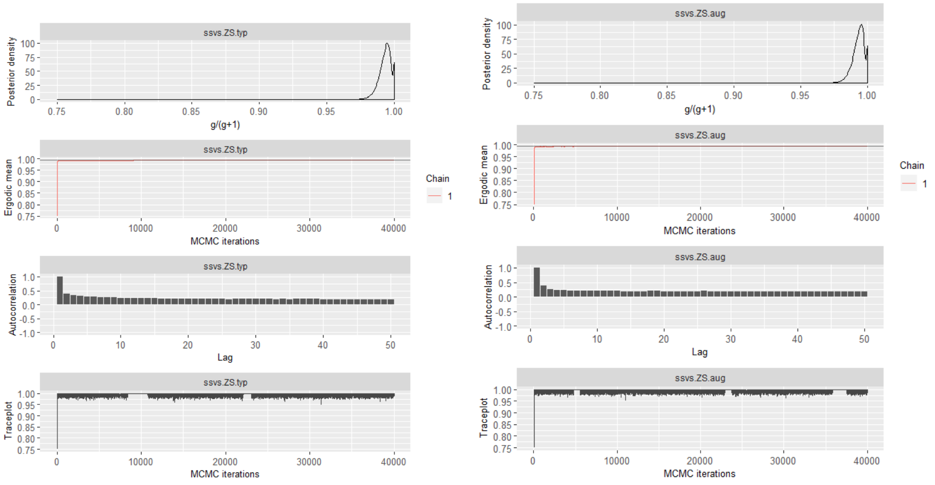
Fig. C.7 Posterior density plots, ergodic mean plots, autocorrelation plots and traceplots of the shrinkage factor $\frac{g}{g+1}$ for typical and augmented logistic regression methods ssvs$^{\mathrm{R}}$.hyp.typ and ssvs$^{\mathrm{R}}$.hyp.aug with hyper-g prior.



(a) gvs$^{\mathrm{R}}$.hyp.typ.

(b) gvs$^{\mathrm{R}}$.hyp.aug.

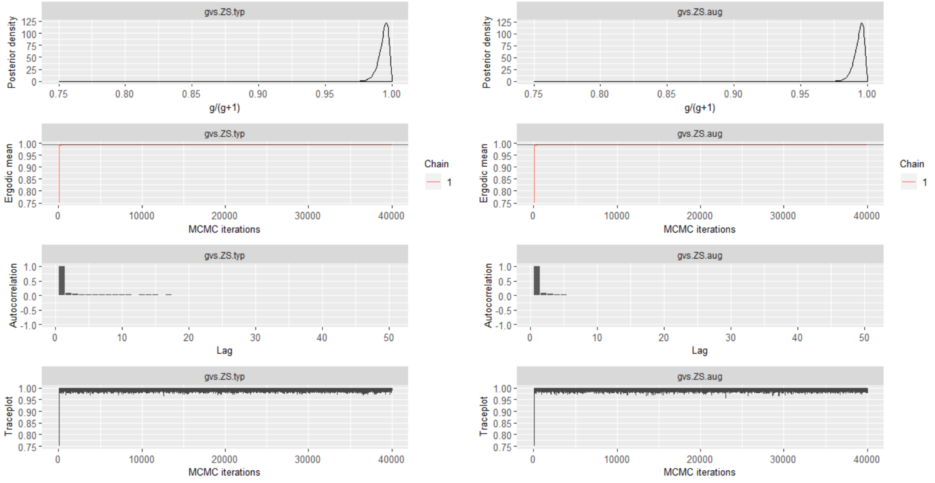
Fig. C.8 Posterior density plots, ergodic mean plots, autocorrelation plots and traceplots of the shrinkage factor $\frac{g}{g+1}$ for typical and augmented logistic regression methods gvs$^{\mathrm{R}}$.hyp.typ and gvs$^{\mathrm{R}}$.hyp.aug with hyper-g prior.



(a) ssvs$^{\mathrm{R}}$.ZS.typ.

(b) ssvs$^{\mathrm{R}}$.ZS.aug.

Fig. C.9 Posterior density plots, ergodic mean plots, autocorrelation plots and traceplots of the shrinkage factor $\frac{g}{g+1}$ for typical and augmented logistic regression methods ssvs$^{\mathrm{R}}$.ZS.typ and ssvs$^{\mathrm{R}}$.ZS.aug with Zellner-Siow prior.



(a) gvs$^{\mathrm{R}}$.ZS.typ.

(b) gvs$^{\mathrm{R}}$.ZS.aug.

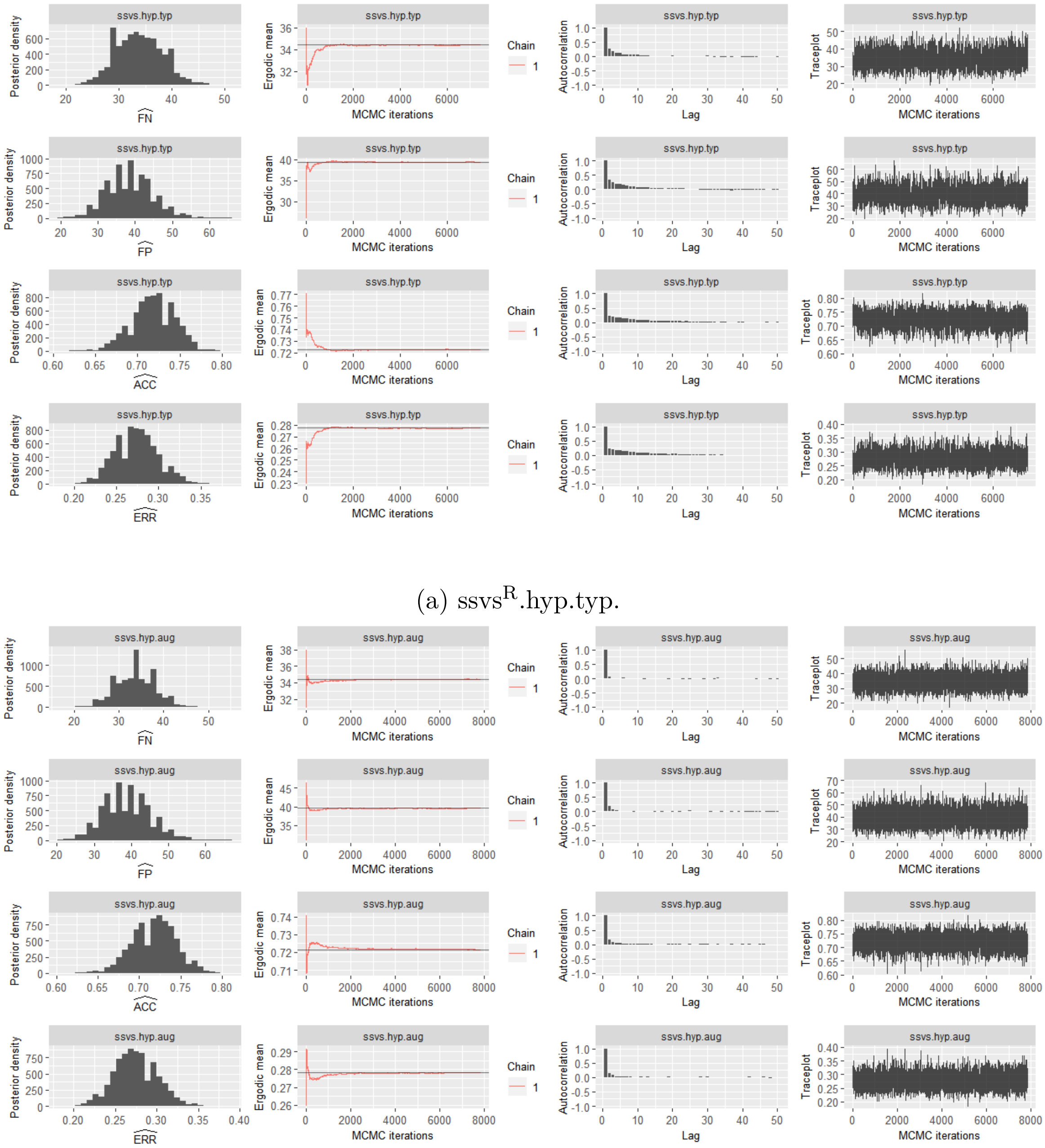
Fig. C.10 Posterior density plots, ergodic mean plots, autocorrelation plots and traceplots of the shrinkage factor $\frac{g}{g+1}$ for typical and augmented logistic regression methods gvs$^{\mathrm{R}}$.ZS.typ and gvs$^{\mathrm{R}}$.ZS.aug with Zellner-Siow prior.

$\mathcal{M}^* : \boldsymbol{X}_1 + \boldsymbol{X}_2 + \boldsymbol{X}_5 + \boldsymbol{X}_6$ **Method**	MAP	$\widehat{FN}$	$\widehat{FP}$	$\widehat{ACC}$	$\widehat{ERR}$
ssvs$^{\mathrm{R}}$.hyp.typ	$\mathcal{M}^*$+$\boldsymbol{X}_7$	**34.465**	**39.385**	**72.236**	**27.763**
ssvs$^{\mathrm{R}}$.hyp.aug	$\mathcal{M}^*$+$\boldsymbol{X}_7$	34.454	39.631	72.148	27.851
gvs$^{\mathrm{R}}$.hyp.typ	$\mathcal{M}^* + \boldsymbol{X}_3 + \boldsymbol{X}_4 + \boldsymbol{X}_7$	34.994	40.093	71.771	28.228
gvs$^{\mathrm{R}}$.hyp.aug	$\mathcal{M}^* + \boldsymbol{X}_3 + \boldsymbol{X}_4 + \boldsymbol{X}_7$	**34.839**	**40.036**	**71.851**	**28.148**
ssvs$^{\mathrm{R}}$.ZS.typ	$\mathcal{M}^* + \boldsymbol{X}_7$	33.972	38.895	72.606	27.393
ssvs$^{\mathrm{R}}$.ZS.aug	$\mathcal{M}^* + \boldsymbol{X}_7$	**33.891**	**38.799**	**72.672**	**27.327**
gvs$^{\mathrm{R}}$.ZS.typ	$\mathcal{M}^* + \boldsymbol{X}_7$	34.963	39.712	71.926	28.073
gvs$^{\mathrm{R}}$.ZS.aug	$\mathcal{M}^* + \boldsymbol{X}_7$	**34.966**	**39.550**	**71.986**	**28.013**
ssvs$^{\mathrm{R}}$.g.typ	$\mathcal{M}^* + \boldsymbol{X}_7$	33.634	38.546	72.864	27.135
ssvs$^{\mathrm{R}}$.g.aug	$\mathcal{M}^* + \boldsymbol{X}_7$	**33.642**	**38.502**	**72.877**	**27.122**
gvs$^{\mathrm{R}}$.g.typ	$\mathcal{M}^*$	33.459	39.566	72.546	27.453
gvs$^{\mathrm{R}}$.g.aug	$\mathcal{M}^*$	**33.417**	**39.566**	**72.562**	**27.437**

Table C.22 Results of false negative $\widehat{FN}$, false positive $\widehat{FP}$, accuracy $\widehat{ACC}$ and missclassification error $\widehat{ERR}$ among all Bayesian variable selection methods under the MAP (column-wise largest value in bold).

$\mathcal{M}^* : \boldsymbol{X}_1 + \boldsymbol{X}_2 + \boldsymbol{X}_5 + \boldsymbol{X}_6$ **Method**	MPM	$\widehat{FN}$	$\widehat{FP}$	$\widehat{ACC}$	$\widehat{ERR}$
ssvs$^{\mathrm{R}}$.hyp.typ	$\mathcal{M}^*$+$\boldsymbol{X}_7$	**34.465**	**39.385**	**72.236**	**27.763**
ssvs$^{\mathrm{R}}$.hyp.aug	$\mathcal{M}^*$+$\boldsymbol{X}_7$	34.454	39.631	72.148	27.851
gvs$^{\mathrm{R}}$.hyp.typ	$\mathcal{M}^* + \boldsymbol{X}_7$	35.500	40.339	71.488	28.511
gvs$^{\mathrm{R}}$.hyp.aug	$\mathcal{M}^* + \boldsymbol{X}_7$	**35.373**	**40.337**	**71.537**	**28.462**
ssvs$^{\mathrm{R}}$.ZS.typ	$\mathcal{M}^* + \boldsymbol{X}_7$	33.972	38.895	72.606	27.393
ssvs$^{\mathrm{R}}$.ZS.aug	$\mathcal{M}^* + \boldsymbol{X}_7$	**33.891**	**38.799**	**72.672**	**27.327**
gvs$^{\mathrm{R}}$.ZS.typ	$\mathcal{M}^* + \boldsymbol{X}_7$	34.963	39.712	71.926	28.073
gvs$^{\mathrm{R}}$.ZS.aug	$\mathcal{M}^* + \boldsymbol{X}_7$	**34.966**	**39.550**	**71.986**	**28.013**
ssvs$^{\mathrm{R}}$.g.typ	$\mathcal{M}^* + \boldsymbol{X}_7$	33.634	38.546	72.864	27.135
ssvs$^{\mathrm{R}}$.g.aug	$\mathcal{M}^* + \boldsymbol{X}_7$	**33.642**	**38.502**	**72.877**	**27.122**
gvs$^{\mathrm{R}}$.g.typ	$\mathcal{M}^* + \boldsymbol{X}_7$	**34.584**	**39.018**	**72.329**	**27.670**
gvs$^{\mathrm{R}}$.g.aug	$\mathcal{M}^* + \boldsymbol{X}_7$	34.597	39.263	72.232	27.767

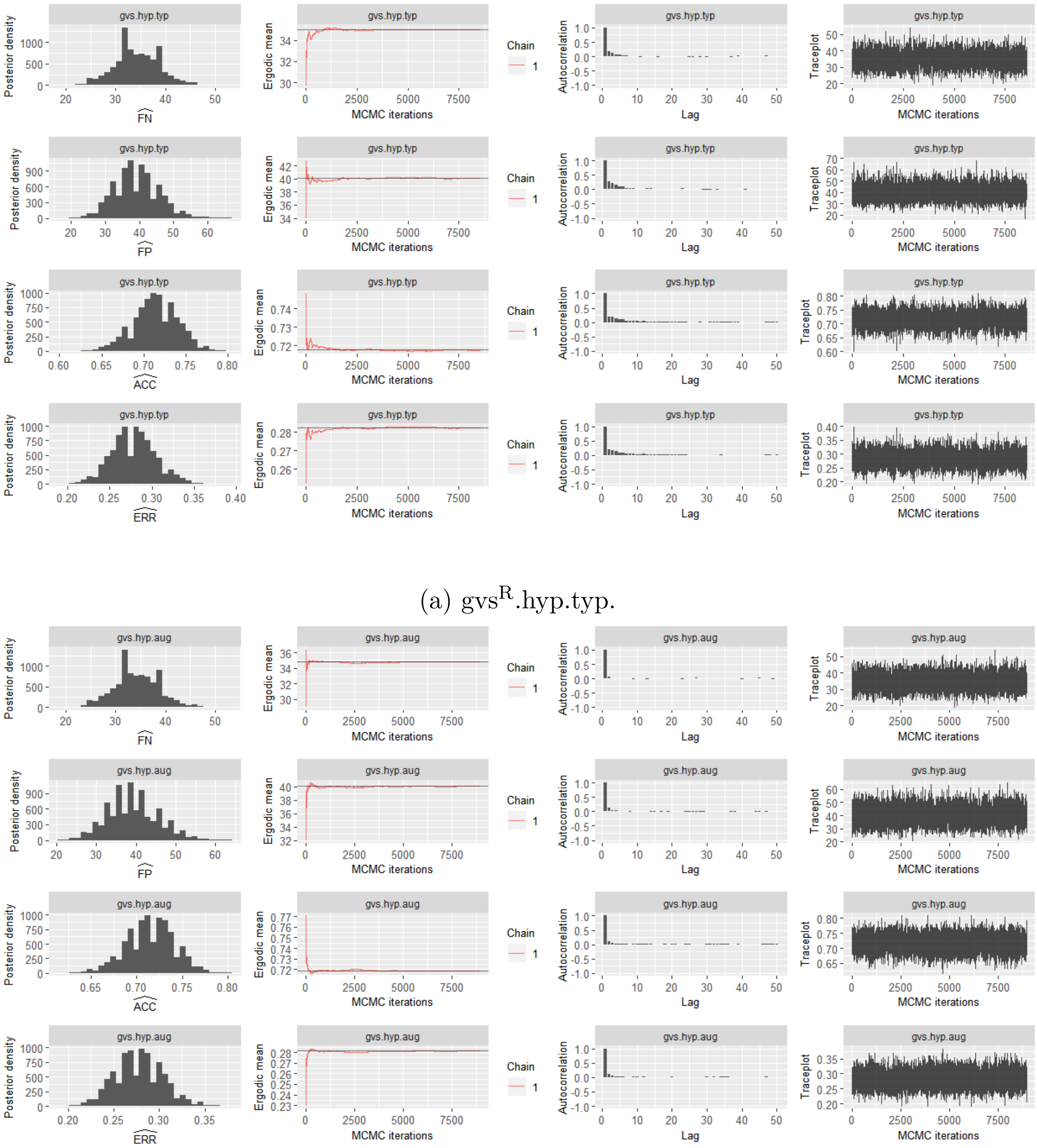
Table C.23 Results of false negative $\widehat{FN}$, false positive $\widehat{FP}$, accuracy $\widehat{ACC}$ and missclassification error $\widehat{ERR}$ among all Bayesian variable selection methods under the MPM (column-wise largest value in bold).



(a) ssvs$^{\mathrm{R}}$.hyp.typ.

(b) ssvs$^{\mathrm{R}}$.hyp.aug.

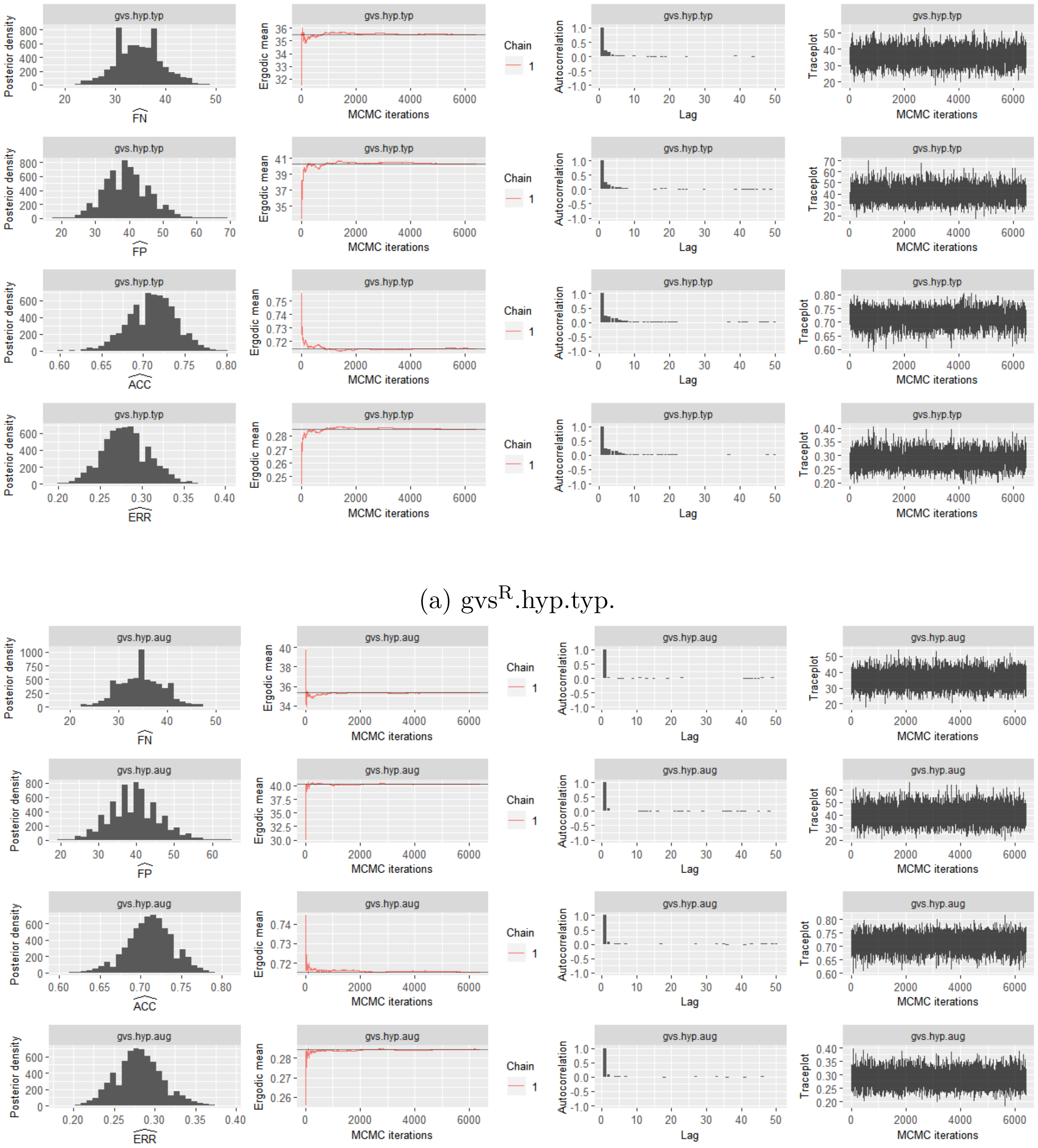
Fig. C.11 Posterior density plots, ergodic mean plots, autocorrelation plots and traceplots of false negative $\widehat{FN}$, false positive $\widehat{FP}$, accuracy $\widehat{ACC}$ and missclassification error $\widehat{ERR}$ for typical and augmented logistic regression methods ssvs$^{\mathrm{R}}$.hyp.typ and ssvs$^{\mathrm{R}}$.hyp.aug with hyper-g priors under the MAP and MPM.



(a) gvs$^{\mathrm{R}}$.hyp.typ.

(b) gvs$^{\mathrm{R}}$.hyp.aug.

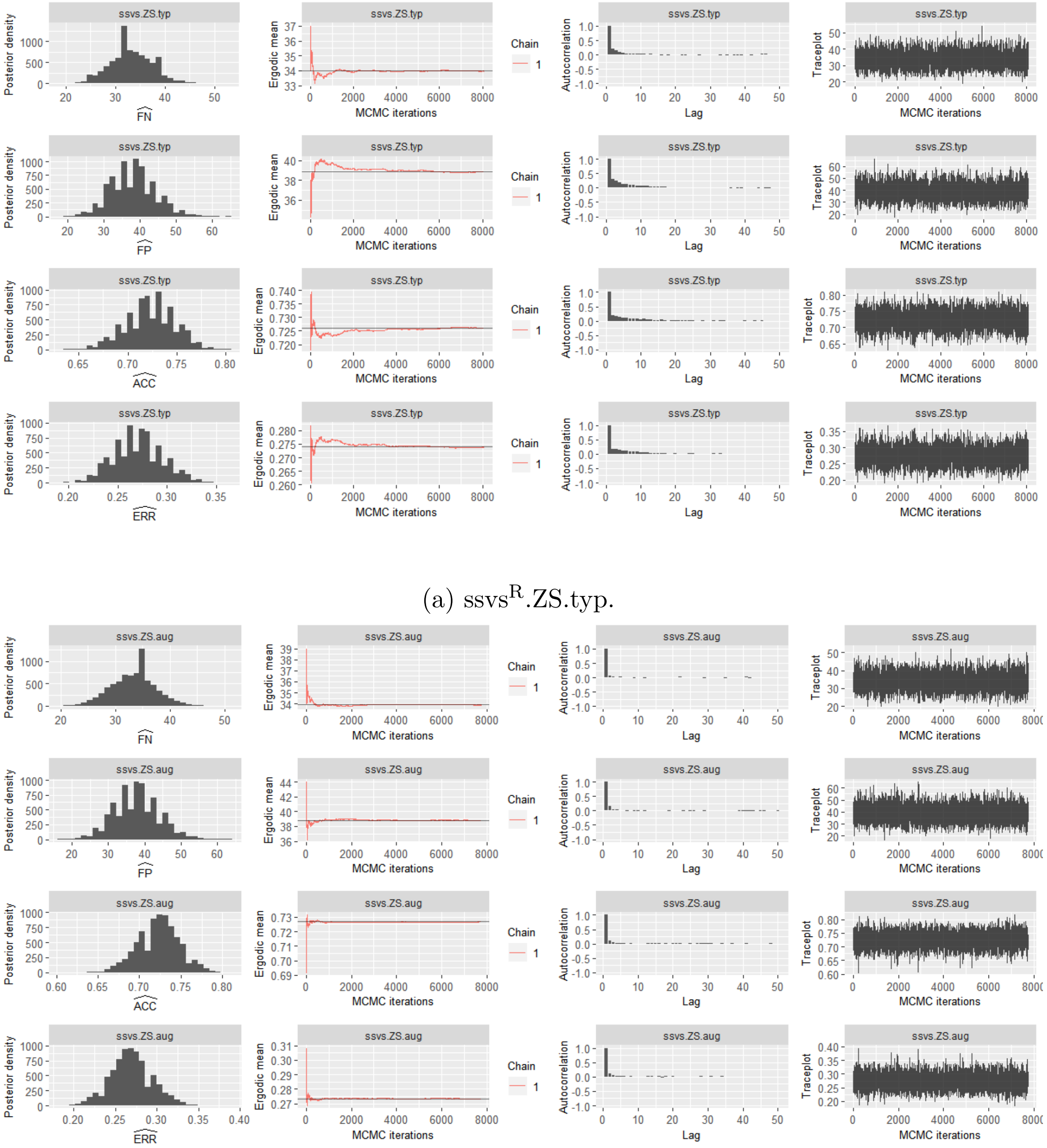
Fig. C.12 Posterior density plots, ergodic mean plots, autocorrelation plots and traceplots of false negative $\widehat{FN}$, false positive $\widehat{FP}$, accuracy $\widehat{ACC}$ and missclassification error $\widehat{ERR}$ for typical and augmented logistic regression methods gvs$^{\mathrm{R}}$.hyp.typ and gvs$^{\mathrm{R}}$.hyp.aug with hyper-g priors under the MAP.



(a) gvs$^{\mathrm{R}}$.hyp.typ.

(b) gvs$^{\mathrm{R}}$.hyp.aug.

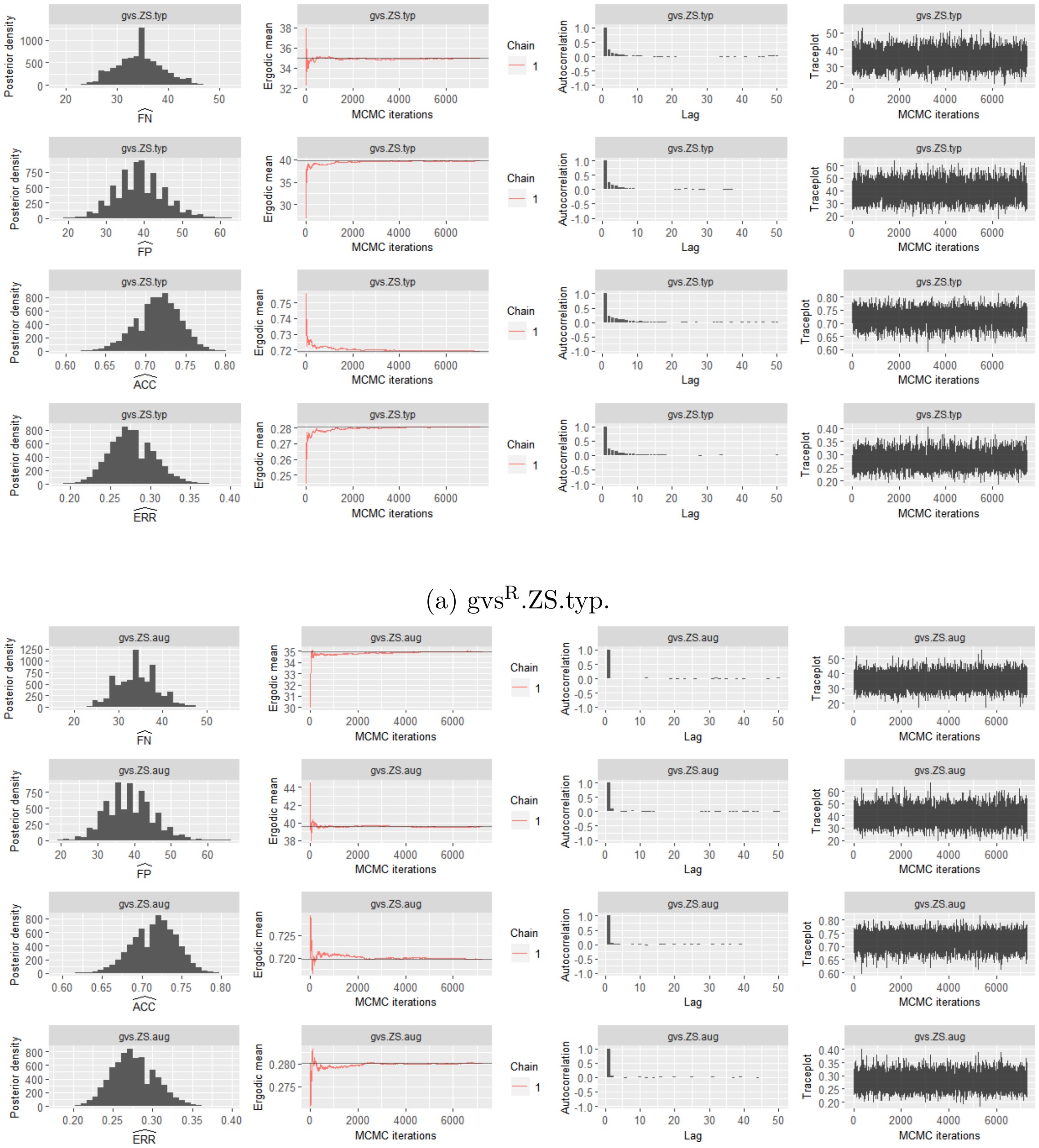
Fig. C.13 Posterior density plots, ergodic mean plots, autocorrelation plots and traceplots of false negative $\widehat{FN}$, false positive $\widehat{FP}$, accuracy $\widehat{ACC}$ and missclassification error $\widehat{ERR}$ for typical and augmented logistic regression methods gvs$^{\mathrm{R}}$.hyp.typ and gvs$^{\mathrm{R}}$.hyp.aug with hyper-g priors under the MPM.



(a) ssvs^{R}.ZS.typ.

(b) ssvs^{R}.ZS.aug.

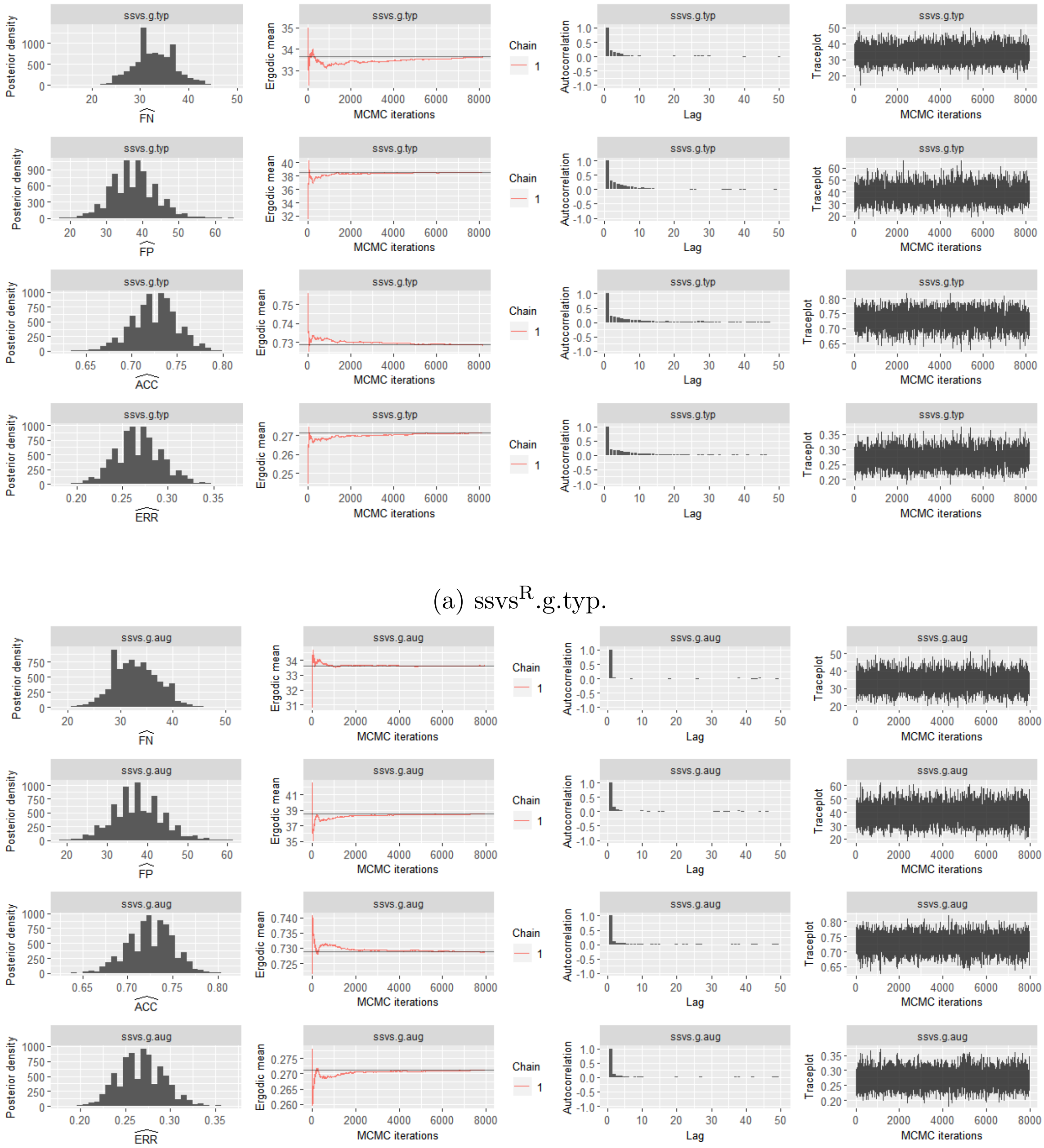
Fig. C.14 Posterior density plots, ergodic mean plots, autocorrelation plots and traceplots of false negative $\widehat{FN}$, false positive $\widehat{FP}$, accuracy $\widehat{ACC}$ and missclassification error $\widehat{ERR}$ for typical and augmented logistic regression methods ssvs^{R}.ZS.typ and ssvs^{R}.ZS.aug with Zellner-Siow priors under the MAP and MPM.



(a) gvs^{R}.ZS.typ.

(b) gvs^{R}.ZS.aug.

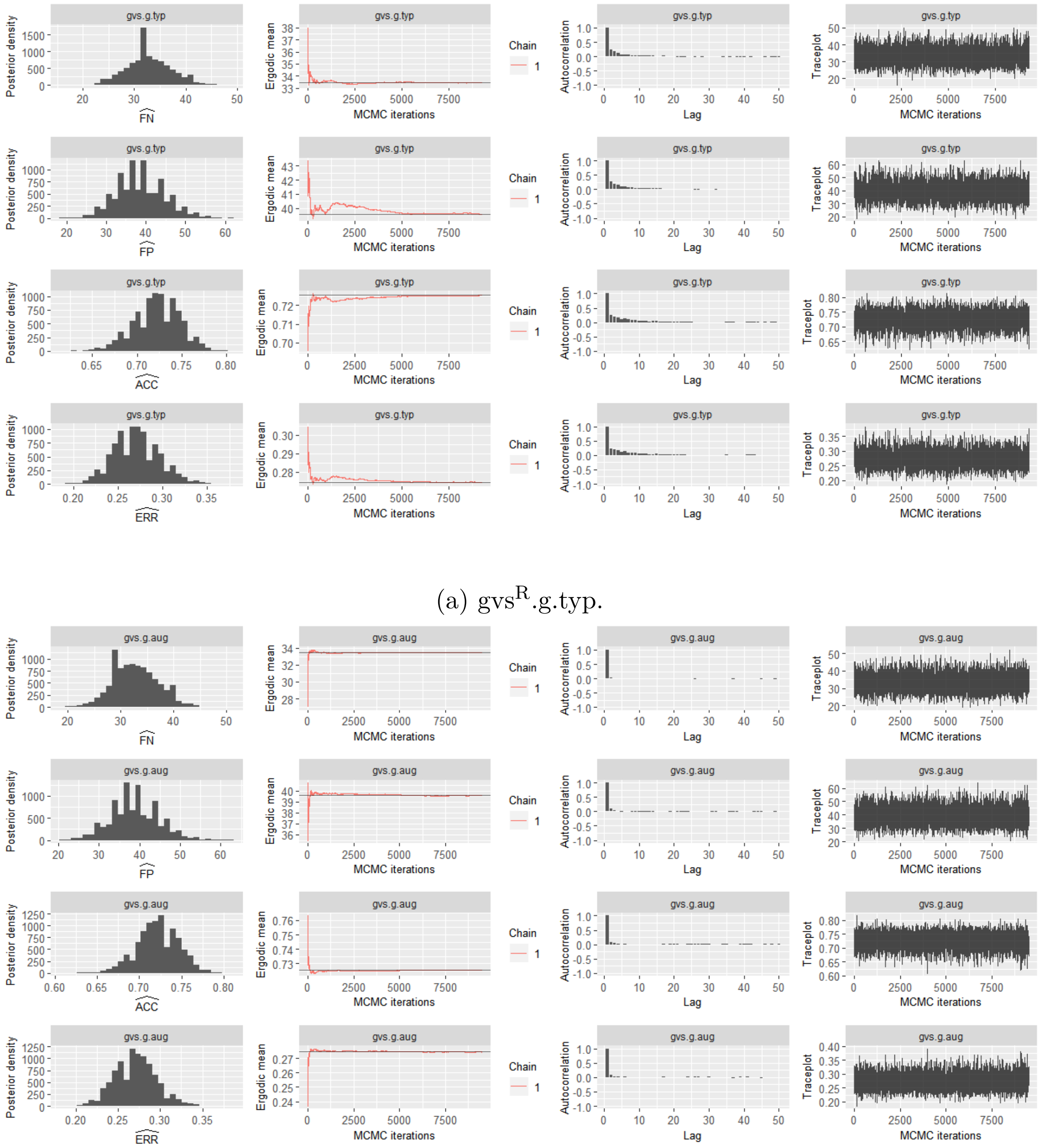
Fig. C.15 Posterior density plots, ergodic mean plots, autocorrelation plots and traceplots of false negative $\widehat{FN}$, false positive $\widehat{FP}$, accuracy $\widehat{ACC}$ and missclassification error $\widehat{ERR}$ for typical and augmented logistic regression methods gvs^{R}.ZS.typ and gvs^{R}.ZS.aug with Zellner-Siow priors under the MAP and MPM.



(a) ssvs$^{\mathrm{R}}$.g.typ.

(b) ssvs$^{\mathrm{R}}$.g.aug.

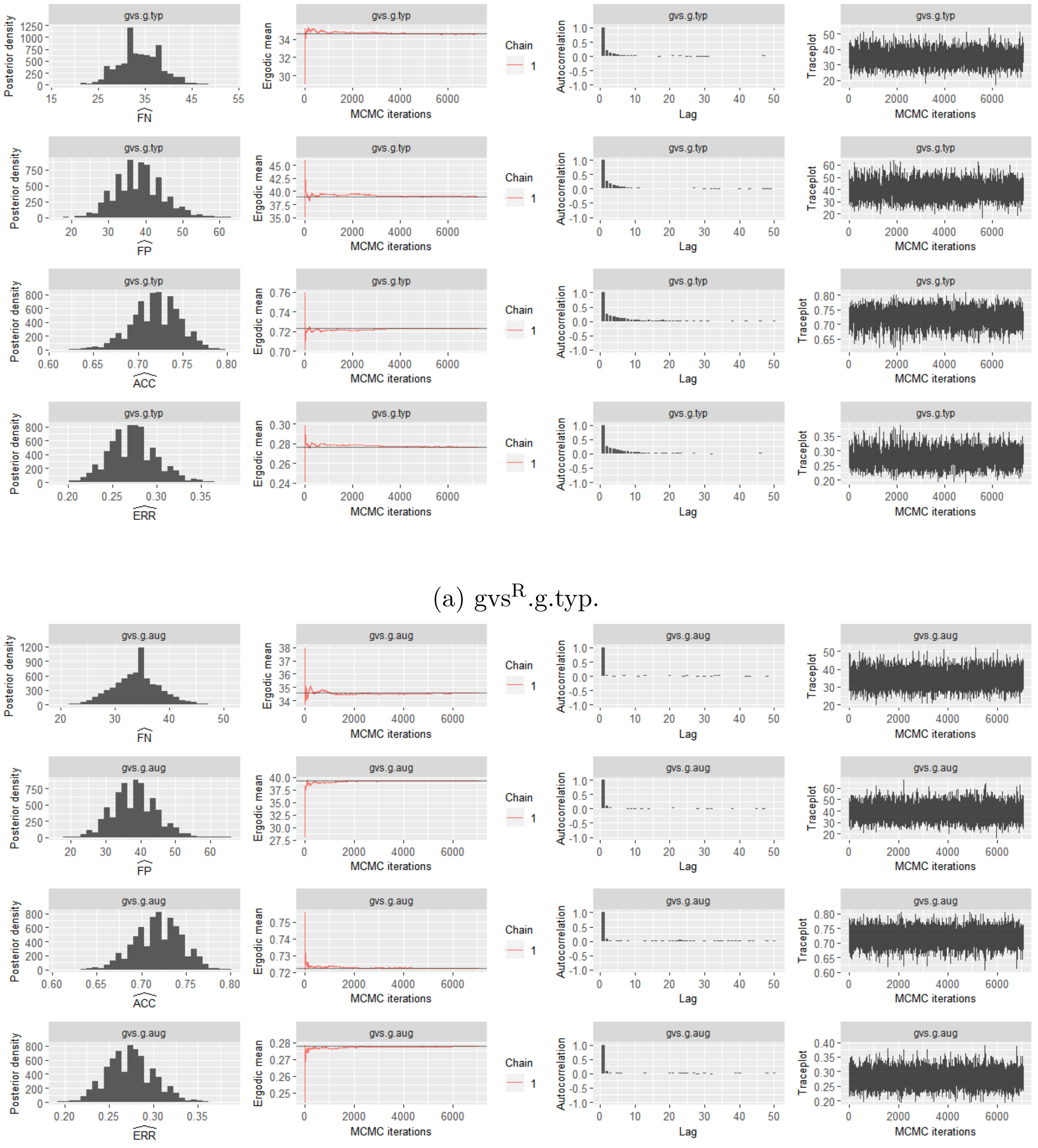
Fig. C.16 Posterior density plots, ergodic mean plots, autocorrelation plots and traceplots of false negative $\widehat{FN}$, false positive $\widehat{FP}$, accuracy $\widehat{ACC}$ and missclassification error $\widehat{ERR}$ for typical and augmented logistic regression methods gvs$^{\mathrm{R}}$.ZS.typ and gvs$^{\mathrm{R}}$.ZS.aug with Zellner-Siow priors under the MAP and MPM.



(a) gvs$^{\mathrm{R}}$.g.typ.

(b) gvs$^{\mathrm{R}}$.g.aug.

Fig. C.17 Posterior density plots, ergodic mean plots, autocorrelation plots and traceplots of false negative $\widehat{FN}$, false positive $\widehat{FP}$, accuracy $\widehat{ACC}$ and missclassification error $\widehat{ERR}$ for typical and augmented logistic regression methods gvs$^{\mathrm{R}}$.g.typ and gvs$^{\mathrm{R}}$.g.aug with g-priors under the MAP.



(a) gvs$^{\text{R}}$.g.typ.

(b) gvs$^{\text{R}}$.g.aug.

Fig. C.18 Posterior density plots, ergodic mean plots, autocorrelation plots and traceplots of false negative $\widehat{FN}$, false positive $\widehat{FP}$, accuracy $\widehat{ACC}$ and missclassification error $\widehat{ERR}$ for typical and augmented logistic regression methods gvs$^{\text{R}}$.g.typ and gvs$^{\text{R}}$.g.aug with g-priors under the MPM.

References

Abramowitz, M. and Stegun, I. (1970). *HANDBOOK OF MATHEMATICAL FUNCTIONS*.

Aijun, Y. and Xinyuan, S. (2010). Bayesian variable selection for disease classification using gene expression data. *Bioinformatics*, 26(2):215–222.

Aijun, Y., Xuejun, J., Liming, X., and Jinguan, L. (2016). Sparse Bayesian variable selection in multinomial probit regression model with application to high-dimensional data classification. *Communications in Statistics - Theory and Methods*, 46(12):6137–6150.

Aitkin, M. (1991). Posterior Bayes Factors. *Journal of Royal Statistical Society*, 53(1):111–142.

Albert, J. H. and Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association'*, 88(422):669—-679.

Allenby, G., Rossi, P., and McCulloch, R. (2005). *Bayesian Statistics and Marketing*. Wiley & Sons.

Arif, M. (2015). Classification of cardiotocograms using random forest classifier and selection of important features from cardiotocogram signal. *Biomaterials and Biomechanics in Bioengineering*, 2(3):173–183.

Ayres-de Campos, D., Bernardes, J., Garrido, A., Marques-de Sa, J., and Pereira-Leite, L. (2000). Sisporto 2.0: a program for automated analysis of cardiotocograms. *Journal of Maternal-Fetal Medicine*, 9(5):311–318.

Bae, K. and Mallick, B. K. (2004). Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics*, 20(18):3423–3430.

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

Bartlett, M. (1957). A Comment on D. V. Lindley's Statistical Paradox. *Biometrika*, 44(3:4):533–534.

Bayarri, M. J., Berger, J. O., Forte, A., and García-Donato, G. (2012). Criteria for bayesian model choice with application to variable selection. *Annals of Statistics*, 40(3):1550–1577.

Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M., and Yakhini, Z. (2000). Tissue classification with gene expression profiles. *Journal of Computational biology*, 7:559–583.

Berger, J. and Pericchi, L. (1996). The Intrinsic Bayes Factor for Model Selection and Prediction. *Annals of Statistics*, 91(433):109–122.

Berger, J. O. and Pericchi, L. R. (2001). Objective Bayesian Methods for Model Selection: Introduction and Comparison. *Lecture Notes-Monograph Series*, 38:135–207.

Bernardo, J. M. (1979). Reference Posterior Distribution for Bayesian Inference (with Discussion). *Journal of Royal Statistical Society*, 41(2):113–147.

Bové, D. S. and Held, L. (2011). Hyper-g priors for generalized linear models. *Bayesian Analysis*, 6(3):387–410.

Bradley, P. S. and Mangasarian, O. L. (1998). Feature Selection via Concave Minimization and Support Vector Machines. *Proceedings of the 15th International Conference on Machine Learning (ICML 1998)*, (6):82–90.

Brown, P. J. and Vannucci, M. (1998). Multivariate Bayesian variable selection and prediction. *Journal of Royal Statistical Society*, 60(3):627–641.

Carlin, B. and Chib, S. (1995). Bayesian Model Choice via Markov Chain Monte Carlo Methods. *Journal of Royal Statistical Society*, 57(3):473–484.

Carlin, B. and Louis, T. (1996). *Bayes and Empirical Bayes Methods for Data Analysis.*

Casella, G. and Moreno, E. (2006). Objective Bayesian variable selection. *Journal of the American Statistical Association*, 101(473):157–167.

Chakraborty, S. (2009). Bayesian binary kernel probit model for microarray based cancer classification and gene selection. *Computational Statistics and Data Analysis*, 53(12):4198–4209.

Chen, M. and Ibrahim, J. (2003). Conjugate priors for generalized linear models. *Statistica Sinica*, 13(2):461–476.

Chen, M. H., Huang, L., Ibrahim, J. G., and Kim, S. (2008). Bayesian variable selection and computation for generalized linear models with conjugate priors. *Bayesian Analysis*, 3(3):585–614.

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

Chipman, H., George, E. I., and McCulloch, R. E. (2001). The Practical Implementation of Bayesian Model Selection. *IMS Lecture Notes - Monograph Series*, 38:65–116.

Choi, H. M., Hobert, J. P., et al. (2013). The polya-gamma gibbs sampler for bayesian logistic regression is uniformly ergodic. *Electronic Journal of Statistics*, 7:2054–2064.

Choi, H. M. and Román, J. C. (2017). Analysis of Polya-Gamma gibbs sampler for Bayesian logistic analysis of variance. *Electronic Journal of Statistics*, 11(1):326–337.

Clyde, M., Ghosh, J., and Littman, M. L. (2011). Bayesian Adaptive Sampling for Variable Selection and Model Averaging. *Journal of Computational and Graphical Statistics*, 20(1):80–101.

Consonni, G., Fouskakis, D., Liseo, B., and Ntzoufras, I. (2018). Prior distributions for objective Bayesian analysis. *Bayesian Analysis*, 13(2):627–679.

Consonni, G. and Veronese, P. (1992). Conjugate priors for exponential families having quadratic variance functions. *Journal of the American Statistical Association*, 87(420):1123–1127.

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

Datta, G., Ghosh, M., Journal, S., Statistical, A., and Dec, N. (1995). Some Remarks on Noninformative Priors. *Journal of the American Statistical Association*, 90(432):1357–1363.

Datta, G. S. and Ghosh, M. (1996). On the Invariance of Noninformative Priors. *Annals of Statistics*, 24(1):141–159.

Davis, P. J. and Rabinowitz, P. (1986). *Methods of Numerical Integration*, volume 70.

Dawid, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika*, 68(1):265 – 274.

Dellaportas, P., Forster, J., and Ntzoufras, I. (2000). Bayesian variable selection using the Gibbs sampler. *Generalized linear models: a Bayesian perspective*, 5:273–286.

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2002). On Bayesian model and variable selection using MCMC. *Statistics and Computing*, 12(1):27–36.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Dempster, Arthur P and Laird, Nan M and Rubin, Donald B. *Journal of the Royal Statistical Society: Series B (statistical methodology)*, 39(1):1–38.

Dettling, M. (2004). BagBoosting for tumor classification with gene expression data. *Bioinformatics*, 20(18):3583–3593.

Dey, D., Ghosh, S., and Mallick, B. (1999). *Generalized Linear Models : A Bayesian Perspective*. Marsel Decker.

DiCiccio, T., Kass, R., and Wasserman, L. (2006). Computing Bayes factors by combining simulation and asymptotic approximations. *Journal of the Royal Statistical Society*, 92(439):903–915.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(457):77–86.

Dunson, D. (2006). *Efficient Bayesian model averaging in factor analysis.* Isds discussion paper, Duke University.

Durante, D., Canale, A., and Rigon, T. (2018). A nested expectation–maximization algorithm for latent class models with covariates. *Statistics and Probability Letters*, 146:97–103.

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

Forte, A. (2014). *Objective Bayes Criteria for Variable Selection.* PhD thesis, University of Valencia.

Foster, D. and George, E. (1994). The Risk Inflation Criterion For Multiple Regression. *The Annals of Statistics*, 22(4):1947–1975.

Fouskakis, D. and Ntzoufras, I. (2012). Power-Expected-Posterior Priors for Variable Selection in Gaussian Linear Models. *Technical Notes*, pages 1–36.

Fouskakis, D. and Ntzoufras, I. (2013). Computation for intrinsic variable selection in normal regression models via expected-posterior prior. *Statistics and Computing*, 23(4):491–499.

Fouskakis, D., Ntzoufras, I., and Perrakis, K. (2018). Power-Expected-Posterior Priors for Generalized Linear Models. *Bayesian Analysis*, 13(3):721–748.

Frühwirth-Schnatter, S. (2016). Bayesian Inference in the Multinomial Logit Model. *Austrian Journal of Statistics*, 41(1):27–43.

Gelfand, A. and Dey, D. (1994). Bayesian Model Choice : Asymptotics and Exact Calculations. *Journal of Royal Statistical Society*, 56(3):501–514.

George, E. and Foster, D. (2000). Calibration and Empirical Bayes Variable Selection. *Biometrika*, 87(4):731–747.

George, E. I. and McCulloch, R. E. (1993). Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Ghosh, J., Herring, A. H., and Siega-Riz, A. M. (2011). Bayesian Variable Selection for Latent Class Models. *Biometrics*, 67(3):917–925.

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

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–527.

Gordy, M. B. (1998). A generalization of generalized beta distributions.

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

Gupta, M. and Ibrahim, J. G. (2009). An Information Matrix Prior for Bayesian Analysis in Generalized Linear Models with High Dimensional Data. *Statistica Sinica*, 19:1641–1663.

Gustafson, P. and Lefebvre, G. (2008). Bayesian multinomial regression with class-specific predictor selection. *Annals of Applied Statistics*, 2(4):1478–1502.

Guyon, I., Weston, J., and Barnhill, S. (1992). Gene Selection for Cancer Classification using Support Vector Machines. *Mach. Learn.*, 46:389–422.

Hansen, M. H. and Yu, B. (2003). Minimum description length model selection criteria for generalized linear models. *Lecture Notes - Monograph Series*, 3(21):145–163.

Held, L., Bové, D. S., and Gravestock, I. (2015). Approximate Bayesian Model Selection with the Deviance Statistic. *Statistical Science*, 30(2):242–257.

Hoeting, J. J. A., Madigan, D., Raftery, A. E. A., and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical science*, 14(4):382–401.

Holmes, C. C. and Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1(1 A):145–168.

Holmes, C. C. and Mallick, B. K. (1998). Bayesian radial basis functions of variable dimension. *Neural Computation*, 10(5):1217–1233.

Ibrahim, J. G. and Chen, M.-H. (2000). Power Prior Distributions for Regression Models. *Statistical Science*, 15(1):46–60.

Ibrahim, J. G. and Laud, P. (1991). On Bayesian Analysis of Generalized Linear Models Using Jeffreys ' s. *Journal of the American Statistical Association*, 86(416):981–986.

Jaynes, E. (2003). *Probability Theory.* Cambridge University Press.

Jeffreys, H. (1961). *Theory of Probability (3rd edition).*

Kamath, R. and Kamat, R. (2016). Random forest modelling for cardiotocography data: A case study on machine learning with sparkr. *RESEARCH JOURNAL OF PHARMACEUTICAL BIOLOGICAL AND CHEMICAL SCIENCES*, 7(6):584–590.

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

Kass, R. E. and Wasserman, L. (1995). A Reference and Bayesian Test for Nested the Schwarz Hypotheses Criterion and its Relationship to the Schwarz Criterion. *Journal of the American Statistical Association*, 90(431):928–934.

Kuo, L. and Mallick, B. K. (1998). Variable Selection for Regression Models. *Sankhyā: The Indian Journal of Statistics, Series B*, 60(1):65–81.

Kyeong, E. L., Naijun, S., Edward, R. D., Martina, V., and Bani, K. M. (2003). Gene selection: A bayesian variable selection approach. *Bioinformatics*, 19(1):90–97.

Leng, C., Tran, M. N., and Nott, D. (2014). Bayesian adaptive Lasso. *Annals of the Institute of Statistical Mathematics*, 66(2):221–244.

Ley, E. and Steel, M. F. (2012). Mixtures of g-priors for Bayesian model averaging with economic applications. *Journal of Econometrics*, 171(2):251–266.

Li, Q. and Lin, N. (2010). The Bayesian elastic net. *Bayesian Analysis*, 5(1):151–170.

Li, Y. and Clyde, M. A. (2013). *Bayesian Hierarchical Models for Model Choice*. PhD thesis, Duke University.

Li, Y. and Clyde, M. A. (2018). Mixtures of g-Priors in Generalized Linear Models. *Journal of the American Statistical Association*, 113(524):1828–1845.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423.

Linderman, S., Johnson, M. J., and Adams, R. P. (2015). Dependent multinomial models made easy: Stick-breaking with the polya-gamma augmentation. In *Advances in Neural Information Processing Systems*.

Lindley, D. (1957). A Statistical Paradox. *Biometrika*, 44(1):187–192.

Madigan, D. and Raftery, A. E. (1994). Model Selection and Accounting for Model Uncertainty in Graphical Models using Occam's Window. *Journal of the American Statistical Association*, 89(428):1535–1546.

Mallick, B. K., Ghosh, D., and Ghosh, M. (2003). Bayesian Classification of Tumors Using Gene Expression Data. *Journal of Royal Statistical Society B*, 67(2):219–234.

Marin, J. M. and Robert, C. P. (2014). *Bayesian Essentials with R - 2nd Edition*.

Maruyama, Y. and George, E. I. (2011). Fully bayes factors with a generalized g-prior. *Annals of Statistics*, 39(5):2740–2765.

Mavridis, D. and Ntzoufras, I. (2014). Stochastic search item selection for factor analytic models. *British Journal of Mathematical and Statistical Psychology*, 67(2):284–303.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032.

Moreno, E. and Girón, F. J. (2008). Comparison of Bayesian objective procedures for variable selection in linear regression. *Test*, 17(3):472–490.

Morris, C. (1983). Natural Exponential Families with Quadratic Variance Functions : Statistical Theory. *Annals of Statistics*, 11(2):515–529.

Naylor, J. and Smith, A. (19). Application of a Method for the Efficient Computation of Posterior Distributions. *Applied Statistics*, 31(3):214–225.

Nguyen, D. V. and Rocke, D. M. (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, 18(1):39–50.

Ntzoufras, I., , P., and Forster, J. J. (2003). {MCMC} Variable and Link Determination for Generalized Linear Models. 111(65):165–180.

Ntzoufras, I. (1999). *Aspects of Bayesian Model and Variable Selection Using MCMC*. PhD thesis, Athens university of economics and business.

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

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

O'Hagan, A. (1995). Fractional Bayes Factors for Model Comparison. *Journal of Royal Statistical Society. Series B Methodological*, 57:99–138.

Pan, W. (2002). A comparative reviewofstatistical methods for discovering differentially-expressed genes in replicated microarrayexperiments. *Bioinformatics*, 18(4):546–554.

Panagiotelis, A. and Smith, M. (2008). Bayesian identification, selection and estimation of semiparametric functions in high-dimensional additive models. *Journal of Econometrics*, 143(2):291–316.

Park, T. and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Paroli, R. and Spezia, L. (2006). VARIABLE SELECTION IN A CLASS OF BAYESIAN TIME SERIES MODELS. *Quaderni-Dipartimento*.

Pereira, A., Salgado, F., Reis, L. P., and Faria, B. M. (2016). Classification model for cardiotocographies. In *2016 11th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–6. IEEE.

Pérez, J. M. and Berger, J. O. (2002). Expected-Posterior Prior Distributions for Model Selection. *Biometrika*, 89(3):491–511.

Perrakis, K. and Ntzoufras, I. (2015). Stochastic Search Variable Selection (SSVS). *Wiley StatsRef: Statistics Reference Online*, pages 1–6.

Perrakis, K. and Ntzoufras, I. (2018). Bayesian variable selection using the hyper-g prior in WinBUGS. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(6):1–13.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using Polya-Gamma latent variables. *Journal of the American Statistical Association'*, 108:1339–1349.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association*, 92(437):179–191.

Ripley, B. (1996). *Pattern recognition and neural networks.* Cambridge University Press.

Roberts, G. O. and Rosenthal, J. S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16(4):351–367.

Ročková, V. and George, E. I. (2014). EMVS: The EM Approach to Bayesian Variable Selection. *Journal of the American Statistical Association*, 109(506):828–846.

Salzer, H. E. and Zucker, R. (1952). Table of the zeros and weight factors of the first fifteen Laguerre polynomials. *Bulletin of the American Mathematical Society*, 55(10):1004–1013.

Savage, L. (1954). The Foundations of Statistical Inference. *John Wiley*, 629.

Scott, J. G. and Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Annals of Statistics*, 38(5):2587–2619.

Sha, N., Vannucci, M., Tadesse, M. G., Brown, P. J., Dragoni, I., Davies, N., Roberts, T. C., Contestabile, A., Salmon, M., Buckley, C., and Falciani, F. (2004). Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. *Biometrics*, 60(3):812–819.

Smith, A. and Spiegelhalter, D. (1980). Bayes Factors and Choice Criteria for Linear Models. *Journal of the Royal Statistical Society*, 42(2):213–220.

Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, 75(2):317–343.

Spiegelhalter, D. and Smith, A. (1982). Bayes Factors for Linear and Log-Linear Models with Vague Prior Information. *Journal of the Royal Statistical Society*, 44(3):377–387.

Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. (2003). *WinBUGS User Manual.*

Stamey, T. A., Kabalin, J. N., McNeal, J. E., Johnstone, I. M., Freiha, F., Redwine, E. A., and Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. Radical prostatectomy treated patients. *Journal of Urology*, 141(5):1076–1083.

Strawderman, W. E. (1971). Proper Bayes Minimax Estimators of the Multivariate Normal Mean. *The Annals of Mathematical Statistics*, 42(1):385–388.

Tanner, M. and Wong, W. H. (1987). The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*, 82(398):528–540.

Tierney, L., Kass, R. E., Kadane, J. B., and Kass, E. (1989). Fully Exponential Laplace Approximations to Expectations and Variances of Nonpositive Functions. *Journal of the American Statistical Association*, 84(407):710–716.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc.Natl.Acad.Sci.USA*, 98(9):5116–5121.

Verdinelli, I. and Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *Journal of the American Statistical Association*, 90(430):614–618.

Wang, X. and George, E. (2007). Adaptive Bayesian criteria in variable selection for generalized linear models. *Statistica Sinica*, 17(3):667–690.

Yau, P., Kohn, R., and Wood, S. (2003). Bayesian variable selection and model averaging in high-dimensional multinomial nonparametric regression. *Journal of Computational and Graphical Statistics*, 12(1):23–54.

Zellner, A. (1986). "On Assessing Prior Distributions and Bayesian Regression Analysis With g-Prior distributions.".

Zellner, A. and Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. *Bayesian Statistics 1*, 31(1):585–603.

Zhou, X., Wang, X., and Dougherty, E. (2006). Multi-class cancer classification using multinomial probit regression with Bayesian gene selection. *IEE Proc-Syst.Biol.*, 153:70–78.

Zucknick, M. and Richardson, S. (2014). Mcmc algorithms for bayesian variable selection in the logistic regression model for large-scale genomic applications. *arXiv preprint arXiv:1402.2713*.