Bayesian Variable Selection for Linear and Generalized Linear Models

Paraskevopoulou Chrysoula

February 5, 2019

Contents

Abstract:

Model selection is a statistical procedure concerning the comparison of various models with respect to their ability to describe a particular set of data. It sustains a field of undiminished interest in both Classical and Bayesian Statistics. The present thesis focuses entirely on the Bayesian approach of Model Selection, which includes a variety of analytical and computational methods. The employment of MCMC techniques has been of vital importance regarding the development of computational Model Selection methods in the Bayesian framework, therefore, the fundamental principles and the most popular algorithms are explicitly outlined. We attempted to make a thorough overview of the most widely known Bayesian methods for Model Selection, which include the derivation of Bayes' factors, the BIC and the DIC criteria and the L-Measure, all of which may be effective, but often prohibitively time-consuming when implemented on high-dimensional models.

Gibbs Variable Selection methods sustain a class of computational methods based on the Gibbs Sampler and their development was motivated by the imperative necessity to speed up and, if possible, automize the process of model selection for complex models. The member of this class of methods studied in the present thesis is Stochastic Search Variable Selection, originally introduced by E. I. George and R. I. MacCullogh in 1993. We explicitly present the theoretical foundation of the method, as well as its implementation on the Normal Linear model and the Logit and Probit models. Additionally, the last chapter includes a real data example of the implementation of SSVS in a logistic regression model.

Chapter 1

Basics of Bayesian Statistics

1.1 Fundamental principles and notation

The cornerstone of Bayesian statistical theory is a well-known and rather simple probabilistic theorem called *Bayes' Theorem* or *Bayes' Rule* presented below for probabilities of events:

Theorem 1 (Bayes' Theorem)**.** *Let us consider two possible outcomes A and B. Then, the conditional probability of A given B, which expresses the probability of A taking place, provided that B has occurred, is equal to:*

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \propto P(B|A) \cdot P(A)$$

Furthermore, if A is such that $A = A_1 \cup A_2 \cup \cdots \cup A_n$, whereas $A_i \cap A_j = \varnothing$ then the conditional probability of $A_i, i = 1, \ldots, n$ given B, is:

$$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{P(B)} \propto P(B|A_i) \cdot P(A_i),$$

$$\textit{where} \qquad P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i).$$

1.2 General perception

1.2.1 Prior distributions

The fundamental difference between Bayesian and Classical Statistics lies in the way an unknown parameter $\boldsymbol{\theta}$ is perceived and handled. In

the Bayesian paradigm, $\boldsymbol{\theta}$ is treated as a ***random variable*** and not as an unknown constant quantity and, therefore, a probability distribution has to be assigned to it. This density function is called ***prior distribution*** and it is denoted as $\boldsymbol{\pi(\theta)}$. It is important to note that the prior distribution $\boldsymbol{\pi}$ is set by the statistician *before* any kind of analysis of the data takes place. In order to determine the optimal distributional form, the statistician takes into consideration the pre-existing beliefs concerning the data and "weighs" their credibility.

Priors are classified, according to the influence they should have on the data, as **informative**, **weakly informative** or **diffused** and **non-informative** or **flat**. The use of informative prior distributions leads to an active involvement of prior beliefs in the analysis, like having a "second source of information", apart from the data, figuratively speaking. On the opposite hand, flat priors seem to have zero-impact on the outcome. A frequently used rationale for their use is "to let the data speak for themselves" (*Bayesian Data Analysis*, Gelman, Carlin et. al, 2014). Weakly informative priors lie in the middle between the two previous classes in terms of influence; the information provided could perhaps be considered as a subtle hint for the analysis, which affects the results, but in no case attempts to "fully capture one's complete scientific knowledge about the studied parameter" (Gelman et al., 2014). Examples and further elaboration about the above classes of distributions will be provided in the following chapters.

1.2.2 Extracting information from the data: The posterior distribution

In the Bayesian framework, the first step in the analysis is setting a probabilistic model that describes in the best possible way the phenomenon under study. Then, we place a prior distribution on the parameter we wish to conduct inference about. The notation used to denote the likelihood of the data is $\boldsymbol{f(data|\theta)}$.

The prior distribution is then combined with the likelihood of the data to update our knowledge about the parameters based on the information contained in the likelihood function. This is achieved by the derivation and the further study of a distribution that summarises the update due to the data. This distribution is called the ***posterior distribution*** and it is obtained by applying Bayes' Theorem, as it is expressed for probability distributions. More specifically, by using Lynch's (2007) more simple and

clarified notation, the posterior distribution, denoted as $\boldsymbol{f(\theta|data)}$, is obtained by the following fraction:

$$f(\theta|data) = \frac{f(data|\theta) \cdot \pi(\theta)}{f(data)}.$$

The denominator of the fraction is the ***marginal distribution*** or ***evidence*** of the data and it is derived from the integral:

$$f(data) = \int_{\Theta} f(data|\theta) \cdot \pi(\theta) d\theta,$$

where Θ is the parameter space of θ.

In mathematical notation, the posterior distribution of $\boldsymbol{\theta}$, assuming that it is a continuous random variable, satisfies the condition

$$\int_{\Theta} f(\theta|data) d\theta = 1 \quad .$$

The analytic calculation of the integral

$$\int_{\Theta} f(data|\theta) \cdot \pi(\theta) d\theta$$

is challenging and as the dimension of the parameter space Θ increases, it turns into a complicated and time consuming process that usually demands extensive integration skills. In order to avoid these intensive calculations, the posterior distribution is determined *up to a normalising constant*, which means that we only need to compute a function, to which the posterior is *proportional to*. Consequently, the marginal distribution of the data does not have to be computed and we are restricted to its expression via proportionality. We should note that we are completely justified to omit the denominator, since $\int_{\Theta} f(data|\theta) \cdot \pi(\theta) d\theta \propto 1$ is a constant regarding $\boldsymbol{\theta}$ and therefore we do not lose any substantial information concerning its posterior distribution.

However, the calculation of the posterior up to a constant proportionality is not sufficient; we have to be more specific in order to be able to make any sort of conclusions (at this point, we cannot even get the basic descriptive characteristics, such as the mean, the median or the variance). Depending on the functional form we have arrived at, the following two options are available to us: we can either determine analytically the posterior density distribution, or we will turn to simulation methods, based on the thinking that by obtaining a large number of

draws from the unknown posterior, we will be able to figure out its main features and properties.

Analytic Computation of the posterior: Conjugate Prior Distributions

The analytic computation of a posterior distribution becomes much more simple and in some cases even possible, by the use of functions belonging to the family of **conjugate distributions**, for which we provide the following definition based on the notation we have previously used.

Definition 1.2.1. Suppose we wish to study a parameter $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^k$, with $k \in \mathbb{N}$ and we have derived the likelihood $f(data|\boldsymbol{\theta})$ of the data, which belongs to a class of probability densities denoted by $\mathcal{F}$. We have also placed a prior on $\boldsymbol{\theta}$ denoted as $\pi(\boldsymbol{\theta})$, that belongs to a parameterized family of distributions, denoted as Π. Π is said to be **conjugate** or **closed under sampling** for $\mathcal{F}$, if for every prior $\pi \in \Pi$, the posterior distribution $f(\boldsymbol{\theta}|data)$ also belongs to Π for every $f \in \mathcal{F}$.

Conjugate prior distributions constitute a very useful mathematical device, precisely due to the property described in the definition. The fact that the analysis always leads to a posterior that belongs to the same distributional family, is a great computational advantage, because basic computational skills are required and most importantly, the functional form of the posterior is always known. What is more, if the likelihood function belongs to the exponential family, the posterior can be derived in a more straightforward way, according to the following proposition, which actually states that each distribution of this particular family has a "natural conjugate prior distribution" (Gelman et al., 2014). We note that the form of an exponential-family distribution we will hereby provide, serves solely the purpose of a rough, first understanding of the context of the proposition and will not be used in the next chapters.

Proposition 1. *Suppose that $\theta \in \Theta \subset \mathbb{R}$ and the likelihood of the data denoted as $f_{X|\theta}(x|\theta)$ belongs to the exponential family.*
Then the likelihood referring to one observation x_i will be given by

$$f_{X|\theta}(x|\theta) = h(x_i)e^{\theta \cdot x_i - \psi(\theta)}$$

Consequently, the likelihood corresponding to a random sample, $X =$

$(x_1, ..., x_n)$, *is*

$$f_{X|\theta}(x|\theta) = \prod_{i=1}^{n} h(x_i) e^{\theta \cdot \sum_{i=1}^{n} x_i - n\psi(\theta)}.$$

The conjugate prior $\pi(\theta)$ *will have the form*

$$\pi(\theta) \equiv \psi(\theta|\mu, \lambda) = K(\mu, \lambda) \cdot e^{\theta\mu - \lambda\psi(\theta)},$$

where $\mu, \lambda \in \mathbb{R}$ *are called hyperparameters and they are constants and so is the term* $K(\mu, \lambda)$ *. The posterior distribution will of course belong in the same distributional family with the prior and in fact, it will be* $\psi(\cdot)$ *with different hyperparameters. Specifically,* $f(\theta|x) \equiv \psi(\theta|\mu + \sum_{i=1}^{n} x_i, \lambda + n)$.

In Table 1.1, we briefly mention some of the most typical cases of conjugate analyses, to better demonstrate the use of conjugate priors in practice. The form of the posterior distributions can be easily verified.

Likelihood	Conjugate Prior	Posterior Distribution
$X \sim Bin(n, \theta)$	$\theta \sim Beta(p, q)$	$\theta \vert X \sim Beta(p + X, q + n - X)$
$X_1, ..., X_n \sim Geom(\theta)$	$\theta \sim Beta(p, q)$	$\theta \vert \boldsymbol{X} \sim Beta(p + n, q + \sum_{i=1}^{n} X_i - n)$
$X_1, ..., X_n \sim Poisson(\theta)$	$\theta \sim Gamma(p, q)$	$\theta \vert \boldsymbol{X} \sim Gamma(p + \sum_{i=1}^{n} X_i, q + n)$
$X_1, ..., X_n \sim N(\theta, \tau^{-1}), (\tau \quad \text{known})$	$\theta \sim N(b, c^{-1})$	$\theta \vert \boldsymbol{X} \sim N\left(\frac{cb + n\tau\bar{X}}{c + n\tau}, \frac{1}{c + n\tau}\right)$

Table 1.1: Standard conjugate priors and corresponding posterior distributions

It is of great importance to note that we should not be tempted by the mathematical convenience of conjugate analysis and in each case, easily place a conjugate prior on the studied parameter, unless we are convinced that the particular distribution expresses in a satisfactory way of representing our uncertainty regarding that specific parameter. Otherwise, we ran a very serous risk of misleading our analysis, due to the subjectivity inserted through the prior.

Extensive research is still conducted in order for this subjectivity to be reduced to the minimum extent and various suggestions have been introduced. The most simple approach to make a conjugate prior less informative is by properly adjusting the values of the hyperparameters, so that the variance of the distribution is large. A typical example would be the Normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 10000$. We shall also briefly refer to the cases of Binary and Poisson data as two

other examples, where it seems that both conjugacy and objectivity are achieved (Kerman, 2006).

In the study of Binary data, two forms of Beta distributions are the most popular and widely used: the uniform $Beta(1,1)$ and the Jeffrey's prior $Beta\left(\frac{1}{2},\frac{1}{2}\right)$. The former, expresses prior ignorance by assigning equal weights to all possible parameter values and it is very commonly used in Bayesian textbooks. The latter, as its name clearly states, is derived by applying Jeffrey's Rule on the binomial likelihood of the data and it is recommended particularly for reasons of objectivity by both Bernardo (1979) and Berger(2006).

The conjugate prior for a Poisson model will be obtained from the Gamma family of distributions. Kerman mentions Gamma distributions of the form $Gamma(\epsilon,\epsilon)$, with $\epsilon \to \infty$, but arrives at the conclusion that the scale-free $Gamma\left(\frac{1}{3},0\right)$ is the optimal candidate, despite the fact that it is improper, since the posterior will be proper.

1.3 Markov Chain Monte Carlo methods

Numerous are the cases, in which the posterior distribution can only be identified up to a normalizing constant. This specific problem is referred to as ***intractability*** and it arises very frequently in Bayesian inference. Before the evolution of computational methods, statisticians would turn to analytical techniques, (the most well known of which is the ***Normal*** and the ***Laplace approximation***), which would indeed handle the problem by providing an asymptotic approximation of the posterior distribution, but at the rather high cost of time consuming and complicated computations. Also, as the dimension of the problem increases, the whole process requires much more time and even more sophisticated computational skills.

The prospect of struggling with mathematically and computationally advanced procedures in order to conduct inference in the Bayesian paradigm, when Classical Statistics methods provide quick and effective results, was one of the main arguments of the opponents of Bayesian Statistics against it. Statisticians and scientists who would engage in statistical procedures in their research, have indeed been discouraged from employing Bayesian methods for inference.

The scenery of statistical inference changed dramatically in the early nineties, due to the evolution of stochastic simulation methods and their successful implementation in Bayesian statistics. It was the use of computers that made this implementation so successful, mainly because of their capacity to perform a large number of complicated computations in a few moments.

The fundamental idea upon which an MCMC method is built, is to consider the distribution we wish to study, usually referred to as the ***"target distribution"***, as the ***equilibrium distribution*** of a Markov chain. The procedure we follow in order to construct this chain requires a solid mathematical background, regarding the knowledge of Probability Theory with special focus on asymptotic theorems of Stochastic Processes. The computer is the tool that enables us to produce a large number of draws from the chain automatically and extremely quickly. More specifically, the speed of the computer relieves us form the tedious task of repeatedly performing the intensive, arithmetic calculations required to build the chain. What is more, the fact that we can rapidly obtain an adequately large number of draws justifies the use of asymptotic properties, thus enhancing the robustness of our conclusions. Consequently, stochastic simulation methods constitute a statistical, computational tool, which can both save us time and provide credibility. Finally, stochastic simulation methods in general, constitute a very active field of research, mainly regarding the improvement of the performance of the algorithmic part, by reaching the maximum speed and effectiveness.

In this chapter, we will attempt a brief review of the Monte Carlo Integration method, the Gibbs and Metropolis-Hastings algorithms and also the Random Walk Sampler, which are the MCMC methods that are employed in the present thesis for the study of the posterior distribution of random variables. An important note we have to make regarding their structure is that the Markov chain formed in each case has as stationary distribution the *joint distribution* of interest.

1.3.1 Monte Carlo Integration

As we have previously mentioned, quite often arises the need to compute marginal distributions, either because they are normalizing con-

stants in the posterior distribution, or because they have their own role in the inference. ***Monte Carlo Integration*** is a simulation method, by which we can compute integrals of the form

$$I_1 = \int t(\boldsymbol{\theta}) \cdot \pi(\boldsymbol{\theta}) d\boldsymbol{\theta},$$

where $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^k$, whereas $t(\cdot), \pi(\cdot) : \mathbb{R}^k \to \mathbb{R}^l$. The idea of *Monte Carlo Integration*, also called the *simple Monte Carlo* estimator, is based on the well-known theorem known as "The Law of Large Numbers":

Theorem 2. ***(The Law of Large Numbers)*** *Let $X_1, ..., X_n$ be a sequenece of independent random variables with common means $E(X_i) = \boldsymbol{\theta}$ and variances $Var(X_i) = \sigma^2$, $i = 1, ..., n$. If $\bar{X_n} = \frac{1}{n} \sum^{i=1} nX_i$, that is the sample mean, then,*

$$E\big(\bar{X_n} - \boldsymbol{\theta}\big) = \frac{\sigma^2}{n} \to 0, \qquad as \quad n \to \infty.$$

What the theorem suggests, is that the sample averages will converge to the population mean with rate $\frac{1}{n}$.
If we study again the integral I_1, we will come to realize that it can be regarded as the expression of a mean and more specifically:

$$I_1 = E_\pi[t(\boldsymbol{\theta})]$$

.
The Monte Carlo Integration Method, provides a natural estimator for I_1, which is obtained thusly:

- Since the above expression suggests that the distribution of $\boldsymbol{\theta}$ is $\pi(\cdot)$, we obtain s draws $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(s)} \sim \pi$.
- We calculate $\hat{I_1} = \frac{1}{s} \sum_{i=1}^{s} t\big(\theta^{(i)}\big)$.

In order to conduct inference, we need to calculate the marginal likelihood or evidence of the data from an integral of the form $I_2 = \int f(\boldsymbol{x}|\boldsymbol{\theta}) \cdot \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$, where $\pi(\cdot)$ is the prior distribution placed on $\boldsymbol{\theta}$.

We can obtain the simple Monte Carlo estimator of I_2, by once again simulating s draws $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(s)} \sim \pi$ and then calculate the sum $\hat{I_2} = \frac{1}{s} \sum_{i=1}^{s} f\big(x|\theta^{(i)}\big)$. Monte Carlo integration has advantage over analytical methods, because it can be implemented as easily in low and high-dimensional problems. Finally, the estimators, denoted as $\hat{I}$, have good frequentist properties, as they are unbiased and strongly consistent estimators of the corresponding integrals I.

1.3.2 Using Metropolis-Hastings Algorithms: A Review of the Generic Metropolis-Hastings Algorithm

As their name clearly states, Markov Chain Monte-Carlo methods are ideas of simulation based on the theory of Markov Chains. Markov Chains constitute a category of stochastic processes that have interesting and appealing transition and limit properties. In order to better comprehend and explain the algorithms that will be used, a very brief review of the most basic theorems and properties of the Markov Chains is provided below.

Fundamentals of the Markov Chain Theory

Definition 1.3.1. Let $\theta_0, ..., \theta_n, \quad n \in \mathbb{N}$, be a sequence of random variables with state space S. Then we refer to $(\theta_n)_{n\geq 0}$ as a Markov Chain if the following property is satisfied:

$$P(\theta_{n+1} = i_{n+1}|\theta_n = i_n, \theta_{n-1} = i_n, ..., \theta_0 = i_0) = P(\theta_{n+1} = i_{n+1}|\theta_n = i_n). \tag{1.1}$$

We refer to $P(\theta_{n+1} = i_{n+1}|\theta_n = i_n), \quad n \in \mathbb{N}$, as the transition probabilities, whereas the probability distribution $P(x, y) = P(\theta_{n+1} = x|\theta_n = y), \quad n \in \mathbb{N}$, is called the transition kernel. Also, we will be using the notation $P^m(x, y) = P(\theta_{n+1} = y|\theta_0 = x)$ for the transition probability from state x to state y over m steps.

Definition 1.3.2. A distribution π is referred to as the **stationary distribution** of the Markov Chain $(\theta_n)_{n\geq 0}$ with transition probabilities $P(x, y)$ if, when it is set as the initial distribution of θ_0, satisfies the system of equations:

$$\begin{aligned} &\sum_{x\in S} \pi(x)P(x, y) = \pi(y), \qquad y \in S, \\ &\sum_{z\in S} \pi(z) = 1. \end{aligned} \tag{1.2}$$

The distribution π is also often called the *equilibrium distribution.*

If the stationary distribution exists, then it also has the property:

$$lim_{m\to\infty} P^m(x, y) = \pi(y)$$

upon which, the whole concept of the Metropolis-Hastings algorithm is built.

Definition 1.3.3. A Markov chain $(\theta_n)_{n\geq 0}$ with state space S, transition probabilities $P(x,y), \quad x,y \in S$, and stationary distribution π is called reversible if:

$$\pi(x)P(x,y) = \pi(y)P(y,x), \quad \text{for all} \quad x,y \in S$$

The idea of the Metropolis-Hastings algorithm is to consider the distribution of the parameter $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^{k+1}$, $f(\boldsymbol{\theta})$, from which we wish to sample and which is usually called the *target distribution*, as the stationary distribution of a reversible Markov chain. In order to achieve that, we need to come up with the appropriate transition kernel $P(\theta,\phi)$, so that the balance equation

$$f(\theta)P(\theta,\phi) = f(\phi)P(\phi,\theta), \quad \theta,\phi \in \Theta,$$

is satisfied. The transition kernel $P(\theta,\phi)$ can be further analysed as the product of a transition kernel q and a probability distribution a:

$$P(\theta,\phi) = q(\theta,\phi)\cdot a(\theta,\phi), \qquad if \quad \theta \neq \phi.$$

Consequently, $P(\theta,\theta) = 1 - \int q(\theta,\phi)\cdot a(\theta,\phi)d\phi$.

The purpose of the function q is to propose the candidate value ϕ to which the chain will move, whereas a is a probability that will ensure reversibility by expressing the probability whether the chain will move or not, through the following equation, proposed by Hastings: the probability that the chain will move to ϕ or equivalently, the acceptance probability of the value ϕ will be:

$$a(\theta,\phi) = min\left\{1, \frac{f(\phi)q(\theta,\phi)}{f(\theta)q(\phi,\theta)}\right\}. \tag{1.3}$$

Implementation of the Generic MH algorithm to obtain $f(\theta|y,X)$

- Set an initial value $\boldsymbol{\theta}^{(0)}$.
  The process to get the j-th value of $\boldsymbol{\theta}$, that is $\boldsymbol{\theta^{(j)}}$ for $j = 1,2,...,T$, is the following loop:
- Get a candidate value from the proposal distribution q, so that $\boldsymbol{\theta^{can}} \sim q(\theta|\boldsymbol{\theta^{(j-1)}})$.

- Calculate the probability $a = min\left\{1, \frac{f(\boldsymbol{\theta}^{(j-1)}|y,X)q(\boldsymbol{\theta}^{can}|\boldsymbol{\theta}^{(j-1)})}{f(\boldsymbol{\theta}^{can}|y,X)q(\boldsymbol{\theta}^{(j-1)}|\boldsymbol{\theta}^{can})}\right\}$.
- Set $\boldsymbol{\theta}^{(j)} = \boldsymbol{\theta}^{can}$ with probability a; with probability $1 - a$ set $\boldsymbol{\theta}^{(j)} = \boldsymbol{\theta}^{(j-1)}$.

Discarding of initial generated values and comments on convergence diagnostics

- The number T mentioned above, is the total number of iterations of the algorithm and consequently, the total number of observations from $f(\boldsymbol{\theta}|y, X)$ we are supposed to get.
- The idea is that the reversible Markov chain generated by the algorithm will converge to the posterior distribution of $\boldsymbol{\theta}$. This results to the dismissal of a number $t \leq T$ of generated values and keeping the ones we get after convergence is achieved. The number t is called the ***burn-in period***, it is specified by the statistician conducting the inference and there can be no "right" choice.

  Nevertheless, there are some quick, informal ways of monitoring the convergence of the algorithm. A common choice is to make a plot of the *ergodic mean*, which is the mean of the sample, calculated until the current iteration. By inspecting the plot of the ergodic means (which will be T in total), we are trying to determine the iteration after which the generated values are very close arithmetically, giving an impression of stability, which will be demonstrated as a roughly straight, horizontal line. This is considered to be the point at which the convergence has been achieved and so we discard all the simulated values right before that particular iteration. Another alternative are the ***trace plots***, which are plots of the iterations versus the generated values. Periodicities or tendencies among the generated values are indications of convergence.
- In general, the algorithm is considered effective if convergence is achieved quickly. Theoretically, the reversible Markov chain constructed with this technique is bound to converge to the target distribution, but this is supposed to happen after infinite iterations. Consequently, it has to be ensured that convergence will be achieved within the number of iterations we are able to perform. Also, after discarding the burnin values, we wish to have a rather large sample of the target distribution, in order to conduct accurate inference. Therefore, benefiting from the fastness of the computers, we should produce long chains, meaning that a large number of it-

erations should be performed.

The speed of convergence depends on the candidate distribution q. A poor choice will result to a slow-converging algorithm and an inadequate investigation of the support of the target distribution, which will cloud our understanding and compromise our inference. Therefore, it is of major importance, but usually also a challenging task, to find the optimal candidate distribution and several techniques have been developed aiming to determine the appropriate q in each case.

- The very important advantage of the MH algorithms is that we do not have to know the precise form of $f(\boldsymbol{\theta}|y, X)$. Since the acceptance probability is computed by a fraction, the complicated integral of (3.4) is simplified and only the term of (3.5) needs to be computed at each iteration.

The Random Walk Sampler

As the name of the algorithm reveals, the fundamental idea of its development was the Markovian (k+1-dimensional) Random Walk, since we consider that the candidate $\boldsymbol{\theta^{can}}$ and the current value $\boldsymbol{\theta^{(j-1)}}$ of the j-th step are related through the following property:

$$\boldsymbol{\theta^{can}} = \boldsymbol{\theta^{(j-1)}} + \boldsymbol{u}_{j-1}, \qquad \boldsymbol{u}_{j-1} \sim N_{k+1}(0, \Sigma), \qquad j = 1, ..., n. \quad (1.4)$$

Consequently, at the j-th step, $\boldsymbol{\theta^{can}} \sim N_{k+1}\big(\boldsymbol{\theta^{(j-1)}}, \Sigma\big)$. Thus, the proposal distribution q will actually be a symmetric distribution and more specifically: $q\big(\boldsymbol{\theta^{can}}|\boldsymbol{\theta^{(j-1)}}\big) = q\big(\big|\boldsymbol{\theta^{can}} - \boldsymbol{\theta^{(j-1)}}\big|\big)$. Consequently, $q\big(\boldsymbol{\theta^{can}}|\boldsymbol{\theta^{(j-1)}}\big) = q\big(\boldsymbol{\theta^{(j-1)}}|\boldsymbol{\theta^{can}}\big)$ and, so, the acceptance probability cancels down to the fraction: $a = min\Big\{1, \frac{f(\boldsymbol{\theta^{can}}|\boldsymbol{y},X)}{f(\boldsymbol{\theta^{(j-1)}}|\boldsymbol{y},X)}\Big\}$.

The most popular choices of proposal distributions are the multivariate Uniform and the multivariate Normal.We will focus on the latter and so the proposal distribution at the j-th step of the algorithm will be $q \equiv N_{k+1}\big(\boldsymbol{\theta^{(j-1)}}, \Sigma\big)$. The matrix Σ is the main issue, as it affects directly the convergence rate of the algorithm. Specifically, supposing we choose Σ to be diagonal, then the positive values Σ_{ii} determine how close the proposed and the current values of each coefficient θ_i will be. Small values yield high acceptance probabilities but slow convergence, due to the fact that the candidate values $\boldsymbol{\theta^{can}}$ will be close to each other and so the algorithm will keep exploring the same, small area of the parameter space, dictated by the proposal distribution. On the other

hand, by placing high values on Σ_{ii}, we are constructing an algorithm with very low acceptance rates, resulting to the repetition of the same values, since the candidate values will be often rejected. In the end, once again a small area of the parameter space will have been explored and the sample of the posterior distribution will contain high autocorrelations. Therefore, we must proceed cautiously by "trial and error", and we have to first monitor the acceptance rate of the algorithm and if it is not adequate, ***tune*** the algorithm appropriately by adjusting the values of Σ. The "adequacy" of the acceptance rate is subjective, but usually a rate of roughly 25% is considered appropriate. We can tolerate a somewhat lower percentage if we are exploring high-dimensional spaces, whereas in small dimensions, it is advised for the acceptance rate to lie between 10% and 40%. We should note that the tuning process is far from trivial and can get very time consuming as the dimensionality of the parameter space grows. Finally, the performance of the proposal distribution can be assessed by running the chain several times and monitoring the acceptance rates.

The Independence Sampler

The *Independence Sampler* is another member of the Metropolis-Hastings algorithmic family, the special characteristic of which is that at the j-th step of the algorithm, the proposal distribution q is set to be *independent* from $\boldsymbol{\theta^{(j-1)}}, j = 1, ..., n$. In mathematical notation, this is expressed as: $q(\boldsymbol{\theta^{can}}|\boldsymbol{\theta^{(j-1)}}) = q(\boldsymbol{\theta^{can}})$.

Based on the above, the acceptance probability of the candidate value at each step j, as given in (3.8), takes the form:

$$a = min\left\{1, \frac{f\big(\boldsymbol{\theta^{(j-1)}}|y, X\big)q(\boldsymbol{\theta^{can}})}{f(\boldsymbol{\theta^{can}}|y, X)q\big(\boldsymbol{\theta^{(j-1)}}\big)}\right\}, \qquad j = 1, 2..., n. \tag{1.5}$$

Once again, the convergence properties of the algorithm depend on q, which has to diverge as little as possible from the target/posterior distribution. Consequently, the challenging part of the process is to find the suitable proposal density. If we choose the "trial and error" option, then we will have to test various proposal densities by running the chain using each one of them as q, monitoring the acceptance probabilities, inspecting the trace plots and then decide whether we have to tune the algorithm or not. Unless we can roughly figure out the form of the posterior distribution, this endeavour will turn out to be very time consuming and as the number of dimensions increases, so does the complexity of

the posterior distribution and thus, the task becomes even more cumbersome. So, if we relied solely on that technique, we would prefer the Random Walk alternative.

However, various computational and analytical methods have been developed in order to approximate the main descriptive statistics of the posterior distribution and improve the performance of q. In general, a successful proposal distribution for the Independence Sampler is the multivariate Normal with parameters the posterior mode $\tilde{\boldsymbol{\theta}}$ and the inverse of the curvature at the posterior mode, denoted as $\boldsymbol{C}(\tilde{\boldsymbol{\theta}})$.

Once again, the computation of the posterior mode can be performed analytically or computationally. A version of the latter option, known as the *Bayesian Iterative Weighted Least Squares* algorithm will be used in the present thesis and it will be directly implemented in the logistic regression model.

The Gibbs Sampler

The Gibbs Sampler, originally introduced by Geman and Geman (1984), is an MCMC method, which is very useful when the studied parameter $\boldsymbol{\theta}$ is considered to be a *multivariate* random variable; in other words, the parameter space denoted as Θ is considered to be a subsection of $\mathbb{R}^k, \quad k \in \mathbb{N}$.

Apart from the Markovian Theory, the fundamental idea behind the construction of this particular algorithmic scheme, is the statistical concept of *conditional conjugacy*. We assume independent priors for each coordinate of the parameter vector $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)$. So, we would a priory consider that $\theta_1 \sim \pi_1, ..., \theta_k \sim \pi_k$, where $\pi_1(\cdot), ..., \pi_k(\cdot) : \mathbb{R} \to \mathbb{R}$, whereas the joint prior distribution of $\boldsymbol{\theta}$, will be:

$$\boldsymbol{\pi}(\boldsymbol{\theta}) = \prod_{i=1}^{k} \pi_i(\theta_i).$$

Supposing that we wish to conduct inference on a vector of independently distributed components $\boldsymbol{\theta} = (\theta_0, \theta_1, ..., \theta_k)$ and that we have obtained the likelihood of the model under study, denoted by $p(\boldsymbol{y}|\boldsymbol{\theta}) \equiv$

$p(\boldsymbol{y}|\theta_0, \theta_1, ..., \theta_k)$, then by applying Bayes' Theorem, we would get:

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\boldsymbol{\theta}) \cdot \boldsymbol{\pi(\theta)} \\ &= p(\boldsymbol{y}|\theta_0, \theta_1, ..., \theta_k) \cdot \pi_0(\theta_0) \cdots \pi_k(\theta_k). \end{aligned}$$

Quite often, the posterior distribution of $\boldsymbol{\theta}$ turns out to be very costly to directly simulate draws from, or even intractable.

However, conjugacy allows us to easily determine the *conditional posterior distribution* of each θ_i, denoted by $p(\theta_i|\boldsymbol{y}, \theta_0, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_k), \quad i = 0, 1, ..., k$.
Indeed,

$$p(\theta_i|\boldsymbol{y}, \theta_0, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_k) \propto p(\boldsymbol{y}|\boldsymbol{\theta}) \cdot \pi(\theta_i|\theta_0, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_k)$$

But, since we have assumed independent priors, we can conclude that

$$\pi(\theta_i|\theta_0, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_k) = \pi_i(\theta_i)$$

and consequently, we deduce that

$$p(\theta_i|\boldsymbol{y}, \theta_0, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_k) \propto p(\boldsymbol{y}|\boldsymbol{\theta}) \cdot \pi_i(\theta_i)$$

The Algorithm

1. Consider an arbitrary initial vector $\boldsymbol{\theta^{(0)}} = \left(\theta_0^{(0)}, \theta_1^{(0)}, ..., \theta_k^{(0)}\right)$, where

$$\begin{aligned} \theta_0^{(0)} &\sim \pi_0 \\ \theta_1^{(0)} &\sim \pi_1 \\ \theta_2^{(0)} &\sim \pi_2 \\ &\vdots \\ \theta_k^{(0)} &\sim \pi_k \end{aligned}$$

2. Use the current vector of $\boldsymbol{\theta}$.
   - Simulate from the conditional posterior distribution $\theta_0^{(1)}$ from the conditional posterior distribution $p(\theta_0|\boldsymbol{y}, \theta_1^{(0)}, \theta_2^{(0)}, ..., \theta_k^{(0)})$.
   - Simulate from the conditional posterior distribution $\theta_1^{(1)}$ from the conditional posterior distribution $p(\theta_1|\boldsymbol{y}, \theta_0^{(1)}, \theta_2^{(0)}, ..., \theta_k^{(0)})$.

- Simulate from the conditional posterior distribution $\theta_2^{(1)}$ from the conditional posterior distribution $p(\theta_2|\boldsymbol{y}, \theta_0^{(1)}, \theta_1^{(1)}, \theta_3^{(0)}..., \theta_k^{(0)})$.
  $\vdots$

- Simulate from the conditional posterior distribution $\theta_k^{(1)}$ from the conditional posterior distribution $p(\theta_k|\boldsymbol{y}, \theta_0^{(1)}, \theta_1^{(1)}, \theta_2^{(1)}, ..., \theta_{k-1}^{(1)})$.

Then, $\boldsymbol{\theta^{(1)}} = \left(\theta_0^{(1)}, \theta_1^{(1)}, ..., \theta_k^{(1)}\right)$.

3. Return to step 2. using as the current value of $\boldsymbol{\theta}$, the $k+1$-dimensional vector obtained in the previous step.

Once again, we can monitor the convergence of the algorithm using the ergodic means plot mentioned before for each coordinate $\theta_i, i = 0, 1, ..., k$ and set an appropriate total number of iterations T, so that after the burn-in period, denoted by t, a sufficiently large sample is available.
The idea is that, as the total number of iteration increases, the formed chain approaches its equilibrium distribution (Gamerman and Lopes, 2006), which is the desired posterior distribution and so,the set of draws defined as: $\left\{\boldsymbol{\theta_i} : i \in \{t+1, ..., T\}\right\}$, can be regarded as a sample from $p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y})$.

A basic fact that has to be pointed out, is that the formed chain is indeed Markovian, since in order to obtain $\boldsymbol{\theta^{(j+1)}}$, we rely solely on $\boldsymbol{\theta^{(j)}}$ and thus, there is probabilistic dependence on the exact previous state of the chain.

As far as the scanning over the components of $\boldsymbol{\theta}$ in each iteration, as well as the derivation of the sample are concerned, various methods have been proposed, mainly for optimization reasons. For instance, instead of the deterministic order suggested above, Roberts and Sahu (1997) consider a scan, in which at each iteration a permutation of the indicator set $0, 1, 2, \ldots, k$ is selected randomly and dictates the order the components $\theta_i, i = 0, 1, \ldots, k$ will be visited. Alternatively, Zeger and Ibrahim (1991) suggested a scheme, according to which, some components are visited only on every jth iteration, where $j \in \mathbb{N}$ is a finite and fixed value.

The way by which we chose to obtain a sample of size n from the posterior distribution of $\boldsymbol{\theta}$ is to perform $t+n$ $k+1$-dimensional generations from one single chain and keep the last n draws, which will all have as marginal distribution the stationary distribution of the chain. Although ergodic theorems ensure that the obtained sample is valid, we still run the risk of failing to acknowledge high autocorrelations between the sample values, because n may not be an adequately large sample size. This possibility makes sense, since each value $\boldsymbol{\theta^{(l)}}$ strongly depends on the previous $\boldsymbol{\theta^{(l-1)}}$. Such an incident could lead to an incomplete exploration of the parameter space, because the chain will move very slow and in the same areas and, as a result, we could end up with inferences that would not offer a sufficient view regarding the posterior distribution of the studied parameter. Bearing that option in mind, a statistician should definitely set n as a large number, in order to enhance the validity of the inference. Of course, safety in this case may come at the cost of time, depending on the amount of computations required. In an attempt to avoid the hazard of a misleading sample due to chain autocorrelation, one could derive an n-sized sample by keeping the generated value after every mth iteration,$m \in \mathbb{N}$, where m is fixed. I.e. if we considered the sample as a set denoted by S, then $S = \left\{\boldsymbol{\theta^{(t+m)}}, \boldsymbol{\theta^{(t+2m)}}, \ldots, \boldsymbol{\theta^{(t+nm)}}\right\}$. Thus, we produce a sample of independent and hence, not autocorrelated values, all drawn from a single chain, at the cost, however, of $t+nm$ generations in total, which could once again be high, depending on the integers n and m.

1.4 Bayesian Inference and Model Selection

1.4.1 Summarizing the information

In the Bayesian framework, inference is conducted by **studying the posterior distribution**. That task is performed in steps, the combination of which, enlightens our view concerning the properties and the peculiarities of the distribution of the studied parameter.

Location Parameters, Spread Parameters and Measures of Association Between Two or More Variables

Supposing we wish to study a random variable denoted by $\boldsymbol{\theta} \in \mathbb{R}^m$. After obtaining its posterior distribution (analytically or computationally), our study begins by deriving the means vector, for which the usual notation is $\bar{\boldsymbol{\theta}}$ and the covariance matrix $\boldsymbol{\Sigma}$. If the posterior distribution

has the form of a known distribution, the elements above are directly available, thanks to properties of the Probability Theory. In any different case, we have to estimate them, using the formulas bellow. Let $S = \{\boldsymbol{\theta^{(1)}}, ..., \boldsymbol{\theta^{(n)}}\}$ be a sample from the posterior distribution of $\boldsymbol{\theta}$. Then,

- the posterior mean is estimated by:

$$\bar{\boldsymbol{\theta}} \equiv \begin{bmatrix} \bar{\theta_1} \\ \bar{\theta_2} \\ \vdots \\ \bar{\theta_m} \end{bmatrix} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\theta^{(i)}}$$

- whereas, the posterior covariance matrix, $\boldsymbol{\Sigma} = \Big(\sigma_{k,l}\Big)$, with $k, l \in \{1, ..., m\}$, is calculated based on the formula:

$$\sigma_{kl} = \frac{Cov(\theta_k, \theta_l)}{\sqrt{Var(\theta_i) \cdot Var(\theta_j)}} = \frac{\frac{1}{n} \sum_{j=1}^{n} \Big(\theta_k^{(j)} - \bar{\theta_k}\Big) \cdot \Big(\theta_l^{(j)} - \bar{\theta_l}\Big)}{\sqrt{\frac{1}{n^2} \sum_{j=1}^{n} \Big(\theta_k^{(j)} - \bar{\theta_k}\Big)^2 \cdot \Big(\theta_l^{(j)} - \bar{\theta_l}\Big)^2}}$$

  The elements in the main diagonal, that is $\Big(\sigma_{ii}\Big)_{i=1,...m}$, correspond to the estimated *variances* of each coordinate θ_i, $i = 1, ..., m$, whereas $\Big(\sigma_{ij}\Big)_{i \neq j} = Cov\Big(\theta_i, \theta_j\Big)$.

Especially in the multivariate case, in which a graph of the posterior distribution cannot be obtained if the parameter space is of dimension over 3, each component of the studied parameter can be studied individually by first deriving the ***marginal posterior distributions*** and then the statistical features mentioned above. Additionally, we can also obtain the central tendency measures of the median and the mode and the spread parameters of range and interquantile range. Likewise, in the case of a posterior distribution of known form, we can obtain these parameters in a straightforward way. Otherwise, if a sample of the posterior distribution is available, trivial computations are required.

More specifically, bearing in mind the relative definitions from the Statistical Theory, after *sorting the observations of the sample in ascending order*, we can deduce that:

- the median of each covariate,$\theta_i \in \mathbb{R}, \quad i = 1, ..., m$ is either $\theta_i^{[(n+1)/2]}$, if n is an odd number, or $\frac{\theta_i^{[n/2]}+\theta_i^{[n/2+1]}}{2}$, if n is an even number
- a value $\theta_i^{[j]}, j \in \{1, ...n\}$ is identified as mode if $f\left(\theta_i^{[j]}|\boldsymbol{y}\right) \leq f\left(\theta_i^{[k]}|\boldsymbol{y}\right)$, for every $k \in \{1, ...n\} - \{j\}$. We note that the value of mode may not be unique.
- the range is described as the difference $\theta_i^{[1]} - \theta_i^{[n]}$, where $\theta_i^{[1]}$ and $\theta_i^{[n]}$ are the smallest and largest values of the random variable θ_i in the obtained sample

We should note that statistical software like R or WinBugs, have special functions, which by using as input the set of the sampling values, can provide all the statistical features mentioned above.

Highest Posterior Density Regions

The numerical summaries mentioned previously cannot offer any measure of accuracy. In an attempt to overcome this lack of information, the idea of *credibility intervals* (when the parameter space $\Theta \subset \mathbb{R}$) and *credibility regions* (when $\Theta \subset \mathbb{R}^k, k \in \mathbb{N}$) was developed. Initially, our goal is to determine an area, denoted by $\mathcal{C}(\alpha) \subset \Theta$, in which the studied parameter $\boldsymbol{\theta}$ will lie with probability $1 - \alpha$, *based on the posterior distribution*. This idea could be formally summarised by the definition provided below:

Definition 1.4.1. Suppose that the posterior distribution of the parameter $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^{k+1}$ is $p(\boldsymbol{\theta}|\boldsymbol{y})$. Then $\mathcal{C}(\alpha) \subset \Theta$ is called a $100\%(1-\alpha)$ ***credible region***, if:

$$\int_{\mathcal{C}(\alpha)} p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta} = 1 - \alpha.$$

The region defined above, however, is not unique and so, the question that automatically rises is which one of the areas satisfying this condition should be chosen?

The optimal choice is considered to be the area with the smallest width, which will inevitably also be a *high density* area. This amounts to an additional constraint, according to which, $\mathcal{C}(\alpha)$ has to be of the form:

$$\mathcal{C}(\alpha) = \left\{ \boldsymbol{\theta} : p(\boldsymbol{\theta}|\boldsymbol{y}) \geq \gamma \right\}.$$

Thus, a highest posterior density region is determined.

It should be noted, that there is still a possibility of more than one areas, satisfying the two, required, conditions. This, for instance, could occur if the posterior distribution has more than two similar peaks.

Highest posterior density regions are a familiar concept from Classic Statistics, although there are fundamental differences, which will be clarified below. First, in the Bayesian framework $\boldsymbol{\theta}$ is treated as a random variable, to which, a probability distribution is assigned, whereas the region $\mathcal{C}(\alpha)$, is defined as a region with *constant* boundaries. In the case of the traditional confidence region, however, its boundaries are the random variables, $\boldsymbol{\theta}$ is an unknown constant and $100\%(1-\alpha)$ amounts to the possibility, that the constructed space $\mathcal{C}(\alpha)$ will include the researched parameter (usually 95% or 99%).

Another very important remark is that Bayesians, contrary to Classical statisticians, usually do not rely on the study of the HPDIs when it comes to hypothesis testing. Instead, the test of two or more hypotheses versus one another is performed by assigning a model to each hypothesis and then, by selecting the optimal model. The hypothesis corresponding to the selected model is considered to be supported by the data. This view is the most celebrated, but there is still debate over the matter. For instance, Robert and Marin (2006), employ highest posterior density regions in the search of a more parsimonious Normal, linear model.

In the absence of an analytic posterior distribution, an informal, computational way to estimate a credibility region of 95% probability would be to select a sample of 1000 observations, after the burn-in period of an MCMC algorithm, rank them with respect to the frequency of appearance and then consider $\mathcal{C}(\alpha)$ to be the area formed between the 25th and 975th drawn value.

In any case, it is strongly recommended, that one should first plot a graph from simulated values of the posterior distribution, which will provide an initial, informal view of the distributional behaviour of the parameter. Thus, insight to the form and the uniqueness of a credibility region or interval will be obtained.

1.4.2 Model Selection

Based on the different statistical approaches, the methods of model selection presented in this thesis could be roughly classified to three basic categories:

1. **Methods that point out as the optimal model, the one, which is best**
   "supported by the data"
   In this particular framework, we reach this conclusion by calculating either the *Bayes factor* or the *marginal likelihood of the data* for each candidate model. The derivation of both is based on the posterior distribution of the studied parameter under each model.

   Bayes factor: Supposing we wish to compare two models M_1 and M_2, regarding a parameter $\boldsymbol{\theta} = (\theta_0, ..., \theta_k)$. Since the present dissertation focuses on the task of variable selection in regression models, each one of the various, compared models suggests a subset of indicators $\theta_i, i = 0, ..., k$, which corresponds to a more economic selection of explanatory variables for the data at hand.

   So let Θ_1 and Θ_2 stand for the vector of parameters suggested by M_1 and M_2 respectively. The *Bayes factor* is defined as the ratio:

$$BF_{12} = \frac{f(data|\Theta_1)}{f(data|\Theta_2)},$$

   where

$$f(data|\Theta_i) = \int_{\Theta_i} f(\Theta_i|data) \cdot \pi(\Theta_i) d\boldsymbol{\theta_i},$$

   with $f(\cdot)$ and $\pi(\cdot)$ being the posterior and prior distribution of Θ_i respectively. What is interesting about the Bayes factor, is that it allows us to measure the information contained in M_1, when it is compared to M_2. A value that is a lot greater than 1, indicates that the information in M_1 offers us a better insight and therefore, it should be preferred over M_2.

   Marginal likelihood of the model: After the posterior distributions of two possible parameter selections Θ_1 and Θ_2 have been derived, we could, in a way, rely on the data to point us to the right direction, in order to make the best choice. An obvious way of doing that, is to obtain the posterior probability of each model, given the data and then, select the one with the maximum marginal

likelihood.

The posterior probability of M_1, given the data, is analytically calculated by Bayes' Theorem as follows:

$$P(M_1|data) = \frac{P(data|M_1) \cdot P(M_1)}{P(data|M_1) \cdot P(M_1) + P(data|M_2) \cdot P(M_2)}$$

where $P(data|M_1)$ and $P(data|M_2)$ are the *marginal likelihoods* of each model respectively. The computation of the two latter probabilities could be challenging, depending on the dimension of the parameter space studied each time. It would require to solve an integral of the following form:

$$\int_{\Theta_i} f(data|\Theta_i) \cdot \pi(\Theta_i) d\boldsymbol{\theta_i}$$

where $f(data|\Theta_i)$ is the likelihood of the data, according to the model structure and $\pi(\cdot) : \mathbb{R}^m \to \mathbb{R}^n$ is the prior distribution assigned on Θ_i, with m being the dimension of the parameter space.

2. **Methods that can be regarded as "self-consistency" or "predictive" checks**
They are mainly graphical tests or tests measuring discrepancies between *replicated data* and the observed data, via an appropriate statistical function. By the term "replicate", we refer to the data generated directly from the studied model. The credibility of the particular model is judged by the compatibility of the replicated values with the observed values.

Graphical checks
They basically involve the plotting of histograms of the replicated data, which, if they are similar to the one of the available data, indicate a good performance of the suggested model. Additionally, small discrepancies of the main statistical features (mean, mode, variance) of the samples, indicate that we are dealing with a plausible model.

The Bayesian p-value
Another important, although controversial, tool is the *Bayesian p-value*, which expresses the possibility of a sort of discrepancy, determined by the researcher, depending on the nature of the data.

A definition will be provided in the next chapter, which will clarify that there is a distinction between the Classic and the Bayesian perspective. Bayesians in general, tend to give a lot less credit to p-value as a tool of inference. It is often employed as an additional feature, but it is seldom used on its own.

The L Measure

The idea that led to the development of the L Measure is similar to the notion described above; it is the notion that the more compatible to the data at hand, $\boldsymbol{y}$, an arbitrary vector of ***predicted values*** $\boldsymbol{z} = (z_1, ..., z_n)$, under the candidate model, turns out to be, the more plausible that specific model would be.

The L Measure is a Bayesian test statistic, by which this "compatibility" between observed and predicted values can be assessed by formal, statistical means. The main idea for its definition is that small discrepancies between $\boldsymbol{y}$ and $\boldsymbol{z}$, indicate a good performance of the model.

Before proceeding to the functional forms suggested for the L Measure, we need to clarify that by $\boldsymbol{z}$, we refer to a vector of predictions, not to replicated values; this means that the sampling distribution of $\boldsymbol{z}$ is the predictive distribution and therefore, it is obtained by the following integral:

$$f(\boldsymbol{z}|\boldsymbol{y}, M) = \int L(\boldsymbol{z}|\boldsymbol{\theta}, M) \cdot f(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta},$$

with $L(\cdot|\boldsymbol{\theta}, M)$ denoting the likelihood under a model M.

Two of the first statisticians, who introduced and researched the idea of the L Measure, were Joseph G. Ibrahim and Laud. They originally considered as an adequate test statistic the *expected squared Euclidean distance* between $\boldsymbol{y}$ and $\boldsymbol{z}$, expressed as

$$L_{IL} = E\Big[(\boldsymbol{y} - \boldsymbol{z})^T(\boldsymbol{y} - \boldsymbol{z})\Big].$$

It can be shown that the former expression can also be written as

$$L_{IL} = \sum_{i=1}^{n} \Big\{ Var(z_i|\boldsymbol{y}) + \big(E[z_i|\boldsymbol{y}] - \boldsymbol{y}\big)^2 \Big\}.$$

Both the expectation and the variation are derived with respect to the *posterior predictive distribution* $f(\boldsymbol{z}|\boldsymbol{y}, M)$.

Ibrahim and Laud (1994) have assigned weight 1 on both the variation and the expectation of the squared difference between a predicted and an observed value. Ibrahim, Chen and Sinha (2001) pointed out that this decision is not theoretically justified and also, that by assigning a weight $\nu \in (0, 1)$ on the second term (which can be regarded as a sort of bias), one can gain "greater flexibility in the tradeoff between bias and variance".
Thus, they finally proposed the functional form below for the L Measure:

$$L_{IL} = \sum_{i=1}^{n} \left\{ Var(z_i|\boldsymbol{y}) + \nu \cdot \left(E[z_i|\boldsymbol{y}] - \boldsymbol{y} \right)^2 \right\} \tag{1.6}$$

The question that arises now is whether there is an "optimal" value for ν and how that is determined. Bibliography provides no definite answer until now. Ibrahim, Chen and Sinha, after a brief discussion over the subject and based on the results of the inference conducted on three models, as an example, they conclude that in the cases of linear and logistic regression, $\nu = \frac{1}{2}$ appears to be a generally good choice. The same value will be assigned on ν in the present paper as well. Consequently, the L Measure will be calculate by the following formula:

$$L_{IL} = \sum_{i=1}^{n} \left\{ Var(z_i|\boldsymbol{y}) + \frac{1}{2} \cdot \left(E[z_i|\boldsymbol{y}] - \boldsymbol{y} \right)^2 \right\} \tag{1.7}$$

The model with the smallest value of the L Measure statistic is the one selected among the possible candidate models.

Last but not least, we should note that an important property of the L Measure is that it very little affected by the prior definition and it is tolerant with improper priors as well. This property is the basic argument, used to justify why it should be preferred over Bayes' factors and Posterior Model probabilities for the task model assessment and comparison. We recall that these two methods are very sensitive to prior distributions, in the sense that a misplaced prior distribution would result to misleading Bayes' factors and posterior model probabilities. Additionally, an improper prior distribution would rule out these methods as options for model selection.

3. **Information Criteria**
Probably a common practice for model selection in both the Classical and the Bayesian approach. Among the many criteria, we

choose to confine our analysis to the *Bayesian Information Criterion* (abbreviated as BIC) and the *Deviance Information Criterion* (DIC).

The Bayesian Information Criterion
BIC is computed for a candidate model M_l as follows:

$$BIC := -2logL\Big(\boldsymbol{y}|\hat{\Theta}_l\Big) + d_l \cdot log(n),$$

where:

- $\hat{\Theta}_l$ is the vector of parameters that maximises the likelihood of the data under model M_l
- $L\Big(\boldsymbol{y}|\hat{\Theta}_l\Big)$ is the likelihood of the data under the structure defined by M_l and also, with $\hat{\Theta}_l$ as the parameter vector
- d_l is the dimension of the parameter space in M_l
- n is the dimension of the response variable $\boldsymbol{y}$.

The original derivation of BIC is based on a method for Bayesian model comparison introduced by Schwarz in 1978 and presents two very appealing computational properties. It is clear that the functional form that defines BIC is the log-likelihood of the model (penalized by the quantity $d_l \cdot log(n)$) and none of the assigned prior distributions. Hence, it constitutes a convenient comparison criterion when the specification of the prior is controversial. What is more, it is not affected by improper or non-conjugate priors in general.

Another interesting feature is that it is closely related to the Bayes factor of the model, due to its connection to the Schwarz criterion. In fact, it can be used to roughly approximate the log-Bayes factor under a wide family of prior distributions (Ntzoufras, 2009), which includes the exponential family. This statement is justified by the following property of the Schwarz criterion, calculated for the comparison of two arbitrary models M_1 and M_2.

The Schwarz criterion, according to Ntzoufras' *Bayesian Modeling using WinBugs* (2009) is derived thusly:

$$S_{12} = logL\Big(\boldsymbol{y}|\hat{\Theta}_2, M_2\Big) - logL\Big(\boldsymbol{y}|\hat{\Theta}_1, M_1\Big) - \frac{1}{2}(d_2 - d_1) \cdot log(n)$$

Based on the definition of BIC, we can deduct that

$$S_{12} = -\frac{1}{2}\Bigg\{ BIC(M_1) - BIC(M_2) \Bigg\}.$$

The property that interests us is that ,the Bayes factor B_{12} is connected asymptotically to the quantity S_{12}:

$$\frac{S_{12} - log(B_{12})}{log(B_{12})} \to 0, \quad \text{as} \quad n \to \infty.$$

Bearing in mind the latter equation, we can conclude that

$$-2log(B_{12}) \approx BIC(M_1) - BIC(M_2).$$

In addition to that, once the BIC of a model m is obtained, we can also approximate the posterior probability of that particular model as follows:

$$P\Big(m|\boldsymbol{y}\Big) \approx \frac{\exp\Big(-1/2 \cdot BIC(m)\Big)}{\sum_{\acute{m}\in\mathcal{M}} \exp\Big(-1/2 \cdot BIC(\acute{m})\Big)},$$

where $\mathcal{M}$ is the set of all the examined models.

An elaborate sketch of the derivation of BIC, as well as various information regarding the range of its use, are provided in Neath and Cavanaugh's paper *"The Bayesian information Criterion:background, information and application"*(2011). An interesting feature pointed out in the paper is the *consistency* of the Bayesian Information Criterion, which is defined as an asymptotic property, by which the selected model will "converge with probability one to the most parsimonious model that is closest to the true model (as measured by the Kullback-Leibler information)", even if the model that generated the observed data is not among the candidate models.

The Deviance Information Criterion

DIC is another very popular Bayesian criterion, originally introduced by Spiegelhalter in 2002. It is constructed as a trade-off between model fit and model complexity (Yong Li,Jun Yu, Tao Zeng, 2017) and it examines whether the data, replicated under a specific model, can predict the observed data in a "satisfactory" way. We will elaborate on the above statement, by defining the quantities,

that express these notions in the Bayesian framework.

Firstly, by the term *deviance*, we actually refer to the statistical function, defined as

$$D(\boldsymbol{\theta}) := -2lnL(\boldsymbol{y}|\boldsymbol{\theta}),$$

where $L(\boldsymbol{y}|\boldsymbol{\theta})$ expresses the likelihood of the data at hand, with respect to the vector of coefficients, denoted by $\boldsymbol{\theta}$, determined by the model under study.
Secondly, we introduce a frequently used, Bayesian measure of model fit, which is the *posterior expectation of the deviance*, expressed as

$$\overline{D(\boldsymbol{\theta})} := E_{\boldsymbol{\theta}|\boldsymbol{y}}\left[D(\boldsymbol{\theta})\right] = -2\int_{\Theta} lnL(\boldsymbol{y}|\boldsymbol{\theta}) \cdot f(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta},$$

with $\mathcal{B}$ indicating the parameter space of the vector of coefficients. The more plausible the model is, the larger the log-likelihood gets, resulting to a smaller value of $\overline{D(\boldsymbol{\theta})}$.

DIC is calculated based on the formula bellow:

$$DIC := \overline{D(\boldsymbol{\theta})} + p_D.$$

The term denoted by p_D is a measure of model complexity, known as the *effective number of parameters*. It is derived as follows:

$$p_D := \overline{D(\boldsymbol{\theta})} - D\big(\bar{\boldsymbol{\theta}}(\boldsymbol{y})\big) = \overline{D(\boldsymbol{\theta})} + 2lnL(\boldsymbol{y}|\bar{\boldsymbol{\theta}}(\boldsymbol{y})),$$

where $\bar{\boldsymbol{\theta}}(\boldsymbol{y}) \equiv \bar{\boldsymbol{\theta}} = \int_{\Theta} \boldsymbol{\theta} \cdot f(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}$ is the posterior mean of the coefficient vector. The last term is the deviance, evaluated at the posterior mean; that is $D\big(\bar{\boldsymbol{\theta}}\big) = -2lnL(\boldsymbol{y}|\bar{\boldsymbol{\theta}})$.The complexity of the model is in fact used to "penalize" the DIC score of the model, since the more complex it is, the larger this score gets.The optimal model is the one corresponding to the smallest value of the Deviance Information Criterion.

We should note that for the specific task of variable selection, this criterion can be a justified choice, due to the choice of the penalty-function, affecting the deviance. As a more specific example, consider a model with a generally suitable structure, that is, an

informative form, concerning the interpretation of the data, but of a very high dimension. Naturally, such a model with a large number of explanatory variables, would have a large likelihood, which is equivalent to a small deviance score. However, the penalty imposed by DIC would be large, since that model is complex. Hence, it would not be preferred over another, nested perhaps model, that includes fewer variables, provided that the likelihood is not significantly reduced.
The paper *"Bayesian measures of model complexity and fit"* (Spiegelhalter, et.al.), provides the original justification and an extensive examination of the Deviance Information Criterion. Also, *"Deviance Information Criterion for Bayesian Model Selection: Justification and Variation"*, authored by Li, Yun and Zeng in 2017, provides a very informative study for DIC and for that, the symbolism and terminology suggested, have been used in the present paper as well.

Chapter 2

Bayesian Analysis of the Multiple Normal Linear Regression Model

General Notation

The Bayesian approach to both univariate and multivariate normal linear models is of great interest and has been extensively studied in a number of scientific fields, especially in the field of econometrics (Kooper 2003) due to their significant applications. In general, the linear regression model suggests a relationship between the *dependent* or *observed variable*, y, and a set of k *explanatory variables* $x_1, ..., x_k$, which is of the form:

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \varepsilon_i, \qquad i = 1, ..., n, \quad n, k \in \mathbb{N}. \tag{2.1}$$

We denote as y_i the ith observation of the dependent variable and as x_{ij} the ith observation of the jth explanatory variable. We will assume that for every observation y_i there is an observation x_{ij} and also a quantity ε_i, $i = 1, ..., n$, $j = 1, ..., k$.
It is important to explain the meaning and nature of every variable that is used in the model:
-The vector y is the quantity of interest and refers to *a priori* available observations. It is a continuous random variable.
-The x_j variables are the factors considered to have a possible influence on y and have to be continuous. The observations of y are considered fixed, whereas the explanatory variables can be either fixed or stochastic. However, their values are given to us by the experimenter.
-The parameters $\beta_0, \beta_1 ..., \beta_k$ are called *regression coefficients*, they are

unknown and their estimation is a major goal of the entire analysis. In the Bayesian paradigm the parameters are considered to be random variables and, therefore are assigned prior density functions or distributions. -The quantities ε_i are error terms usually called *disturbancies* .Their existence is inevitable, since the formula suggested by the model is only an approximation of the actual relationship between y and x_j and it is certain that the values $\beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}$ and y_i are in fact unequal. Therefore, we are justified to use the concept of the equation only if we conclude the errors which could be the result of false measurements, the absence of influential factors from the model etc. Their values are unknown and consequently they are treated as random variables. We should note that the assumptions made regarding the distribution of the disturbancies define the likelihood of the model and the whole process of the analysis.

The normal linear model

The most widely used version of linear models is the *normal linear model* in which we assume that the disturbancies follow a normal distribution and consequently y_i will also have a normal distribution given x_{ij} with a mean that is a linear combination of x_{ij}:

$$E[y_i|\beta_0, \beta_1 ..., \beta_k, x_{i1}, ..., x_{ik}] = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}, \qquad i = 1, ..., n, \quad j = 1, ..., k. \tag{2.2}$$

2.1 The ordinary linear model

Assumptions

The *ordinary linear model* is the simplest version of a normal linear model and it is defined by the assumption that the disturbancies are homoscedastic, independent and identically distributed. That is :
1. $\varepsilon_i \sim N(0, \sigma^2) \qquad i = 1, 2, .., n.$
2. For every i, j the random variables ε_i and ε_j are independent from one another for every $i \neq j$.
In the next chapters we will discuss deviations from the above assumptions.

Using matrices
Since most of the times we are working with a rather large amount of data, it is accustomed to use vectors and matrices for computational reasons and for convenience. These are the vectors and matrices we will

be using from now on for the variables mentioned above:

- We consider the vector $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$ which contains the values of the dependent variable y.
- The values of each explanatory variable x_j are placed as columns in a $n \times (k+1)$ matrix named X the first column of which, is in fact a unit vector: $\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix}$. If these values are random variables, X is a stochastic matrix.
- There is also the regression coefficients vector $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$.
- Finally, we create a $n \times 1$ vector with the values of the disturbancies: $\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$.

From now on,the equation of the ordinary linear model will be the following:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \quad , \quad \boldsymbol{\varepsilon} \sim N_n(0, \sigma^2 I_n), \tag{2.3}$$

where $N_n(0, \sigma^2 I_n)$ is the multivariate normal distribution with mean a $n \times 1$ vector of zeros and I_n is the $n \times n$ identity matrix. From the form of the model and our previous assumptions, we can also assume that $\boldsymbol{y|X, \beta, \sigma} \sim N_n(\boldsymbol{X\beta}, \sigma^2 I_n)$.
The vector $\boldsymbol{\beta}$ as well as the value of σ^2 are considered unknown and therefore, they are treated as random variables and the inference problem is the estimation of the vector $\theta = (\beta_0, \beta_1, ..., \beta_k, \sigma^2)$.

Formal justification of the form of conditioning

Before we proceed to the full analysis, we need to clarify the conditioning when the explanatory variables are also random variables and so

X is a stochastic matrix. In this case there is also the distribution of X, $p(X|\psi)$ where ψ is an unknown parameter similar to θ and, therefore, we have a likelihood of the form $f(y, X|\theta, \psi)$ and a joint prior distribution $p(\theta, \psi)$. Although this seems to influence the whole concept of the regression, this is not the case, because we assume that the distribution of X offers no information to the conditional modelling of y given X; in other words we assume *prior independence* between the parameters θ and ψ. As a result, we can write $p(\theta, \psi) = p(\theta)p(\psi)$ and the posterior distribution can be written as $p(\theta, \psi|X, y) = p(\psi|X)p(\theta|X, y)$. Hence, we can write the second factor as $p(\theta|X, y) \propto p(\theta)p(y|X, \theta)$, which is the form of conditioning we will be using. Of course, when the values of X are chosen; meaning they are fixed and known, there is no parameter ψ to consider.

2.1.1 First-step analysis of the ordinary normal linear regression using Jeffreys' non-informative prior distribution for $\boldsymbol{\beta}$ and τ

The prior

The first step of our analysis is to determine a prior distribution for the variables $\boldsymbol{\beta}$ and σ^2. We assume in advance that $\beta_0, \beta_1, ..., \beta_k$ and σ^2 are independent.

Supposing that nothing is known about the vector of parameters $\boldsymbol{\beta}$, we can make the general assumption that $\beta_i \in (-\infty, +\infty)$ for each $i = 0, 1, ..., k$. Based on this, we can state that $p(\beta_i) \propto c_i$, where c_i is a constant or, even more simply, that $p(\beta_i) \propto 1$ for each i. Consequently, we will be using

$$p(\boldsymbol{\beta}) \propto 1 \tag{2.4}$$

as the non-informative prior of the vector of parameters.

We have to note that the suggested prior for $\boldsymbol{\beta}$ is improper, because $\int_{\mathbb{R}^{k+1}} \cdots \int p(\boldsymbol{\beta}) d\boldsymbol{\beta} \neq 1$. However, this consists no problem in our further analysis, since the respective posterior is proper.

In the case of the variance σ^2, we choose a suitable non-informative prior to represent our ignorance, thinking in a similar way. Obviously, $\sigma^2 \in (0, +\infty)$, which is equivalent to $\log(\sigma^2) \in (-\infty, +\infty)$. Once again,

we can use the non-informative prior that satisfies the condition:

$$p(\log(\sigma^2)) \propto 1. \tag{2.5}$$

Now, we can deduce the prior of σ^2 based on the change-of-variable technique:

$$p(\sigma^2) = p(\log(\sigma^2))\Big|\frac{d\log(\sigma^2)}{d\sigma^2}\Big| \propto 1 \cdot \frac{1}{\sigma^2}.$$

At this point, we will follow a rather common practice of transforming the variance σ^2 to the new variable $\tau = \frac{1}{\sigma^2}$ called *precision*. Using the same technique and based on the fact that $\sigma^2 = \dfrac{1}{\tau}$ we have that:

$$p(\tau) = p\Big(\frac{1}{\tau}\Big)\left|\frac{d\frac{1}{\tau}}{d\tau}\right| \propto \tau\frac{1}{\tau^2} = \frac{1}{\tau}. \tag{2.6}$$

Finally, the joint non-informative prior distribution of $\boldsymbol{\beta}$ and τ will be:

$$p(\boldsymbol{\beta}, \tau) \propto \frac{1}{\tau}. \tag{2.7}$$

Obtaining the posterior distributions

First, we will determine the posterior distribution of $\boldsymbol{\beta}$, conditional on the precision τ; that is $p(\boldsymbol{\beta}|\tau, y)$ and then we will obtain the marginal distribution of τ;that is $p(\tau|y)$.

The analytical form of the likelihood of the model is:

$$p(\boldsymbol{y}|\boldsymbol{\beta}, \tau) = (2\pi)^{-\frac{n}{2}}(\tau)^{\frac{n}{2}} \exp\Big[-\frac{\tau}{2}(\boldsymbol{y} - \boldsymbol{X\beta})^T(\boldsymbol{y} - \boldsymbol{X\beta})\Big]. \tag{2.8}$$

After applying Bayes' Rule and considering that we are interested only in the terms that include $\boldsymbol{\beta}$ and τ, we get the following expression:

$$p(\boldsymbol{\beta}, \tau|\boldsymbol{y}) \propto (\tau)^{\frac{n}{2}-1} \exp\Big[-\frac{\tau}{2}(\boldsymbol{y^T y - y^T X\beta - \beta^T X^T y + \beta^T X^T X\beta})\Big]. \tag{2.9}$$

We set $b = (X^T X)^{-1}X^T y$, namely the least squares estimate, as a quantity that provides help to calculate and identify the necessary posterior distributions. We are going to insert b in (1.7) by observing that:

$$y - X\beta = y - Xb + Xb - X\beta = y - Xb + X(b - \beta). \tag{2.10}$$

Using (1.8) we can rewrite (1.7) as:

$$
\begin{aligned}
&p(\boldsymbol{\beta},\tau|\boldsymbol{y}) \propto (\tau)^{\frac{n}{2}-1}\exp\Big[-\frac{\tau}{2}\Big([(y-Xb)^T+(b-\beta)^TX^T][(y-Xb)+X(b-\beta)]\Big)\Big]\\
&=(\tau)^{\frac{n}{2}-1}\exp\Big[-\frac{\tau}{2}\Big[(y-Xb)^T(y-Xb)+(y-Xb)^TX(b-\beta)\\
&\cdot\exp(b-\beta)^TX^T(y-Xb)+(b-\beta)^TX^TX(b-\beta)\Big]\Big]\\
&=(\tau)^{\frac{n}{2}-1}\exp\Big[-\frac{\tau}{2}\Big[(y-Xb)^T(y-Xb)+(b-\beta)^TX^TX(b-\beta)\Big]\Big].
\end{aligned}
\tag{2.11}
$$

Because of the form of b, the crossproduct $(y-Xb)^TX(b-\beta)$ is equal to zero. We set $s^2=\dfrac{(y-Xb)^T(y-Xb)}{n-k-1}$ and thus (1.9) can finally take the form below:

$$
\begin{aligned}
&p(\boldsymbol{\beta},\tau|\boldsymbol{y}) \propto (\tau)^{\frac{n}{2}-1}\exp\Big[-\frac{\tau}{2}(n-k-1)s^2\Big]exp\Big[-\frac{\tau}{2}(b-\beta)^TX^TX(b-\beta)\Big]\\
&=(\tau)^{\frac{n-k-1}{2}-1}\exp\Big[-\frac{\tau}{2}(n-k-1)s^2\Big](\tau)^{\frac{k+1}{2}}exp\Big[-\frac{\tau}{2}(b-\beta)^TX^TX(b-\beta)\Big].
\end{aligned}
\tag{2.12}
$$

Observing the final result,we now can use the exponential terms to factorize the joint posterior probability as: $p(\boldsymbol{\beta},\tau|\boldsymbol{y})=p(\boldsymbol{\beta}|\tau,\boldsymbol{y})p(\tau|\boldsymbol{y})$, where $p(\boldsymbol{\beta}|\tau,\boldsymbol{y}) \propto (\tau)^{\frac{k+1}{2}}exp\Big[-\frac{\tau}{2}(b-\beta)^TX^TX(b-\beta)\Big]$ and therefore we can assume that $\boldsymbol{\beta}|\tau,\boldsymbol{y}\sim N_{k+1}(b,\dfrac{1}{\tau}(X^TX)^{-1})$.
Also, since $p(\tau|\boldsymbol{y}) \propto (\tau)^{\frac{n-k-1}{2}+1}\exp\Big[-\frac{\tau}{2}(n-k-1)s^2\Big]$, we can conclude that $\tau|\boldsymbol{y}\sim Inv-\chi^2(n-k-1,s^2)$.

We must note that the prior we used is improper and, therefore, we need to make sure that the posterior distribution is proper. In this case, the joint posterior distribution $p(\boldsymbol{\beta},\tau|\boldsymbol{y})$ has to be proper; this condition is true when $n>k$ and X is of full rank. In other words, it is necessary that the number of observations is larger than the number of the model parameters and the explanatory variables are linearly independent. The marginal posterior distribution of $\boldsymbol{\beta}$, that is $p(\boldsymbol{\beta}|\boldsymbol{y})$, can be computed by the following integral:

$$
\begin{aligned}
p(\boldsymbol{\beta}|\boldsymbol{y}) &= \int_0^\infty p(\boldsymbol{\beta},\tau|\boldsymbol{y})d\tau\\
&\propto \int_0^\infty (\tau)^{\frac{n}{2}-1}\exp\Big[-\frac{\tau}{2}(n-k-1)s^2\Big]\exp\Big[-\frac{\tau}{2}(b-\beta)^TX^TX(b-\beta)\Big]d\tau.
\end{aligned}
\tag{2.13}
$$

By carefully observing the last term, we can recognise the main body of a $\Gamma\left[\frac{n}{2}, \frac{(n-k-1)s^2 + (b-\beta)^T X^T X (b-\beta)}{2}\right]$. Consequently, we can perform the integration, deriving the following result:

$$p(\boldsymbol{\beta}|\boldsymbol{y}) \propto \frac{\left(\frac{(n-k-1)s^2 + (\beta-b)^T X^T X (b-\beta)}{2}\right)^{-\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)}$$

$$\propto \left((n-k-1)s^2\right)^{-\frac{n}{2}} \left[1 + \frac{(b-\beta)^T X^T X (b-\beta)}{(n-k-1)s^2}\right]^{-\left(\frac{n-k-1}{2} + \frac{k+1}{2}\right)}$$

$$\propto \left[1 + \frac{(b-\beta)^T X^T X (b-\beta)}{(n-k-1)s^2}\right]^{-\left(\frac{n-k-1}{2} + \frac{k+1}{2}\right)}.$$

We can conclude that $\boldsymbol{\beta}|\boldsymbol{y} \sim t_{n-k-1}\left(b, \frac{X^T X}{(n-k-1)s^2}\right)$. However, the marginal distribution of $\boldsymbol{\beta}$ is rarely used since it is more convenient to sample the marginal posterior distribution of $\boldsymbol{\beta}$ via the Gibbs sampler using the conditional distribution of $\boldsymbol{\beta}$.

Checking the fit of the model

Before we proceed to a more thorough examination of the model parameters, we should check the credibility of that particular model. In other words, we need to confirm the consistency of our prior assumptions and the general fit of the model to the data, since only then we are justified to continue our analysis. A way to do that is to use the *Bayesian p-value*, which will provide a first hint of the suitability of the model.We should mention that the Bayesian p-value, despite its popularity, is a point of controversy for statisticians: it is either approved and recommended or faced with scepticism. Each side, nevertheless, has famous supporters. Since we will be based on Gelman's definition of the posterior predictive p-value, in order to comprehend it, we need to clarify the meaning of what we will refer to us a *replicated value*:
By $\boldsymbol{y^{rep}}$ we refer to *replicated data*, which are the observations we would get, instead of $\boldsymbol{y}$, if the suggested model was true. These values would be obtained, if we used the original matrix X as input to the model we

would get, after our posterior inference. The main concept behind their use is that if we are dealing with a credible model, they should be close to the corresponding values of $\boldsymbol{y}$ as measured by a test quantity.

Usually, the replicated values are simulated after the posterior distributions of the unknown parameters of the model are obtained. The vector of the replicated values will have dimension $n \times 1$ and we will refer to it by $\boldsymbol{y^{rep}} = \left(y_1^{rep}, y_2^{rep}, ..., y_n^{rep}\right)$. The steps of the simulation of the replicated and the predicted values are the same:

- Fist, we draw the precision $\tau^{(1)}$ as a value of the distribution $Inv-\chi^2(n-k-1, s^2)$.
- Then we draw $\boldsymbol{\beta}^{(1)} \sim t_{n-k-1}\Big(b, \frac{X^T X}{(n-k-1)s^2}\Big)$.
- Baring in mind that $y_1^{rep} \sim N(X[1,\cdot]\boldsymbol{\beta}^{(1)}, \tau^{(1)})$, we draw y_1^{rep} as a value of $N(X[1,\cdot]\boldsymbol{\beta}^{(1)}, \tau^{(1)})$.
- We repeat the same steps for the rest of the $y_j^{rep}, \quad j = 2, .., n$.

We introduce the definition of the Bayesian p-value according to Gelman ("*Bayesian Data Analysis*" (2014)).

Definition 2.1.1. (*p-value*)
Let θ be the vector of unknown parameters. Then, given a test quantity $T(y, \theta)$ or $T(y)$, the posterior predictive p-value is defined as the probability:

$$ppv = P\Big[T(y^{rep}, \theta) \geqslant T(y, \theta)\Big|y\Big] = \iint \mathbb{I}_{T(y^{rep},\theta)\geqslant T(y,\theta)} p(y^{rep}|\theta, y)p(\theta|y)dy^{rep}d\theta. \tag{2.14}$$

The Bayesian p-value is the probability that the data, generated under the assumption that the studied model is true, could be more extreme than the observed data, as measured by the test quantity. As far as the test quantity is concerned, its formula is based entirely on the nature of the data and it has to be carefully constructed in order to capture in the best possible way the discrepancies between the observed and the replicated data. Since the functional form of T is determined by the statistician and the values of y are available, $T(y, \theta)$ can be calculated immediately and will get an arithmetic value. Furthermore, as it is noted in the definition, T can be a function of both the unknown parameter and the data.

A simplified interpretation of the posterior predictive p-value would be that it is the percentage of extreme discrepancies of the kind we have just described. The general notion is that if the model generally fits, then extreme discrepancies between the values of T for y and y^{rep} should occur rarely. That means that p-values that are greater or smaller than the percentages we set as boundaries (usually 0.95 as the highest value and 0.05 as the lowest) would indicate misfit of the model.

In our case,the vector of the unknown parameters is $\theta = \left(\boldsymbol{\beta}, \tau\right)$ and so (1.14) will take the following form:

$$ppv = \iint \mathbb{I}_{T(y^{rep},\boldsymbol{\beta},\tau)\geqslant T(y,\boldsymbol{\beta},\tau)} p(y^{rep}|\boldsymbol{\beta},\tau,y)p(\boldsymbol{\beta},\tau|y)dy^{rep}d\boldsymbol{\beta}d\tau \quad (2.15)$$

2.1.2 Inference for the coefficient vector of β

The most common inferential problem in classical statistics is the attempt to reduce the number of explanatory variables by assessing the statistical significance of the coefficients associated with each one of them. In classical statistics,these conclusions are drawn based on the p-values and the confidence intervals of each coefficient. In the Bayesian paradigm, p-values have a different use and are used to check the goodness of fit as we will display in a following section.

Inference via Point Estimators
As it has been shown previously, $\boldsymbol{\beta} \sim t_{k+1}(b, s^2(X^TX)^{-1}, n-k)$, which yields that $\beta_j|\boldsymbol{y} \sim t(b_j, s^2((X^TX)^{-1})_{jj}, n-k)$ for $j \in \{0, ..., k\}$. Since the properties of the Student's t-distribution have been extensively studied in Probability Theory, the main location and variance parameters can be easily derived, via already available formulas. More specifically, for the posterior distribution of each coefficient parameter, denoted by β_j, the same value corresponds to the descriptive statistics of mean, mode and median and that is $\overline{\beta_j}|\boldsymbol{y} = b_j$. Also, the variance is determined by the next formula: $Var(\beta_j|\boldsymbol{y}) = \frac{n-k}{n-k-2}s^2\left(\left(X^TX\right)^{-1}\right)_{jj}$. In addition to that, any statistical language can easily plot a t-distribution, thus providing a visual image, which contributes to the general perception of

the posterior distribution. It is equally easy to obtain the interquartile range.

Using Highest Posterior Density Regions for inference regarding hypothesis testing

Highest posterior density regions are a rather quick, intuitive statistical tool, which is frequently employed by Bayesians strictly for inference and not as a model selection technique. Regarding the task of parameter selection in particular, they can provide useful insight concerning hypothesis tests of the form $H_0 : \beta_j = 0$ versus $H_1 : \beta_j 0$ for $j = 0, 1, 2, ..., k$, by cross-examining the behaviour of the posterior distribution, under each hypothesis.
We recall that $\boldsymbol{\beta} \sim t_{k+1}(b, s^2(X^T X)^{-1}, n-k)$.
Consequently,$\beta_j|\boldsymbol{y} \sim t(b_j, s^2((X^T X)^{-1})_{jj}, n-k)$.
Since the posterior distribution of $\beta_j|\boldsymbol{y}$ is known for every j,we can obtain 95% or 99% HPDIs for each one of them using the properties of the t-distribution and study the position of 0 in the formed area. Thus, we can get a hint for the contribution of a specific variable to the interpretation of the data. We clarify that by obtaining a 95% HPDI for a β_j that does not include zero, we can state that $P\Big(\beta_j \neq 0\Big) = 95\%$. By displaying the 95% or 99% HPDI for every β_j, $j \in 1, 2, .., k$ as well as the posterior mean and the 25% and the 95% quantiles, which can be easily simulated by R, we can *informally* justify a model comparison between the two models, corresponding to H_0 and H_1 respectively.

It has to be pointed out that, according to the Bayesian approach, HPDIs cannot be used as a method for model selection, which constitutes an obvious contrast to the framework of Classical statistics.

Quite often, we are dealing with more complicated, multiple restrictions such as $H_0 : \beta_1 + \beta_2 = -1$ versus $H_1 : \beta_1 + \beta_2 \neq -1$. These constraints are expressed, using matrix notation. The general formula we use is $R\beta = w$, where the dimensions of the matrix R and the vector m depend on the number of constraints.We provide an example:
Suppose we have three explanatory variables and we want to test the hypotheses

$$\begin{aligned} H_0 : \beta_1 + \beta_2 - \beta_3 &= 0 \\ \beta_3 - \beta_2 &= -1 \end{aligned}$$

versus

$$H_1 : \beta_1 + \beta_2 - \beta_3 \neq 0$$
$$\beta_3 - \beta_2 \neq -1 \quad ,$$

then

$$R = \begin{bmatrix} 0 & 1 & 1 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix}$$

and

$$w = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$$

As we observe, if we have l constraints and k coefficients, then R is a $l \times k+1$ matrix and w an l-dimensional vector.

The following theorem refers to a particular feature of the t-distribution, with a very helpful computational use.

Theorem 3. *Let W be a k-dimensional vector following the multivariate $t(\mu, V, \nu)$ distribution and A be an $m \times k$ non-stochastic matrix with $rank(A) = m$. Then $AW \sim t(A\mu, AWA^T, \nu)$.*

Based on the above theorem, given that $\boldsymbol{\beta} \sim t_{k+1}(b, s^2(X^TX)^{-1}, n-k)$, we may assume that $R\boldsymbol{\beta} \sim t_l(Rb, s^2R(X^TX)^{-1}R^T, n-k)$. Since we are studying a random variable vector, we will, automatically, be exploring a multidimensional subspace. This amounts to the derivation of *confidence regions*, denoted by D, which are in fact hypersurfaces, usually hyperellipsoids, with the following property:

$$\int_D p(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{y}) d\boldsymbol{\beta} = 1 - \alpha.$$

Usually, $\alpha = 0,05$ or $\alpha = 0,1$. We will prefer the first value throughout this thesis.

Consequently, our confidence region will satisfy the property:

$$\int_D p(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{y}) d\boldsymbol{\beta} = 0,95,$$

and since we are dealing with a multivariate t-distribution, we know that the vectors that satisfy this constraint are the vectors β with $p(\beta|\boldsymbol{X}, \boldsymbol{y})$. We can state that the null hypothesis is favoured by the data, if the condition

$$\frac{(\boldsymbol{Hb - w})^{\boldsymbol{T}}(\boldsymbol{H^T(X^TX)^{-1}H(Hb - w)}}{\boldsymbol{ls^2}} < \boldsymbol{F_{1-\alpha,l,n-k}}$$

is satisfied. However, we are restricted to inference and we do not proceed to decision making, by relying on the particular outcome.

2.1.3 Model Selection

We will focus on the comparison between models embodying the various coefficient combinations. For that purpose, we could employ the Bayesian Information Criterion, as well as the Deviance Information Criterion.

We recall that the likelihood function is expressed as:

$$p(\boldsymbol{y}|\boldsymbol{\beta},\tau) \equiv L(\boldsymbol{y}|\boldsymbol{\beta},\tau) = (2\pi)^{-\frac{n}{2}}(\tau)^{\frac{n}{2}} \exp\Big[-\frac{\tau}{2}(\boldsymbol{y}-\boldsymbol{X\beta})^T(\boldsymbol{y}-\boldsymbol{X\beta})\Big].$$

Therefore, the log-likelihood takes the form:

$$lnL((\boldsymbol{y}|\boldsymbol{\beta},\tau) = -\frac{n}{2}ln(2\pi) + \frac{n}{2}ln\tau - \frac{\tau}{2}\cdot(\boldsymbol{y}-\boldsymbol{X\beta})^T(\boldsymbol{y}-\boldsymbol{X\beta}).$$

In both information criteria, the first step would be to maximize the *deviance*, which was defined as the function: $D(\boldsymbol{\beta}) = -2lnL((\boldsymbol{y}|\boldsymbol{\beta},\tau)$, which is equivalent to minimizing the log-likelihood . Since our interest lies on the impact of the coefficient vector, we will consider that we need to minimize $lnL((\boldsymbol{y}|\boldsymbol{\beta},\tau)$ over the parameter space of $\boldsymbol{\beta}$, as this is formed by each studied model.

Supposing we wish to do so for an arbitrary model M_i, with corresponding parameter space denoted by $\mathcal{B}_i$, for the particular model structure, we can achieve this by analytical and computational methods. As a computational method, we could employ the widely-used Newton-Raphson algorithm. However, the vector $\hat{\boldsymbol{\beta}} \in \mathcal{B}_i$ that minimizes the log-likelihood can as well be determined as the vector that satisfies the following equation:

$$\frac{\partial lnL(\boldsymbol{y}|\boldsymbol{\beta},\tau)}{\partial \boldsymbol{\beta}} = \mathbf{0}$$

The solution would be the vector $\hat{\boldsymbol{\beta}} = \left(X^TX\right)^{-1}X^T\boldsymbol{y}$, namely the ***least squares estimator*** of $\boldsymbol{\beta}$. If $\hat{\boldsymbol{\beta}}$ indeed minimizes the log-likelihood function, then the condition below has to be satisfied as well:

$$\frac{\partial^2 lnL((\boldsymbol{y}|\boldsymbol{\beta},\tau)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} > \mathbf{0},$$

which can be more simply expressed as the restrain that the matrix $X^T X$ has to be positive-definite. Having derived the required vector $\hat{\boldsymbol{\beta}}$, the Bayesian Information Criterion and the Deviance Information Criterion can be easily calculated and suggest the optimal model. It is noted that the Bayes factors, as well as the posterior probabilities of each model, could be approximated by the corresponding BIC scores.

2.1.4 The posterior predictive distribution

Since the predictive performance of a model is often of major concern, another important part of posterior inference is to compute the predictive distribution. By $\boldsymbol{y^{pr}}$ we refer to the *predicted data*; that is an $l \times 1$-dimensional vector of the observations we would obtain by using a new $l \times (k+1)$ matrix of explanatory variables, X_{new}, in the studied model, which has first been updated with the posterior distributions of the unknown variables. We should note that X_{new} contains totally new data, which are not involved in the derivation of the posterior distributions.

Analytic Computation

It is possible to obtain an analytic form of the posterior predictive distribution by computing the following integral:

$$p(\boldsymbol{y^{pr}}|\boldsymbol{y}) = \int_B \cdots \int \int_\tau p(\boldsymbol{y^{pr}}|\boldsymbol{y}, \boldsymbol{\beta}, \tau) \cdot p(\boldsymbol{\beta}, \tau|\boldsymbol{y}) d\boldsymbol{\beta} d\tau \tag{2.16}$$

where $\boldsymbol{y^{pr}}|\boldsymbol{y}, \boldsymbol{\beta}, \tau \sim N_l(X_{new}\beta, \frac{1}{\tau} I)$ and so we can set

$$p(\boldsymbol{y^{pr}}|\boldsymbol{y}, \boldsymbol{\beta}, \tau) \propto (\tau)^{\frac{l}{2}} \exp\Big[-\frac{\tau}{2} (\boldsymbol{y^{pr}} - \boldsymbol{X_{new}\beta})(\boldsymbol{y^{pr}} - \boldsymbol{X_{new}\beta}) \Big]$$

Simulation via an MCMC algorithm

The computation of (1.15) can be proved to be very challenging and for that reason, it is preferred to simulate the posterior predictive distribution via the MCMC algorithm. By $\boldsymbol{y^{pr}} = \big(y_1^{pr}, ..., y_l^{pr}\big)$, we will be referring to the vector of predictions.
The steps we take for the simulation are the following:

- We get the arithmetic values of the quantities b and s^2 as we have determined them previously.
- We draw $\tau^{(1)} \sim Inv - \chi^2(n-k-1, s^2)$ which is the posterior distribution of τ.
- We draw $\boldsymbol{\beta^{(1)}} \sim t_{n-k-1}\Big(b, \frac{X^T X}{(n-k-1)s^2}\Big)$.
- Baring in mind that $y_1^{pr}|y, \beta^{(1)}, \tau^{(1)} \sim N(X_{new}[1,\cdot]\beta^{(1)}, \tau^{(1)})$, we draw y_1^{pr} as a value from $N(X_{new}[1,\cdot]\beta^{(1)}, \tau^{(1)})$.
- We draw $\tau^{(2)} \sim Inv - \chi^2(n-k-1, s^2)$.
- We draw $\boldsymbol{\beta^{(2)}} \sim t_{n-k-1}\Big(b, \frac{X^T X}{(n-k-1)s^2}\Big)$.
- And y_2^{pr} will be obtained as a value from $N(X_{new}[2,\cdot]\beta^{(2)}, \tau^{(2)})$.

We repeat the same steps l times in total; that is the number of lines of X_{new} and thus, we get the predicted values under the model.

2.2 Using conjugate prior distributions for β and τ

The use of conjugate prior distributions is very helpful with the computational part, since the posterior distributions will belong to the same distributional family as the priors. Consequently, integrating and recognising distributions is much simpler than in the case of non-informative priors. Usually, independence between $\boldsymbol{\beta}$ and τ is assumed and therefore we will assign the following conjugate priors to each one of them:

- $\boldsymbol{\beta} \sim N_{k+1}(b_0, \tau^{-1}T_0)$, where b_0 is a $k+1$-dimensional vector and T_0 is a $k+1 \times k+1$ positive matrix, the form of which depends on the assumptions we make for $\beta_j, j = 0, 1, ..., k$. The vector and the matrix are defined by us and so they are fixed. Here, we will assume that the β_js are independent and, therefore, uncorrelated and so T_0 will be diagonal. The jth diagonal element is the variance of $\beta_j, j = 0, 1, ..., k$. In the more general case, which includes correlations between the β_js, T_0 will be a positive, symmetric matrix and its non-diagonal elements are these correlations.

- $\tau \sim \Gamma\Big(\frac{\nu_0}{2}, \frac{\nu_0}{[2\tau_0]}\Big)$, where τ_0 is a prior belief about the precision and by ν_0 we express how strong this belief is.

Although conjugate priors are referred to as informative, we can adjust the provided information by setting the appropriate precision. More

specifically, if information about the mean and the variance is given by the experimenter, then we set a prior distribution that satisfies these conditions. If we are based solely on our own beliefs concerning a parameter, but we strongly believe that they are close to the truth, then we set a small value for the precision of the prior distribution of that particular parameter, making the variance smaller. Thus, our belief will strongly influence our results and we should always take this fact into consideration. On the opposite hand, if we are sceptical or not so certain about our guess, then by setting a large precision, the prior distribution will affect less the information provided by the data. Such distributions are often called *diffuse priors*. For instance, if we are convinced that the explanatory variable X_j is not needed in our model, we can use as prior distribution of β_j a Normal distribution with mean 0 and precision about 0,5. If, however, we think that we put the analysis under risk, we can use the same Normal distribution with $\tau = 100$. The value of τ can increase or decrease depending on our certainty.

2.2.1 Obtaining the posterior distributions

The analytic formulas of the prior density functions mentioned above are the following:

$$p(\boldsymbol{\beta}) = (2\pi)^{-\frac{k+1}{2}} \tau^{\frac{k+1}{2}} |T_0|^{-\frac{1}{2}} \exp\big[-\frac{1}{2}(\boldsymbol{\beta} - b_0)^T \tau T_0^{-1} (\boldsymbol{\beta} - b_0)\big],$$

$$p(\tau) = \frac{\big(\frac{\nu_0}{[2\tau_0]}\big)^{\frac{\nu_0}{2}}}{\Gamma\big(\frac{\nu_0}{2}\big)} \tau^{\frac{\nu_0}{2}-1} \exp\bigg(-\frac{\nu_0}{[2\tau_0]}\tau\bigg).$$

The likelihood function of the model has already been expressed in (1.6). After applying Bayes' Theorem we have the relationship:

$$p(\boldsymbol{\beta}, \tau | y) \propto \tau^{\frac{\nu_0+n+k+1}{2}-1} \exp\bigg[-\frac{\tau}{2}\big[(y - X\beta)^T (y - X\beta) - (\beta - b_0)^T T_0^{-1} (\beta - b_0)\big] - \frac{\nu_0 \tau}{[2\tau_0]}\bigg].$$

With conjugate prior distributions, the marginal distributions of $\boldsymbol{\beta}$ and τ can be analytically computed. In fact, the computation of the conditional posterior distribution of $\tau|\boldsymbol{y}$ is straightforward. If we examine carefully the joint posterior distribution of τ and $\boldsymbol{\beta}$ and bearing in mind that there is a dependence between $\boldsymbol{\beta}$ and τ, due to the conditional form of the prior distribution of the first, we can rewrite the above distribution as a function of τ:

$$p(\boldsymbol{\beta}, \tau|y) \propto \tau^{\frac{\nu_0+n+k+1}{2}-1} \exp\left[-\frac{\tau}{2} \cdot A\right] \tag{2.17}$$

Consequently, we can conclude that $\tau|\boldsymbol{y} \sim \Gamma\left(\frac{n+k+1+\nu_0}{2}, \frac{A}{2}\right)$, where $A \in \mathbb{R}$, since $A = (y - X\beta)^T(y - X\beta) - (\beta - b_0)^T T_0^{-1}(\beta - b_0) + \frac{\nu_0}{[\tau_0]}$.

In order to obtain the marginal distribution of $\boldsymbol{\beta}$, we need to perform the following computation:

$$p(\boldsymbol{\beta}|\boldsymbol{y}) = \int_0^\infty p(\boldsymbol{\beta}, \tau|\boldsymbol{y}) d\tau. \tag{2.18}$$

In order to be able to deduce and recognise the form of the distribution, we will once again use the least squares estimate $b = (X^T X)^{-1} X^T y$, as well as the quantity $s^2 = \frac{(y - Xb)^T(y - Xb)}{n - k - 1}$.
The joint posterior distribution can be rewritten and will have the following form:

$$p(\boldsymbol{\beta}, \tau|\boldsymbol{y}) \propto \tau^{\frac{n+k+1+\nu_0}{2}} exp\Bigg\{ \frac{-\tau}{2}\Big[s^2(n-k-1) + (b-\boldsymbol{\beta})^T X^T X (b-\boldsymbol{\beta})$$

$$+ (\boldsymbol{\beta} - b_0)^T T_0^{-1}(\boldsymbol{\beta} - b_0) + \frac{\nu_0}{[\tau_0]}\Big]\Bigg\}$$

In order to be able to compute (1.19). we will replace the joint distribution in the integral with the above form. Also, the following additive calculations are necessary to complete the task:

$$(b-\boldsymbol{\beta})^T X^T X (b-\boldsymbol{\beta}) + (\boldsymbol{\beta} - b_0)^T T_0^{-1}(\boldsymbol{\beta} - b_0) = (\boldsymbol{\beta} - \mu)^T T^{-1}(\boldsymbol{\beta} - \mu) + Q,$$

where:

- $Q = b^T X^T X b + b_0^T T_0^{-1} b_0 - \mu^T T^{-1} \mu$
- $T = X^T X + T_0^{-1}$
- $\mu = T(X^T X + T_0^{-1})$

Consequently, (1.19) will be:

$$p(\boldsymbol{\beta}|\boldsymbol{y}) \propto \int_0^\infty \tau^{\frac{n+k+1+\nu_0}{2}} exp\Bigg\{ \frac{-\tau}{2}\Big[s^2(n-k-1) + (\boldsymbol{\beta} - \mu)^T T^{-1}(\boldsymbol{\beta} - \mu) + Q + \frac{\nu_0}{[\tau_0]}\Big]\Bigg\} d\tau$$

$$\propto \Big[(\boldsymbol{\beta} - \mu)^T T^{-1}(\boldsymbol{\beta} - \mu) + N\Big]^{\frac{-n+k+1+\nu_0}{2}}, \Big]$$

where $N = s^2(n-k-1) + Q + \frac{\nu_0}{[\tau_0]}$.
We can conclude that $\boldsymbol{\beta|y} \sim t_{k+1}(\mu, T)$ with $\nu_0 + n$ degrees of freedom.

Inference via Point Estimators and Highest Posterior Density Regions

In order to obtain HPDIs for each β_j or the highest posterior density region of the vector β, it is necessary to have the marginal posterior distributions. Based on the previous analysis, we can assume that $\beta_j \sim t(\mu_j, (X^T X + T_0^{-1}[\tau_0]^{(-1)})_{jj})$ and since it is a known distribution, we can easily obtain a 95% HPDI for $j = 0, 1, ..., k$. By checking whether zero is included in each one of these HPDIs, we get an indication, whether the null hypothesis $H_0 : \beta_j = 0$ is favoured by the data or not.

Generally, we can examine any hypothesis of the form $H_0 : \beta_j = a$ versus $H_1 : \beta_j \neq a$ for every j and for any real number a, by simply observing whether a is included in the HPDI. For more complicated forms of the null hypothesis such as $H_0 : R\beta = w$ or $H_0 : R\beta > w$, we use the Highest Posterior Density Regions. Based on the theorem, we can assume that $R\beta \sim t_l(R\mu, R^T(X^T X + T_0^{-1})R)$ and then, once again H_0 would appear plausible, if the condition

$$\frac{(\boldsymbol{Rb - w})^{\boldsymbol{T}} \boldsymbol{R}^{\boldsymbol{T}} (\boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{X} + \boldsymbol{T}_{\boldsymbol{0}}^{-\boldsymbol{1}})(\boldsymbol{Rb - w})}{\boldsymbol{l s}_{\boldsymbol{0}}^{\boldsymbol{2}}} < \boldsymbol{F}_{\boldsymbol{1-\alpha, l, n-k}}$$

is satisfied. The reader is reminded that α denotes the level of statistical significance, as this is specified by either the statistician or the experimenter.

2.2.2 Model Selection

Using Bayes' Factors

Another advantage of using conjugate priors is that it allows us to compute analytically the Bayes' Factor. Consequently, the comparison between two models can be conducted in this way as well. Specifically, we can test the hypothesis $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$ by comparing the full model $M1$ with $M0$, which is the model without the explanatory variable X_j; that is the model where $\beta_j = 0$, $j = 1, 2, ..., j$. What we need to do is compute the marginal likelihoods under each model and then compute the ratio:

$$BF = \frac{p(y|M1)}{p(y|M0)} \tag{2.19}$$

If $BF > 1$, then we have evidence against $M0$, which means that X_j is rather significant, otherwise, it can be omitted.
Marginal likelihoods are generally difficult to compute and usually an analytic form cannot be obtained, since we are dealing with complicated integrals. However, due to conjugate priors these integrals are much simpler as we will demonstrate below.

We need to point out that $p(y|M1)$ is the *marginal likelihood or evidence* of the model M_1 and it is obtained by integrating the joint poste-

rior distribution $p(y, \boldsymbol{\beta}, tau|M_1)$ over $\boldsymbol{\beta}$ and τ as follows:

$$
\begin{aligned}
p(y|M1) &= \int\int\underset{B}{\cdots}\int p(y, \boldsymbol{\beta}, tau|M_1)d\boldsymbol{\beta}d\tau = \int\int\underset{B}{\cdots}\int p(y|\boldsymbol{\beta}, \tau)p(\boldsymbol{\beta}|\tau)d\boldsymbol{\beta}d\tau \\
&= (2\pi)^{-\frac{n+k+1}{2}}|T_0|^{-\frac{1}{2}}\frac{\left(\frac{\nu_0}{2[\tau_0]}\right)^{\frac{\nu_0}{2}}}{\Gamma\left(\frac{\nu_0}{2}\right)}\times \\
&\times\int\int\underset{B}{\cdots}\int \exp\Big\{-\frac{\tau}{2}(y-X\beta)^T(y-X\beta)+(\beta-b_0)^T T_0^{-1}(\beta-b_0)+\frac{\nu_0}{2[\tau_0]}\Big\}\tau^{\frac{\nu_0+n+k-1}{2}}d\boldsymbol{\beta}d\tau \\
&= (2\pi)^{-\frac{n+k+1}{2}}|T_0|^{-\frac{1}{2}}\frac{\left(\frac{\nu_0}{2[\tau_0]}\right)^{\frac{\nu_0}{2}}}{\Gamma\left(\frac{\nu_0}{2}\right)}\times \\
&\times\int\int\underset{B}{\cdots}\int \exp\Big\{-\frac{\tau}{2}(\beta-b_1)^T\Sigma_0^{-1}(\beta-b_1)+\frac{\nu_0}{[\tau_0]}\tau^{\frac{\nu_0+n+k+1}{2}}d\boldsymbol{\beta}d\tau\Big\} \\
&= (2\pi)^{-\frac{n+k+1}{2}}|T_0|^{-\frac{1}{2}}\frac{\left(\frac{\nu_0}{2[\tau_0]}\right)^{\frac{\nu_0}{2}}}{\Gamma\left(\frac{\nu_0}{2}\right)}\frac{\Gamma\left(\frac{\nu_0+n+k+1}{2}\right)}{2} \\
&\times\int\underset{B}{\cdots}\int\Big[\frac{\nu_0}{[\tau_0]}+(\beta-b_1)^T\Sigma_0^{-1}(\beta-b_1)\Big]^{-\frac{\nu_0+n+k+1}{2}}d\boldsymbol{\beta} \\
&= (2\pi)^{-\frac{n+k+1}{2}}|T_0|^{-\frac{1}{2}}\frac{\left(\frac{\nu_0}{2[\tau_0]}\right)^{\frac{\nu_0}{2}}}{\Gamma\left(\frac{\nu_0}{2}\right)}\frac{\Gamma\left(\frac{\nu_0+n+k+1}{2}\right)}{2}\frac{\Gamma(\nu 0+n)(\nu 0+n)^{\frac{k+1}{2}}\pi^{\frac{k+1}{2}}|\Sigma_0|^{\frac{1}{2}}}{\Gamma\left[\frac{\nu_0+n+k+1}{2}\right]} \\
&= 2^{-\frac{n+k+3}{2}}\pi^{-\frac{n}{2}}|T_0|^{-\frac{1}{2}}\frac{\left(\frac{\nu_0}{2[\tau_0]}\right)^{\frac{\nu_0}{2}}}{\Gamma\left(\frac{\nu_0}{2}\right)}\Gamma(\nu 0+n)(\nu 0+n)^{\frac{k+1}{2}}|\Sigma_0|^{\frac{1}{2}}
\end{aligned}
\tag{2.20}
$$

where similarly:

$$
\begin{aligned}
\Sigma_0^{-1} &= X^T X + T_0^{-1} \\
b_1 &= \Sigma_0\big(X^T y + T_0^{-1}b_1\big)
\end{aligned}
$$

We derive that:

$$p(y|M0) = 2^{-\frac{n+k+2}{2}}\pi^{-\frac{n}{2}}|T_1|^{-\frac{1}{2}}\frac{\left(\frac{\nu_0}{2[\tau_0]}\right)^{\frac{\nu_0}{2}}}{\Gamma\left(\frac{\nu_0}{2}\right)}\Gamma(\nu 0+n)(\nu 0+n)^{\frac{k}{2}}|\Sigma_1|^{\frac{1}{2}}$$

According to $M1$, β_j is excluded, consequently b_{0j}, the jth column and row of T_0 and X_j also have to be excluded.
By Σ_1 and T_1 we denote the modified matrices Σ_0 and T_0.
After our calculations, we can obtain a result for (1.12) in the form of the following formula:

$$BF = 2^{-\frac{1}{2}}\frac{|T_1|^{\frac{1}{2}}}{|T_0|^{\frac{1}{2}}}(\nu_0+n)\frac{|\Sigma_0|^{\frac{1}{2}}}{|\Sigma_1|^{\frac{1}{2}}}$$

It is rather obvious that Bayes' Factor is sensitive to our prior assumptions, especially those concerning the variance of the variables, whereas the actual data, influence the outcome solely through the terms Σ_0 and Σ_1. This is a fact that has been long pointed out by Bayesian statisticians and therefore it is very important to choose the prior distributions with caution, as unreasonable priors would lead to unreasonable inference. Another disadvantage of this method, is that Bayes' Factor usually cannot be defined when we use improper prior distributions. Finally, when we are dealing with very small values of BF, the use of the logarithm of BF (noted as LBF) because it is a more stable form, by which we also gain more computational precision.

Model Comparison based on the Posterior Probabilities of the candidate models

Another more intuitive way of comparing models is to compare their posterior probabilities and consider the optimal model to be the one with the highest posterior probability, assuming that this would be the model which is best supported by the data. Every possible combination of the parameters will be matched to a different model, as we did before. Consequently, we end up with the parameters of the model M_j, where $M_j : P(M_j|y) = \max_{i\in\mathcal{M}} P(M_i|y)$. The posterior probability of each model, using Bayes' Theorem, is given by:

$$P(M_j|y) = \frac{P(y|M_j)P(M_j)}{\sum_{i\in\mathcal{M}} P(y|M_j)P(M_j)}$$

Once again, the outcome is affected by the prior distributions and also by our initial convictions concerning the plausibility of each model, which are expressed by the prior probabilities $P(M_j)$ for each j. If we prefer not to favour any model, then we can choose all of the models to have equal prior probabilities. Consequently, the term that interests us is $P(y|M_j)$, the calculation of which, has been shown previously.

Having conducted the model comparison, based on the above criteria, we could skip calculating the information criteria of BIC and DIC, since the latter are often regarded as equivalent means of model selection. In fact, if it is possible for the Bayes factor and the posterior model probabilities to be derived in a straightforward way, they are generally preferred over information criteria in Bayesian Statistics.

2.2.3 The posterior predictive distribution

Let X_{new} be a $l \times k+1$ matrix containing new observations, for which we need to predict the values of the dependent variable, denoted by $\boldsymbol{y^{pr}}$, as in the previous section. It is known that $\boldsymbol{y^{pr}}|\boldsymbol{y}, \boldsymbol{\beta}, \tau \sim N_l\Big(X_{new}\boldsymbol{\beta}, \frac{1}{\tau}I\Big)$.

Analytic Computation
We can obtain the posterior predictive distribution by calculating the following integral:

$$p(\boldsymbol{y^{pr}}|\boldsymbol{y}) = \int_B \cdots \int \int_\tau p(\boldsymbol{y^{pr}}|\boldsymbol{y}, \boldsymbol{\beta}, \tau) \cdot p(\boldsymbol{\beta}, \tau|\boldsymbol{y}) d\boldsymbol{\beta} d\tau$$

Obtaining the predicted values via simulation
It is a common practice to get the predicted values via simulation, once we have derived the posterior distributions of the unknown parameters. We will follow the same procedure we have described in the previous section, when we referred to the predictive values with the slight difference that in this case we will be using only marginal posterior distributions:

- We get the arithmetic values of the quantities A, b and s^2 as we have determined them previously.
- We draw $\tau^{(1)} \sim \Gamma\left(\frac{n+k+1+\nu_0}{2}, \frac{A}{2}\right)$ which is the posterior distribution of τ.

- We draw $\boldsymbol{\beta^{(1)}} \sim t_{k+1}(\mu, T)$.
- Baring in mind that $y_1^{pr}|y, \beta^{(1)}, \tau^{(1)} \sim N(X_{new}[1,\cdot]\beta^{(1)}, \tau^{(1)})$, we draw y_1^{pr} as a value from $N(X_{new}[1,\cdot]\beta^{(1)}, \tau^{(1)})$.
- We draw $\tau^{(2)} \sim \Gamma\left(\frac{n+k+1+\nu_0}{2}, \frac{A}{2}\right)$.
- We draw $\boldsymbol{\beta^{(2)}} \sim t_{k+1}(\mu, T)$.
- And y_2^{pr} will be obtained as a value from $N(X_{new}[2,\cdot]\beta^{(2)}, \tau^{(2)})$.

We repeat the same steps l times in total; that is the number of lines of X_{new} and thus, we get the predicted values under the model.

Chapter 3

Bayesian Inference for Generalized Linear Models

3.1 Introduction

3.1.1 Motivation

Linear models are appealing due to the straightforward and rather intuitive interpretation of the relationship between the response and the explanatory variables and because they can be handled relatively easily in terms of computation and inference. As mentioned before, the main assumption that has to be made is that the mean of the response variable $\boldsymbol{y}$ is a linear function of the explanatory variables $x_1, x_2, ..., x_k$. In mathematical notation that is expressed as:

$$\boldsymbol{E}[\boldsymbol{y}|\boldsymbol{X}] = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k. \tag{3.1}$$

We also assume that the model is described by the formula:

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2), \quad i = 1, ..., n, \quad n, k \in \mathbb{N}. \tag{3.2}$$

However, these very conditions that make them easy to handle, also ensure their limited applicability, since they are very restrictive and rather seldom satisfied. For instance, linear models fail to describe adequately the relationship between $\boldsymbol{y}$ and $\boldsymbol{X}$ when $\boldsymbol{y}$ is a binary or dichotomous variable describing the occurrence or non-occurrence of an event and consequently $\boldsymbol{y} \in \{0, 1\}^n$. Another example is the case of $\boldsymbol{y}$ being a vector of counts, as it would be if y_i represented the number of customers arriving in a bank at the i-th specified time period, which yields that $y_i \in \mathbb{N} \quad i = 1, ..., n$. We can conclude that linear models cannot be an

option for a very wide area of applications and, therefore, alternative options need to be investigated.

3.1.2 Basic Assumptions of the GLM Class

Generalized Linear Models, often referred to us GLMs, enable us to conduct a statistical analysis when the support of the response variable is $\mathbb{R}_+$ or $\mathbb{N}$. The appeal of GLMs is due to the really wide range of applications they provide (clinical, environmental researches etc), but also because they sustain the familiar concept of linearity.
We could say that the fundamental idea of the GLM theory is expressed mathematically by the following statement:

$$y|X, \beta \sim \boldsymbol{f}(x_i^T \beta), \tag{3.3}$$

where $\boldsymbol{y} = (y_1, y_2, ..., y_n)^T$ is the $1 \times n$ vector of the dependent variable, $\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix}$ is a $n \times (k+1)$ matrix and $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)^T$ is the $(k+1) \times 1$ vector of the regression coefficients. The special restriction we impose upon function $\boldsymbol{f}$ is that it must belong to the *exponential family* of distributions.

The exponential family

Definition 3.1.1. Let $\boldsymbol{y} = \big(y_1, ..., y_n\big) \in \mathbb{R}^n$ denote an n-dimensional random variable. The distribution f of the n-dimensional random variable $\boldsymbol{y}$ given the *natural parameter* vector $\boldsymbol{\theta} = \big(\theta_1, ..., \theta_k\big) \in \Theta \subset \mathbb{R}^k$ and dispersion parameter ϕ, belongs to the exponential family if:

- The support of the distribution, $S = \Big\{ y \in \mathbb{R}^n : f(y|\boldsymbol{\theta}, \phi) > 0 \Big\}$, is independent of θ and ϕ.
- The density function of each y_i may be written in the form:

$$f(y_i|\boldsymbol{\theta}, \phi) = exp\Big\{ \frac{\boldsymbol{b(\theta)} \cdot \boldsymbol{T}(y_i) - B(\boldsymbol{\theta})}{a(\phi)} \Big\} \cdot c(y_i, \phi)$$

  for known real functions $a(\cdot), b(\cdot), B(\cdot), T(\cdot)$ and $c(\cdot)$. If $b(\cdot)$ is the identity function, the density function is in *canonical* form. We can always convert the distribution to canonical form, by defining

a transformed parameter $\boldsymbol{b} = b(\theta)$ and considering the following transformation of (2.3):

$$f(y_i|\boldsymbol{b},\phi) = exp\Big\{\frac{\boldsymbol{b}\cdot\boldsymbol{T}(y_i) - B(\boldsymbol{b})}{a(\phi)}\Big\}\cdot c(y_i,\phi) \tag{3.4}$$

An important remark about ϕ is that, depending on the nature of the problem, it can be either a vector, like in overdispersed Poisson models, or a scalar, which is common for every y_i, like in the case of a normal linear model with homoscedastic error terms, where $\phi = \sigma^2$. The GLMs we are going to focus on in the following chapters are the Logit, Probit models and they constitute cases of models, where $\phi = 1$ for every $y_i \quad i = 1, 2, ..., n$. We consider that in these models there is no dispersion parameter and so, our inference is restricted to $\boldsymbol{\theta}$.

Important properties of the exponential family

- In (3.4) $b(\boldsymbol{\theta})$, is referred to as the ***canonical link***.

- From (3.4) we can deduce the mean and the variance of y_i through the following equations:

$$E(T_j(y_i)|\boldsymbol{b}) = \frac{dB(\boldsymbol{b})}{db_j},$$
$$Var(T_j(y_i)|\boldsymbol{b}) = \frac{d^2B(\boldsymbol{b})}{d^2b_j}\cdot a(\phi), \qquad i = 1, 2, ..., n, \quad j = 1, ..., k.$$

The exponential family is a wide class of functions including the most important and frequently encountered distributions: the Normal, the Binomial, the Poisson and the Multinomial distribution.

Basic Specifications of a GLM

A *generalized linear model* is specified by three functions:

- The *linear predictor* denoted as $\eta = X\beta$.

- A *link function*, usually denoted as g, by which the mean $\mu = E\Big[\boldsymbol{y}|X,\beta\Big]$ of y is related to the *systematic component* X through the equation:

$$g(\mu) = \eta \tag{3.5}$$

  For identifiability reasons, g is a one-to-one function and therefore, it is also invertible. Usually, the canonical link is chosen to be the link function, but in fact there is no "optimal" choice, which means that it is up to the statistician to decide its form. Thus,(2.5) can be written in the form:

$$\mu = g^{-1}(\eta) = g^{-1}\big(X\beta\big) \tag{3.6}$$

- The density function f of the *random component* $\boldsymbol{y}$, which as mentioned before has to belong to the exponential family of distributions and it can also depend on the dispersion parameter ϕ.

So, supposing that $y_1, y_2, ..., y_n$ represent a random sample from the distribution f with mean $\tilde{\mu} = (\mu_1, .., \mu_n)$ and a dispersion parameter ϕ, then the joint distribution would be:

$$f\Big(\boldsymbol{y}|\tilde{\mu},\phi\Big) = \prod_{i=1}^{n} f\Big(y_i|\mu_i,\phi\Big). \tag{3.7}$$

Since the response variable is related to the explanatory variables through its mean according to (2.6), it is preferred to reparameterize the conditional density function via (2.6), so that it is directly a function of $\mu_i, i = 1, 2, .., n$. Thus, the dependence on $\boldsymbol{\beta}$ and X through the linear predictor is clarified and also, it is a useful form to apply Bayes' Theorem and determine the posterior distributions. Consequently, the likelihood function will be given by:

$$f(\boldsymbol{y}|X,\beta) = \prod_{i=1}^{n} f\Big(y_i|X,\beta,\phi\Big). \tag{3.8}$$

We should note at this point that, based on all the above, normal linear models also belong to the GLM family. Indeed, the distribution of $\boldsymbol{y}$ is Normal, $\boldsymbol{\theta} \equiv \boldsymbol{\mu} = (\mu_1, ..., \mu_n)$ and the dispersion parameter is $\phi = \sigma^2$, whereas the identity function is used as the link function.

In the chapters that follow, we will study Bayesian inference on two of the most common cases of GLMs; these would be the logit and probit models.

Chapter 4

Logistic Regression

4.1 First-step analysis of the logistic regression model

4.1.1 The likelihood of the model

Logistic regression is performed when the response variable $\boldsymbol{y}$ is binary or binomial. In the binary case, y_is refer to the occurrence or non-occurrence of an incident with different success probabilities p_i and therefore $y_i \sim Bern(p_i)$, for each $i \in \{1, 2, ..., n\}$. Our intention is to conduct inference on the vector $\boldsymbol{p} = (p_1, p_2, ..., p_n)$. First, however, we need to associate the data with the vector $\boldsymbol{p}$ via a formal mathematical expression.
The Bernoulli distribution has the form: $p(y_i|p_i) = p_i^{y_i} \cdot (1-p_i)^{1-y_i}$, with $E\Big[y_i|p_i\Big] = p_i$. As a first step, we express the probability function in the exponential family canonical form:

$$\begin{aligned} f(y_i|p_i) &= p_i^{y_i} \cdot (1-p_i)^{1-y_i} \\ &= exp\{y_i \log(p_i) + (1-y_i)\log(1-p_i)\} \\ &= exp\Bigg\{y_i \log\bigg(\frac{p_i}{1-p_i}\bigg) - \log\bigg(\frac{1}{1-p_i}\bigg)\Bigg\}. \end{aligned}$$

We can easily get that $b_i \equiv b(p_i) = \log\Big(\frac{p_i}{1-p_i}\Big)$ and $B(p_i) = \log\Big(\frac{1}{1-p_i}\Big)$, which yields that $B(b_i) = \log(1+e^b)$, whereas ϕ is considered a constant equal to 1.
Also, we can obtain the following equations:

$$E\big[y_i|b_i\big] = \frac{dB(b)}{db_i} = \frac{e^{b_i}}{1+e^{b_i}}, \tag{4.1}$$

$$Var\big[y_i|b_i\big] = \frac{d^2B(b)}{db_i^2} = \frac{e^{b_i}}{\big(1+e^{b_i}\big)^2} = \Big(1 - \frac{e^{b_i}}{1+e^{b_i}}\Big)\frac{e^{b_i}}{1+e^{b_i}} = p_i(1-p_i). \tag{4.2}$$

The canonical link function is $g(\mu_i) = g(p_i) = \log\Big(\frac{p_i}{1-p_i}\Big)$, known as the ***logit function***. Consequently, we will set:

$$\begin{aligned} \log\left(\frac{p_i}{1-p_i}\right) = X_i^T\boldsymbol{\beta} \Leftrightarrow \frac{p_i}{1-p_i} &= \mathrm{e}^{X_i^T\boldsymbol{\beta}} \\ \Leftrightarrow p_i = \frac{\mathrm{e}^{X_i^T\boldsymbol{\beta}}}{1+\mathrm{e}^{X_i^T\boldsymbol{\beta}}}, &\qquad i = 1,2,..,n. \end{aligned} \tag{4.3}$$

The likelihood function, denoted as $L(\boldsymbol{y}|\boldsymbol{p})$ would be:

$$\begin{aligned} L(\boldsymbol{y}|\boldsymbol{p}) &= \prod_{i=1}^{n} f(y_i|p_i) \\ &= \prod_{i=1}^{n} p_i^{y_i} \cdot (1-p_i)^{1-y_i}. \end{aligned}$$

Using (3.1), we can express the likelihood involving directly the linear predictor:

$$L(\boldsymbol{y}|X,\boldsymbol{\beta}) = \prod_{i=1}^{n}\left(\frac{\mathrm{e}^{X_i^T\boldsymbol{\beta}}}{1+\mathrm{e}^{X_i^T\boldsymbol{\beta}}}\right)^{y_i}\left(\frac{1}{1+\mathrm{e}^{X_i^T\boldsymbol{\beta}}}\right)^{1-y_i}. \tag{4.4}$$

Finally, equations (4.1) and (4.2) will be transformed likewise

4.1.2 Setting a prior distribution and obtaining the posterior distribution

We will focus on the most frequently used prior for the vector of parameters $\boldsymbol{\beta}$, which is the multivariate Normal $N_{k+1}(\boldsymbol{\mu}, \boldsymbol{C})$. As we have already discussed before, the mean and the covariance matrix are adjusted properly depending on the quality of the prior information that is available to us. Specifically, if the mean and the covariance matrix are set in such a way that the Normal distribution best fits the prior information. For instance, if we are confident about the source of this information, then we can set small variances as the diagonal values of $\boldsymbol{C}$. Furthermore, if there are indications of correlations between the explanatory variables,we can add non-diagonal values to $\boldsymbol{C}$. In the absence of prior information, $\boldsymbol{\mu}$ is set as a $k+1$-dimensional zero vector, whereas the

corresponding $\boldsymbol{C}$ can have large diagonal values. So, if $\boldsymbol{\beta} \sim N_{k+1}(0, \boldsymbol{C})$, the prior distribution will have the form:

$$\pi(\boldsymbol{\beta}) = (2\pi|C|)^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2}\boldsymbol{\beta}^T C^{-1}\boldsymbol{\beta}}. \tag{4.5}$$

According to Bayes' Theorem, the full form of the posterior distribution of $\boldsymbol{\beta}$ will be:

$$\begin{aligned} f(\boldsymbol{\beta}|\boldsymbol{y}, X) &= \frac{L(\boldsymbol{y}|X, \boldsymbol{\beta}) \cdot \pi(\boldsymbol{\beta})}{\int \cdots \int_B L(\boldsymbol{y}|X, \boldsymbol{\beta}) \cdot \pi(\boldsymbol{\beta}) d\boldsymbol{\beta}} \\ &= \frac{\prod_{i=1}^n \left(\frac{\mathrm{e}^{X_i^T\boldsymbol{\beta}}}{1+\mathrm{e}^{X_i^T\boldsymbol{\beta}}} \right)^{y_i} \left(\frac{1}{1+\mathrm{e}^{X_i^T\boldsymbol{\beta}}}^{1-y_i} \right) \cdot (2\pi|C|)^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2}\boldsymbol{\beta}^T C^{-1}\boldsymbol{\beta}}}{\int \cdots \int_B \prod_{i=1}^n \left(\frac{\mathrm{e}^{X_i^T\boldsymbol{\beta}}}{1+\mathrm{e}^{X_i^T\boldsymbol{\beta}}} \right)^{y_i} \left(\frac{1}{1+\mathrm{e}^{X_i^T\boldsymbol{\beta}}}^{1-y_i} \right) \cdot (2\pi|C|)^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2}\boldsymbol{\beta}^T C^{-1}\boldsymbol{\beta}} d\boldsymbol{\beta}}, \end{aligned} \tag{4.6}$$

where $B = \mathbb{R}^{k+1}$ is the parameter space of $\boldsymbol{\beta}$. The integral of the denominator is most of the times very troublesome to compute and in the particular case cannot be computed analytically. For that reason, instead of equalities and since the integral is a normalising constant, we proceed with proportionalities and so we will be using that

$$f(\boldsymbol{\beta}|\boldsymbol{y}, X) \propto \prod_{i=1}^n \left(\frac{\mathrm{e}^{X_i^T\boldsymbol{\beta}}}{1+\mathrm{e}^{X_i^T\boldsymbol{\beta}}} \right)^{y_i} \cdot \mathrm{e}^{-\frac{1}{2}\boldsymbol{\beta}^T C^{-1}\boldsymbol{\beta}}. \tag{4.7}$$

Our major obstacle for the inference is that, contrary to the normal linear model case, we cannot identify the joint posterior distribution of the parameters, neither compute it analytically. That fact also excludes Gibbs sampler, as it requires the full, analytic form of the distributions used.

Approximation methods have been developed as a solution to this problem, such as the ***Laplace approximation method***, which is the oldest and probably most known method. Another, equally popular method is the ***Normal approximation method***. The popularity of these two approaches lies mainly on the theoretical robustness and the accuracy of the results they produce. However, both of these qualities come at the cost of complex and rather sophisticated mathematical computations, which tend to intensify as the dimension of the problem increases.

In the present thesis, we choose to focus on the MCMC algorithms and specifically Metropolis-Hastings algorithms that come as a rather simpler and very effective alternative approach to the same problem, with an equally robust, theoretical background. MCMC processes are under constant study and development and they have proven to be extremely beneficiary in terms of Bayesian inference, especially in high-dimensional cases, in which the already cumbersome, computational part of the approximating methods becomes extremely difficult and time consuming. As we have mentioned before, the most important part of Bayesian inference lies on studying the form of the posterior density function of the parameter of interest and this is what we expect to achieve by simulating it via the Metropolis-Hastings algorithms, that have already been presented in the introduction of chapter 2.

Simulating the posterior distribution via Gamerman's Metropolis-Hastings IWLS algorithm

Instead of using the standard form of the Metropolis-Hastings algorithm in order to approximate $f(\boldsymbol{\beta}|\boldsymbol{y}, X)$, we will employ a slightly changed version, originally introduced by David Gamerman in the paper *Sampling from the posterior distribution in generalized mixed linear models*(1997). The modification suggested by Gamerman, considers the form of the proposal distribution and can be summarized as follows:
Suppose we have performed $j-1$ iterations and we are at the step of the loop in which a new value for $\boldsymbol{\beta}$ is suggested. The *current value*, $\boldsymbol{\beta}^{(j-1)}$, is used as input for a single iteration of the Bayesian Iterative Least Squares Algorithm. Thus, a vector denoted by $\boldsymbol{b}^{(j)}$ is obtained, which is considered an approximation of the posterior mode and the proposal distribution will be a multivariate Normal of the form $N(\boldsymbol{b}^{(j)}, C_j)$.

The great advantage of this procedure is that there is no need of tuning the proposal distribution, the convergence of the algorithm is quicker and after it is reached, we have additionally obtained the posterior mode, after the last iteration of the IWLS algorithm.

An explicit scheme of Gamerman's algorithm, as well as a review of the Bayesian IWLS algorithm are provided in the following pages.

Bayesian Computation of the Posterior Mode Using an IWLS scheme

We will present a Bayesian, computational way to obtain the posterior mode, which incorporates the Iterative Weighted Least Squares (IWLS) algorithm. Prior to that, we shall briefly review the main concept and algorithmic form of the IWLS algorithm from the view of Classic Statistics.

The IWLS algorithm is very often employed in Classical Statistics, in order to derive the Maximum Likelihood estimator, $\boldsymbol{\beta_{ML}}$, of the coefficients vector $\boldsymbol{\beta}$ in a GLM. Bearing in mind the notation of paragraph 2.1.2, we will hereby outline the generic steps of the algorithm.

1. Set an initial value $\boldsymbol{\beta^{(0)}}$.

2. Consider the random variables

$$\begin{aligned} Z_i^{(0)} &:= \eta_i^{(0)} + (y_i - \mu_i) \cdot g'(\mu_i) \\ &= X_i^T \boldsymbol{\beta^{(0)}} + (y_i - \mu_i) \cdot g'(\mu_i), \quad i = 1, 2, ..., n, \end{aligned}$$

   and the $n \times n$ diagonal matrix $\boldsymbol{W}\big(\boldsymbol{\beta^0}\big)$, with elements the *weights* $w_{ii}\big(\boldsymbol{\beta^0}\big) := 1/g'(\mu_i), \quad i = 1, 2, ..., n.$

3. Consider that $\boldsymbol{Z^{(0)}} \sim N_n\big(X\boldsymbol{\beta^{(0)}}, \boldsymbol{W^{-1}}\big(\boldsymbol{\beta^0}\big)\big)$ and perform linear regression on the corresponding model. Thus, we obtain $\boldsymbol{\beta^{(1)}}$.

4. Monitor the quantity $||\boldsymbol{\beta^{(1)}} - \boldsymbol{\beta^{(0)}}||$.
5. Iterate the process of steps 2.-4. using the current value of $\boldsymbol{\beta}$.

The loop described in step 5. stops at the j-th iteration, if $||\boldsymbol{\beta^{(j)}} - \boldsymbol{\beta^{(j-1)}}||$ is a sufficiently small value according to the researcher and $\boldsymbol{\beta_{ML}} = \boldsymbol{\beta^{(j)}}$, $j \in \mathbb{N}$.

According to the Bayesian paradigm, we place a multivariate Normal prior on $\boldsymbol{\beta}$ and so we consider that $\boldsymbol{\beta} \sim \boldsymbol{N_{k+1}(b_0, C_0)}$. At the j-th step of the algorithm, we perform Bayesian linear regression on the model $\boldsymbol{Z}\big(\boldsymbol{\beta^{(j-1)}}\big) \sim \boldsymbol{N_n}\big(\boldsymbol{X\beta, W}\big(\boldsymbol{\beta^{(j-1)}}\big)\big)$, where we consider that a priori $\boldsymbol{\beta} \sim N_{k+1}(b_0, C_0)$. Also, $\boldsymbol{\beta^{(j-1)}}$ is a simulated value already available to us, by which the quantities $Z_i\big(\beta^{(j-1)}\big)$ and $w_{ii}\big(\beta^{(j-1)}\big), \quad i = 1, ..., n$ are derived from the same equations as in the classic approach, where of

course we use $\boldsymbol{\beta^{(j-1)}}$ as the current value of $\boldsymbol{\beta}$. Thus, we obtain the n-dimensional vector $\boldsymbol{Z}\big(\boldsymbol{\beta^{(j-1)}}\big)$ and the $n \times n$ diagonal matrix $W\big(\beta^{(j-1)}\big)$. Up to a normalizing constant, the above distributions can be written as:

$$f(\boldsymbol{\beta}) \propto exp\Big\{ \big[\boldsymbol{\beta} - \boldsymbol{b_0}\big]^{\boldsymbol{T}} \boldsymbol{C_0^{-1}} \big[\boldsymbol{\beta} - \boldsymbol{b_0}\big] \Big\},$$

$$L\Big(\boldsymbol{Z}(\boldsymbol{\beta^{(j-1)}})|\boldsymbol{X}\Big) \propto exp\Big\{ \big[\boldsymbol{Z}(\boldsymbol{\beta^{(j-1)}}) - \boldsymbol{X\beta}\big]^{\boldsymbol{T}} \boldsymbol{W}(\boldsymbol{\beta^{(j-1)}}) \big[\boldsymbol{Z}(\boldsymbol{\beta^{(j-1)}}) - \boldsymbol{X\beta}\big] \Big\}.$$

Bayes' Theorem yields that:

$$\begin{aligned}
&f\Big(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{Z}(\boldsymbol{\beta^{(j-1)}})\Big) \propto f(\boldsymbol{\beta}) \cdot L\Big(\boldsymbol{Z}(\boldsymbol{\beta^{(j-1)}})|\boldsymbol{X}\Big) \\
&\propto exp\Big\{ \boldsymbol{\beta^T} \Big[\boldsymbol{X^T W}(\boldsymbol{\beta^{(j-1)}})\boldsymbol{X} + \boldsymbol{C_0^{-1}}\Big]\boldsymbol{\beta} - \boldsymbol{\beta^T}\Big[\boldsymbol{X^T W}(\boldsymbol{\beta^{(j-1)}})\boldsymbol{Z}(\boldsymbol{\beta^{(j-1)}}) + \boldsymbol{C_0^{-1}}\boldsymbol{b_0}\Big] \Big\} \\
&\cdot exp\Big\{ -\Big[\boldsymbol{Z}(\boldsymbol{\beta^{(j-1)}})^{\boldsymbol{T}} \boldsymbol{W}(\boldsymbol{\beta^{(j-1)}})\boldsymbol{X} + \boldsymbol{b_0^T C_0^{-1}}\Big]\boldsymbol{\beta} \Big\} \\
&\propto exp\Big\{ (\boldsymbol{\beta} - \boldsymbol{b_1})^{\boldsymbol{T}} \boldsymbol{C_1^{-1}} (\boldsymbol{\beta} - \boldsymbol{b_1}) \Big\},
\end{aligned}$$

where

$$\begin{aligned}
C_j &= \Big[X^T W(\beta^{(j-1)})X + C_0^{-1}\Big]^{-1} \\
b_j &= \Big[X^T W(\beta^{(j-1)})X + C_0^{-1}\Big] \cdot \Big[Z(\beta^{(j-1)})^T W(\beta^{(j-1)})X + b_0 C_0^{-1}\Big].
\end{aligned}$$

We set $\boldsymbol{\beta^{(j)}} = \Big[\boldsymbol{X^T W}(\boldsymbol{\beta^{(j-1)}})\boldsymbol{X} + \boldsymbol{C_0^{-1}}\Big] \cdot \Big[\boldsymbol{Z}(\boldsymbol{\beta^{(j-1)}})^{\boldsymbol{T}} \boldsymbol{W}(\boldsymbol{\beta^{(j-1)}})\boldsymbol{X} + \boldsymbol{b_0 C_0^{-1}}\Big]$.

We repeat the steps described above until we diagnose the convergence of the algorithm.

We have shown that the posterior distribution of $\boldsymbol{\beta}|\boldsymbol{y}$ is a multivariate normal density and, consequently, due to the symmetry, the posterior mode, noted as $\hat{\boldsymbol{\beta}}$, equals to the mean of the sample we get after the burn-in period, which is trivial to calculate. Also, the precision matrix of the posterior distribution will be the inverse of the curvature at the posterior mode ; that is $\Big[\boldsymbol{X^T W}(\hat{\boldsymbol{\beta}})\boldsymbol{X} + \boldsymbol{C_0^{-1}}\Big]^{\boldsymbol{-1}}$. In other words, asymptotically $\boldsymbol{\beta}|\boldsymbol{y} \sim \boldsymbol{N_{k+1}}\Big(\hat{\boldsymbol{\beta}}, \Big[\boldsymbol{X^T W}(\hat{\boldsymbol{\beta}})\boldsymbol{X} + \boldsymbol{C_0^{-1}}\Big]^{\boldsymbol{-1}}\Big)$.

Since we are studying logistic regression models, we should clarify that in the above calculations :

$$\mu_i = p_i$$

and we bear in mind as well that

$$g(\mu_i) \equiv g(p_i) = \log\left(\frac{p_i}{1-p_i}\right), \quad i = 1, .., n.$$

Gamerman's Independent Sampler Algorithm

Dani Gamerman uses the Independent Sampler scheme to obtain the posterior distribution of $\boldsymbol{\beta}$, suggesting the following modification, which concerns solely the proposal distribution q:
At the point of the loop where the current value of the regression coefficients vector $\boldsymbol{\beta^{(j-1)}}$ has already been obtained and the next value $\boldsymbol{\beta^{can}}$ has to be proposed, a single iteration of the Bayesian IWLS is performed. Thus, we derive the $k+1$-dimensional vector
$\boldsymbol{b_j} = \left[\boldsymbol{X^T W(\beta^{(j-1)})X + C_0^{-1}}\right] \cdot \left[\boldsymbol{Z(\beta^{(j-1)})^T W(\beta^{(j-1)})X + b_0 C_0^{-1}}\right]$
and the $(k+1)\times(k+1)$ diagonal matrix $\boldsymbol{C_j} = \left[\boldsymbol{X^T W(\beta^{(j-1)})X + C_0^{-1}}\right]^{\boldsymbol{-1}}$, which we respectively use as the mean and the covariance matrix of a Gaussian distribution. This is the proposal distribution for the *current* step of the algorithm. More specifically, $\boldsymbol{q\big(\beta^{can}|\beta^{(j-1)}, y\big) \equiv N_{k+1}(b_j, C_j)}$.

The candidate value is accepted with probability

$$p = \min\left\{\frac{f(\beta^{can}|y)\cdot q\big(\beta^{(j-1)}|\beta^{can}, y\big)}{f\big(\beta^{(j-1)}|y\big)\cdot q\big(\beta^{can}|\beta^{(j-1)}, y\big)}, 1\right\},$$

where by $f(\cdot|\cdot)$, we refer to the posterior distribution. The term $q\big(\beta^{(j-1)}|\beta^{can}, y\big)$ in the numerator is the proposal of the reverse move and has to be handled with cautiousness, because we have to perform the calculations of a single iteration of the IWLS using β^{can} as an operating value and then obtain $b_{can} = \left[X^T W(\beta^{can})X + C_0^{-1}\right] \cdot \left[Z(\beta^{can})^T W(\beta^{can})X + b_0 C_0^{-1}\right]$ and $C_{can} = \left[X^T W(\beta^{can})X + C_0^{-1}\right]^{-1}$. Then, we use the formula of $N_{k+1}(b_j, C_j)$.

Gamerman's Independent Sampler has several appealing properties:

- There is no need of tuning the algorithm and as a result, there is no difficulty with the convergence and we are spared the cumbersome process of searching for the proper candidate generator;

- The proposal distribution q of each iteration of the algorithm is reasonably close to the target distribution, that is the posterior distribution of $\boldsymbol{\beta}$. Therefore, the acceptance probabilities are high, which indicates that the support of the posterior is more thoroughly explored;
- The proposed algorithm has a wide area of applications beyond the common GLMs dealing with logistic or Poisson regression. In the relative paper, it is used for inference for mixed effects and nested random effects models, whereas the issue of non-normal priors for $\boldsymbol{\beta}$ is also discussed.

4.1.3 Inference and Model Selection

In order to assess a logistic regression model or compare it with another, we will use the Deviance Information Criterion and the L Measure. The form of the posterior distribution of $\boldsymbol{\beta}$, $f(\boldsymbol{\beta}|\boldsymbol{y}, X)$ prevents us from using Bayes' factors or the posterior probability of the models under comparison.

Using the Deviance Information Criterion
We recall that the score of the Deviance Information Criterion is given as:

$$\begin{aligned} DIC &= \overline{D(\boldsymbol{\beta})} + p_D \\ &= 2\overline{D(\boldsymbol{\beta})} + 2lnL(\boldsymbol{y}|\bar{\boldsymbol{\beta}}(\boldsymbol{y})) \\ &= -4\int_{\mathcal{B}} lnL(\boldsymbol{y}|\boldsymbol{\beta}) \cdot f(\boldsymbol{\beta}|\boldsymbol{y}) d\boldsymbol{\beta} + 2lnL(\boldsymbol{y}|\bar{\boldsymbol{\beta}}(\boldsymbol{y})), \end{aligned} \tag{4.8}$$

where $\bar{\boldsymbol{\beta}}(\boldsymbol{y}) = \int_{\mathcal{B}} \boldsymbol{\beta} \cdot f(\boldsymbol{\beta}|\boldsymbol{y}) d\boldsymbol{\beta}$. Also, from (4.2), we can easily derive that

$$lnL(\boldsymbol{y}|\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \cdot X_i^T \boldsymbol{\beta} - ln\Big(1 + e^{X_i^T \boldsymbol{\beta}}\Big).$$

The two integrals involved in (4.6) can be estimated via Monte Carlo integration. More specifically, supposing that $S = \left\{\boldsymbol{\beta}^{(1)}, ..., \boldsymbol{\beta}^{(n)}\right\}$ is a sample of simulated values from the posterior distribution, after the initial values have been discarded, then the simple Monte Carlo estimator

for the posterior mean $\bar{\boldsymbol{\beta}}(\boldsymbol{y})$ would be:

$$\begin{aligned}\bar{\boldsymbol{\beta}}(\boldsymbol{y}) &= \int_{\mathcal{B}} \boldsymbol{\beta} f(\boldsymbol{\beta}|\boldsymbol{y}) d\boldsymbol{\beta} \\ &\approx \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\beta}^{(i)},\end{aligned} \tag{4.9}$$

whereas $\overline{D(\boldsymbol{\beta})}$ can be estimated based on the following approximation of the corresponding integral:

$$\begin{aligned}&\int_{\mathcal{B}} lnL(\boldsymbol{y}|\boldsymbol{\beta}) \cdot f(\boldsymbol{\beta}|\boldsymbol{y}) d\boldsymbol{\beta} \\ &\approx \frac{1}{n}\sum_{i=1}^{n} lnL(\boldsymbol{y}|\boldsymbol{\beta}^{(i)}).\end{aligned} \tag{4.10}$$

Using the L Measure method

Suppose we need to compare two models M_1 and M_2. In order to use the L Measure, an n-dimensional vector of predicted values is needed. Let this be denoted by $\boldsymbol{z} = \big(z_1, ..., z_n\big)$. The sampling distribution of $\boldsymbol{z}$, under each model, can be obtained via Monte Carlo integration.

Also, let $S_1 = \big(\boldsymbol{\beta^{(1)}}, ..., \boldsymbol{\beta^{(I)}}\big)$ be a sample from the posterior distribution of $\boldsymbol{\beta}$ regarding model M_1 and $S_2 = \big(\boldsymbol{\beta^{(1)}}, ..., \boldsymbol{\beta^{(J)}}\big)$ be a sample from the posterior distribution of $\boldsymbol{\beta}$ regarding model M_2. The sampling distribution of $\boldsymbol{z}$, that is the predictive distribution, under model M_1, can be approximated via Monte Carlo integration, based on the formula below:

$$f\Big(\boldsymbol{z}|\boldsymbol{y}, M_1\Big) \approx \frac{1}{I}\sum_{j=1}^{I} L(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\beta^{(j)}}), \quad \text{with} \quad \boldsymbol{\beta^{(1)}}, ..., \boldsymbol{\beta^{(I)}} \in S_1.$$

Likewise, we can obtain the distribution of $\boldsymbol{z}$, under model M_2:

$$f\Big(\boldsymbol{z}|\boldsymbol{y}, M_2\Big) \approx \frac{1}{J}\sum_{j=1}^{J} L(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\beta^{(j)}}), \quad \text{with} \quad \boldsymbol{\beta^{(1)}}, ..., \boldsymbol{\beta^{(J)}} \in S_2.$$

We will compute the L Measure, corresponding to each one of the candidate models by the equation (1.7) and then select the one, corresponding to the lowest value.

Since logistic regression models belong to the class of GLMs, our endeavour becomes much easier with the assistance of several statistical properties:

1. The full conditional posterior distribution of the prediction vector $\boldsymbol{z}$, under a specific model and given a vector of predictors $\boldsymbol{\beta^{(j)}}$, is the likelihood of that model, as given in (4.4), with the difference that a sample S_m from the posterior distribution $f(\boldsymbol{\beta}|\boldsymbol{y}, M)$, is obtained via Gamerman's Independent Sampler Algorithm.

   In mathematical notation, the property above can be written as:

$$L(\boldsymbol{z}|\boldsymbol{\beta^{(j)}}, \boldsymbol{y}, M) = \prod_{i=1}^{n} \left(\frac{\mathrm{e}^{X_i^T \boldsymbol{\beta}^{(j)}}}{1 + \mathrm{e}^{X_i^T \boldsymbol{\beta}^{(j)}}} \right)^{z_i} \left(\frac{1}{1 + \mathrm{e}^{X_i^T \boldsymbol{\beta}^{(j)}}} \right)^{1 - z_i}, \quad (4.11)$$

   with $\boldsymbol{\beta^{(j)}} \in S_m$ and $\boldsymbol{z} \sim f(\boldsymbol{z}|\boldsymbol{y}, M)$.

2. By the ***Law of Double Expectation***, we can get an equivalent expression for the expectation $E[z_i|\boldsymbol{y}]$ in (1.6):

$$E[z_i|\boldsymbol{y}] = E_{\boldsymbol{\beta}|\boldsymbol{y}}\Big[E\big[z_i|\boldsymbol{\beta}, \boldsymbol{y}\big]\Big]. \quad (4.12)$$

3. Recalling equations (4.1) and (4.2) and bearing in mind that in a logistic regression model

$$\theta_i = log\big(\frac{p_i}{1 - p_i}\big) = X_i^T \beta$$

   and also the distribution of each $z_i|\boldsymbol{\beta}, \boldsymbol{y}, \quad i = 1, 2, ..., n$, we can conclude that:

$$E\big[z_i|\boldsymbol{\beta}, \boldsymbol{y}\big] = b'(\theta_i) = \frac{e^{\theta_i}}{1 + e^{\theta_i}} = \frac{e^{X_i^T \beta}}{1 + e^{X_i^T \beta}}, \quad (4.13)$$

$$Var\big[z_i|\boldsymbol{\beta}, \boldsymbol{y}\big] = b''(\theta_i) = \frac{e^{\theta_i}}{(1 + e^{\theta_i})^2} = \frac{e^{X_i^T \boldsymbol{\beta}}}{(1 + e^{X_i^T \boldsymbol{\beta}})^2}. \quad (4.14)$$

   Consequently, (4.12) can be rewritten as follows:

$$E[z_i|\boldsymbol{y}] = E_{\boldsymbol{\beta}|\boldsymbol{y}}\Big[b'(\theta_i)\Big] = E_{\boldsymbol{\beta}|\boldsymbol{y}}\Big[\frac{e^{X_i^T \beta^{(i)}}}{1 + e^{X_i^T \beta^{(i)}}}\Big] \approx \sum_{k=1}^{s} \frac{1}{s} \left[\frac{e^{X_i^T \beta^{(k)}}}{1 + e^{X_i^T \beta^{(k)}}} \right]. \quad (4.15)$$

   The sum in the last equation is the simple Monte Carlo estimator for the expectation, with respect to the posterior distribution of

$\boldsymbol{\beta}$, whereas the set $\left(\beta^{(1)}, ..., \beta^{(s)}\right)$ is a sample from the distribution $f(\boldsymbol{\beta}|\boldsymbol{y})$. In addition to that, there is also an alternative expression for the variance $Var(z_i|\boldsymbol{y})$. Since we know that

$$Var(z_i|\boldsymbol{y}) = E[z_i^2|\boldsymbol{y}] - E^2[z_i|\boldsymbol{y}], \tag{4.16}$$

based on (4.15), we can rewrite the squared expectation as $E^2[z_i|\boldsymbol{y}] \approx \left(\sum_{k=1}^{s} \frac{1}{s}\left[\frac{e^{X_i^T\beta^{(k)}}}{1+e^{X_i^T\beta^{(k)}}}\right]\right)^2$. By making the observation below, based on the Law of Double Expectation and the general definition of the variance:

$$\begin{aligned}
E[z_i^2|\boldsymbol{y}] &= E_{\boldsymbol{\beta}|\boldsymbol{y}}\Big[E\big[z_i^2|\boldsymbol{\beta},\boldsymbol{y}\big]\Big] \\
&= E_{\boldsymbol{\beta}|\boldsymbol{y}}\Big[Var\big[z_i|\boldsymbol{\beta},\boldsymbol{y}\big] + \Big(E\big[z_i|\boldsymbol{\beta},\boldsymbol{y}\big]\Big)^2\Big] \\
&= E_{\boldsymbol{\beta}|\boldsymbol{y}}\Big[b''(\theta_i) + \big(b'(\theta_i)\big)^2\Big] \\
&= E_{\boldsymbol{\beta}|\boldsymbol{y}}\left[\frac{e^{X_i^T\boldsymbol{\beta}}}{(1+e^{X_i^T\boldsymbol{\beta}})^2} + \left(\frac{e^{X_i^T\boldsymbol{\beta}}}{1+e^{X_i^T\boldsymbol{\beta}}}\right)^2\right] \\
&= E_{\boldsymbol{\beta}|\boldsymbol{y}}\left[\frac{e^{X_i^T\beta}}{1+e^{X_i^T\beta}}\right] \\
&\approx \frac{1}{s}\sum_{k=1}^{s}\left[\frac{e^{X_i^T\beta^{(k)}}}{1+e^{X_i^T\beta^{(k)}}}\right].
\end{aligned}$$

So, we conclude that (4.16) can be approximated by the form

$$Var(z_i|\boldsymbol{y}) \approx \frac{1}{s}\sum_{k=1}^{s}\left[\frac{e^{X_i^T\beta^{(k)}}}{1+e^{X_i^T\beta^{(k)}}}\right] - \left(\sum_{k=1}^{s}\frac{1}{s}\Big[\frac{e^{X_i^T\beta^{(k)}}}{1+e^{X_i^T\beta^{(k)}}}\Big]\right)^2. \tag{4.17}$$

Now, we can calculate (1.7) and obtain the L Measure criterion for $M1$:

$$\begin{aligned}
L_{IL}^{\mathbf{1}} \approx &\sum_{i=1}^{n}\left\{\frac{1}{I}\sum_{k=1}^{I}\left[\frac{e^{X_i^T\beta^{(k)}}}{1+e^{X_i^T\beta^{(k)}}}\right] - \left(\sum_{k=1}^{I}\frac{1}{I}\Big[\frac{e^{X_i^T\beta^{(k)}}}{1+e^{X_i^T\beta^{(k)}}}\Big]\right)^2\right\} \\
&+ \frac{1}{2}\sum_{i=1}^{n}\left\{\frac{1}{I}\sum_{k=1}^{I}\left(\frac{e^{X_i^T\beta^{(k)}}}{1+e^{X_i^T\beta^{(k)}}} - \boldsymbol{y}\right)^2\right\}, \quad \text{with} \quad \beta^{(1)}, ..., \beta^{(I)} \in S_1.
\end{aligned}$$

Likewise, the L Measure for $M2$ can be derived by:

$$L_{IL}^{\mathbf{2}} \approx \sum_{i=1}^{n} \left\{ \frac{1}{J} \sum_{k=1}^{J} \left[\frac{e^{X_i^T \beta^{(k)}}}{1 + e^{X_i^T \beta^{(k)}}} \right] - \left(\sum_{k=1}^{J} \frac{1}{J} \Big[\frac{e^{X_i^T \beta^{(k)}}}{1 + e^{X_i^T \beta^{(k)}}} \Big] \right)^2 \right\}$$
$$+ \frac{1}{2} \sum_{i=1}^{n} \left\{ \frac{1}{J} \sum_{k=1}^{J} \left(\frac{e^{X_i^T \beta^{(k)}}}{1 + e^{X_i^T \beta^{(k)}}} - \boldsymbol{y} \right)^2 \right\}, \quad \text{with} \quad \beta^{(1)}, ..., \beta^{(J)} \in S_2.$$

Chapter 5

The Probit Model

5.1 Introduction

The treatment of binary, dichotomous data, that need to somehow be linked to a set of explanatory parameters through their mean, is not restricted solely to logistic regression models. Suppose we are dealing with a series of outcomes describing the occurrence or not occurrence of an incident; in mathematics notation, we have n random binary variables, $y_i \in \{0,1\}$, such that $y_i = \begin{cases} 1 & \text{with probability} \quad p_i \\ 0 & \text{with probability} \quad 1-p_i \end{cases}$. Once again, it is obvious that $y_i \sim Bern(p_i)$, consequently the probability mass function will be:

$$\begin{aligned} f(y_i|p_i) &= p_i^{y_i}(1-p_i)^{1-y_i} \\ \mu_i = E(y_i) &= p_i, \qquad i = 1, ..., n. \end{aligned}$$

Our priority is to associate each mean $\mu_i = p_i$ with the corresponding linear predictor $\eta_i = \boldsymbol{X_i^T\beta}$, where $\boldsymbol{X = (X_1, X_2, ..., X_n)}$ is the already known design matrix and $\boldsymbol{\beta}$ is the vector of the coefficients. In the previous section, the canonical link was defined as the link function. In the Probit model case we use as link function (usually referred to as the probit link) the ***cumulative distribution of the Standard Normal distribution***, noted as $\Phi(\cdot)$ and so we set:

$$\begin{aligned} &\Phi(p_i) = \boldsymbol{X_i^T\beta} \\ &\Leftrightarrow p_i = \Phi^{-1}\big(\boldsymbol{X_i^T\beta}\big). \end{aligned} \tag{5.1}$$

We note that our link function is well defined, since, as a cumulative distribution, $\Phi(\cdot) : \mathbb{R} \to [0,1]$ and therefore the quantities $\Phi^{-1}\big(\boldsymbol{X_i^T\beta}\big)$

can be regarded as possibilities for every $\boldsymbol{X_i}$, $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$.
So now, each coordinate y_i of the random component $\boldsymbol{y} = (y_1, ..., y_n)$ is described by $y_i \sim Bern\big(\Phi^{-1}\big(\boldsymbol{X_i^T\beta}\big)\big)$ and by substituting p_i in the corresponding probability mass function we have that:

$$f(y_i|\boldsymbol{X}, \boldsymbol{\beta}) = \big[\Phi^{-1}\big(X_i^T\beta\big)\big]^{y_i}\big[1 - \Phi^{-1}\big(X_i^T\beta\big)\big]^{1-y_i}, \quad i = 1, ..., n.$$

Finally, the probability mass function of the probit model, since we consider the observations to be independent, is:

$$\begin{aligned} L(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{X}) &= \prod_{i=1}^{n} f(y_i|\boldsymbol{X}, \boldsymbol{\beta}) \\ &= \prod_{i=1}^{n}\big[\Phi^{-1}\big(X_i^T\boldsymbol{\beta}\big)\big]^{y_i}\big[1 - \Phi^{-1}\big(X_i^T\boldsymbol{\beta}\big)\big]^{1-y_i}. \end{aligned} \tag{5.2}$$

5.2 Inference for the Probit Model

5.2.1 Prior Specification and First-step Analysis

Our interest obviously lies on deriving the posterior distribution of the coefficient parameter vector $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)$ and thus, link the data of the design matrix $\boldsymbol{X}$ to the mean of the response variable $\boldsymbol{y}$. We place a proper prior on $\boldsymbol{\beta}$ and so consider that $\boldsymbol{\beta} \sim N_{k+1}(b_0, C_0)$, where b_0 is a $k \times 1$-dimensional column vector and C_0 a $(k+1) \times (k+1)$ diagonal matrix. The values we set upon the mean vector and the covariance matrix depend on the quality of the prior information. If we decide that the provided piece of information is unreliable or with little importance and consequently, we are in a state of ignorance, we can set $b_0 = 0$ and place large values (i.e. $C_{0ii} = 10000$)on the non-zero elements of C_0.
By applying Bayes' Theorem, we get that:

$$\begin{aligned} &f(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}) \propto f(\boldsymbol{\beta}) \cdot L(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{X}) \\ &\propto exp\Bigg\{-\frac{1}{2}(\boldsymbol{\beta} - b_0)^T C_0^{-1}(\boldsymbol{\beta} - b_0)\Bigg\} \cdot \prod_{i=1}^{n}\left[\Phi^{-1}\big(X_i^T\boldsymbol{\beta}\big)\right]^{y_i}\left[1 - \Phi^{-1}\big(X_i^T\boldsymbol{\beta}\big)\right]^{1-y_i}. \end{aligned}$$

Once again, the posterior distribution is known up to a normalising constant and it is intractable, since it does not correspond to any known distribution. Therefore, we have to turn to MCMC methods in order to obtain it.

We could naively implement the algorithms mentioned above, but we are going to take advantage of the form of the link function and introduce an innovative computational trick, which,in the end, allows us to use a Gibbs sampler in order to extract the distribution of $\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{y}$.

5.3 Data Augmentation in the Probit Model

Data augmentation (Tanner and Wong,1987) is a sophisticated, stochastic technique, which is employed by statisticians as a way to overcome the problems of missing data or the intractability of likelihood distributions. In the Bayesian paradigm, both problems are frequently faced and particularly in the probit model, we are dealing with the latter at the course of our analysis.

The fundamental idea upon which the whole process is built, is to introduce an n-dimensional vector of ***latent variables or data*** denoted as $\boldsymbol{z} = (z_1, ..., z_n)$ and conduct inference on $(\boldsymbol{\beta}, \boldsymbol{z})$. It is noted, that there are no specific criteria regarding when $\boldsymbol{z}$ is considered to be a variable vector or a vector of data, since in the Bayesian context, both are handled as random variables. So, after adding $\boldsymbol{z}$ as a random variable vector, the task of extracting the posterior distribution of $(\boldsymbol{\beta}, \boldsymbol{z})$, although the dimension of the problem has increased, has become less complicated.

Contrary to $f(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X})$, the posterior distribution of the new parameter vector, that is $f(\boldsymbol{\beta}, \boldsymbol{z}|\boldsymbol{y}, \boldsymbol{X})$, demands less intensive computational methods. As a matter of fact, it will be shown that a sample from the joint posterior distribution can be obtained via the Gibbs Sampler.

5.3.1 A latent data-based expression of the Probit Model

The introduction of the auxiliary variable vector $\boldsymbol{z}$, enables us to express the probit regression model as a Normal linear model, with $\boldsymbol{z}$ being the "unobserved" vector of responses:

$$z_i = \boldsymbol{X_i^T}\boldsymbol{\beta} + e_i, \qquad where \quad e_i \sim N(0,1), \quad i = 1, ..., n. \tag{5.3}$$

The link to the observed data, $\boldsymbol{y}$ is described by the following argument:

$$y_i = \begin{cases} 1, & \text{if and only if} \quad z_i > 0, \\ 0, & \text{if and only if} \quad z_i \leq 0 \end{cases}.$$

Let us consider the above setup before proceeding any further. The latent data z_i, as they have been defined according to the concept of a Normal linear model, are continuous random variables and $z_i \sim N\big(\boldsymbol{X_i^T\beta}, 1\big)$ for every $i \in \{1, ..., n\}$. However, the dichotomous responses y_i are still associated with the linear predictor $\boldsymbol{\eta = X\beta}$, because the values 0 or 1 are assigned, depending on whether the corresponding z_i is positive or not. The probability of each z_i falling above the threshold of 0 is expressed thusly:

$$\begin{aligned} P(y_i = 1) = P(z_i > 0) &= 1 - P(z_i \leq 0) \\ &= 1 - P\big(z_i - \boldsymbol{X_i^T\beta} \leq -\boldsymbol{X_i^T\beta}\big) \\ &= 1 - \Phi\big(-\boldsymbol{X_i^T\beta}\big) = \Phi\big(\boldsymbol{X_i^T\beta}\big), \quad \text{because} \quad z_i - \boldsymbol{X_i^T\beta} \sim N(0,1). \end{aligned}$$

Also, $P(y_i = 0) = 1 - P(y_i = 1) = 1 - \Phi\big(\boldsymbol{X_i^T\beta}\big) = \Phi\big(-\boldsymbol{X_i^T\beta}\big)$.

Our inference will be conducted on the model:

$$\boldsymbol{z \sim N_n(X\beta, 1)}$$

with the restriction that

$$\begin{cases} z_i > 0, & \text{if} \quad y_i = 1, \\ z_i \leq 0, & \text{if} \quad y_i = 0, \qquad i = 1, ...n. \end{cases}$$

and we place again a proper prior on the coefficients vector

$$\boldsymbol{\beta \sim N_{k+1}(b_0, C_0)}. \tag{5.4}$$

What is special about the above model is that the values z_i follow a **truncated Normal distribution**, $z_i \sim TN\big(X_i^T, 1\big)$, with 0 as the point of truncation, whereas the truncated area of the distribution is determined by the corresponding values of y_i thusly:

- if $y_i = 0$, the distribution is truncated above 0,
- if $y_i = 1$, the distribution is truncated below 0.

By making the simple observation that $\{z_i, i = 1, ..., n\} = \{i : y_i = 1\} \cup \{i : y_i = 0\}$, the likelihood of the model in (4.4), conditional on $\boldsymbol{y}$,$\boldsymbol{\beta}$, can be expressed as

$$L(\boldsymbol{z}|\boldsymbol{\beta},\boldsymbol{y}) =$$
$$= \prod_{i:y_i=1} exp\Big\{-\frac{1}{2}\big(z_i - \boldsymbol{X_i^T\beta}\big)^2\Big\} I[z_i > 0] \cdot \prod_{i:y_i=0} exp\Big\{-\frac{1}{2}\big(z_i - \boldsymbol{X_i^T\beta}\big)^2\Big\} I[z_i < 0]. \tag{5.5}$$

We will attempt to obtain the posterior distribution of $\boldsymbol{\beta}$ computationally, via a Gibbs sampler. In order to do that, we need to make the remark that the distribution of the auxiliary variables vector, $\boldsymbol{z}$, conditional only on $\boldsymbol{\beta}$ is a usual multivariate Normal, $N_n(\boldsymbol{X\beta}, I_n)$.
We can easily apply Bayes' Rule and thus, derive that:

$$\begin{aligned}
&f(\boldsymbol{\beta}|\boldsymbol{z}) \propto L(\boldsymbol{z}|\boldsymbol{\beta}) \cdot f(\boldsymbol{\beta}) \\
&\propto exp\Big\{-\frac{1}{2}(\boldsymbol{z - X\beta})^T I_n(\boldsymbol{z - X\beta})\Big\} \cdot exp\Big\{-\frac{1}{2}(\boldsymbol{\beta} - b_0)^T C_0^{-1}(\boldsymbol{\beta} - b_0)\Big\} \\
&\propto exp\Big\{-\frac{1}{2}\Big[\beta^T\big(X^TX + C_0^{-1}\big)\beta - \beta^T\big(X^Tz + C_0^{-1}b_0\big) - \big(z^TX + b_0^TC_0^{-1}\big)\beta\Big]\Big\} \\
&\cdot exp\Big\{-\frac{1}{2}\big(z^Tz + b_0^TC_0^{-1}b_0\big)\Big\} \\
&\propto exp\Big\{-\frac{1}{2}(\beta - b_1)^T\Sigma^{-1}(\beta - b_1)\Big\}.
\end{aligned} \tag{5.6}$$

Consequently, $\boldsymbol{\beta}|\boldsymbol{z} \sim \boldsymbol{N_{k+1}(b_1, \Sigma)}$, where $b_1 = \big(X^TX + C_0^{-1}\big)^{-1}\big(X^Tz + C_0^{-1}b_0\big)$ and $\Sigma = \big(X^Tz + C_0^{-1}b_0\big)^{-1}$.

Knowing and having fully identified the distributions of the variables $\boldsymbol{z}|\boldsymbol{\beta},\boldsymbol{y}$ and $\boldsymbol{\beta}|\boldsymbol{z}$, we can obtain the posterior distribution of $\boldsymbol{\beta}|\boldsymbol{z},\boldsymbol{y}$ via the following Gibbs sampler, according to Scott M.Lynch (2007):

1. Initialize the parameter vector $\boldsymbol{\beta}$ by setting as a starting value $\boldsymbol{\beta^{(0)} \sim N_{k+1}(b_0, C_0)}$.
2. Use the current value of $\boldsymbol{\beta}$ to simulate $z_i|\boldsymbol{\beta},\boldsymbol{y} \sim TN(X_i^T\beta, 1) \quad i \in 1, ..., n$ and consider the vector $\boldsymbol{z} = (z_1, ..., z_n)$.
3. Use $\boldsymbol{\beta}$ and $\boldsymbol{z}$ of Step.2 to calculate b_1 and C_1 and then, simulate $\boldsymbol{\beta}|\boldsymbol{z},\boldsymbol{y} \sim N_{k+1}(b_1, C_1)$.
4. Return to Step.2 .

As usual, we iterate the algorithm until convergence is achieved.

However, in the above procedure, we need to perform the somewhat challenging simulation of the truncated Normal distribution in Step.2.

We present two possible simulation methods in order to deal with the problem.

5.3.2 Simulating Truncated Gaussians

By truncating a Normal distribution, we bound the values that the normally distributed random variable can take within an interval of the form $[a, b] \subset (-\infty, +\infty)$, whereas the plot of the truncated distribution can be obtained by "cutting off" the sides below a and above b from the plot of the corresponding Normal. Although it is easy to comprehend the form of a truncated Gaussian, an effective simulation turns out to be as trivial as expected.

The "naive" approach

The "naive approach" as it was referred to by Robert (1995), and Lynch (2007) as well afterwards, is the idea we would instinctively follow in our attempt to simulate a truncated Normal:
Suppose we have already iterated the Gibbs sampler j times and we have obtained a parameter vector $\beta^{(j)}$. At this point, based on the fact that $z_i|\beta, y_i \sim TN(X_i^T\beta, 1)$ and depending on whether $y_i = 1$ or $y_i = 0$, we would keep simulating values from $N(X_i^T\beta, 1)$ until, we got $z_i > 0$ or $z_i < 0$ respectively. The flaw in that idea is that it can be proved to be extremely time consuming if the desired z_i is an outlier for $N(X_i^T\beta, 1)$. This would occur in two occasions:

- If $X_i^T\beta >> 0$ and $y_i = 0$, then we would accept a simulated value from $N(X_i^T\beta, 1)$ and set it as z_i only if it is negative. However, since the variance is 1, a large number of failed simulations will take place, before a negative value is obtained.
- In the exact opposite case, where $X_i^T\beta << 0$, but $y_i = 1$, then due to the larger probability mass that negative values would have, a positive outcome would be rear. Therefore, once again, numerous simulations would take place, until a positive value would appear.

Considering that a simulation from an appropriately truncated Gaussian has to be obtained in each step of the Gibbs sampler, until convergence is diagnosed, we realize that despite the simplicity of the algorithm, we run the risk of engaging in a cumbersome and very slow process. Therefore, this approach is not usually recommended.

The Inversion Sampling approach

Lynch (2007) recommends the Inversion Sampling approach as the most rapid and straightforward way, since there is no need of monitoring the drawn values in order to accept or reject them. In other words, the acceptance rate is 100%. The fundamental idea upon which this method is built is the following lemma:

Lemma 4. *Let u be a random variable such that $u \sim U(0,1)$ and $F(\cdot) : (-\infty, +\infty) \to [0,1]$ be a cumulative distribution, corresponding to a density function, denoted as f. Then, if we consider the random variable $F^{-1}(u)$, we get that $F^{-1}(u) \sim f$.*

Proof
We will show that the cumulative distribution of $F^{-1}(u)$ is F.
Indeed, we can easily observe that

$$P\big(F^{-1}(u) \leq x\big) = P(u \leq F(x)) = F(x).$$

As we have mentioned before, we need to simulate values from Gaussians of the form $N(X_i^T, 1)$, that are truncated below or above 0 if $y_i = 1$ or $y_i = 0$, respectively. We shall describe a method of simulating the side of a $N(X_i^T, 1), \quad i \in 1, ..., n$, below zero. That is the case when for that particular i, $y_i = 0$.

We will use Lynch's notation and denote the cumulative distribution of $N(X_i^T, 1)$ as $\Phi_{\mu_i,1}(\cdot)$, where $\mu_i = X_i^T$. First, we should consider that we cannot directly use the inverse of the cumulative distribution, $\Phi^{-1}_{\mu_i,1}$, because we will get values in $(-\infty, +\infty)$, whereas we are restricted in $(-\infty, 0)$. Mathematically, this restriction is expressed as $-\infty < z < 0$. Suppose that we do set $z_i = \Phi^{-1}_{\mu_i,1}(u)$, where $u_i \sim U(0.1)$. Then, our restriction takes the form: $-\infty < \Phi^{-1}_{\mu_i,1}(u_i) < 0$, which is equivalently expressed as $\Phi_{\mu_i,1}(-\infty) < u_i < \Phi_{\mu_i,1}(0)$.

Consequently, by setting $u_i \sim U(0, \Phi_{\mu_i,1}(0))$, the problem is solved, since the drawn z_is will always belong in an interval of the form: $\Big(\Phi^{-1}_{\mu_i,1}(0), \Phi^{-1}_{\mu_i,1}(\Phi_{\mu_i,1}(0))\Big)$, which is the interval $(-\infty, 0)$. In order to obtain values from $N(X_i^T, 1)$, truncated above 0, we simply consider that for each $i = 1, ..., n$, $u_i \sim U(\Phi_{\mu_i,1}(0), 1)$.

5.4 Model Selection

Due to the fact that the posterior distribution $f(\boldsymbol{\beta}|\boldsymbol{y})$ can only be simulated and because the use of latent variables, according to many authors, should prevent us from using the Bayesian and the Deviance Information Criterion, we will employ the L Measure in order to compare two arbitrary probit models.

Let $S_1 = \Big(\beta^{(1)}, ..., \beta^{(I)}\Big)$ and $S_2 = \Big(\beta^{(1)}, ..., \beta^{(J)}\Big)$ denote two samples from the posterior distribution of $\boldsymbol{\beta}$, that is $f(\boldsymbol{\beta}|\boldsymbol{y})$, under model M_1 and M_2 respectively, whereas $\boldsymbol{z} = \big(z_1, ..., z_n\big)$ shall denote a vector of future observations. We recall, that the L Measure is given by (1.7):

$$L_{IL} = \sum_{i=1}^{n} \Big\{ Var(z_i|\boldsymbol{y}) + \frac{1}{2} \cdot \big(E[z_i|\boldsymbol{y}] - \boldsymbol{y}\big)^2 \Big\}.$$

We will employ the *Law of Double Expectation*, as we did in the case of the logistic regression, to benefit from some special properties of the Generalized Linear Models and thus, significantly simplify the computational part. First, we should consider the following equations, which will allow us to express the conditional expectation of an observation, with the probit link. We already know from (4.3) that:

$$E[y_i|\boldsymbol{\beta}] = p_i = \frac{e^{X_i^T\boldsymbol{\beta}}}{1 + e^{X_i^T\boldsymbol{\beta}}}$$

and also from (5.1) that:

$$p_i = \Phi^{-1}\big(X_i^T\boldsymbol{\beta}\big).$$

Consequently, we can make the following assumptions:

$$E[y_i|\boldsymbol{\beta}] = \Phi^{-1}\big(X_i^T\boldsymbol{\beta}\big) \tag{5.7}$$

and that

$$e^{X_i^T\boldsymbol{\beta}} = \frac{\Phi^{-1}\big(X_i^T\boldsymbol{\beta}\big)}{1 - \Phi^{-1}\big(X_i^T\boldsymbol{\beta}\big)}. \tag{5.8}$$

Since from (4.14) we have deducted that $Var(y_i|\boldsymbol{\beta}) = \frac{e^{X_i^T\boldsymbol{\beta}}}{\left(1+e^{X_i^T\boldsymbol{\beta}}\right)^2}$, the last equation yields that:

$$Var(y_i|\boldsymbol{\beta}) = 1 - \Phi^{-1}\big(X_i^T\boldsymbol{\beta}\big). \tag{5.9}$$

Also, we need to bear in mind that $E[z_i|\boldsymbol{y},\boldsymbol{\beta}] = \Phi^{-1}\big(X_i^T\boldsymbol{\beta}\big)$. By the *Law of double expectation*, as this is expressed in (4.12), the expectation of a future observation with respect to the data at hand, $\boldsymbol{y}$, and by applying the Monte Carlo integration, takes the following form:

$$\begin{aligned} E[z_i|\boldsymbol{y}] &= E_{\boldsymbol{\beta}|\boldsymbol{y}}\Big[E\big[z_i|\boldsymbol{\beta},\boldsymbol{y}\big]\Big] = E_{\boldsymbol{\beta}|\boldsymbol{y}}\Big[\Phi^{-1}\big(X_i^T\boldsymbol{\beta}\big)\Big] \\ &\approx \frac{1}{s}\sum_{k=1}^{s}\Big[\Phi^{-1}\big(X_i^T\boldsymbol{\beta^{(k)}}\big)\Big]. \end{aligned} \tag{5.10}$$

Following similar steps with the ones required to transform (4.16) to (4.17), it can be shown that the variance of a predicted value z_i with respect to $\boldsymbol{y}$ can be written as:

$$Var(z_i|\boldsymbol{\beta}) \approx \frac{1}{s}\sum_{k=1}^{s}\Phi^{-1}\big(X_i^T\boldsymbol{\beta^{(k)}}\big) - \left(\frac{1}{s}\sum_{k=1}^{s}\Phi^{-1}\big(X_i^T\boldsymbol{\beta^{(k)}}\big)\right)^2, \tag{5.11}$$

where $S_m = \Big(\boldsymbol{\beta^{(1)}},...,\boldsymbol{\beta^{(s)}}\Big)$ is a sample of simulated values from the posterior distribution of $\boldsymbol{\beta}$ under a model m.

The values we need to compare, in order to select the model with the lowest score of the L Measure are the following:

$$\begin{aligned} L_{IL}^{\mathbf{1}} \approx &\sum_{i=1}^{n}\left\{\frac{1}{I}\sum_{k=1}^{I}\left[\frac{1}{I}\sum_{k=1}^{I}\Phi^{-1}\big(X_i^T\boldsymbol{\beta^{(k)}}\big) - \left(\frac{1}{I}\sum_{k=1}^{I}\Phi^{-1}\big(X_i^T\boldsymbol{\beta^{(k)}}\big)\right)^2\right]\right\} \\ &+ \frac{1}{2}\sum_{i=1}^{n}\left\{\frac{1}{I}\sum_{k=1}^{I}\left(\frac{1}{I}\sum_{k=1}^{I}\Big[\Phi^{-1}\big(X_i^T\boldsymbol{\beta^{(k)}}\big)\Big] - \boldsymbol{y}\right)^2\right\}, \quad \text{with} \quad \beta^{(1)},...,\beta^{(I)} \in S_1. \end{aligned}$$

and

$$\begin{aligned} L_{IL}^{\mathbf{2}} \approx &\sum_{i=1}^{n}\left\{\frac{1}{J}\sum_{k=1}^{I}\left[\frac{1}{J}\sum_{k=1}^{J}\Phi^{-1}\big(X_i^T\boldsymbol{\beta^{(k)}}\big) - \left(\frac{1}{J}\sum_{k=1}^{J}\Phi^{-1}\big(X_i^T\boldsymbol{\beta^{(k)}}\big)\right)^2\right]\right\} \\ &+ \frac{1}{2}\sum_{i=1}^{n}\left\{\frac{1}{J}\sum_{k=1}^{J}\left(\frac{1}{J}\sum_{k=1}^{J}\Big[\Phi^{-1}\big(X_i^T\boldsymbol{\beta^{(k)}}\big)\Big] - \boldsymbol{y}\right)^2\right\} \quad \text{with,} \quad \beta^{(1)},...,\beta^{(J)} \in S_2. \end{aligned}$$

Chapter 6

Stochastic Search Variable Selection

Motivation

Given a standard model structure, a major statistical task is to investigate the existence of a more parsimonious version of it, that will adequately interpret the data at hand. This procedure is in other words described as "variable selection" and amounts to the detection and removal of explanatory variables form the original, "full model". In the Bayesian framework, in order to perform variable selection, we need to compare the submodels corresponding to each one of the possible variable combinations. More specifically, for a given matrix $\boldsymbol{X} = (\boldsymbol{X_1}, ..., \boldsymbol{X_k})$ of explanatory variables, where $\boldsymbol{X_i} \in \mathbb{R}^n$ for each $i = 1, ..., k$, there will also be a corresponding vector of coefficients, denoted by $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k) \in \mathbb{R}^{k+1}$. The reader should recall that the first coordinate β_0 is the intercept of the model. The statement that "the variable X_i can be excluded from the model" is equivalent to setting the corresponding coefficient coordinate β_i equal to zero. Consequently, the candidate models are defined and distinguished from one another by the specific subset $\mathcal{B}$ of predictors, based on which, they are built. It is clarified that $\mathcal{B} \subset B$, where $B \subset \mathbb{R}^{k+1}$ denotes the set of predictors dictated by the full model.

Although it is very clear which models need to be compared and even though, depending on the form of the model, a statistician can employ one or more of the various methods mentioned in the previous chapters, in order to perform the necessary model comparisons, another very important aspect of the variable selection problem is the **total number** of these comparisons.

The simplest example of a case of variable selection would be a normal linear model. More specifically, suppose we wish to study a model of the following linear structure, with the usual assumptions:

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \epsilon_i, \quad \text{where} \quad \epsilon_i \sim N(0, \tau^{-1}),$$

for every $i \in \{1, 2, ..., n\}$. Initially, our interest lies in conducting inference for the vector of predictors, $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)$, each coordinate of which is considered to be a random variable. Once we gain a more solid perspective of the general statistical performance of the model, as well as for the behaviour of $\boldsymbol{\beta}$, it is required to find out whether we can afford to reduce its dimension without affecting its performance. To this purpose, we have to compare **all** the submodels, which can be derived by setting one or more predictors equal to zero. As a result, we are led to the study of 2^k submodels, which will all have to be compared with each other, so that the optimal, candidate submodel can be selected out.

It is obvious that the particular task requires the use of computers and can easily become very time-consuming. A linear model of just four explanatory variables would generate $2^5 = 32$ candidate models, that would have to be compared by couples. The reader should also bear in mind that an initial suggested model is usually of a higher dimension. This is due to the fact that a large number of explanatory variables provides a level of safety, considering the informative character of a model. Additionally, a researcher may be compelled to do so, due to the complexity of the phenomenon under study.

Thus, we reach the conclusion that, although the methods of model comparison mentioned in the previous chapters are consistent and not hard to implement, they seem to fail to serve the purpose of variable selection, in the sense that the statistician will have to engage in the cumbersome process of multiple model comparisons. Evidently, the search of a more parsimonious model comes at a rather high price in terms of time and resources by the traditional methods of the Bayesian paradigm. That fact sustained for a long time a major drawback of Bayesian Statistics, especially since it is easily carried out in the framework of Classical Statistics.

Bayesian Variable Selection and Model Averaging have been introduced as techniques to overcome the obstacle of the cumbersome process of multiple comparisons. Their use is at the current moment extensive and they are considered both efficient and very practical methods, with the additional advantage that they have a solid probabilistic background and they are based on intuitional, statistical thinking.

Due to the significance of the problem, various methods of Variable Selection have been introduced over the past few years. In their paper " A Review of Bayesian Variable Selection Methods:What, How and Which", R.B. O'Hara and M.J. Sillanpaam make an informative review of the most popular methods used currently: Kuo and Mallick, Gibbs Variable Selection (GVS), Stochastic Search Variable Selection (SSVS), adaptive shrinkage with Jereys prior or a Laplacian prior, and reversible jump MCMC. The present thesis shall focus solely on the study of the SSVS method, originally introduced by Edward I. George and Robert E. MacCullogh in 1993, and its application on Normal Linear models and Generalized Linear models.

6.1 Stochastic Search Variable Selection in Normal Linear Regression

Let us review the general structure and assumptions of a Normal Linear Regression model. The data available to us is the $n-$dimensional vector of independent variables, denoted by $\boldsymbol{y} = (y_1, ..., y_n) \in \mathbb{R}^n$ and the matrix of the explanatory variables, $\boldsymbol{X} = (\mathbf{1}, X_1, ..., X_k) \in \mathbb{R}^{n\times(k+1)}$. There is also the vector of predictors, with the notation $\boldsymbol{\beta} = (\beta_0, ..., \beta_k) \in \mathbb{R}^{k+1}$. We assume that the vector of the mean values of $\boldsymbol{y}$ can be approached as a linear combination of the explanatory variables, by the following formula:

$$\begin{aligned} &E[\boldsymbol{y}] = \boldsymbol{\beta}^T \boldsymbol{X} + \boldsymbol{\epsilon} \\ &\iff E[y_i] = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + \epsilon_i, \end{aligned}$$

where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \tau^{-1} I_n)$ is the vector of random errors.

Having conducted inference for that particular model, our efforts are turned to the detection of a "promising" subset of predictors $\boldsymbol{X*} = \left(X_1^*, X_2^*, ..., X_q^*\right)$, where $q \leq k+1$, by which another model, of similar performance to the full model, is built. The new model will be henceforth denoted by M^*. In order to find out which coefficients are non-zero in M^*, we consider a vector of latent variables, denoted by $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, ..., \gamma_k)$, with $\gamma_j \in \{0, 1\}$ for every $j = 0, 1, ..., k$. The idea is to use the posterior distribution of $\boldsymbol{\gamma}$ in the following way: the selected coefficients will be indicated by the unit coordinates of $\boldsymbol{\gamma}$, whereas the zero coordinates will correspond to the coefficients that can be set equal to zero.

6.1.1 Building a hierarchical model

In order to implement this idea, George and Mac Cullogh suggested that "the original regression model should be embedded in a larger hierarchical model", with the key feature that the prior assigned to each coordinate of $\boldsymbol{\beta}$ would be the following mixture of two Gaussians, with respect to $\boldsymbol{\gamma}$:

$$(\beta_j|\gamma_j) \quad \sim \quad (1-\gamma_j)N(0,\tau_j^{-1}) + \gamma_i N(0, c_j\tau_j^{-1}), \qquad j = 0, 1, ..., k. \tag{6.1}$$

Thus, we successfully establish a strong dependence between $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Since $\boldsymbol{\gamma}$ is an unknown vector of variables, a prior distribution has to be placed on it. Assuming prior independence between its coordinates, the simplest choice would be to assign a Bernoulli distribution on each one of them and so we consider that:

$$\gamma_j = \begin{cases} 1, & \text{with porbability} \quad p_j, \\ 0, & \text{with probability} \quad 1-p_j, \qquad j = 0, 1, ...k. \end{cases}$$

Each probability $p_i \in [0, 1]$ is defined by the statistician, reflecting his belief about the importance of the corresponding variable X_i for the interpretation of the data. Hence, a large prior probability expresses the belief that X_i is a very valuable explanatory variable, whereas with a small p_i, the incident ($\gamma_i = 1$) rarely occurs, implying that the information provided by X_i is of little significance for inference. Thus, we end up with the following hierarchical structure:

$$\begin{aligned} &(y_i|\boldsymbol{\beta}, \tau^2, \boldsymbol{\gamma}) \sim N\big(X_i^T\boldsymbol{\beta}, \tau^{-1}\big), \qquad i = 1, ..., n, \\ &(\beta_j|\gamma_j) \sim (1-\gamma_j)N(0,\tau_j^{-1}) + \gamma_j N(0, c_j\tau_j^{-1}), \\ &\tau \sim G\Big(\frac{\nu}{2}, \frac{\lambda}{[2\nu]}\Big), \\ &\gamma_j \sim Bern(p_j), \qquad j = 0, 1, ..k. \end{aligned} \tag{6.2}$$

In matrix notation the model is written as follows:

$$\begin{aligned} &\big(\boldsymbol{y}|\boldsymbol{\beta}, \tau, \boldsymbol{\gamma}\big) \sim N_n\Big(\boldsymbol{X}\boldsymbol{\beta}, \tau^{-1}I_n\Big), \\ &\big(\boldsymbol{\beta}|\boldsymbol{\gamma}\big) \sim N_{k+1}\Big(\mathbf{0}, D_\gamma^T\Sigma D_\gamma\Big), \\ &\tau \sim G\Big(\frac{\nu}{2}, \frac{\lambda}{[2\nu]}\Big), \\ &\gamma_j \sim Bern(p_j), \qquad j = 0, 1, ..k, \end{aligned} \tag{6.3}$$

where Σ is the prior correlation matrix for the vector $\boldsymbol{\beta}$ and $D_\gamma := diag\big(\alpha_0\tau_0^{-1}, \alpha_1\tau_1^{-1}, ..., \alpha_k\tau_k^{-1}\big)$, with $\alpha_j = 1$, if $\gamma_j = 0$, whereas $\alpha_j =$

c_j, if $\gamma_j = 1$. We note that the precision of each Gaussian density in the conditional prior of $\beta_j, j = 0, 1, ..., k$, that is τ_j^2 is irrelevant to the precision τ of the response variables.

Furthermore, although we chose to set τ independent from $\boldsymbol{\gamma}$, George and MacCullogh in their original paper attempted to suggest ways of setting appropriate values λ_γ and ν_γ in order to incorporate dependence on $\boldsymbol{\gamma}$. Nevertheless, they point out that the fact that $\boldsymbol{y}$ depends on $\boldsymbol{\gamma}$ only through $\boldsymbol{\beta}$ is a common feature in hierarchical modelling, which in our case simplifies the computational procedure.

6.1.2 Setting τ_j and c_j

Contrary to the precision of the explanatory variables, that is τ, which is a random variable and therefore a prior distribution is placed on it, τ_j and c_j have fixed values, prespecified by the statistician. Nevertheless, these values need to be set with caution, so that the purpose of the SSVS model structure is served. More specifically, we need to ensure that the following two properties will hold:

- By getting $\gamma_j = 0$ *a posteriori*, we want to assume that β_j can be "safely" set equal to zero, for every $j = 0, 1, ..., k$. Since, if $\gamma_j = 0$, then $(\beta_j|\gamma_j) \sim N(0, \tau_j^{-1})$, such an assumption can be justified only if the value of τ_j^{-1} is small, thus causing the Gaussian distribution to be clustered around 0. As a result, the probability $P\Big(|\beta_j| \leq \epsilon\Big)$, for an appropriately small value $\epsilon > 0$, will be sufficiently large to permit setting $\beta_j = 0$.
- If γ_j turns out to be equal to 1, this yields that $(\beta_j|\gamma_j) \sim N(0, c_j\tau_j^{-1})$. If $\gamma_j = 1$, we want to be led to the conclusion that $\beta_j \neq 0$. For that reason, the quantity $c_j\tau_j^{-1}$ has to be large enough, so that only regions with substantially large, non-zero values have high density, meaning that a zero or "close to zero" value for β_j is a very rare and probably unlikely incident.

The problem we are facing is how the product $c_j\tau_j^{-1}$ is affected by the values we set, since a very large c_j and a very small τ_j^{-1} could cancel each other out and thus, lead to an inappropriate value of $c_j^2\tau_j^2$. Should this occur, then the assumption that $\beta_j \neq 0$, if $\gamma_j = 1$, is unjustified. In other words, we need to ensure that if the data support $\gamma_j = 1$ over $\gamma_j = 0$, then the incident $\big(\beta_j = 0\big)$ is very unlikely. So, we are challenged to come up with an efficient combination of values for c_j and τ_j.

MacCullogh and George's remarks and suggestions

Mac Cullogh and George attempted to suggest a form of criterion that could guide us to a safe choice of values, which is based on the following thinking:
They considered the *intersection point* of $N\big(0, \tau_j^{-1}\big)$ and $N\big(0, c_j\tau_j^{-1}\big)$, which can be expressed as $c_j t_j$. Based on this expression, they concluded that $t_j = \sqrt{2(lnc_j)c_j^2/(c_j - 1)}$. They observed that the intersection point has the following property: "The density of $N\big(0, c_j\tau_j^{-1}\big)$ is larger that the density of $N\big(0, \tau_j^{-1}\big)$ if and only if $|\beta_j| > c_j t_j$." Additionally, they pointed out that c_j is "the ratio of the heights of the two Gaussians at 0. Thus, they concluded the very interesting interpretation of c_j as "the prior odds that X_j should be excluded when β_j is very close to 0".

Furthermore, t_j can be regarded as the ***statistic-threshold*** with a very particular use:
If the outcome $\gamma_j = 1$ leads to a posterior value of β_j, such that $|\beta_j| > c_j t_j$, then we get that $P\big(\gamma_j = 1\big) > P\big(\gamma_j = 0\big)$. This means that "the variable X_j has an increased probability to be involved in the model. This yields that small values for t_j tend to favour more complex models, whereas large values of t_j would point to more parsimonious models.

Finally, in the paper "The Practical Implementation of Bayesian Model Selection", which was published in 2001 and included a review of the SSVS, Chipman, George and MacCullogh noted that any τ_j and c_j satisfying the property: $ln\big(c_j\tau_j/\tau_j\big)/\big[\tau_j^{-1} - c_j^{-1}\tau_j^{-1}\big] = t_j^2$, for a given t_j, are considered appropriate choices. They additionally advised that c_j should be set to a value less than 10.000.

6.1.3 Extracting the best subsets through $f(\gamma|y)$ using the Gibbs Sampler

In order to avoid the cumbersome process of calculating 2^{k+1} posterior probabilities for every model derived from the possible combinations of predictors, SSVS uses the hierarchical structure described above and then the Gibbs Sampler to get a sequence of observations from the posterior distribution of $\boldsymbol{\gamma}$, that is $f(\boldsymbol{\gamma}|\boldsymbol{y})$. We collect the values we will get from the algorithm after the burn-in period in order to consider a sample from $f(\boldsymbol{\gamma}|\boldsymbol{y})$, denoted by

$$\boldsymbol{S_\gamma} = \big\{\boldsymbol{\gamma_1}, ..., \boldsymbol{\gamma_L}\big\}, \quad \text{where} \quad L \in \mathbb{N}*, \tag{6.4}$$

and we rely on it to indicate which is the optimal subset of predictors. First, we shall describe the structure of the Gibbs sampler.

The Gibbs Sampler

1. We initialize $\boldsymbol{\beta}, \boldsymbol{\gamma}$ and τ by setting $\boldsymbol{\beta^{(0)}} = \left(\boldsymbol{X^T X}\right)^{-1}\boldsymbol{X^T Y}$ and $\tau = \frac{(\boldsymbol{y}-\boldsymbol{X\beta^0})^{\boldsymbol{T}}(\boldsymbol{y}-\boldsymbol{X\beta^{(0)}})}{n-k-1}$, which are the least squares estimators for the coefficient vector and the precision respectively, whereas we set $\boldsymbol{\gamma^{(0)}} = (1, 1, ..., 1)$.

2. At the j-th iteration of the algorithm, we simulate $\boldsymbol{\beta^{(j)}}$ from its conditional distribution $f\left(\boldsymbol{\beta^{(j)}}|\boldsymbol{y}, \tau^{(j-1)}, \boldsymbol{\gamma^{(j-1)}}\right) \equiv N(\mu_{\gamma^{(j-1)}}, \Sigma_{\gamma^{(j-1)}})$, where

$$\begin{aligned} \mu_{\gamma^{(j-1)}} &= \tau^{(j-1)}\Sigma_{\gamma^{(j-1)}}\left(X^T X\right)\hat{\beta}_{LS} \quad \text{and} \\ \Sigma_{\gamma^{(j-1)}} &= \left[(D_\gamma R D_\gamma)^{-1} + \tau^{-1}(X^T X)\right]^{-1}. \end{aligned} \tag{6.5}$$

The result above is derived by Bayes' Theorem, since

$$f(\boldsymbol{\beta^{(j-1)}}|\boldsymbol{\gamma^{(j-1)}}) \propto \exp\left\{\frac{\boldsymbol{\beta^{(j-1)}}^T(D_\gamma R D_\gamma)^{-1}\boldsymbol{\beta^{(j-1)}}}{2}\right\} \quad \text{and}$$

$$f(\boldsymbol{y}|\boldsymbol{\beta^{(j-1)}}, \boldsymbol{\gamma^{(j-1)}}, \tau^{(j-1)}) \propto \left(\tau^{(j-1)}\right)^{n/2} \exp\left\{\frac{(y - X\beta^{(j-1)})^T\tau^{(j-1)}(y - X\beta^{(j-1)})}{2}\right\}$$

and so by applying Bayes' Theorem, we get that

$$\begin{aligned} f(\boldsymbol{\beta^{(j)}}|\boldsymbol{y}, \boldsymbol{\gamma^{(j-1)}}, \tau^{(j-1)}) &\propto f(\boldsymbol{y}|\boldsymbol{\beta^{(j)}}, \boldsymbol{\gamma^{(j-1)}}, \tau^{(j-1)}) f(\boldsymbol{\beta^{(j)}}|\boldsymbol{\gamma^{(j-1)}}) \\ &\propto \exp\left\{\frac{\left(\boldsymbol{\beta^{(j)}} - \mu_{\gamma^{(j-1)}}\right)^T\Sigma^{-1}_{\gamma^{(j-1)}}\left(\boldsymbol{\beta^{(j)}} - \mu_{\gamma^{(j-1)}}\right)}{2}\right\}. \end{aligned}$$

Again by applying Bayes' Theorem and using $\boldsymbol{\beta^{(j)}}$, we can derive the conditional distribution of $\tau^{(j)}$, that is $f(\tau^{(j)}|\boldsymbol{y}, \boldsymbol{\beta^{(j)}}, \boldsymbol{\gamma^{(j-1)}})$, since

$$\begin{aligned} f(\tau^{(j)}|\boldsymbol{y}, \boldsymbol{\beta^{(j)}}, \boldsymbol{\gamma^{(j-1)}}) &\propto f(\boldsymbol{y}|\boldsymbol{\beta^{(j)}}, \boldsymbol{\gamma^{(j-1)}}, \tau^{(j)}) f(\boldsymbol{\tau^{(j)}}) \\ &\propto \left(\tau^{(j)}\right)^{n/2} \exp\left\{\frac{\left(y - X\beta^{(j)}\right)^T\tau^{(j)}\left(y - X\beta^{(j)}\right)}{2}\right\}\left(\tau^{(j)}\right)^{\lambda/2-1} \exp\left\{-\frac{\lambda}{[2\nu]}\tau^{(j)}\right\} \\ &= \left(\tau^{(j)}\right)^{\frac{n+\lambda}{2}-1} \exp\left\{-\tau^{(j)}\left[\frac{\left(y - X\beta^{(j)}\right)^T\left(y - X\beta^{(j)}\right) + \lambda/[\nu]}{2}\right]\right\}. \end{aligned}$$

Evidently, $\left(\tau^{(j)}|\boldsymbol{y}, \boldsymbol{\beta^{(j)}}, \boldsymbol{\gamma^{(j-1)}}\right) \sim \Gamma\left(\frac{n+\lambda}{2}, \frac{\left(y-X\beta^{(j)}\right)^T\left(y-X\beta^{(j)}\right)+\lambda/[\nu]}{2}\right)$.

3. The vector $\boldsymbol{\gamma^{(j)}}$ is sampled *componentwise* and consecutively according to the following scheme:
Let us assume that we are at the point of the algorithm at which a total number of $m \leq k+1$ coordinates have been updated for the j-th time. We choose γ_i from the set of the coordinates that have not yet been upgraded at the j-th iteration of the Gibbs Sampler - we do not necessarily follow any particular order for the selection of i - and we sample $\boldsymbol{\gamma_i^{(j)}}$ from $f\Big(\gamma_i^{(j)}|\boldsymbol{y}, \gamma_{S_i^{(j-1)}}^{(j-1)}, \boldsymbol{\beta^{(j)}}, \tau^{(j)}\Big)$, where $S_i^{(j-1)} = \Big(\gamma_l^{(k)}\Big)_{k,l}$, with $l \in \{0, 1, ..., i-1, i+1, ..., k\}$ and $k \in \{j-1, j\}$. We note that the values of k, mathematically express whether the coordinate γ_i has been updated for the j-th time or not. What is more, due to the structure of the model, there is nondependence on $\boldsymbol{y}$ and $\tau^{(j)}$ (George and MacCullogh, 1993) and therefore $f\Big(\gamma_i^{(j)}|\boldsymbol{y}, \gamma_{S_i^{(j-1)}}^{(j-1)}, \boldsymbol{\beta^{(j)}}, \tau^{(j)}\Big) = f\Big(\gamma_i^{(j)}|\gamma_{S_i^{(j-1)}}^{(j-1)}, \boldsymbol{\beta^{(j)}}\Big)$, which simplifies the computational procedure.

The distribution $f\Big(\gamma_i^{(j)}|\gamma_{S_i^{(j-1)}}^{(j-1)}, \boldsymbol{\beta^{(j)}}\Big)$ is a Bernoulli distribution with success probability $P\Big[\gamma_i^{(j)} = 1|\gamma_{S_i^{(j-1)}}^{(j-1)}, \boldsymbol{\beta^{(j)}}\Big] = \dfrac{a}{a+b}$, where

$$\begin{aligned} a =& f\Big[\boldsymbol{\beta^{(j)}}|\gamma_{S_i^{(j-1)}}^{(j-1)}, \gamma_i^{(j)} = 1\Big] \cdot P\Big[\gamma_i^{(j)} = 1\Big] \\ =& f\Big[\boldsymbol{\beta^{(j)}}|\gamma_{S_i^{(j-1)}}^{(j-1)}, \gamma_i^{(j)} = 1\Big] \cdot p_i, \end{aligned} \tag{6.6}$$

whereas

$$\begin{aligned} b =& f\Big[\boldsymbol{\beta^{(j)}}|\gamma_{S_i^{(j-1)}}^{(j-1)}, \gamma_i^{(j)} = 0\Big] \cdot P\Big[\gamma_i^{(j)} = 0\Big] \\ =& f\Big[\boldsymbol{\beta^{(j)}}|\gamma_{S_i^{(j-1)}}^{(j-1)}, \gamma_i^{(j)} = 0\Big] \cdot \big(1 - p_i\big). \end{aligned} \tag{6.7}$$

We perform the same procedure until we have finally obtained $\boldsymbol{\gamma^{(j)}}$.

Diebolt and Robert (1994) proved that the sequence

$$\boldsymbol{\gamma^{(1)}}, \boldsymbol{\gamma^{(2)}}, ..., \boldsymbol{\gamma^{(M)}}, \quad \text{where} \quad M \in \mathbb{N}^*,$$

that is formed by the Gibbs Sampler is a homogeneous ergodic Markov chain that converges geometrically to its equilibrium distribution $f(\boldsymbol{\gamma}|\boldsymbol{y})$. The convergence is more rapid if $f(\boldsymbol{\gamma}|\boldsymbol{y})$ is peaked, having as a result the probability mass not to be scattered, but concentrated in a small area. If this occurs, it is beneficiary for the purpose of model selection, since the most "valuable" and informative explanatory variables would be few and they would stand out as those, corresponding to the predictors with the highest posterior probability. More specifically, a tabulation of the obtained sample S_L can show us the frequency of the incident $(\gamma_i = 1)$, for every $i = 0, 1, ..., k$. Thus, we can derive a vector $\left(\boldsymbol{\gamma^*_{q_1}}, \boldsymbol{\gamma^*_{q_2}}, ..., \boldsymbol{\gamma^*_{q_l}}\right)$, with $\{q_1, q_2, ..., q_l\} \subset \{0, 1, ..., k\}$ of the coordinates that are most frequently equal to 1, indicating the predictors that most likely have non-zero values. The optimal, more parsimonious model is the one built by the corresponding explanatory variables.

6.2 Stochastic Search Variable Selection for Logistic Regression

Given a set of binary data $\boldsymbol{y} = (y_1, ..., y_n)$, where we assume that $y_i \sim Bern(p_i), p_i \in [0, 1]$ for every $i = 1, 2, ..., n$, a vector of explanatory variables $\boldsymbol{X} = (X_1, ..., X_k)$ and the coefficient vector $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)$, we wish to conduct inference regarding the unknown vector of parameters $\boldsymbol{p} = (p_1, ..., p_n)$. We link $\boldsymbol{p}$ with the vector of explanatory variables $\boldsymbol{X}$, assuming the following connection between them:

$$\begin{aligned} & log\left(\frac{p_i}{1-p_i}\right) = X_i^T\boldsymbol{\beta} \\ \iff\ & p_i = \frac{e^{X_i^T\boldsymbol{\beta}}}{1+e^{X_i^T\boldsymbol{\beta}}}, \\ & \text{where} \quad p_i = E[y_i|p_i], \quad \text{for every} \quad i = 1, 2, ..., n. \end{aligned} \tag{6.8}$$

We have already elaborated on the mathematical and computational procedures we follow in the Bayesian framework in previous chapters. Admittedly, logistic regression is a commonly used model structure in studies in econometrics, genetics, biology etc., the majority of which include a very large number of explanatory variables, thus creating very complex models. The computational difficulties that arise and the fact that even with the assistance of computers the assessment of the data becomes time-consuming and cumbersome, justify the expansion of the SSVS method on the generalised linear models. It was a necessity to

avoid the overwhelming amount of comparisons between the 2^p models that had to be examined, so that more parsimonious submodels could be proposed for the interpretation of the data. MacCullogh, George and Tsay (1995) proceeded to the implementation of SSVS on logistic regression models and soon after that, SSVS was adopted by a variety of scientists in statistical research. A very characteristic example is the paper of Swarz et. al. (2006), which employed SSVS in the study of Gene Mapping, thus introducing the "SSGS" method for Gene Mapping. Furthermore, SSVS for generalised linear models captured the interest of many statisticians, who engaged in the research of computational optimization, implementation potentials and comparison to other techniques of Bayesian variable selection. The present thesis, reviews results and remarks from the papers of Dellaportas and Smith (1993), Ntzoufras, Foster and Dellaportas (2000), Chipman, George and MacCullogh (2001) and Swartz, M.D., Yu, R.K., Shete,S. (2008).

The hierarchical model we consider is built in a similar way as the one in (6.2), which was used for normal linear regression. Naturally, the likelihood of the data is expressed by (4.4) and the random variables of the model are $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. More specifically, the conditional distributions describing the variables currently are as follows:

- The likelihood of the data, with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$; that is

$$L(\boldsymbol{y}|X,\boldsymbol{\beta},\boldsymbol{\gamma}) = \prod_{i=1}^{n}\left(\frac{\mathrm{e}^{X_i^T\boldsymbol{\beta}}}{1+\mathrm{e}^{X_i^T\boldsymbol{\beta}}}\right)^{y_i}\left(\frac{1}{1+\mathrm{e}^{X_i^T\boldsymbol{\beta}}}\right)^{1-y_i}.$$

- The conditional prior placed on $\boldsymbol{\beta}$ with respect to $\boldsymbol{\gamma}$ is again expressed as a mixture of two Normal distributions; that is $(\beta_j|\gamma_j) \sim (1-\gamma_j)N(0,\tau_j^{-1}) + \gamma_j N(0, c_j\tau_j^{-1})$, where τ_j and c_j are set, as described in section 6.1.2, for every $j = 0, 1, ..., k$. In matrix notation, the prior can be expressed thusly:

$$\big(\boldsymbol{\beta}|\boldsymbol{\gamma}\big) \sim N_{k+1}\Big(\mathbf{0}, D_\gamma^T \Sigma D_\gamma\Big),$$

  where Σ is the prior correlation matrix for the vector of $\boldsymbol{\beta}$ and $D_\gamma := diag\big(\alpha_0\tau_0^{-1}, \alpha_1\tau_1^{-1}, ..., \alpha_k\tau_k^{-1}\big)$, with $\alpha_j = 0$, if $\gamma_j = 0$ whereas $\alpha_j = c_j$, if $\gamma_j = 1$.
- Each one of the coordinates of $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, ..., \gamma_k)$ gets values from $\{0, 1\}$, according to a Bernoulli distribution, where $P\big[\gamma_j = 0\big] = 1 - P\big[\gamma_j = 1\big] = 1 - p_j$, with $p_j \in [0, 1]$ for every $j = 0, 1, ..., k$. In mathematical notation, $\gamma_j \sim Bern(p_j), \quad j = 0, 1, ..., k$.

Consequently, the hierarchical model we will be studying has the following structure:

$$\begin{aligned} &L(\boldsymbol{y}|X,\boldsymbol{\beta},\boldsymbol{\gamma}) = \prod_{i=1}^{n}\left(\frac{\mathrm{e}^{X_i^T\boldsymbol{\beta}}}{1+\mathrm{e}^{X_i^T\boldsymbol{\beta}}}\right)^{y_i}\left(\frac{1}{1+\mathrm{e}^{X_i^T\boldsymbol{\beta}}}\right)^{1-y_i}, \\ &(\beta_j|\gamma_j) \sim (1-\gamma_j)N(0,\tau_j^{-1}) + \gamma_j N(0, c_j\tau_j^{-1}), \\ &\gamma_j \sim Bern(p_j), \qquad j = 0,1,..k. \end{aligned} \tag{6.9}$$

6.2.1 Drawing from the full conditional distribution of $(\boldsymbol{\beta}|\boldsymbol{\gamma},\boldsymbol{y},\boldsymbol{X})$

The extra difficulty in applying the SSVS in the logistic regression model is the intractability of the conditional posterior distribution of $\boldsymbol{\beta}$. More specifically, by applying Bayes' Theorem, we are once more led to expression (4.7); that is

$$f(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{\gamma},X) \propto \prod_{i=1}^{n}\left(\frac{\mathrm{e}^{X_i^T\boldsymbol{\beta}}}{1+\mathrm{e}^{X_i^T\boldsymbol{\beta}}}\right)^{y_i} \cdot \mathrm{e}^{-\frac{1}{2}\boldsymbol{\beta}^T(D_\gamma R D_\gamma)^{-1}\boldsymbol{\beta}}.$$

Since the above functional form cannot be identified, we have to simulate a sample from the conditional posterior distribution of $\boldsymbol{\beta}$ using a slightly more sophisticated algorithm. We will follow a computational procedure proposed by Lynn Kuo and Bani Mallick in their paper "Variable Selection for Regression Models", which is mostly known as the ***"Metropolis-within-Gibbs"*** algorithm.

Employing the Metropolis-within-Gibbs algorithm for SSVS

The algorithmic scheme known as "Metropolis-within-Gibbs" is a hybrid MCMC procedure, originally introduced by Muller(1991, 1994) and can be regarded as a combination of Gibbs and M-H. It is often employed for the study of complex, high-dimensional models and it is basically the Gibbs algorithm with a step, that is an embedded loop, where a Metropolis-Hastings step is performed.

Returning to the hierarchical logit model, we are facing the major obstacle that $f(\boldsymbol{\beta}|\boldsymbol{\gamma})$ is intractable, and, as a result, it cannot be simulated by the Gibbs Sampler. For that particular purpose, at the step of the algorithm where a value from $f(\boldsymbol{\beta}|\boldsymbol{\gamma},\boldsymbol{y})$ has to be drawn, this is achieved via the Independence Sampler. More specifically, suppose that

we are at the point where the $i+1$ iteration of the algorithm is initiated, meaning that $\boldsymbol{\beta^{(i)}}$ and $\boldsymbol{\gamma^{(i)}}$ are available and it is the point the iterations of the M-H loop are performed. Let us assume that the total number of iterations is $I \in \mathbb{N}$. Mallick and Kuon describe that the K-th iteration, $1 \leq K \leq I$ goes through the following steps:

- Since we are building a Metropolis-Hastings algorithm, and more specifically an Independence Sampler, an appropriate proposal distribution needs to be set. At the K-th iteration, the input would be the vector $\boldsymbol{\beta_K^{(i)}}$ and the proposal distribution is set to be $q = N_{k+1}\Big(\boldsymbol{\beta_K^{(i)}}, c\Sigma_K\Big)$, where Σ_K is the current estimation of the posterior covariance matrix, whereas c is a fixed constant, appropriately adjusted so that a satisfactory staying rate is achieved. This yields, that the acceptance probability of a proposed vector $\boldsymbol{\beta^{(can)}} \sim N_{k+1}\Big(\boldsymbol{\beta_K^{(i)}}, c\Sigma_K\Big)$ is

$$a_K = min\left\{1, \frac{f\big(\boldsymbol{\beta_K^{(i)}}|\boldsymbol{\gamma^{(i)}}, \boldsymbol{y}\big) \cdot q(\boldsymbol{\beta^{(can)}}|\boldsymbol{\beta_K^{(i)}}))}{f\big(\boldsymbol{\beta^{(can)}}|\boldsymbol{\gamma^{(i)}}, \boldsymbol{y}\big) \cdot q(\boldsymbol{\beta_K^{(i)}}|\boldsymbol{\beta^{(can)}})}\right\}.$$

  We should be cautious, when calculating $q(\boldsymbol{\beta^{(can)}})$, because in order to obtain the mean and the covariance matrix, a single iteration of the IWLS has to be performed using as input $\boldsymbol{\beta_K^{(i)}}$.

- We set

$$\boldsymbol{\beta_{K+1}^{(i)}} = \begin{cases} \boldsymbol{\beta^{(can)}} & \text{with porbability} \quad a_K \\ \boldsymbol{\beta_K^{(i)}} & \text{with probability} \quad 1 - a_k \end{cases}$$

- We perform the $K+1$-st iteration, with $\boldsymbol{\beta_{K+1}^{(i)}}$ as input.

This loop has to be repeated, until the target distribution $f(\boldsymbol{\beta^{(i+1)}}|\boldsymbol{y}, \boldsymbol{\gamma^{(i)}}, X)$, which is also the equilibrium distribution of the formed chain, is approached. Mallick and Kuon report that a total number I of iterations between 20 and 50 would suffice. The last vector that is obtained, denoted by $\boldsymbol{\beta_I^{(i)}}$ is set as $\boldsymbol{\beta^{(i+1)}}$, since it can be regarded as a draw from the conditional distribution $f(\boldsymbol{\beta^{(i+1)}}|\boldsymbol{y}, \boldsymbol{\gamma^{(i)}}, X)$. Then, the main algorithm proceeds to the simulation of $\boldsymbol{\gamma^{(i+1)}}$, which will also be described in the following section.

Employing Gamerman's Independence Sampler

Previously, as we described the Metropolis-Hastings part of the algorithm, when the proposal distribution q was set, it was implied that we need to come up with an appropriate value for c, so that the acceptance rate is within the desirable limits. That fact can be proved troublesome and time consuming, since we will probably need to engage in a series of "trial and error" tests, in order to come up with a suitable value. Furthermore, the M-H step is part of a larger algorithm, meaning that there are more complex updates and dependence between the variables, which automatically makes the task much more challenging.

Bearing in mind the introduction of Gamerman's Independence Sampler in section 4.1.2 for the specific task of the simulation of the posterior distribution of $\boldsymbol{\beta}$ in the logit model, instead of having to handle a tuning constant, by employing that particular technique, we end up with an automatic tuning process, which could also lead to more rapid convergence. To be more specific, the modification we propose for the M-H part of the algorithm, given $\boldsymbol{\beta^{(i)}}$ and $\boldsymbol{\gamma^{(i)}}$, is that at each iteration K of the loop, the proposal distribution will also be updated, based on the current value of $\boldsymbol{\beta}$. The modified algorithmic scheme is thoroughly explained below.

Let $\boldsymbol{\beta_K^{(i)}}$ be the current value, used as input, as the $K+1$-st iteration of the M-H loop initiates, where $1 \leq K \leq I$. A single iteration of the IWLS algorithm provides an approximation of the mode of $f(\boldsymbol{\beta^{(i+1)}}|\boldsymbol{\gamma^{(i)}}, \boldsymbol{y}, X)$, which is denoted by $b_K^{(i+1)}$. Based on the methodology and notation of section 4.1.2 , can be expressed as $b_K^{(i+1)}$

$$b_K^{(i+1)} = [X^T W(\beta_K^{(i)})X + (D_{\gamma^{(i)}} R D_{\gamma^{(i)}})^{-1}] \cdot [Z(\beta_K^{(i)})^T W(\beta_K^{(i)})X],$$

where $W(\beta_K^{(i)})$ and $Z(\beta_K^{(i)})$ are calculated as described in section 4.1.2. We also consider the matrix

$$C_K^{(i+1)} = \left[X^T W(\beta_K^{(i)})X + (D_{\gamma^{(i)}} R D_{\gamma^{(i)}})^{-1}\right]^{-1}.$$

Consequently, the proposal distribution at the K-th iteration is $q \equiv N_{k+1}(b_K^{(i+1)}, C_K^{(i+1)})$. The rest part of the algorithm remains the same.

6.2.2 Drawing from the full conditional distribution of $(\boldsymbol{\gamma}|\boldsymbol{\beta}, \boldsymbol{y})$

The update of $\boldsymbol{\gamma}$, with respect to $\boldsymbol{y}$ and the current value of $\boldsymbol{\beta}$ is achieved with the same method as in the case of the SSVS for the linear

regression model. More specifically, at the $i+1$-st step of the main algorithm, using $\boldsymbol{\beta^{(i)}}$, we update γ_i componentwise and consecutively. A very important remark we need to make is that after each the component $\gamma_j, \quad j = 1, 2, ..., k+1$ is updated, we have to accordingly update the matrix D and use it as input for a Gamerman's Independence Sampler, which we iterate 20 times in order to obtain a new value for the coefficient vector, denoted by $\boldsymbol{\beta}_{\boldsymbol{\gamma_j}}^{\boldsymbol{(i+1)}}$. Then, we can proceed to the update of the rest of the components, which is carried out likewise. It is preferable that the update is performed randomly, rather than in a specific, deterministic order. Hence, we arrive at the conclusion that for each component $j \in \{0, 1, ..., k\}$, $\gamma_j^{(i+1)}$ is drawn from the conditional distribution denoted by $f\left(\gamma_j^{(i+1)} | \gamma_{S_j^{(i)}}^{(i)}, \boldsymbol{\beta}_{\boldsymbol{\gamma_j}}^{\boldsymbol{(i+1)}}, \boldsymbol{y}\right)$, where, once again, $S_j^{(i)} = \left(\gamma_m^{(n)}\right)_{m,n}$, with $m \in \{0, 1, ..., j-1, j+1, ..., k\}$ and $n \in \{i, i+1\}$. Based on Bayes' Theorem, we can finally conclude that

$$f\left(\gamma_j^{(i+1)} | \gamma_{S_j^{(i)}}^{(i)}, \boldsymbol{\beta}_{\boldsymbol{\gamma_l}}^{\boldsymbol{(i+1)}}, \boldsymbol{y}\right) \equiv Bern\left(\frac{a_j}{a_j + b_j}\right), \tag{6.10}$$

where

$$\begin{aligned} a_j &= P\left[\gamma_j^{(i)} = 1\right] \cdot L\left(\boldsymbol{y} | \boldsymbol{\beta}_{\boldsymbol{\gamma_l}}^{\boldsymbol{(i+1)}}, \gamma_{S_j^{(i)}}^{(i)}, \gamma_j^{(i)} = 1, X\right) \\ &= p_i \cdot L\left(\boldsymbol{y} | \boldsymbol{\beta}_{\boldsymbol{\gamma_l}}^{\boldsymbol{(i+1)}}, \gamma_{S_j^{(i)}}^{(i)}, \gamma_j^{(i)} = 1, X\right) \end{aligned} \tag{6.11}$$

and

$$\begin{aligned} b_j &= P\left[\gamma_j^{(i)} = 0\right] \cdot L\left(\boldsymbol{y} | \boldsymbol{\beta}_{\boldsymbol{\gamma_l}}^{\boldsymbol{(i+1)}}, \gamma_{S_j^{(i)}}^{(i)}, \gamma_j^{(i)} = 0, X\right) \\ &= (1 - p_i) \cdot L\left(\boldsymbol{y} | \boldsymbol{\beta}_{\boldsymbol{\gamma_l}}^{\boldsymbol{(i+1)}}, \gamma_{S_j^{(i)}}^{(i)}, \gamma_j^{(i)} = 0\right). \end{aligned} \tag{6.12}$$

By $\gamma_l^{(i+1)}$ we refer to the last updated component, which is not necessarily $\gamma_{j-1}^{(i+1)}$, since the components are updated randomly, whereas $\boldsymbol{\beta}_{\boldsymbol{\gamma_l}}^{\boldsymbol{(i+1)}}$ denotes the corresponding simulated value of $\boldsymbol{\beta}$, derived from Gamerman's Independence Sampler.

6.2.3 The SSVS Algorithm for the Logit model

The algorithmic scheme used for the construction of the chain follows the steps described below, which are iterated until convergence is reached:

1. Initialize $\boldsymbol{\gamma^{(0)}} = (1, ..., 1)$ and $\boldsymbol{\beta^{(0)}} = \boldsymbol{\beta}_{LS} = \boldsymbol{(X^T X)^{-1} X^T Y}$.

2. At the $i+1$-th iteration of the main loop we use as input $\left(\boldsymbol{\gamma^{(i)}}, \boldsymbol{\beta^{(i)}}\right)$ and perform the following steps:

   (a) Use the current value of $\boldsymbol{\beta}$ and the current version of matrix D to update a single component of $\boldsymbol{\gamma_j}$, by drawing $\gamma_j^{(i+1)} \sim f\left(\gamma_j^{(i+1)} | \gamma^{(i)}_{S_j^{(i)}}, \boldsymbol{\beta^{(i+1)}_{\gamma_l}}, \boldsymbol{y}\right) \equiv Bern\left(\frac{a_j}{a_j + b_j}\right)$, where $S_j^{(i)} = \left(\gamma_m^{(n)}\right)_{m,n}$, with $m \in \{0, 1, ..., j-1, j+1, ..., k\}$ and $n \in \{i, i+1\}$, whereas a_j and b_j are calculated according to (6.11) and (6.12).

   (b) Depending on whether $\gamma_j^{(i+1)} = 0$ or 1, set $D(j,j) = \tau_j$ or $D(j,j) = c_j \cdot \tau_j$ respectively.

   (c) (***Metropolis-within-Gibbs***)
   Perform 20 iterations of Gamerman's Independence Sampler, using as input $\boldsymbol{\beta^{(i+1)}_{\gamma_l}}$ and the current vector of $\boldsymbol{\gamma}$, containing the updated components, as well as those that have not been yet updated. Thus, we obtain a new vector $\boldsymbol{\beta^{(i+1)}_{\gamma_j}}$, which is used as the current value for $\boldsymbol{\beta}$.

   (d) Return to step a).

3. (***Metropolis-within-Gibbs***)
   Perform 50 iterations of Gamerman's Independence Sampler, using as input the last updated value of $\boldsymbol{\beta}$, which has been derived in step c), after the last component of $\boldsymbol{\gamma}$ was updated and as prior covariance matrix, the matrix D, as this is formed after all the components of $\boldsymbol{\gamma}$ have been updated. Set the 50-th value obtained by the algorithm as $\boldsymbol{\beta^{(i+1)}}$.

4. Return to step 2.

6.3 Stochastic Search Variable Selection for the Probit model

In the case of the Probit model, which was described in chapter 5, the response variable $\boldsymbol{y}$ has binary components, each one of which, informs us whether an event occurs ($y_i = 1$) or not ($y_i = 0$). Consequently, $y_i \sim Bern(p_i), \quad i = 1, 2, ..., n$ and as a result, the likelihood of the data is expressed as:

$$f(\boldsymbol{y}) = \prod p_i^{y_i} (1 - p_i)^{1 - y_i},$$
$$\text{whereas} \quad E[y_i] = p_i.$$

The vector of means, denoted by $E[\boldsymbol{y}] = \boldsymbol{p} = (p_1, p_2, ..., p_n)$ is linked to the linear predictor $\boldsymbol{\eta} = \boldsymbol{X\beta}$ through (5.1) :

$$\boldsymbol{p} = \Phi^{-1}(\boldsymbol{X\beta}).$$

Hence, the likelihood of the model with respect to $\boldsymbol{\beta}$ is expressed by (5.2) as follows:

$$L(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{X}) = \prod_{i=1}^{n} \left[\Phi^{-1}\left(X_i^T\boldsymbol{\beta}\right)\right]^{y_i} \left[1 - \Phi^{-1}\left(X_i^T\boldsymbol{\beta}\right)\right]^{1-y_i}.$$

Since the above form leads to the intractable form of $f(\boldsymbol{\beta}|\boldsymbol{y}, X)$ in section 5.1.2, it has been noted that the inference for the Probit model can be simplified with the method of **data augmentation**. We consider an additional n-dimensional vector of *latent variables*, denoted by $\boldsymbol{z} = (z_1, ..., z_n)$, which enables us to express the original model as a special case of a Normal linear model, according to relation (5.3), which leads to the structural form of (5.4):

$$\boldsymbol{z} \sim \boldsymbol{N_n(X\beta, 1)},$$

with the restriction that

$$\begin{cases} z_i > 0 & \text{if} \quad y_i = 1 \\ z_i \leq 0 & \text{if} \quad y_i = 0 \qquad i = 1, ...n \end{cases}$$

and we place again a proper prior on the coefficients vector

$$\boldsymbol{\beta} \sim \boldsymbol{N_{k+1}(b_0, C_0)}$$

We will perform Stochastic Search Variable Selection in the Probit model, according to the papers of Lee et.al. ("Gene Selection: a Bayesian variable selection approach",2003) and Chang, Chen and Chin ("Bayesian Variable Selection for Probit Models with Componentwise Gibbs", 2014). The latter paper describes two methods of SSVS, one of which, is the approach of the first paper. What is more, Lee's approach of SSVS using a Gibbs Sampler is given the title "SSVS-Lee" by Chang et.al.

6.3.1 Building a hierarchical model for SSVS for the Probit model

The SSVS approach for the Probit model followed in the present thesis, was originally employed by Lee et.al. in 2003 for the specific purpose of

Gene selection, as the title of the corresponding paper testifies, and it is based on the SSVS technique, as this was introduced by MacCullogh and George (1993). It is important to point out the fundamental elements of Lee's SSVS, which are the following:

- We use the technique of data augmentation in the same way as we did for the simple inference. This means that we consider the vector of auxiliary variables $\boldsymbol{z} = (z_1, ..., z_n)$ and for each component we assume that

$$\begin{cases} (z_i|\boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{\gamma}) \sim TN_0\big(X_i^T\boldsymbol{\beta}, 1\big), & \text{iff} \quad y_i = 0, \\ (z_i|\boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{\gamma}) \sim TN_1\big(X_i^T\boldsymbol{\beta}, 1\big), & \text{iff} \quad y_i = 1, \end{cases}$$

  where by $TN_0\big(X_i^T\boldsymbol{\beta}, 1\big)$, we refer to the Normal distribution $N\big(X_i^T\boldsymbol{\beta}, 1\big)$, with truncation above 0, whereas $TN_1\big(X_i^T\boldsymbol{\beta}, 1\big)$ stands for $N\big(X_i^T\boldsymbol{\beta}, 1\big)$, with truncation below 0.

- Based on the general scheme of SSVS, we consider the vector of latent variables $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, ..., \gamma_k)$, where $\gamma_j \in \{0, 1\}$ for every $j = 0, 1, ..., k$, the purpose of which, is to indicate whether the corresponding predictors are involved in the model or not. In the two previous applications we placed the following conditional prior on each $\beta_j, \quad j = 0, 1, ..., k$:

$$(\beta_j|\gamma_j) \sim (1 - \gamma_j)N(0, \tau_j^{-1}) + \gamma_j N(0, c_j\tau_j^{-1}).$$

  The quantities τ_j and c_j were adjusted appropriately, so that if $\gamma_j = 0$, then the resulting Gaussian $N(0, \tau_j^{-1})$ would be such, that we could "safely" assume that $\beta_j = 0$, whereas if $\gamma_j = 1$, based on $N(0, c_j\tau_j^{-1})$, we could set $\beta_j \neq 0$. Lee's approach is slightly bolder, since if we deduce at any part of the analysis that $\gamma_j = 0$, then we straightly set $\beta_j = 0$, which sustains an important structural differentiation. Furthermore, given the vector of $\boldsymbol{\gamma}$, Lee chooses to consider the reduced versions of both $\boldsymbol{\beta}$ and the matrix of explanatory variables $\boldsymbol{X}$, by omitting the components of $\boldsymbol{\beta}$ and the columns of $\boldsymbol{X}$, which correspond to zero indicators. Consequently, with each update of $\boldsymbol{\gamma}$, we are led to a different coefficient vector, denoted by $\boldsymbol{\beta}_\gamma$ and a different matrix $\boldsymbol{X}_\gamma$. Also, the prior placed on $\boldsymbol{\beta}$ with respect to $\boldsymbol{\gamma}$ is not a mixture of two Gaussians, but a multivariate Normal of the form $(\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma}) \sim N_l(0, \boldsymbol{c}^T R\boldsymbol{c})$, where $1 \leq l \leq k + 1$ is the dimension of $\boldsymbol{\beta}_\gamma$ and $\boldsymbol{c}_\gamma = (c_1, c_2, ..., c_l) \in \mathbb{R}^l_{>0}$ is a vector of fixed values. First, we specify the full vector $\boldsymbol{c}_{\boldsymbol{\gamma}=1} = (c_0, c_1, ..., c_k)$ and after every update of $\boldsymbol{\gamma}$, we consider that $\boldsymbol{c}_\gamma$ consists of all those components that correspond to a non-zero γ_i.

- The posterior distribution of $\boldsymbol{\gamma}$, that is $f(\boldsymbol{\gamma}|\boldsymbol{Z},\boldsymbol{y})$ is obtained via simulation and more specifically by a Gibbs Sampler. An innovative modification to the conditioning of the sampling distribution of $\boldsymbol{\gamma}$ is made, due to which, the required calculations are less intensive and the algorithm is performed more rapidly. Specifically, instead of using $f(\boldsymbol{\gamma}|\boldsymbol{\beta},\boldsymbol{z},\boldsymbol{y})$, $\boldsymbol{\beta}$ is integrated out and so, $\boldsymbol{\gamma}$ is sampled from $f(\boldsymbol{\gamma}|\boldsymbol{z},\boldsymbol{y}) \equiv f(\boldsymbol{\gamma}|\boldsymbol{z})$.

Last but not least, several remarks need to be made before proceeding to the algorithmic scheme of the Gibbs Sampler. When Chang, Chen and Chi describe SSVS-Lee, they use a componentwise Gibbs Sampler for both $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, whereas in the paper of Lee et.al., only $\boldsymbol{\gamma}$ is sampled componentwise.

Regarding the prior placed on $(\boldsymbol{\beta}|\boldsymbol{\gamma})$, both Lee et.al. and Chang et.al. propose to set $R = (X^TX)^{-1}$ and $c_0 = c_1 = ... = c_k = c > 0$. As a result, we consider that $(\boldsymbol{\beta}|\boldsymbol{\gamma}) \sim N_l(0, c(X^TX)^{-1})$. What is more, Lee et.al. state that c has to vary between 10 and 100 according to the research of Kohn and Smith (1996) and they set $c = 100$, thus making the prior of $\boldsymbol{\beta}_\gamma$ given $\boldsymbol{\gamma}$ considerably less informative than the likelihood of the data.

Before presenting the steps of the algorithm, we present the hierarchical model we will be studying, whereas the computational scheme is outlined in the following section. The augmented model we are using is:

$$\begin{cases} (z_i|\boldsymbol{\beta}_\gamma,\boldsymbol{\gamma},\boldsymbol{y}) \sim TN_0\big(X_i\boldsymbol{\beta}_\gamma,1,\big), & \text{iff} \quad y_i = 0, \\ (z_i|\boldsymbol{\beta}_\gamma,\boldsymbol{\gamma},\boldsymbol{y}) \sim TN_1\big(X_i\boldsymbol{\beta}_\gamma,1,\big), & \text{iff} \quad y_i = 1, \end{cases}$$
$$(\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma}) \sim N_l\big(0, c(X_\gamma^TX_\gamma)^{-1}\big),$$
$$\gamma_i \sim Bern(p_i), \quad p_i \in [0,1], \quad i = 0,1,...k. \tag{6.13}$$

6.3.2 The Gibbs Sampler for the SSVS-Lee

Now, we may outline the algorithmic scheme of the Gibbs Sampler. We iterate the following steps until convergence is achieved.

1. We draw $\boldsymbol{\gamma}$ from the distribution $f(\boldsymbol{\gamma}|\boldsymbol{z})$, which is obtained from Bayes' Theorem:
$$f(\boldsymbol{\gamma}|\boldsymbol{z}) \propto f(\boldsymbol{z}|\boldsymbol{\gamma}) \cdot f(\boldsymbol{\gamma}). \tag{6.14}$$

The distribution $f(\boldsymbol{z}|\boldsymbol{\gamma})$ can be calculated up to a normalizing constant by integrating out from $f(\boldsymbol{z}|\boldsymbol{\beta}_\gamma, \boldsymbol{\gamma})$ the variable $\boldsymbol{\beta}_\gamma$):

$$\begin{aligned} f(\boldsymbol{z}|\boldsymbol{\gamma}) &\propto \int \cdots \int f(\boldsymbol{z}|\boldsymbol{\beta}_\gamma, \boldsymbol{\gamma}) \cdot f(\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma}) d\boldsymbol{\beta}_\gamma \\ &\propto \int \cdots \int \exp\Big\{ -\frac{1}{2}\Big[(\boldsymbol{z} - X_{i\gamma}^T\boldsymbol{\beta}_\gamma)^T \cdot (\boldsymbol{z} - X_{i\gamma}^T\boldsymbol{\beta}_\gamma) + \boldsymbol{\beta}_\gamma^T c^{-1}(X_\gamma^T X_\gamma)\boldsymbol{\beta}_\gamma\Big]\Big\} d\boldsymbol{\beta}_\gamma \\ &\propto \exp\Big\{ -\frac{1}{2}\Big[(\boldsymbol{z}^T\boldsymbol{z} - \frac{c}{1+c}\boldsymbol{z}^T X_\gamma)(X_\gamma^T X_\gamma)^{-1} X_\gamma^T \boldsymbol{z}\Big]\Big\}. \end{aligned} \tag{6.15}$$

Now, we can calculate (6.14) up to a normalising constant and derive that:

$$\begin{aligned} &f(\boldsymbol{\gamma}|\boldsymbol{z}) \propto f(\boldsymbol{z}|\boldsymbol{\gamma}) \cdot f(\boldsymbol{\gamma}) \\ &\propto \exp\Big\{ -\frac{1}{2}\Big[(\boldsymbol{z}^T\boldsymbol{z} - \frac{c}{1+c}\boldsymbol{z}^T X_\gamma)(X_\gamma^T X_\gamma)^{-1} X_\gamma^T \boldsymbol{z}\Big]\Big\} \cdot \prod p_i^{\gamma_i}(1-p_i)^{1-\gamma_i}. \end{aligned} \tag{6.16}$$

Since we sample $\boldsymbol{\gamma}$ componentwise and consecutively, we will be using the probabilities

$$\begin{aligned} &P\Big[\gamma_i = 1|\gamma_{(-i)}, \boldsymbol{z}\Big] \propto P\Big[\boldsymbol{z}|\gamma_{(-i)}, \gamma_i = 1\Big] \cdot P\Big[\gamma_i = 1\Big] \\ &\propto \exp\Big\{ -\frac{1}{2}\Big[(\boldsymbol{z}^T\boldsymbol{z} - \frac{c}{1+c}\boldsymbol{z}^T X_\gamma)(X_\gamma^T X_\gamma)^{-1} X_\gamma^T \boldsymbol{z}\Big]\Big\} \cdot p_i \end{aligned} \tag{6.17}$$

and

$$\begin{aligned} &P\Big[\gamma_i = 0|\gamma_{(-i)}, \boldsymbol{z}\Big] \propto P\Big[\boldsymbol{z}|\gamma_{(-i)}, \gamma_i = 0\Big] \cdot P\Big[\gamma_i = 0\Big] \\ &\propto \exp\Big\{ -\frac{1}{2}\Big[(\boldsymbol{z}^T\boldsymbol{z} - \frac{c}{1+c}\boldsymbol{z}^T X_\gamma)(X_\gamma^T X_\gamma)^{-1} X_\gamma^T \boldsymbol{z}\Big]\Big\} \cdot (1-p_i), \end{aligned} \tag{6.18}$$

where $\gamma_{(-i)} = (\gamma_0, ..., \gamma_{i-1}, \gamma_{i+1}, ..., \gamma_k) \quad i = 0, 1, ..., k.$
Since γ_i is a binary variable, we can deduct that $\big(\gamma_i|\gamma_{(-i)}, \boldsymbol{z}\big) \sim Bern\Big(\frac{a}{a+b}\Big)$, where

$$a = \exp\Big\{ -\frac{1}{2}\Big[(\boldsymbol{z}^T\boldsymbol{z} - \frac{c}{1+c}\boldsymbol{z}^T X_\gamma)(X_\gamma^T X_\gamma)^{-1} X_\gamma^T \boldsymbol{z}\Big]\Big\} \cdot p_i \tag{6.19}$$

and

$$b = \exp\Big\{ -\frac{1}{2}\Big[(\boldsymbol{z}^T\boldsymbol{z} - \frac{c}{1+c}\boldsymbol{z}^T X_\gamma)(X_\gamma^T X_\gamma)^{-1} X_\gamma^T \boldsymbol{z}\Big]\Big\} \cdot (1-p_i). \tag{6.20}$$

2. We sample $\boldsymbol{\beta}_\gamma$ from the conditional distribution $f(\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma}, \boldsymbol{z})$, which can be derived by Bayes' Theorem as follows:

$$\begin{aligned} f(\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma}, \boldsymbol{z}) &\propto f(\boldsymbol{z}|\boldsymbol{\beta}_\gamma, \boldsymbol{\gamma}) \cdot f(\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma}) \\ &\propto \exp\Big\{ -\frac{1}{2}\Big[(\boldsymbol{z} - X_{i\gamma}^T\boldsymbol{\beta}_\gamma)^T \cdot (\boldsymbol{z} - X_{i\gamma}^T\boldsymbol{\beta}_\gamma) + \boldsymbol{\beta}_\gamma^T c^{-1}(X_\gamma^T X_\gamma)\boldsymbol{\beta}_\gamma\Big]\Big\} \\ &\propto \exp\Big\{ -\frac{1}{2}\Big[(\boldsymbol{\beta}_\gamma - \mu_\gamma)^T V_\gamma^{-1}(\boldsymbol{\beta}_\gamma - \mu_\gamma)\Big]\Big\}. \end{aligned} \tag{6.21}$$

   Consequently, $(\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma}, \boldsymbol{z}) \sim N_l\big(\mu_\gamma, V_\gamma\big)$, where $V_\gamma = \dfrac{c}{1+c}\boldsymbol{z}^T X_\gamma)(X_\gamma^T X_\gamma)^{-1}$ and $\mu_\gamma = V_\gamma X_\gamma^T \boldsymbol{z}_\gamma$.

3. Finally, we update the vector of latent variables $\boldsymbol{z}$, by simulating each component z_i as dictated by the hierarchical structure of the model; that is, for each $i = 1, ..., n$ we have that:

$$\begin{cases} (z_i|\boldsymbol{\beta}_\gamma, \boldsymbol{\gamma}, \boldsymbol{y}) \sim TN_0\big(X_i\boldsymbol{\beta}_\gamma, 1, \big), & \text{iff} \quad y_i = 0, \\ (z_i|\boldsymbol{\beta}_\gamma, \boldsymbol{\gamma}, \boldsymbol{y}) \sim TN_1\big(X_i\boldsymbol{\beta}_\gamma, 1, \big), & \text{iff} \quad y_i = 1. \end{cases}$$

   We can draw values from the truncated Gaussians by using the method of Inversion Sampling, described in section 5.3.2.

By monitoring the convergence of the formed Markov chain regarding the vector of indicators $\boldsymbol{\gamma}$, we can estimate the burn-in period and consider the values that are drawn afterwards as a sample from $f(\boldsymbol{\gamma}|\boldsymbol{y})$.

6.4 Proposing the optimal submodel using the Median Probability Model (MPM)

Having obtained the posterior distribution of the latent variable vector $\boldsymbol{\gamma}$, denoted by $f(\boldsymbol{\gamma}|\boldsymbol{y})$, in the Normal Linear, the Logit and the Probit model, the question of which could be the submodel that best fits the data, remains still open. A further study of $f(\boldsymbol{\gamma}|\boldsymbol{y})$ is required in order to be able to extract that answer. The way of study proposed in the present thesis is a widely used method of Bayesian model selection, known as ***The Median Probability Model (MPM)***, which was originally introduced in 2004 by Barbieri and Berger in the paper *Optimal Predictive Model Selection* .

The essence of the Median Probability Model can be easily comprehended by the following two definitions. We should note that Barbieri and Berger (2004) provided these definitions restricted to the case of the normal linear model. However, we assume that they can be adequately extended for the cases of the Logit and Probit models without loss of generosity. Hence, they are given in the following forms in the present thesis.

Definition 6.4.1 (Posterior Inclusion Probability)**.** Let $\mathcal{M}$ denote an arbitrary model, with a response variable denoted by $\boldsymbol{y} \in \mathbb{R}^n$ and an $n \times (k+1)$ matrix of explanatory variables, denoted by $\boldsymbol{X} = \big(\boldsymbol{1}, \boldsymbol{X_1}, ..., \boldsymbol{X_k}\big)$, where $X_i \in \mathbb{R}^n$ for every $i = 1, ..., k$. Additionally, by $\boldsymbol{\beta} = \big(\beta_0, \beta_1 ...., \beta_k\big) \in \mathbb{R}^{k+1}$ we will refer to the vector of unknown predictors, which link the response variable, $\boldsymbol{y}$, to the design matrix, $\boldsymbol{X}$, through the following expression:

$$E\big[\boldsymbol{y}|\boldsymbol{\beta}\big] = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N_n\big(0, I_n\big),$$

where $\boldsymbol{\epsilon}$ stands for the vector or random errors. We consider $\boldsymbol{\gamma} = \big(\gamma_0, \gamma_1, ..., \gamma_k\big) \in \{0,1\}^{k+1}$ to be a vector of indicators. We specify $\mathcal{M}_{\tilde{\boldsymbol{\gamma}}}$, as the submodel built according to the following expression:

$$\mathcal{M}_{\tilde{\gamma}} : E\big[\boldsymbol{y}_{\tilde{\boldsymbol{\gamma}}}|\boldsymbol{\beta}_{\tilde{\boldsymbol{\gamma}}}\big] = \boldsymbol{X}_{\tilde{\boldsymbol{\gamma}}}\boldsymbol{\beta}_{\tilde{\boldsymbol{\gamma}}} + \boldsymbol{\epsilon}_{\tilde{\boldsymbol{\gamma}}},$$

where $\tilde{\boldsymbol{\gamma}} = \big(\tilde{\gamma}_0, \tilde{\gamma}_1 ...., \tilde{\gamma}_k\big) \in \{0,1\}^{k+1}$ and each coordinate $\tilde{\gamma}_j, j = 0, 1, ..., k$ is either 1 or 0 if and only if the explanatory variable $\boldsymbol{X_j}$ is included in the model or not, respectively. It is clarified that $\boldsymbol{X}_{\tilde{\boldsymbol{\gamma}}}$ and $\boldsymbol{\beta}_{\tilde{\boldsymbol{\gamma}}}$ are respectively the design matrix and the coefficient vector corresponding to the non-zero coordinates of $\tilde{\boldsymbol{\gamma}}$.

The *posterior inclusion probability* for variable j is

$$\hat{P}_j := \sum_{\tilde{\boldsymbol{\gamma}} \in \{0,1\}^{k+1} : \tilde{\gamma}_j = 1} P\Big(\mathcal{M}_{\tilde{\boldsymbol{\gamma}}}|\boldsymbol{y}\Big)$$

Definition 6.4.2. The *Median Probability Model*, denoted by $\mathcal{M}_{\boldsymbol{\gamma}^*}$, is defined as the model consisting of those variables whose posterior inclusion probability is <u>at least</u> $1/2$. Let $\boldsymbol{\gamma}^*$ be the corresponding indicator vector, whose non-zero coordinates correspond to the explanatory variables included in the submodel. Then

$$\gamma_j^* = \begin{cases} 1 & \text{if} \quad \hat{P}_j \geq 1/2, \\ 0 & \text{otherwise} \end{cases} \qquad j = 0, 1, ..., k$$

Based on the rather obvious remark that the vector $\boldsymbol{\gamma} = \big(1, 1, ..., 1\big)$ corresponds to the *full model* and then by contemplating on the notion

that there is a "one-to-one correspondence" between an arbitrary submodel, denoted by $\mathcal{M}_{\tilde{\gamma}}$ and the respective vector of indicators, $\tilde{\gamma}$, we can safely assume that the posterior inclusion probability for variable j can be given by:

$$\hat{P}_j \equiv \sum_{\tilde{\gamma}\in\{0,1\}^{k+1}:\tilde{\gamma}_j=1} P\Big(\tilde{\gamma}|\boldsymbol{y}\Big). \tag{6.22}$$

The immediate result that we get is the fact that the Median Probability Model will include the explanatory variables $\boldsymbol{X_j}$ $j = 0, 1, ..., k$,that satisfy the following condition:

$$\boldsymbol{X_j} : \sum_{\tilde{\gamma}\in\{0,1\}^{k+1}:\tilde{\gamma}_j=1} P\Big(\tilde{\gamma}|\boldsymbol{y}\Big) \geq 1/2 \tag{6.23}$$

Consequently, after we have obtained a sample from $f(\boldsymbol{\gamma}|\boldsymbol{y})$, we calculate $\hat{P}_j$ for every $j \in \{0, 1, ..., k\}$ and we check for which js the above condition is satisfied. Then, the optimal choice for a parsimonious submodel is the Median Probability Model, to which we are led to by the process described above.

Chapter 7

Applications to a Simulated Data Example

In this chapter we apply the methods of inference and stochastic variable selection, described in chapters 4 and 6, to simulated data.

7.1 Conducting Inference

We have simulated data as follows. The matrix of explanatory variables is the following 7×6 matrix containing zeros and units:

$$\mathbf{x} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

Each coordinate of the vector of the response variable, $\boldsymbol{y}$, was generated according to a Binomial distribution; that is $y_i \sim Bin(n_i, p_i), i = 1, 2, ..., 7$, with the vector of the trials, n, being specified as $n = (40, 2, 58, 18, 9, 26, 98)$. The corresponding possibility of success for each y_i was calculated by the formula: $p_i = \dfrac{\exp(x_i^T \boldsymbol{\beta})}{1 + \exp(x_i^T \boldsymbol{\beta})}, i = 1, 2, ..., 7$, where the vector of predictors, $\boldsymbol{\beta}$, was specified as: $\boldsymbol{\beta} = (1, 0.2, 0, 0, 0, 0.3)$. Consequently, our model has the following structure: $E[y_i|\boldsymbol{\beta}] = \boldsymbol{x_{i0}} + 0.2\boldsymbol{x_{i2}} + 0.3\boldsymbol{x_{i6}}, \quad i = 1, ...7$.

Firstly, we conducted simple inference on the model, employing Gamerman's Independence Sampler in order to simulate the conditional posterior distribution of $\boldsymbol{\beta}$, that is $f(\boldsymbol{\beta}|\boldsymbol{y})$. We placed the following prior

distribution on $\boldsymbol{\beta}$: $\boldsymbol{\beta} \sim N_6(0, P)$, where P is a 6×6 diagonal matrix, with all its non-zero elements equal to 100. Thus, we assume prior ignorance. We set the algorithm to perform 10000 iterations and afterwards, we obtained the ergodic means plot for each predictor, in order to determine the burn-in period. All 6 ergodic means plots are presented below.

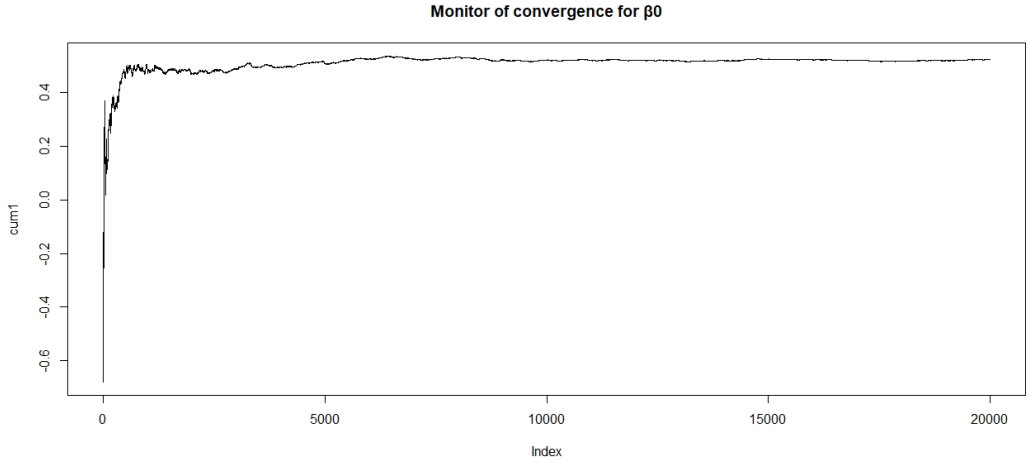

Figure 7.1: Monitoring the convergence of Gamerman's chain for β_0.

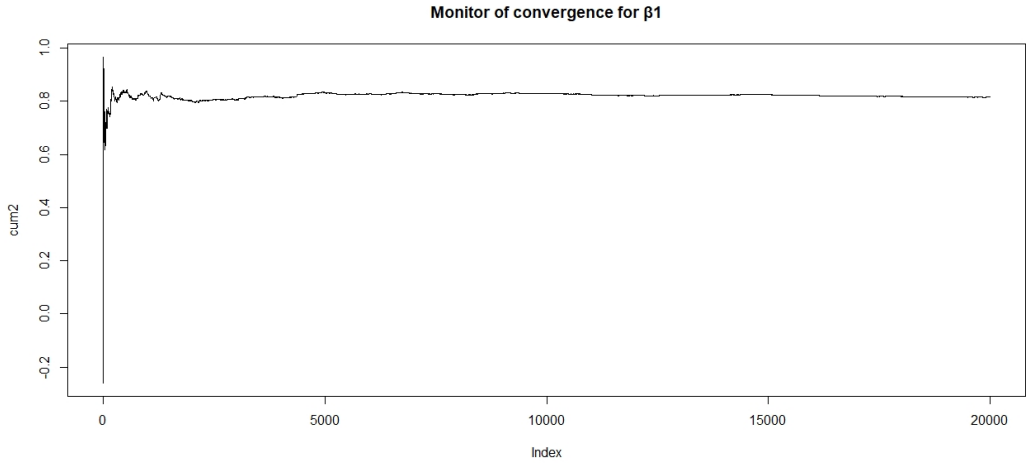

Figure 7.2: Monitoring the convergence of Gamerman's chain for β_1.

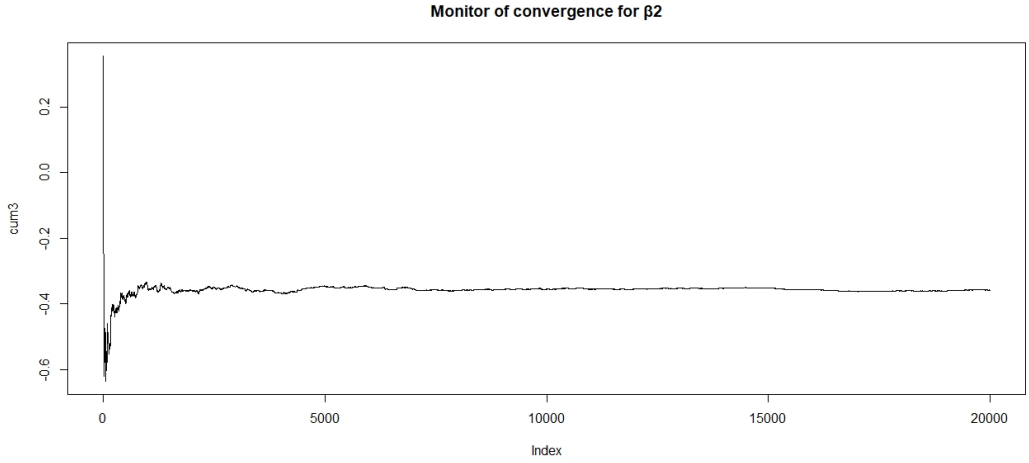

Figure 7.3: Monitoring the convergence of Gamerman's chain for β_2.

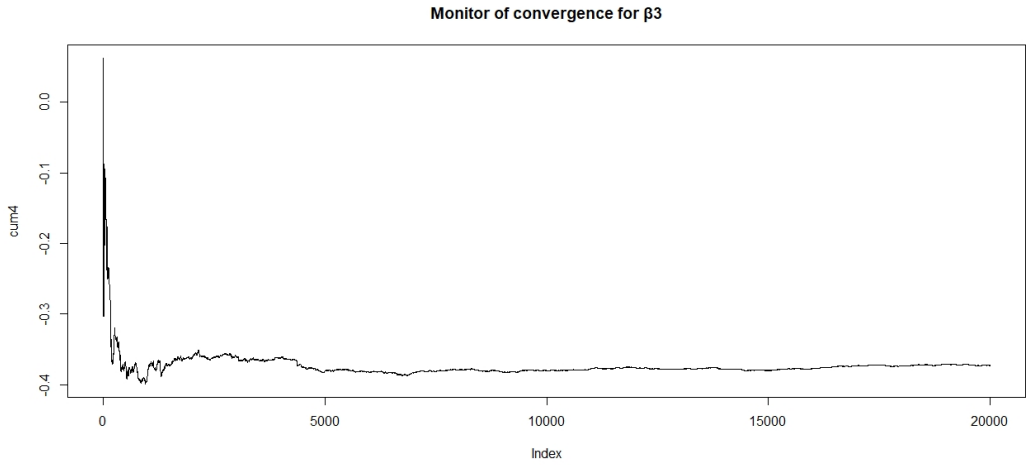

Figure 7.4: Monitoring the convergence of Gamerman's chain for β_3.

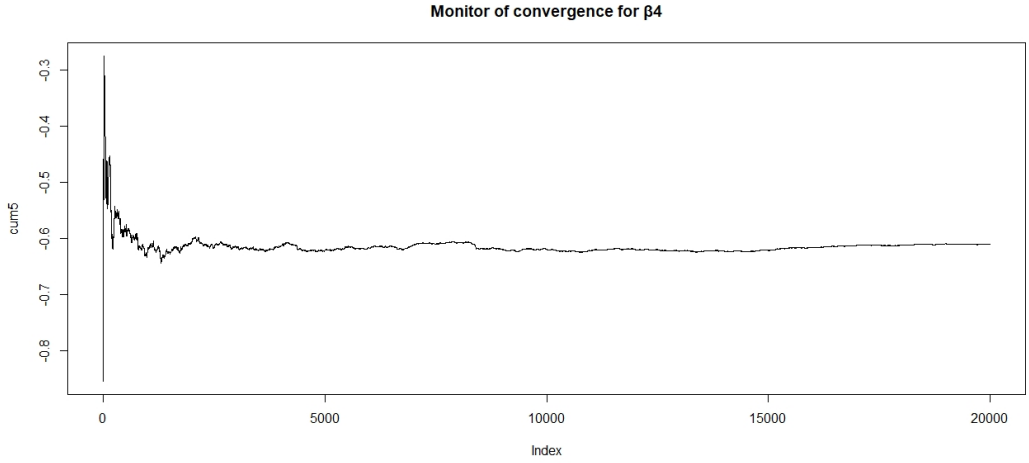

Figure 7.5: Monitoring the convergence of Gamerman's chain for β_4.

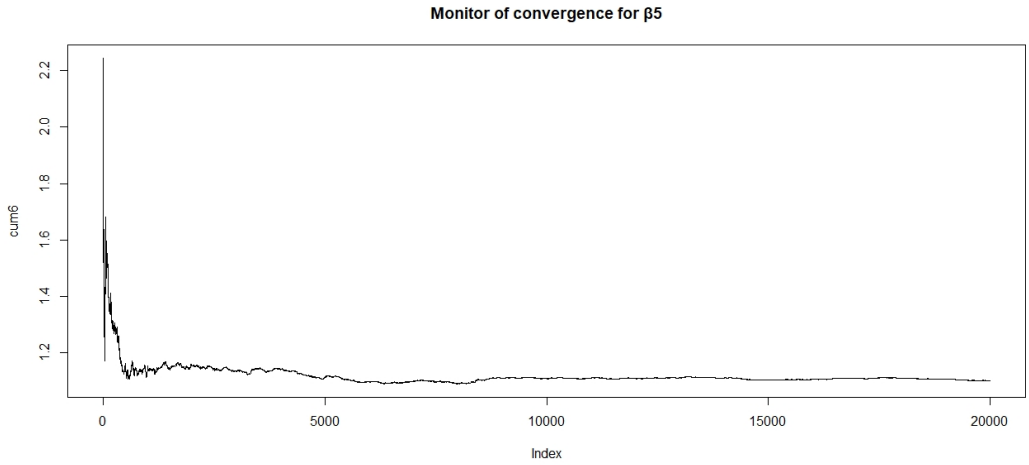

Figure 7.6: Monitoring the convergence of Gamerman's chain for β_5.

We discard the first 16000 iterations, after which, we consider that the formed chain has reached convergence. Thus, we end up with a sample of 4000 draws from $f(\boldsymbol{\beta}|\boldsymbol{y})$. This sample is individually studied for each predictor, by obtaining a histogram of the posterior densities and a table of descriptive statistics, which include the location parameters of the mean and the median and also the spread parameter of the standard deviation for every predictor. The histograms are given in Figures 7.7-7.11.

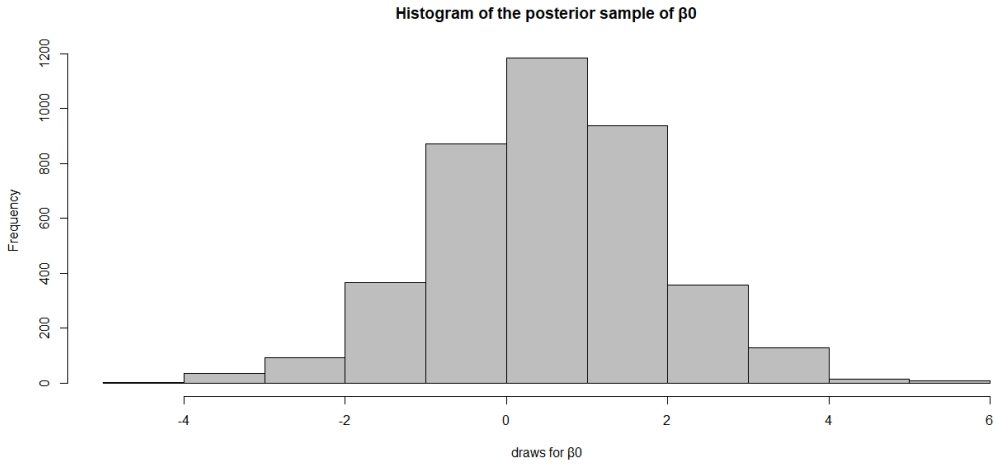

Figure 7.7: Histogram of densities for the posterior sample of β_0.

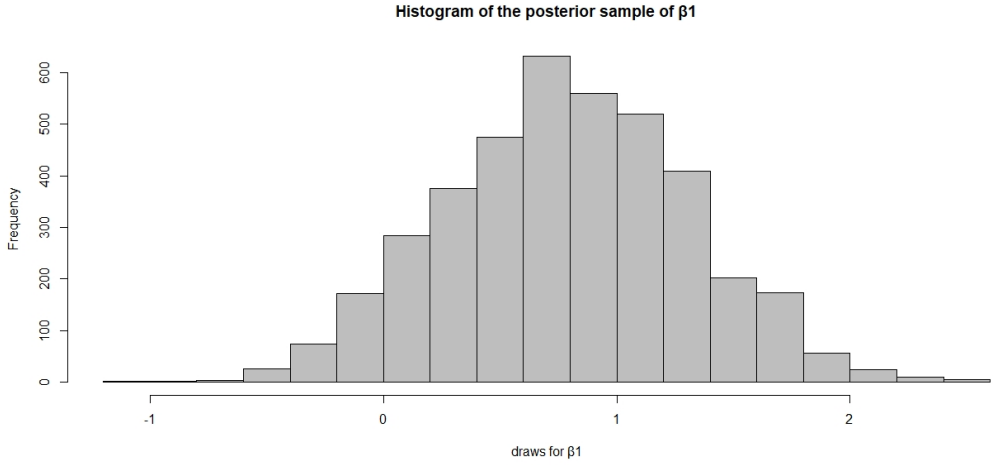

Figure 7.8: Histogram of densities for the posterior sample of β_1.

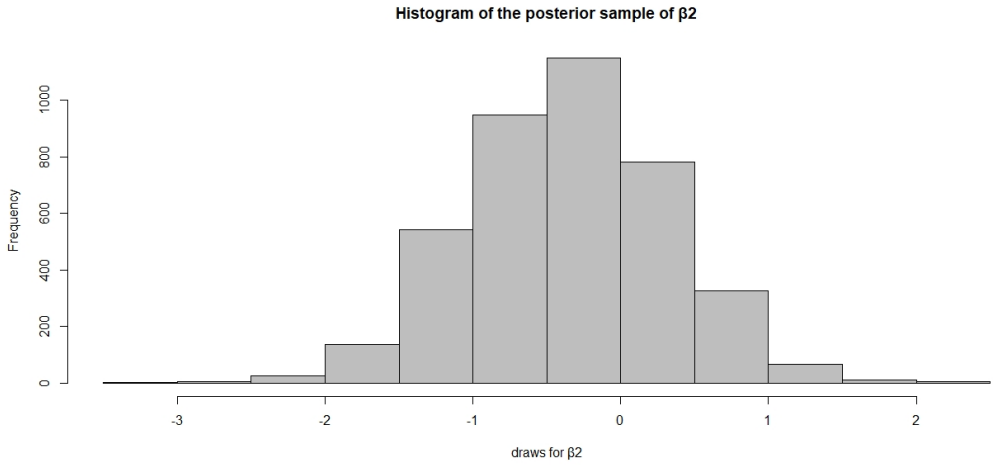

Figure 7.9: Histogram of densities for the posterior sample of β_2.

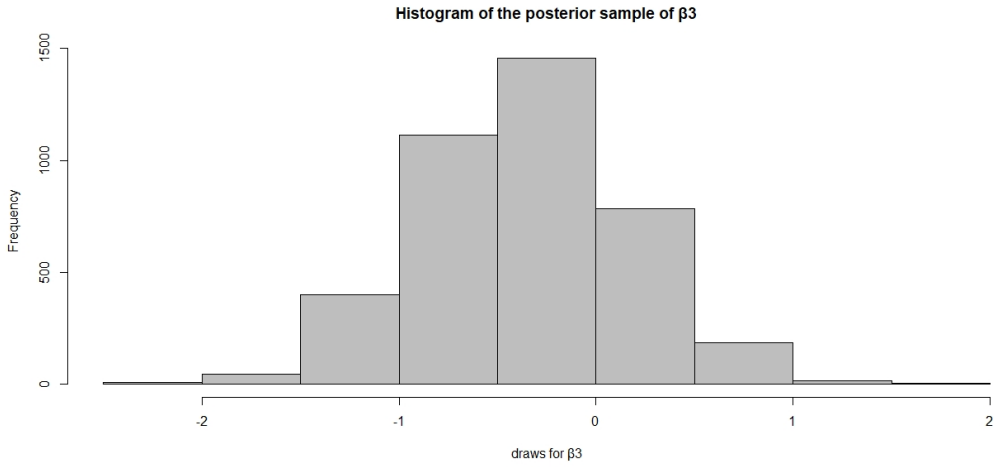

Figure 7.10: Histogram of densities for the posterior sample of β_3.

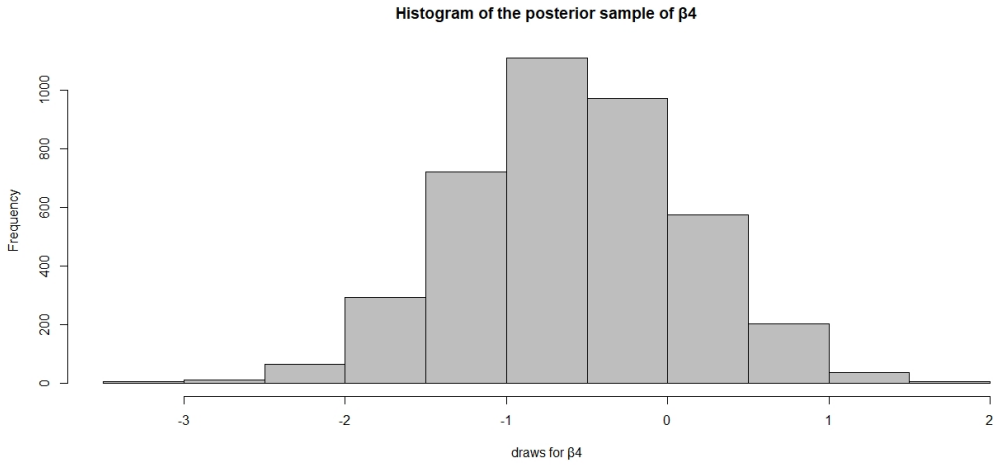

Figure 7.11: Histogram of densities for the posterior sample of β_4.

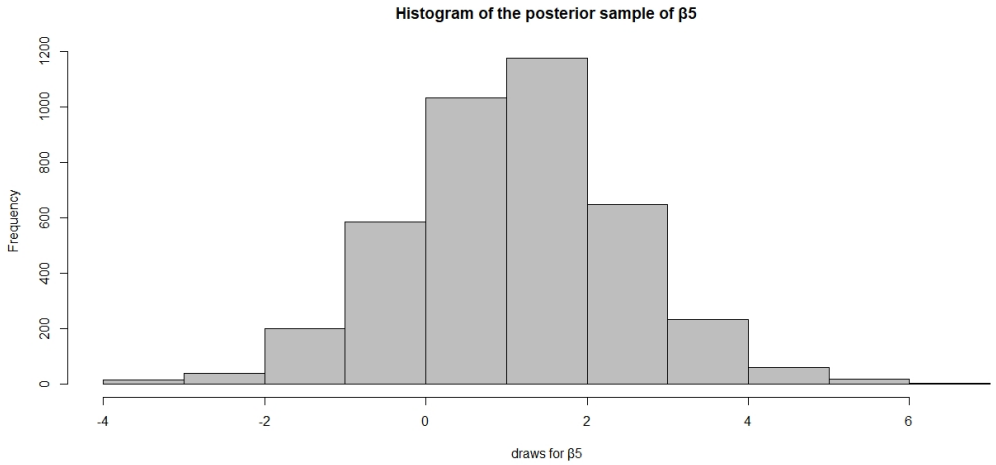

Figure 7.12: Histogram of densities for the posterior sample of β_5.

In Table 7.1, we give the the location and spread parameters mentioned above, as well as the real values of the predictors.

Predictor	Posterior Mean	Posterior Standard Deviation	Posterior Median	Real Value
β_0	0.46	1.46	0.38	1
β_1	0.93	0.6	0.88	0.2
β_2	-0.29	0.81	-0.3	0
β_3	-0.46	0.62	-0.43	0
β_4	-0.74	0.81	-0.71	0
β_5	1.21	1.54	1.34	0.3

Table 7.1: Table of basic location and spread parameters

7.2 Performing SSVS for the proposed Logit model

In this particular example, the search of a more parsimonious model would require $2^6 = 64$ model comparisons. We performed SSVS as it was described in chapter 6, using the following prior distributions:

$$\gamma_j \sim Bern(1/2), \quad j = 1, 2, ..., 6$$
$$\boldsymbol{\beta}|\boldsymbol{\gamma} \sim N_6\big(0, D\big), \quad \text{where} \quad D = diag(10, ..., 10) \in \mathbb{R}^6$$

The initial values used as input for the algorithm are $\gamma^{(0)} = (1, 1, 1, 1, 1, 1)$, since we begin from the full model and $\beta^{(0)} = \big(x^t x\big)^{-1} x^t y$. We performed a total number of 1000 iterations. Since the algorithmic scheme we structured is a computationally intensive, as we will show later, the time required for the statistical software of R to perform the total number of iterations was approximately 40 minutes. However, we may note that the required time is still much less than the time that 64 model comparisons, carried out with the most popular methods of Model, would take. A brief summary of the steps performed at each iteration is the following:

1. A the i-th iteration of the algorithm, we updated γ consecutively and componentwise, by first calculating the probability of each γ_j getting the value 1 and then, by simulating it from the corresponding Bernoulli distribution. In order to do that, we set $\gamma_j^{(i)} = 0$, without changing the values of $\gamma^{(i)}_{(Sj)^{(i)}}, j = 1, 2, ..., 6$. We set $D[j, j] = 0.3$ and with the current version of matrix D, we perform 20 iterations of Gamerman's Independence Sampler, in order to get a draw from $f(\boldsymbol{\beta}|\gamma_j^{(i)} = 0, \gamma^{(i)}_{(Sj)^{(i)}}, y)$. Then, we calculate the amount a_j, based on (6.11). Next, by setting $\gamma_j^{(i)} = 1$ and $D[j, j] = 10$, we perform another 20 iterations of Gamerman's Independence Sampler, in order to get a draw from $f(\boldsymbol{\beta}|\gamma_j^{(i)} = 1, \gamma^{(i)}_{(Sj)^{(i)}}, y)$ and then,

we compute bi, as this is defined in (6.12) . We update component γ_j, by considering that $\gamma_j^{(i)} \sim Bern\big(a/(a+bi)\big)$. We repeat the procedure described above for all the components of γ and we store $\gamma^{(i)}$.

2. Having updated γ, we perform 50 iterations of Gamerman's Independence Sampler, using as prior covariance matrix the current version of D, which reflects which are the non-zero components of the indicator $\gamma^{(i)}$ and as prior mean, the vector $\beta^{(i-1)}$, which we have also stored from the previous iteration.
3. Using the current value $\beta^{(i)}$, we return to step 1.

We monitor the convergence of the chain, by examining the ergodic means plot for each component $\gamma_j, j = 1, 2, ..., 6$. Based on Figures 7.13-7.18 given below, we can presume that convergence is achieved after the first 800 iterations, which are specified as the burn-in period.

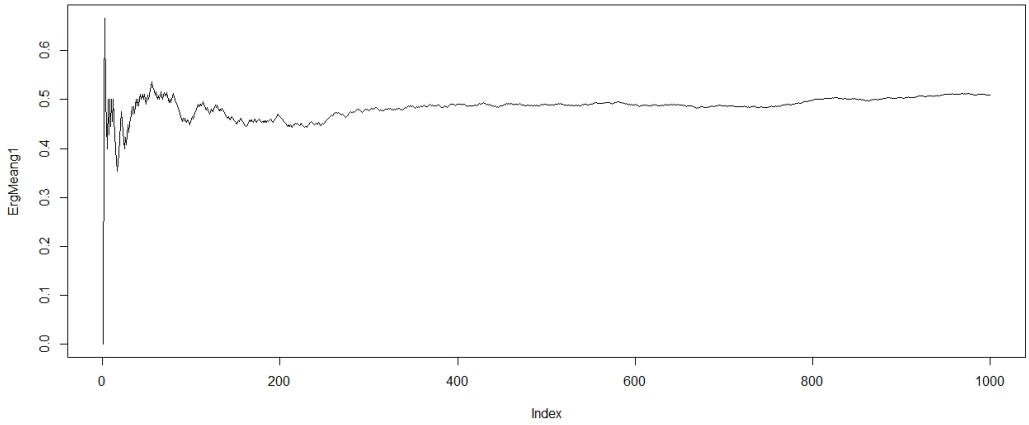

Figure 7.13: Monitoring convergence of the SSVS algorithm for γ_1.

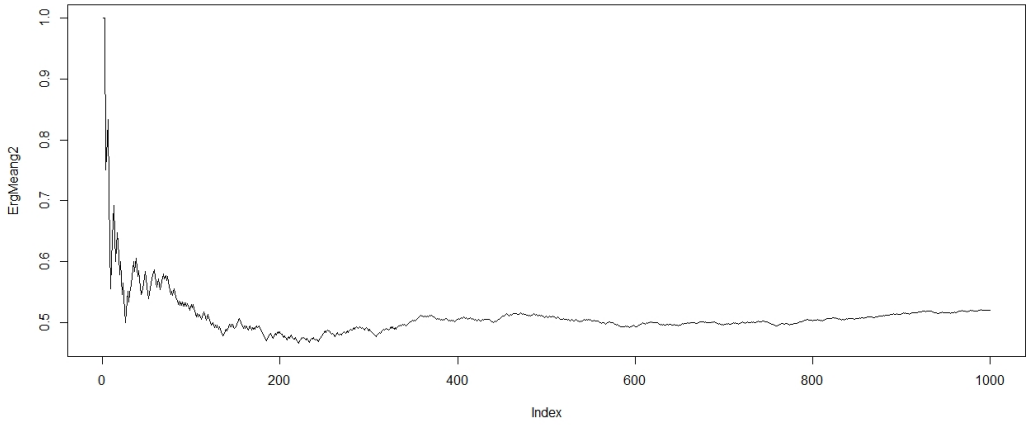

Figure 7.14: Monitoring convergence of the SSVS algorithm for γ_2.

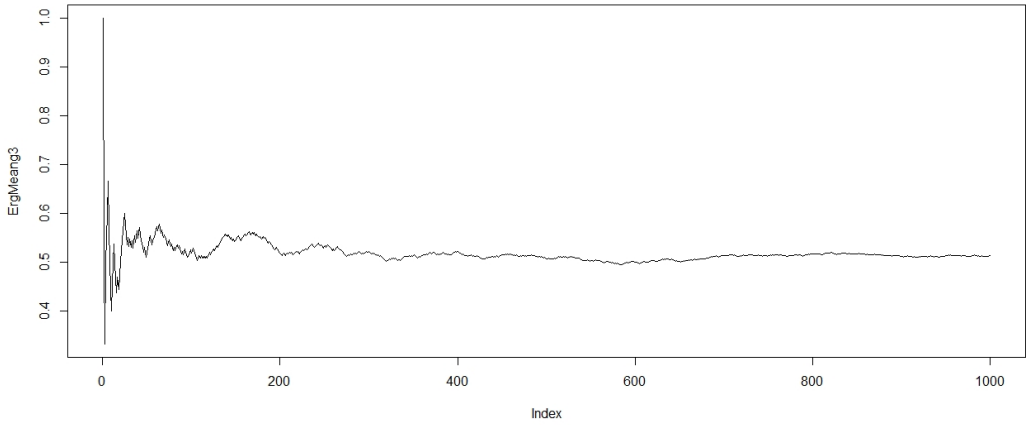

Figure 7.15: Monitoring convergence of the SSVS algorithm for γ_3.

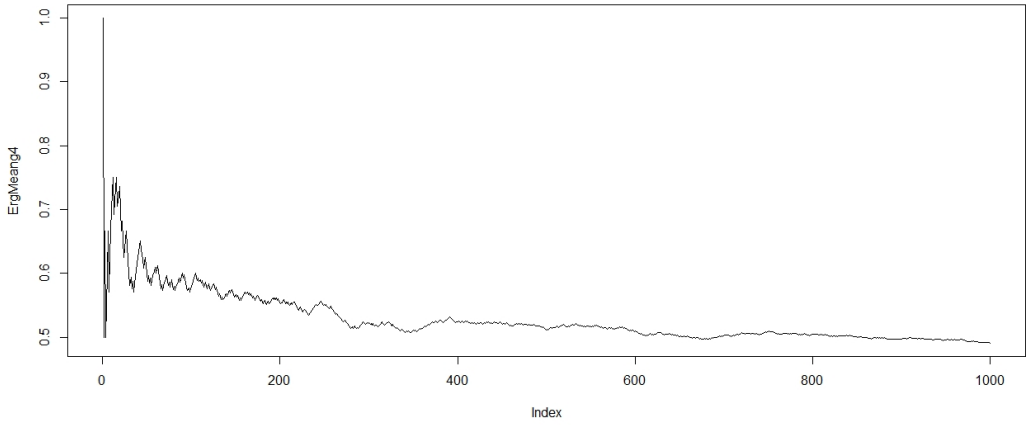

Figure 7.16: Monitoring convergence of the SSVS algorithm for γ_4.

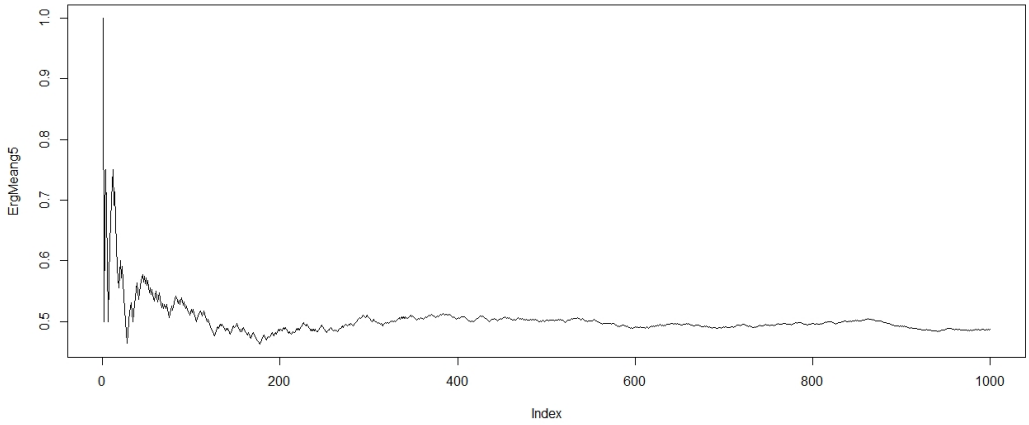

Figure 7.17: Monitoring convergence of the SSVS algorithm for γ_5.

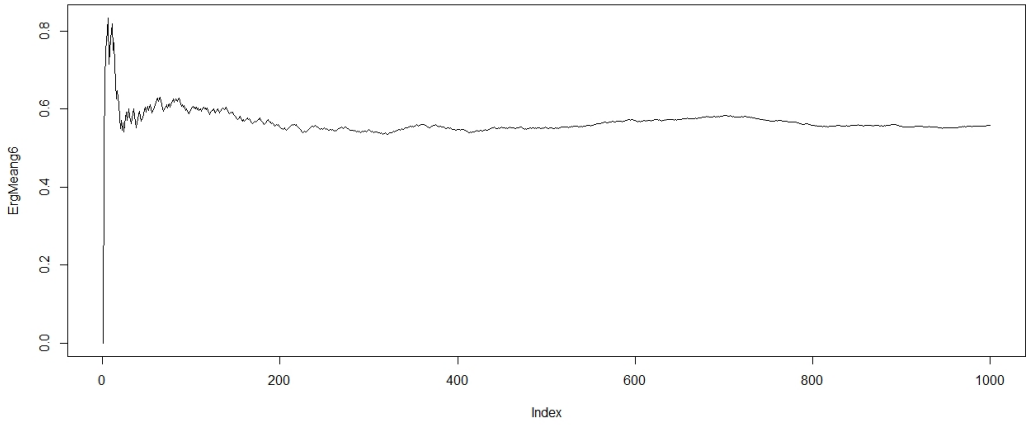

Figure 7.18: Monitoring convergence of the SSVS algorithm for γ_6.

Now, we can specify the Median Probability Model, based on the inclusion probabilities of each variable, which are presented on Table 7.2.

Intercept	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$	$\boldsymbol{X_3}$	$\boldsymbol{X_4}$	$\boldsymbol{X_5}$
0.56	0.57	0.53	0.43	0.44	0.58

Table 7.2: Table of the inclusion probabilities

We arrive at the conclusion that the Median Probability Model is:

$$E[y_i|\boldsymbol{\beta}] = \boldsymbol{\beta_0 x_{0i}} + \boldsymbol{\beta_1 x_{2i}} + \boldsymbol{\beta_2 x_{3i}} + \boldsymbol{\beta_5 x_{6i}}, \quad i = 1, ...7. \tag{7.1}$$

This means that we have excluded variables $\boldsymbol{x_4}$, $\boldsymbol{x_5}$.

7.3 Inference for the Median Probability Model

We conducted inference for the Median Probability model, once again employing Gamerman's Independence Sampler, with 10000 iteration in total. We concluded that convergence was reached after 16000 iterations an thus, the sample from $f(\boldsymbol{\beta}|\boldsymbol{y})$ contained 4000 draws. However, we must point out that now, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_5)$ and the design matrix we used was $\boldsymbol{x}$, after we have erased columns 4 and 5, which corresponded to predictors β_3 and β_4.

We present the histograms of frequencies for each predictor and the corresponding table of location and spread parameters.

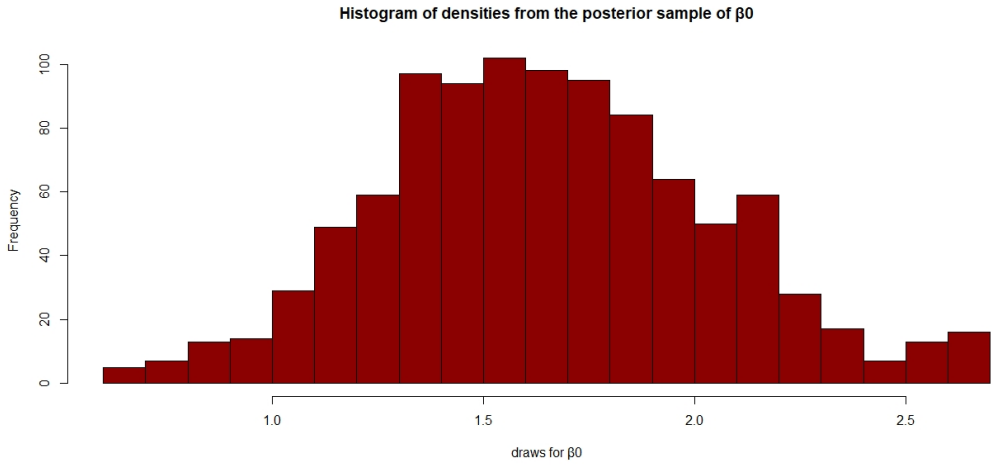

Figure 7.19: Histogram of densities for the posterior sample of β_0 in the Median Probability Model.

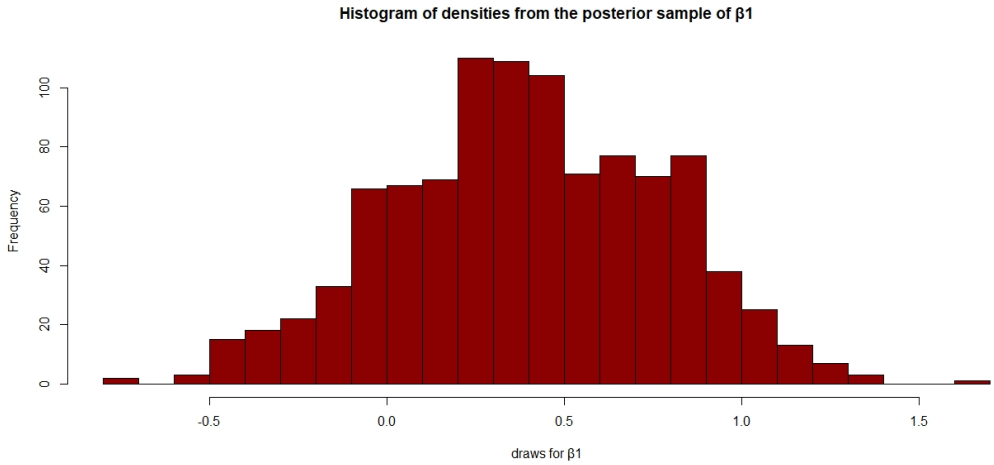

Figure 7.20: Histogram of densities for the posterior sample of β_1 in the Median Probability Model.

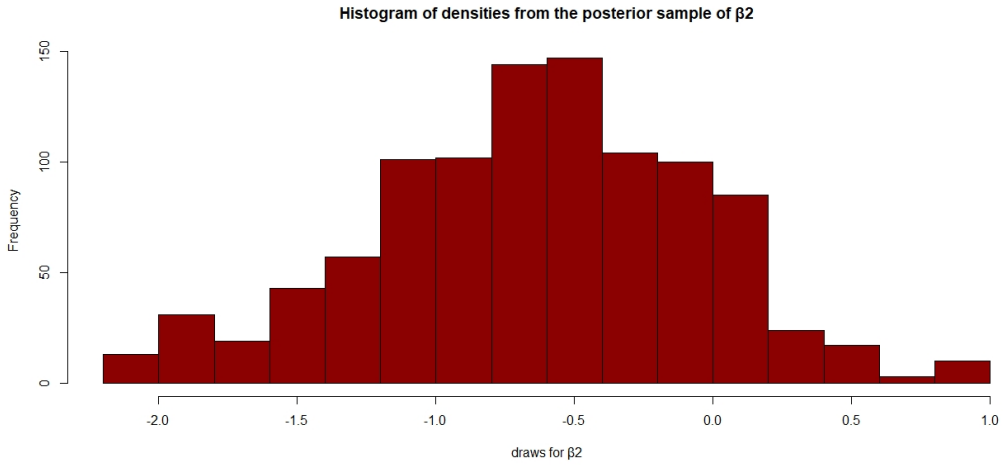

Figure 7.21: Histogram of densities for the posterior sample of β_2 in the Median Probability Model.

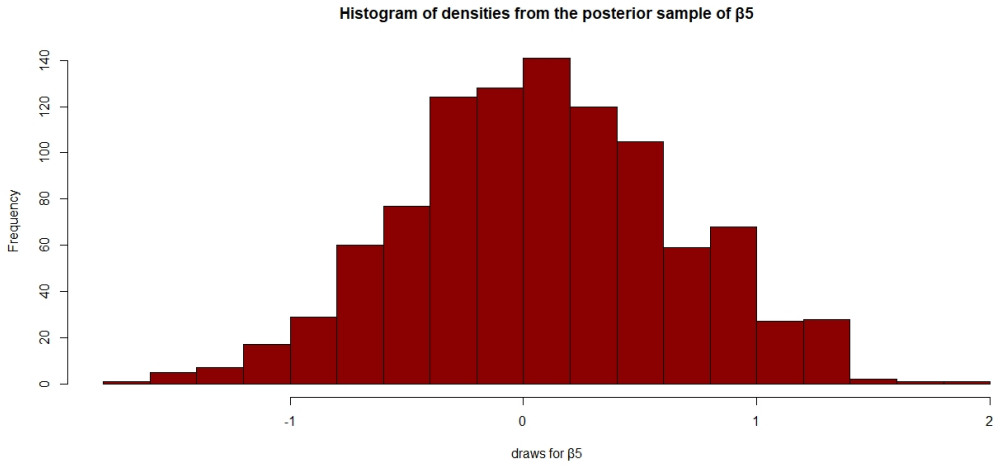

Figure 7.22: Histogram of densities for the posterior sample of β_5 in the Median Probability Model.

Predictor	Posterior Mean	Posterior Standard Deviation	Posterior Median	Real Value
β_0	1.65	0.39	1.65	1
β_1	0.39	0.38	0.39	0.2
β_2	-0.64	0.58	-0.67	0
β_5	0.08	0.58	0.07	0.3

Table 7.3: Table of basic location and spread parameters for the Median Probability Model

We can conclude that the real values of the predictors are more efficiently approximated than in the case of the saturated model, whereas the general fit of the Median Probability Model is also good. What is more, we have succeeded in excluding two of the three unnecessary variables.

Chapter 8

Conclusion

This dissertation has explored the methods of inference and Model Selection in the Bayesian framework for the Normal Linear model, the Logistic model and the Probit model. As far as Model Selection is concerned, a variety of methods has been studied, including the most well known method of the Bayes' Factor, graphical checks, the information criteria of BIC and DIC and the predictive method, called the L-Measure. Despite their efficiency and the robustness of their results, the implementation of the methods mentioned above has proven to be very time-consuming in the study of high dimensional models, due to the fact that they can compare only two models each time. Consequently, when studying a p-dimensional model, a total number of 2^p comparisons have to be performed.

For that reason, Gibbs Variable Selection methods have been developed. The idea upon which this class of methods is built, is to structure a Markov Chain the equilibrium distribution of which, when obtained and properly studied, will point out the models of high posterior probability. This Markov Chain is built via a Gibbs Sampler. The present thesis focuses on one particular technique belonging to that family of methods, called "Stochastic Search Variable Selection (SSVS)" (MacCullogh and George, 1993) and contemplates on its application on the three types of models mentioned above. The implementation on the Normal Linear model is basically the one described in the original paper of MacCullogh and George introducing the method. In the case of the Logit model, where a Metropolis-within-Gibbs step has to be employed within the Gibbs Sampler, we differentiate from the relevant papers, as we employ Gamerman's Independent Sampler, instead of Gilk's Adaptive Rejection Sampling, which is a technique we have used in order to conduct simple inference as well. The whole algorithmic scheme may be computation-

ally intensive, but it is a fully automatic procedure, which requires no tuning and, therefore, it can be regarded as a more user-friendly algorithm with a universal application on the class of Logit models. In order to implement SSVS on the Probit model, we follow the procedure described by Lee et.al (2003)(SSVS-Lee). The structure of the part of SSVS is slightly modified from the original structure proposed by MacCullogh and George(1993), probably in order to achieve faster convergence of the algorithm, but the fundamental principles remain the same. SSVS is a beneficiary technique regarding Model Selection, not only because of the rapidness in producing results, but also because the conclusions we are led to are robust, due to its solid Probabilistic background. Therefore, it can be safely recommended for statistical research and it also already included as a method in statistical software for Bayesian Statistics.

Bibliography

[1] Albert J. H. and Chib S. (1993). *Bayesian Analysis of Binary and Polychotomous Response Data*, Journal of the American Statistical Association **88**: 669-679

[2] Barbieri M.M. and Berger J. O. (2004). *Optimal predictive model selection*. The Annals of Statistics, **32**:870897

[3] Bernardo J. and Smith A. (1994). *Bayesian Theory*, N. York: John Wiley

[4] Bolstad W. M. and Curran J. M. (2016). *Introduction to Bayesian Statistics*, 3rd edition, N. Jersey: John Wiley and Sons LTD

[5] Chang S., Chen R. and Chi Y. (2014). *Bayesian Variable Selections for Probit Models with Componentwise Gibbs Samplers*. Communication in Statistics-Simulation and Computation **45**: 2752-2766

[6] Chen M., Huang L., Ibrahim J. G. and Kim S. (2008). *Bayesian Variable Selection and Computation for Generalized Linear Models with Conjugate Priors*. Bayesian Analysis **3**: 585-614

[7] Chipman H., George E. I., McCulloch R. E., Clyde M., Foster D. P. and Stine R. A. (2001). *The Practical Implementation of Bayesian Model Selection*. Lecture Notes-Monograph Series **38**: 65-134, Institute of Mathematical Statistics

[8] Christensen R., Johnson W., Branscum A. and Hanson T. E. (2011). *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians*, USA: Chapman and Hall

[9] Congdon P. (2003). *Applied Bayesian Modelling*, N. York: John Wiley

[10] Congdon P. (2006). *Bayesian Statistical Modelling*, 2nd edition, (Wiley Series in Probability and Statistics), Chichester, W. Sussex, England: John Wiley and Sons LTD

[11] Gamerman D. (1997). *Sampling from the Posterior Distribution in Generalized Linear Mixed Models*. Statistics and Computing **7**: 57-68

[12] Gamerman D. and Lopes Hedibert F. (2006). *Markov Chain Monte Carlo-Stochastic Simulation for Bayesian Inference*, 2nd edition, N. York: Chapman and Hall/CRC

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

[14] Gelman A. , Carlin J. B. et al. (2014). *Bayesian Data Analysis*, N. York: Chapman and Hall/CRC

[15] Gelman A., Meng X. and Stern H. (1996). *Posterior Predictive Assessment of Model Fitness via Realized Discrepancies*. Statistica Sinica **6**: 733-807

[16] George E. I. and McCulloch R. E. (1993). *Variable Selection via Gibbs Sampling*. Journal of the American Statistical Association **88**: 881-889

[17] George E. I. and McCulloch R. E. (1997). *Approaches for Bayesian Variable Selection*. Statistica Sinica **7**: 339-373

[18] George E. I., McCulloch R. E. and Tsay R. (1995). *Two Approaches to Bayesian Model Selection with Applications*. Bayesian Statistics and Econometrics: Essays in Honor of Arnold Zellner, Amsterdam.

[19] Ghosh J. (2015). *Bayesian model selection using the median probability model*. WIREs Comput Stat **7**: 185-193

[20] Harney H. L. (2003). *Bayesian Inference Data Evaluation and Decisions*, 2nd edition, Switzerland: Springer

[21] Ibrahim J. G., Chen M. and Sinha D. (2001). *Criterion-based Methods for Bayesian Model Assessment*. Statistica Sinica **11**: 419-443

[22] Koch K. R. (2007). *Introduction to Bayesian Statistics*, 2nd edition, Berlin-Heidelberg: Springer-Verlag

[23] Koop G. (2003). *Bayesian Econometrics*, N. York: John Wiley and Sons LTD

[24] Kwon D., Tadesse M. G., Sha N., Pfiffer R. M. and Vannucci M. (2007). *Identifying Biomarkers from Mass Spectrometry Data with Ordinal Outcome*. Cancer Informatics **3**: 19-28

[25] Lee K. E., Sha N., Dougherty E. R., Vannucci M. and Mallick B. K. (2003). *Gene Selection: A Bayesian Variable Selection Approach*. Bioinformatics **19**: 90-97

[26] Li Y., Yu J. and Zeng T. (2017). *Deviance Information Criteria for Bayesian Model Selection: Justification and Variation*, Singapore: SMU Econometrics and Statistics

[27] Lynch M. S. (2007). *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*, N. York: Springer

[28] Mallick B. and Kuo L. (1998). *Variable Selection for Regression Models*. The Indian Journal of Statistics, Series B **60**: 65-81

[29] Marin J. M. and Robert C. P. (2007). *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*, N. York: Springer

[30] McCullagh P. and Nelder J. A. (1989). *Generalized Linear Models, Monographs on Statistics and Applied Probability*, 2nd edition **37**. Cambridge: Chapman and Hall

[31] Neath A. A. and Cavanaugh J. A. (2012) *The Bayesian Information Criterion: Background, Derivation and Applications*. WIREs Computational Statistics **4**, (issue 2):199-203, N. York: John Wiley and Sons LTD

[32] Ntzoufras I., Forster J. J. and Dellaportas P. (2000). *Stochastic Search Variable Selection for Log-Linear Models*. Journal of Statistical Computation and Simulation **68**: 23-37

[33] Ntzoufras I. (2009). *Bayesian Modeling Using WinBUGS*, (Wiley Series in Computational Statistics), N. Jersey: John Wiley and Sons LTD

[34] O Hara R. B. and Sillanpaa M. J. (2009). *A Review of Bayesian Variable Selection Methods: What, How and Which*. Bayesian Analysis **4**: 85-117

[35] Robert C. (2001). *The Bayesian Choice*, 2nd edition, N. York: Springer-Verlag

[36] Spiegelhalter D. J., Best N. G., Carlin B. P. and Van der Linde A. (2002). *Bayesian Measures of Model Complexity and Fit*. J. R. Statistical Society B **64** part 4: 583-639

[37] Swartz M. D., Kimmel M., Mueller P. and Amos C. I. (2006). *Stochastic Search Gene Suggestion: A Bayesian Hierarchical Model for Gene Mapping*. Biometrics **62**: 495-503, International Biometric Society

[38] Zellner A. (1996). *An Introduction to Bayesian Inference in Econometrics*, (Wiley Classics Library), N. York: John Wiley and Sons LTD