Bayesian Shrinkage in Linear Models with an Application to Genetic Networks

Dler Mustafa Khidhr

Thesis submitted to the University of Sheffield for the degree of Doctor of Philosophy in Statistics

School of Mathematics and Statistics
University of Sheffield
Sheffield, U.K.

January 2020

Supervisor: Dr Miguel Juarez

Abstract

Different challenging issues have emerged in recent years regarding the analysis of high dimensional data. In such datasets, the number of observations is much lower than the number of covariates which is problematic in the conventional statistical model. Nowadays, high-dimensional dataset is common in several fields of sciences such as biology, economics, genetics and medicine. For instance, gene expression data is an example of the high-dimensional dataset where the number of genes is larger than the number of samples (patients). In order to tackle the issue of high-dimensional there are many regularization and shrinkage methods that have been proposed and developed to gain a sparse model; these approaches belong either to frequentist penalty methods or Bayesian shrinkage prior. In this thesis, we aim to overcome the high dimensional problem through proposing two alternative novel Bayesian shrinkage prior distributions. The suggested methods are hierarchical Normal inverse Pareto distribution (hNiP) and Rescaled Beta hierarchical Normal inverse Pareto distribution (ReB-hNiP). We consider that both proposed priors are absolutely continuous prior distributions and belong to the family of scale mixture of normal distribution. The proposed Bayesian shrinkage methods have been applied for three different linear models. Particularly, they have been applied and compared with some shrinkage methods based on multivariate Bayesian linear regression model. The proposed methods have also been implemented on both the Bayesian linear regression model with measurement error model a shrinkage and dynamic Bayesian networks with measurement error model. The hyperparameters were selected by using several criteria such as Watanabe-Akaike information criteria (WAIC).

Acknowledgements

First of all, I would like to express a deep acknowledgment to my supervisor (Dr. Miguel Juarez) who accepted the opportunity to supervise my Ph.D. journey. Also, he provided me with constructive advice and criticism as well as supporting and guiding me to write my thesis via useful suggestions. I would like to thank all the staff in the Statistics department who encouraged me to complete my project. Furthermore, I would like to thank Dr. Marta Milo who provided me the real data. I would also wish to express my gratitude to Dr. Keith J Harris who advised me in writing my PhD.

As well, I would like to thank my sponsorship, from the High Ministry of Higher Education/ (KRG) Kurdistan Regional Government, that gave me an opportunity to study my Ph.D. abroad and give financial support during some of the study periods. I would like to thank my family, particularly my parents, who have encouraged me to complete my education throughout my life; also I appreciate members of my family and friends that have provided and supported me, especially my brother and sister which spent much time and effort sorting out my documents in Kurdistan. I would like to thank all the friends that helped me in different ways during my Ph.D. journey, Dler Kaider, Massud Hassan and particularly, Dr. Karwan G Ameen who is professional GP in Sheffield.

Last but not least, I would like to give great thanks to my wife, Ala A Ahmed, who has supported me during my PhD and has had the capability to overcome the difficulties that we faced as well as taking the responsibility to look after our children Dima and Anas for a few years away from her family.

DLER M. KHIDHR

Contents

List of Tables

List of Figures

Chapter 1

High Dimensional and Genetic Data

1.1 Background to High Dimensional Data

Over decades, statistical models have been developed and become complex, especially for high-dimensional data. The issue of high-dimensional data appears in different fields of scientific studies such as economics, social and biological studies. High-dimensionality arises from a dataset when the number of parameters are exceeded by the number of observations so called $p \gg n$ case (Mallick and Yi, 2013). For example, in genetics it is very common, to measure thousands of genes on dozen of people. Practically, only a few possible covariates (observations) really have an effect on the response variable, but the impact of most coefficients is close to zero or very small. Meanwhile, model misspecification has a substantial effect on interpreting the results scientifically. It is crucial belief of analysing any statistical model that have patterns of high-dimensional problem the number important features k (the number of the parameters are non-zero of such as in regression model) should be much lower than all parameters p (Brimacombe, 2014).

It is necessary to use a specific method to overcome the problem high-dimensionality and obtain a sparse model. Consider the model is sparse if a small number of the parameters are not equal to zero and the rest of coefficients are equal to zeros, which means that the model can be described by a few important variables k (features) out of p regression coefficients. X are gene

expression $X \in \mu_{n+p}$ and $Y \in \mathbb{R}^n$ is response variable. Sparse models have been motivated in various ways in many application fields (Aijun et al., 2017). For that purpose, over the last decade, statisticians have proposed diverse methods for dealing with this problem by penalizing the likelihood then obtaining a sparse model and variable selection (Mallick and Yi, 2013). In practice, these methods have been improved in order to tackle a high dimensional problem and variable selection in both the frequentist framework and the Bayesian modelling. For example, Tibshirani (1996) suggested LASSO procedure (Least Absolute Shrinkage and Selection Operator) in term of a frequentist model. LASSO is widely applied in several fields to penalize the least square model in statistical literature. After that, some others shrinkage penalty functions have been proposed and developed in order to overcome the issue of high dimensional datasets such as Elastic Net by Zou and Hastie (2005), (SCAD) Smoothly Clipped Absolute Deviation by Fan and Li (2001), Adaptive LASSO by Zou (2006) and among others in the direction of the non-Bayesian methods. In Section 2.2, we will explain some frequentist penalization methods.

Recently, a Bayesian perspective has provided the power to decrease the complexity associated within coefficients estimation procedures by combining prior information. Thus, the data in this technique in order to do variable selection. A number of shrinkage prior distributions have been suggested as an alternative to frequentist method and these approaches are gradually becoming widespread. In this framework, the Bayesian shrinkage prior also is called Bayesian regularization method (Mallick and Yi, 2013). One family of shrinkage prior has a scale mixture of Normal distribution with mean is equal to zero and variance has positive hierarchical mixing prior distribution. In this case of dealing with Bayesian shrinkage priors, we utilise MCMC algorithm (see Appendices A.1). We are going to discuss more about Bayesian shrinkage prior distributions in Section 2.4. The important part of variable selection for any shrinkage method is how to choose the optimal tuning prior, where each shrinkage approach has at least one tuning parameter which is control amount of shrinkage the coefficients regression Jung (2016). In practice, there are several methods and criteria recommended and developed to select tuning parameter, for instance, cross-validation technique is one of the popular methods used for this purpose (Zou, 2006). In Section 2.5.2, we explain some techniques and criteria used in this thesis in order to to choose the best hyperparameter values.

The goal of the thesis is to propose an alternative Bayesian shrinkage prior distribution and at the same time deal with variable selection in a high dimensional model that involves a hierarchi-

cal Normal scale mixture of inverse Pareto distributions (hNiP) and rescaled beta hierarchical Normal scale a mixtures of inverse Pareto distributions ReB-hNiP. We believe that the proposed Bayesian shrinkage priors is faster in terms of time computing a model and results could be more accurate compared to similar shrinkage priors. The proposed methods are belong to a family of scale mixtures of normal distributions. Moreover, we will address those shrinkage priors distributions into different types of modelling, such as linear regression model as standard statistical model in the Chapter 4 and we compare with some shrinkage methods in measurement error model and dynamic Bayesian networks in Chapter 5. Another challenge we face in this thesis is how to select hyperparameter values of proposed Bayesian shrinkage priors. We utilised sensitivity analysis by thresholding hyperparameter values by using and evaluating criteria called Watanabe-Akaike information criteria (WAIC) method Vehtari and Gelman (2014). Hence, we utilise some other statistical tools so as to select hyperparameters values then check a performance of models based on the proposed shrinkage prior distributions.

1.2 Thesis Outline

This thesis consists of seven Chapters. The first chapter includes an introduction to high-dimensional issues, genetics background and literature and methodology of dynamic Bayesian networks. Then, we display the overview of shrinkage methods in linear models for both frequentist penalization methods and Bayesian shrinkage methods in the second chapter. In Chapter 2 the scale mixture of normal distribution is illustrated as background of Bayesian shrinkage prior. The last section of the Chapter 2 involves an overviews of Bayesian variable selection, hyperparameter choice with model selection illustrated and receiver operating characteristic curves ROC methods described in order to evaluate and analyse shrinkage priors. the last subsection of second chapter is related to posterior predictive checking distribution is illustrated for checking the performance methods and prediction behaviour. In Chapter 3, we introduce two novel Bayesian shrinkage priors which are the hierarchical Normal inverse Pareto distribution (hNiP) those hyperparameters Gamma are distributed and Rescaled Beta hierarchical normal inverse Pareto distributed (ReB-hNiP). In addition, the Normal exponential gamma prior distribution (NEG) is presented due to having similar structure of our proposed methods.

In Chapter 4, calibration on Bayesian linear regression model is addressed using our novel prior distribution which includes hNiP and ReB-hNiP and compare the results with Bayesian and non-Bayesian approaches. After describing the model structure, hyperparameter prior values are selected by different criteria, then the model is run over 100 times and the results analysed and compared, in order to assess the performance of the shrinkage prior distributions. High-dimensional measurement error with linear models are studied in Chapter 5, which starts with an overview of measurement error model. In the next section, high-dimensional measurement error with multivariate linear regression model is addressed values of the Watanabe-Akaike information criteria (WAIC) for selecting tuning parameters. Subsequently, results are analysed by run model over 100 times based on simulated data. Additionally, dynamic Bayesian networks with measurements error model is considered in this chapter, this kind of modelling is based on autoregressive model with measurements error model. Both proposed Bayesian shrinkage prior distributions are addressed for both non-high-dimensional and high-dimensional simulated datasets. Moreover, credible interval (C.I.) and t test utilised as extra tools for variables selection, after that our proposed modelling is compared to with some alternative methods in the literature.

In Chapter 6, a statistical literature review related to cardiovascular diseases CVD is displayed, after that the source of real gene expression data are introduced and then process of choosing sub-sample of real gene expression is presented to reduce the computation time and applied our proposed model on it. Thus, real sample of CVD gene expression level is applied on dynamic Bayesian network with measurements error model. Chapter 7 concludes with our the most important results and future work.

The most common statistical distributions used in this thesis which are the Gamma distribution, Normal distribution $N(x|\mu, \lambda^{-1})$, Inverted Pareto distribution iP$(x|\psi, b)$, Student-t St$(x|\mu, \nu, \lambda)$, Truncated Gamma distribution TrGa$(x|b, \omega, \gamma)$ and Resca led Beta ReB$(x|a, b, e)$, are presented in the appendices.

1.3 Genetic Background

This section provides general information about genetics in order to attract a reader about: what is genetic, gene and biological information. To know how protein synthesis. We are addressing genetic data which plays a circular role in a statistical model, especially gene regularity networks (GRN). Another reason to write this information in here is using real genetic data in this project which is gene expression level. In general, genetics is a science that involves and studies genes and how living things, the way some attributes are passed on from generation to generation, for example as eye-color. Geneticists and statistician are interested in deep searching the genetic data in a different direction because of the familiar that for some genes plays in developing the disease. (Laird and Lange, 2010). In the following subsection, we briefly explain some information about genes and their protein construction.

1.3.1 Genes and DNA

Cells are the construction blocks in all living organisms. All cells execute many processes to survive in life. One of the main components and an essential part of all living cells are proteins (Bani, 2009). Proteins consist of a collection of different chains of amino acids; the sequences of these amino acids play a role in determining the types of proteins inside the cell under the control of genes. Genes are defined as functional pieces of DNA (Deoxyribonucleic Acid) which is a code for proteins. They are simple elements in living organisms, having control over transferring inherited features (Lee, 2007). For example, characteristics such as the colour of eyes are moved from ancestor to next generation via the parents. DNA is found inside the nucleus. The shape of DNA is similar to a double strand helix formed by four different primary types of nucleotide; thymine (T), cytosine (C), guanine (G) and adenine (A). Each Nucleotide is the construction block of DNA (Robinson, 2010). The nucleotide sequences are genetic information used for the development and function of the organism, implemented during the construction of proteins. The same DNA is presented in all cells, but each cell has a different rate of protein expression, thus living cells are different from each other.

1.3.2 Protein Synthesis

Protein synthesis from DNA happens through ribonucleic acid (RNA) production as explained by the central dogma of molecular biology (Causton et al., 2009). There are three essential stages: *transcription*, *splicing* and *translation*. Figure 1.1 is shows all main steps of protein synthesis. The first stage is transcription; the information contained within a gene is copied into messenger RNA (mRNA) by the enzyme RNA polymerase, utilizing the DNA structure as a blueprint. The mRNA translates into protein after it is modified. Sections of the RNA are removed. These are named as introns and the remaining part of the coding sequences named as exons, which are joined to each other which this process is called splicing reactions (Bani, 2009). In *translation*, proteins are synthesized by organelles known as ribosome typically found in the cytoplasm. Ribosomes read the triplets of nucleotide in the mRNA series which are called codons; they specify amino acids into a sequence of rising polypeptides. The protein replicates into a three-dimensional composition when translation is completed. For more details see Causton et al. (2009), Lee (2007) and Bani (2009),

1.3.3 Gene Regulatory Networks

Beside the details we mentioned previously about protein regulation, gene expression is well regulated for controlling the quantity of produced proteins. Transcription factors are a key method for regulating gene expression at transcription level. These factors are proteins attached to a promoter area of a specific gene, which prevents or activates the expression level by promoting the employment of the polymerase of RNA (Hill, 2012; Russe, 2009). This relates to gene regulatory networks. Gene regulatory networks (GRNs) are a dynamic and complex process of protein production between genes. The gene interacts with other gene production. This interaction occurs when the transcription factor from a gene activates another gene and it leads to production. The interaction is not only a product of a gene activating other genes, but also, some genes are prohibiting other gene expression. The relation between gene-gene can be represented by topological networks, in which genes are nodes in graph and interactions between genes are represent the edges or the link between nodes it can be seen in Figure (1.2), which shows an example of a gene regulatory network, it show that gene D is suppressing gene C and

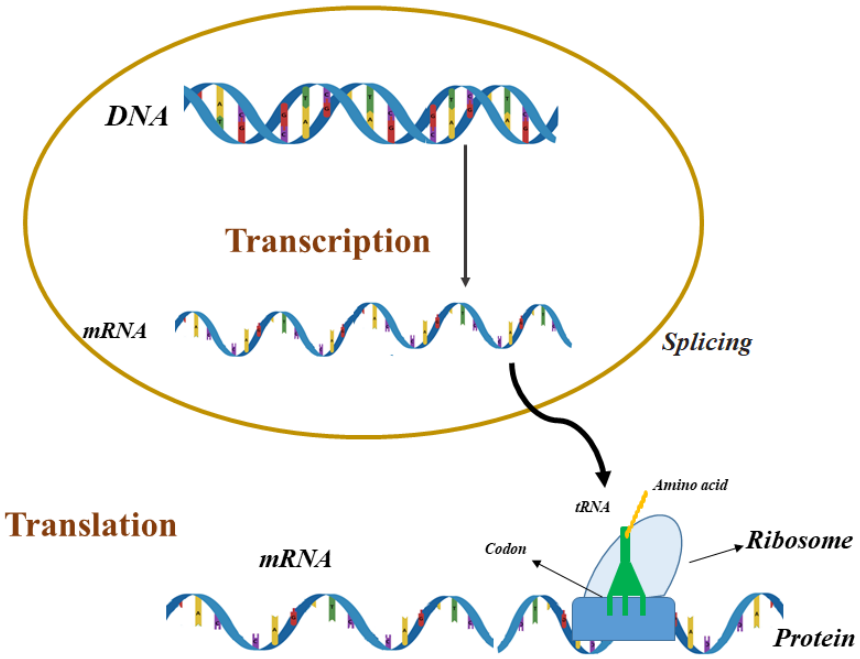


Figure 1.1: The process of synthesized proteins, which starts from the transcription of information in DNA into mRNA and splicing after reactions which leads to removing the intron regions and fuse and kipping exon regions together. Then process the mRNA is translated into protein. This diagram I have recreated based on a similar diagram in Bani (2009).

also gene B is activated gene A. Figure (1.2) I have recreated based on paper for Kuznetsov et al. (2004) after gave me his permission.

1.4 Cardiovascular Diseases Overview

1.4.1 Cardiovascular Diseases

Cardiovascular diseases (CVDs) are a range of disturbances of the blood vessels and heart. CVD consists of some types of heart disease that include heart failure, stroke, angina, con-

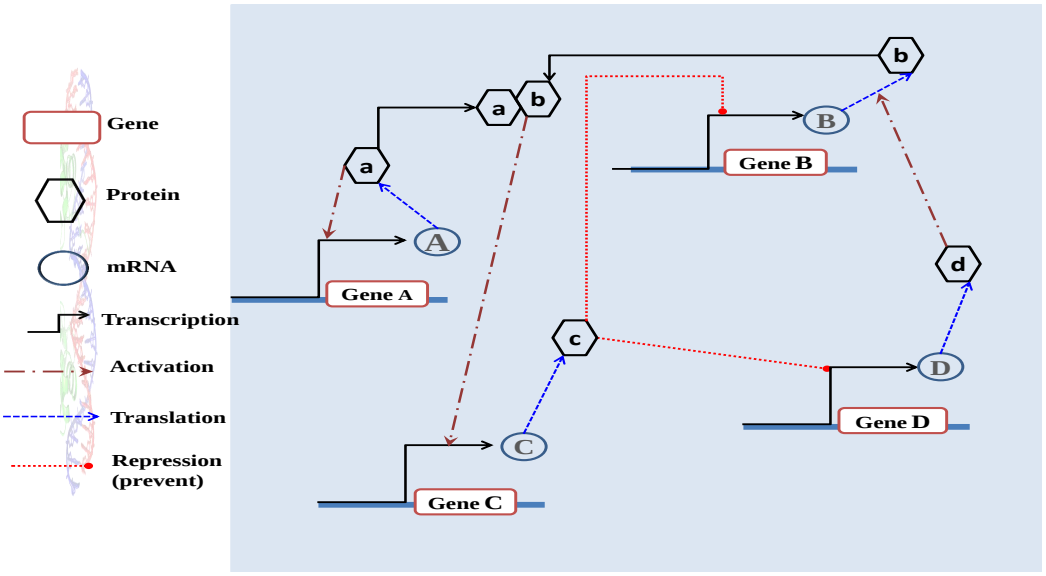


Figure 1.2: An example of a gene regulatory network for four different genes (A, B, C, D). Each gene begins with transcription of the information into mRNA, which is indicated by the black arrows. During protein synthesis, some process happen such as (brown dash-dot line) activation in which some genes activate other genes (positive impact), in contrast, the red dotted line refers to prevention (repression) of some gene by another gene, that is, negative impact).

genital heart disease and heart attack, among others. CVD is also known as circulatory heart disease. According to many heart organizations, CVD is one of the biggest causes of death and morbidity worldwide (National Health Service-NHS, Heart Foundation and NHI). The World Health Organization WHO claimed that about 17.7 million people die from CVDs every year and approximately 31% of deaths worldwide are due to CVDs (WHO, 2017). Moreover, about 570,341 people passed away due to CVDs in 2014 in the UK, and in that year about 28% of deaths in the UK were caused by CVDs and only 15% from coronary heart disease (CHD), which is a type of CVD. The WHO consider CVDs as the biggest type of killer disease. Many people suffer from severe disability due to CVDs and suffer chronically. For more information see the published statistics from the Heart Foundation in Nick et al. (2015).

1.4.2 Literature Review of Cardiovascular Diseases

There are several risk factors for cardiovascular diseases and biomarkers. A biomarker is defined as a feature which is evaluated and measured objectively as an index of pathogenic pro-

cesses, normal biological processes or pharmacologic reactions to a therapeutic involvement of CVDs (Strimbu and Tavel, 2010). They have an impact on the patient by various means, for example, high blood cholesterol, smoking, diabetes, obesity, high blood pressure and history of heart disease and some other biomarkers. The CVDs are indicated by a biomarker and contain a type or level of exposure for an environmental factor, genetic responses to exposures and genetic susceptibility (Vasan, 2006). Single nucleotide polymorphisms (SNPs) are variations of the simple form for DNA structure that occur when a single nucleotide ((A), thymine (T), cytosine (C), or guanine (G)) in the genome sequence is altered. For instance, a SNP may modify the DNA sequence such as ATGGCTAA to AAGGCTAA. Cardiovascular diseases has been evaluated extensively in relation to SNPs (Vasan, 2006).

In gene expression analysis, a fundamental change has happened in the assessment which reflects a paradigm shift from the traditional one-molecule strategy to gene regulatory network assessment (Napoli et al., 2003), which is called gene regulatory networks (GRNs). A few decades ago, with regards to the statistical genetics analysis, the most successful results related to select genes that have an impact on disease progression. Numerous statistical methods for achieving this success were developed, and were mainly effective in analysing diseases such as Huntington disease and heart disease, among others. Statisticians in the genetics field should not simply acknowledge but also directly confront the tremendous complicating factors that may contribute to genetic diseases, as they imposes great challenges for classical statistics (Thornton-Wells et al., 2004).

Shiffman and Porter (2000) argue that, until now, the process of CVDs was not fully understood from the molecular events and psychophysiology with respect to disease condition due to processes of CVDs being complex. Ndiaye et al. (2011) claims that many genetic risks of CVDs are so far unexplained and they defined this claim as dark matter of genetic risk. Furthermore, they stated that gene-gene interactions will play a great role in detecting genes that have not been known until now by using single-locus, a particular position on a chromosome structure (Suzuki et al., 1976).

According to Seo et al. (2006), genetics contributed to complex CVDs such as atherosclerosis. They claimed that a research based on genetic data of CVDs have a beneficial in clinical trial, particular when applying and translating findings relying on gene expression data into clinical

medicine. However, they said is difficult to have the full information about single gene has cause and impact on CVDs and there was not much progress in genetic research with regards to the CVDs field, due to the difficulties the researchers encountered in accessing sufficient information about genetic data from tissues of heart. Lack of access to appropriate human tissue is one of the major challenges in the advance out of the field (Seo et al., 2006).

Health care planning and diagnostics have improved significantly based on genomic analysis because patients suffering from CVDs may change to specific medication that has fewer side effects and can be less costly. Geneticists collect data by three different methods. Firstly, from heart and lung donors, in recent studies, researchers have used gene expression levels exclusively from human aorta to identify genes that predict the extent of atherosclerotic burden in the aortic artery (Seo et al., 2006). However, those people who donated their organs had not necessary been diagnosed with CVDs. Secondly, researchers often utilized animal data for applying statistical modelling, but these types of model considered useless for determining the CVDs risk factors in human disease progression. Despite animal organs being used by researchers, there are some major challenges that have been encountered, and several factors that may contribute to CVDs in humans cannot be found in animals such as smoking, drinking alcohol, drugs and stress (Seo et al., 2006). Thirdly, the obtained sample data from blood is noisy. Blood may carry other diseases than CVDs. Alongside this, Seo et al. (2006) noticed that they can obtain a sample of human tissue during a transplant or placement of left ventricular assist devices (LVADs).

Lusis and Weiss (2010) demonstrated that the fundamental concept of multivariate genetic risk factors results in CVDs. Every genetic and environmental risk factor partly contributed to disease risk. The protein structure is affected by both environmental factors and genetic factors which cause molecular phenotypes. Their review focused on biological networks and the relationship between phenotypes and genetics. Most of the studies are displayed as examples; the data comes from mice or rats. An example used partial correlation coefficients to calculate relationships between the DNA and other metabolic factors of CVDs. They believe that the most powerful approaches to describe the systems biology are dynamic networks because they clarify interactions occurs over time between genes.

1.5 Statistical Methodology for Gene Networks

Genetic data have become one of the most important kinds of data to build statistical models to determine functional genes that cause diseases due to advances in the technology in this field. It is known that during the process of protein synthesis some interactions happen between genes. One of the essential structures to fit and create good models based on this type of genetic data is networks, especially gene regularity networks (GRNs) (Vijesh et al., 2013). We will focus on probabilistic graphical models GM to describe GRNs and we are going to use real gene expression data, explained in chapter 6 for a kind of GRNs model. We are interested in directed graphical models, especially dynamic Bayesian networks (DBNs), that is a type of a graphical model that has become a typical technique for modelling different stochastic time-points. Using DBNs to build models of genetic data is useful for inferring the interaction uncertainties between genes. It is known that a classical regression method for building gene networks is impractical, especially when dealing with the issue of high-dimensional data, $p \gg n$, for more information see Brimacombe (2014). Such a method is infeasible due to the time-consuming nature for a high number of iterations difficult to fit a model. Consequently, researchers have developed numerous computational approaches for modelling by different techniques. The DBN is one of the most popular statistical models for advanced interactions between proteins and genes as described in a gene network (Vijesh et al., 2013).

We will now state some methodology and discus some literature about GRNs in the current section and then dynamic Bayesian networks are explained in detail in the following sections. According to my knowledge of the literature about dynamic Bayesian networks, Friedman et al. (1998) first suggested DBNs as approach for creating gene networks by using gene or protein expression data. Murphy et al. (1999) then developed this concept by utilizing dynamic Bayesian networks to build a model and provide theoretical properties.

Regarding the literature of dynamic networks, there are several kinds of techniques used to create probabilistic network model: the first kind is discrete models as can be found in (Chai et al., 2012; Chen et al., 2012; Ong et al., 2002; Zou and Conzen, 2005). They utilised discrete data over different time-points or they discretised the data, for particular gene expression. The discretised data is typically (gene expression) divided into three different three categories: normal, under-expressed, and over-expressed. Discretisation values are based on the rate of

expression which is significantly greater than, similar to, or lower than control, by placing a threshold on the ratio between control and measured gene expression. For instance, Friedman et al. (2000) assumed a threshold of $2^{0.5}$, that is, if the ratio is higher than $2^{0.5}$ it categorised as over-expressed and if the ratio is less than $2^{-0.5}$ it is categorised as under-expressed. Some of the researchers showed the results according to a conditional probability table (CPT). For example, the model structure of Ong et al. (2002) was based on Dirichlet priors, and also, the expression data were discretised. While Chen et al. (2012) used categorical variables and discrete time-series data in their model. The data they used is longitudinal morphological that study the changes in the human brain.

The second kind is non-parametric or semi-parametric models, for example (David and Wiggins, 2007; Imoto et al., 2002, 2003; Kim et al., 2004; Morrissey et al., 2011; Penfold and Wild, 2011; Sugimoto and Iba, 2004; Yang et al., 2016). In some cases, creating a model under dynamic linearity assumption infeasible for data that has feedback loops within a living system, such as gene expression data (Imoto et al., 2002). In this situation, there are numerous published studies that are interested in using non-linearity behavior of gene regulatory networks as an alternative method to linear models as we stated above, especially with a dynamic Bayesian network structure. The general formula of the non-parametric model is as follows:

$$Y_i^t = F(Y_i^{t-1}) + \epsilon_i^t \tag{1.1}$$

where $i = 1, 2, \ldots, p$, ϵ_i^t is the noise of the model, Y_i^t refers to the expression level at time t. $F(Y_i^{t-1})$ is the function of the non-parametric or semi-parametric model. For instance, Kim et al. (2004) constructed the nonlinear dynamic Bayesian network by setting the conditional densities based on non-parametric additive regression, where interaction between genes is non-linear. Another example is the network model built on utilising semi-parametric regression from the Bayesian perspective by Morrissey et al. (2011). They assumed that the relationship between the variables was nonlinear and could be described by using spline functions. The third kind of dynamic networks is vector auto-regressive processes which are addressed by (Dondelinger et al., 2013; Lèbre, 2009; Morrissey et al., 2011; Opgen-Rhein and Strimmer, 2007; Wit and Abbruzzo, 2015). The main formula for constricting the model, which depends on the autoregressive AR which is a multivariate linear regression model for the time series data of the current value t on the previous value $t - 1$, we will summaries and discuss some

studies later. The last type of modelling for probabilistic networks are hidden Markov models or state space models, which can be found in (Beal et al., 2005; Doshi et al., 2011; Godsey, 2013; Perrin et al., 2003; Tobon-Mejia et al., 2012; Zhu and Wang, 2012). In these types of modelling hidden variable and missing data are take into account. The model of state space comprises of two equations: the first equation of observation and the second equation of state. The equation of observation describes measurement of the state. The observation is usually stochastic, and the data either continuous or discrete. The state equation describes the state process's temporal evolution as a dynamic system. Either through a deterministic or stochastic differential equation or through a stochastic differential equation, the states can be described in continuous time. The state space model is called Markovian if the present state depends only on the previous state. Dynamic linear state-space model which represent in Equation 1.2, X_t is hidden variable (state) at time t, Y_t is observed variable (state), A and C denotes transition matrix and projection matrix respectively. Also, u, v are the error of models and μ_{obs} is measurement adjustment. For instance, Perrin et al. (2003) utilised this kind of DBN and Expectation-Maximization algorithm (EM) which allows in particular to infer hidden variables and to handle missing data. Non-high dimensional gene expression data they used.

$$\begin{aligned} X_{t+1} &= AX_t + u \\ Y_t &= CX_t + \mu_{obs} + v \end{aligned} \tag{1.2}$$

Tobon-Mejia et al. (2012) used mixing DBN and Gaussian Hidden Markov Model structure in order to check the fault diagnosis of computer numerical control of CNC machine. The number of interesting gene network inference approaches is increasing. Scientists have tried to classify them into several groups depending on criteria. For example, Karlebach and Shamir (2008) and En Chai et al. (2014) categorized the network inference approaches into some classes depending on the mathematical modelling for these approaches: the continuous, single-molecule level and logical models. The continuous models are able to take into account the time series data, as well as progressive biological applications for modelling gene networks. DBNs belong to this modelling class. Then The logical models are flexible and can give fundamental knowledge of functionality versus several system's conditions, but they can only able to supply qualitative solution (Naldi et al., 2009). An example of a logical model is the Boolean network where gene is denoted to be 'off' when gene is activated and denoted 'on' if gene deactivated. The last

class of models of the gene expression level places emphasis on the connection among genes stochastically.

In order to explain the link between probabilistic network models and shrinkage priors, we first state some studies and then we succinctly explain our modelling approaches and how it differ from previous research. For instance, Zou and Feng (2009) compared two methods which are commonly used; these methods are DBNs which are based on a simple linear model (Gaussian model) and Granger causality methods (that can identify the causal impact of the series, it can improve the prediction of a series of one-time by integrating knowledge of the second series) (Zou and Feng, 2009). They showed that the result of the former methods outperform the latter, when the data length is short and that the opposite is also true, that is, Granger causality is best if data length is large. Wit and Abbruzzo (2015) proposed a type of dynamic Bayesian network that captures the slow changes amongst genetic regulation over time. An autoregressive model and least absolute deviations (LAD) penalty function is used to shrink the coefficients matrix in the model. The general form of the model for the time variation networks is as follows;

$$Y^t = B^t Y^{t-1} + \epsilon^t \tag{1.3}$$

where ϵ^t is the noise model, Y^t refers to variables (expression level) at time t and B^t refers to the coefficients changing for every single time in the model. Morrissey et al. (2010) studied the interaction amongst variables based on dynamic Bayesian networks and an autoregressive model with repeated measurements. They used a non-Gaussian distribution for measurement error parameter. Also, they utilised two different synthetic data which are come linear and non-linear mode in order to compare them. A spike and slab prior distribution are implemented to assess indicator variables (Morrissey et al., 2011). This prior is a discrete mixture of normal distributions, which a high peak and low variance (spike) and flatter tails with a high variance that belongs to the slab.

One part of my work is similar to Morrissey et al. (2010) in terms of the model structure which includes an autoregressive model plus a measurement error model. However, we use a different prior distribution with regard to penalizing the high-dimensionality in the coefficient vector of the model and identifying the variables at the same time. We are going to discuss the dynamic Bayesian network model based on an autoregressive process with a measurements error model

by using proposed shrinkage prior distributions in Chapter 5.

In the following sections, we explain the background of Bayesian networks which is given to understand the idea of dynamic Bayesian networks. Also, we establish general information about dynamic Bayesian networks.

1.5.1 Bayesian Networks

A Bayesian network (BN) is one of several statistical methods applied for modelling in biological science, particularly in gene expression data. BNs are known by other titles in the literature, such as belief networks and probabilistic networks. Moreover, the BN model is similar to the probabilistic graphical model (GM) family for the spatial case, when we have a directed acyclic graph (DAG), denoted by G (Bani, 2009; Ben-Gal, 2008; Radhakrishnan et al., 2013). Probabilistic network models are divided into two main kinds. The first one is directed acyclic graphs (Bayesian networks), which we will describe in the following paragraphs. The second is called an undirected graphical model. It is different from BNs as the links (edges) between any two variables do not include the directional knowledge. Undirected graphical models are also known as Markov networks (Ben-Gal, 2008). A DAG $G(Y, E)$ in the BN model is composed of two main parts, the first part are nodes that represent a set of random variables $Y_1, Y_2, \cdots, Y_n$ and the second part is a set of the edges which are indicated by E, and model dependence between variables. Commonly, an arrow between the variables represent the edges. The arrow represents that a value taken by node Y_j is based on node Y_i, that is, node Y_i have an effect on Y_j. For example, Figure 1.3 shows node Y_2 is child of node Y_1 which joined by an arrow. In the BN model, the node Y_i (variable) denotes a parent of Y_j and correspondingly Y_j denotes a child of Y_i.

According to the definition of a Bayesian network, it is an acyclic graph that is described by a joint probability distribution for a group of random variables Y and edges E as mentioned above. The joint probability distribution of a set of random variables can be represented by:

$$P(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3, ...Y_n = y_n) = P(y_1, y_2, ..., y_n). \tag{1.4}$$

The joint probability distribution above can be based on probability theory written as:

$$P(y_1, y_2, \ldots, y_n) = P(y_1)P(y_2|y_1)\ldots P(y_n|y_{n-1}). \tag{1.5}$$

$$P(y_1, y_2, ..., y_n) = P(y_1)P(y_2|y_1)P(y_3|y_1, y_2)...P(y_n|y_1, y_2\ldots y_{n-1}). \tag{1.6}$$

The conditional probability for each variable is decomposed based on the Bayesian Network structure as follows:

$$P(y_i|y_1, y_2, ..., y_{n-1}) = \ P(y_i|Pa_i(G)). \tag{1.7}$$

where y_i represents the value of Y_i nodes and $Pa_i(G)$ represents all the parent nodes of the network of Y_i nodes. Also parent nodes can be described by $Pa(Y_i) \subseteq \{Y_1, Y_2, ..., Y_{n-1}\}$. Then the joint probability distributions in Equation 1.6 can be represented as:

$$P(y_1, y_2, ..., y_n) = \prod_i^n \ P\bigg(y_i|Pa_i(G)\bigg). \tag{1.8}$$

In this graphical network, the variable Y_i is independent of its non-descendants given the rest of the parents in DAG. The second component is a collection of numerical parameters that generally describe distributions of conditional probability. The set of the parameters in the network are represented by Θ, which involve $\theta_{y_i|Pa_i(G)}$ for each realization y_i conditioned on $Pa_i(G)$, that is, the group of parents Y_i in G where $\theta \in \Theta$ (Ben-Gal, 2008). The joint distribution based on conditional distribution can be represented by the followings rule as parameter learning:

$$P(y_1, y_2, ..., y_n) = \prod_i^n \bigg(\theta_{y_i|Pa_i(G)}\bigg). \tag{1.9}$$

Inferring gene networks model (GNs) employs Bayesian network is corresponding to classical statistical inference in terms of parameter estimation. The BN inferences includes two typical stages includes structure learning and parameter learning. The first stage is structure learning that concentrates on definition of the (DAG), G introduced earlier in this section. There are various algorithms and scoring functions (statistical criterion such as Bayesian information criterion (BIC) (Schwarz et al., 1978)) proposed to assess and to learn the structure of Bayesian

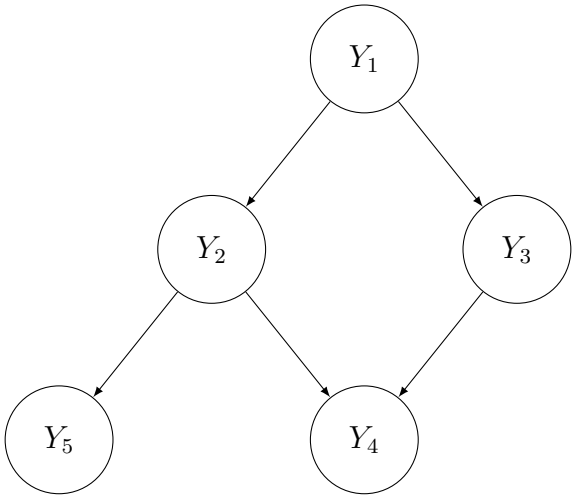


Figure 1.3: Simple toy example of a Bayesian network with five nodes that have a conditionally relation; circles represent the variables (nodes) and edges (arrows) represent the interaction between variables (link). For example, Y_4 and Y_5 are conditionally independent given Y_2.

networks. Popular algorithms are score-based, constraint-based, hybrid structure learning and Bayesian scoring function. One of these algorithm to infer the model structure is constraint-based algorithm which utilizes the outcomes of conditional independence test between nodes based on a statistical test. Score-based approaches construct a structure with the use of conditional independence between variables. This approach attempts to determine a score function which assesses a network structure by reflecting its goodness of fit data and then to determine the graph at the maximum score. Hybrid structure learning seek to learn parameters by combining the constraint-based and score-based methods. They are local search approaches dealing with identifying the local structure and optimal global model, which is constrained by this local information, see Radhakrishnan et al. (2013) for acquiring more knowledge about these algorithms.

The second stage is parameter learning (parameter estimation), involving a specific probability distribution over variables (nodes) in the estimation process that is related to determine each node in the network model.

For understanding the Bayesian network simply, consider the following example. Figure 1.3 shows an example of a BN for five variables Y (nodes). We assume the variables are gene expression levels for five genes. The circles represent the nodes and arrows to indicate the link

between nodes. The probability distribution for variables (genes) depends only on its regulators (parents) in the network. For instance, the expression levels of Y_4 and Y_5 have a relation only via Y_2 as expressed in Figure 1.3. In mathematical terms, a conditional model is factorized into components of the current example is represents by bellow formula, and also both node Y_4 and Y_5 are conditional independent given Y_2. Such a relation leads to building conditional distributions, where each variable depends only on its parents in the network.

$$P(Y_1, Y_2, Y_3, Y_4, Y_5) = P(Y_5|Y_2)\ P(Y_4|Y_3, Y_2)\ P(Y_2|Y_1)\ P(Y_3|Y_1)\ P(Y_1)$$

1.5.2 Dynamic Bayesian Networks

Dynamic Bayesian networks (DBNs) are an extension of Bayesian networks. BNs are based on data from a single time point, while DBNs employ time series data for finding connections amongst variables to create a model (Radhakrishnan et al., 2013). There are some constraints of Bayesian networks for inferring networks. The first constraint is equivalent to the class of Bayesian networks, where the equivalent class is obtained when the structure of model has V-structural. V-structures are the only basic link in which the two non-adjacent nodes are not independent of the third. For example, nodes $Y_4 \rightarrow Y_2 \leftarrow Y_5$ as can see in Figure 1.3 for more information see Husmeier et al. (2005) and Radhakrishnan et al. (2013). This constraint is problematic because it creates issues in determining the causal direction of the interaction between the features in the direct acyclic graph / Bayesian network. The second constraint is that Bayesian networks are built under a static model and it is assumed that the data sample is independent during collection it in the experiment. Therefore, they are not especially convenient for analysing time slice (time points) and gene expression data. In addition, DBNs consist of more than one time slice for all variables. The third constraint of BNs, their acyclic characteristic, creates issues for modelling biological data, particularly for regulatory networks, because the biological and gene regulation have a feedback loops, whereas BNs do not have this feature (Dondelinger et al., 2013). In these kind of situations, the restrictions mentioned above can be overcomes by dynamic Bayesian networks.

We describe the dynamic Bayesian networks model mathematically as we will use it in the current study to handle time series gene expression data. We assume that having R sample

replicates (for example, microarray data for each patient), and each microarray measurement can include expression levels of p genes (variables) and T times. Therefore, Y is a matrix with size $p \times R \times T$. Generally, the array Y has entries Y_{ij}^{t} where $(j = 1, 2 \ldots, p), (i = 1, 2 \ldots, R)$ and $(t = 1, 2, \ldots, T)$. For example, $Y_{4,5}^{2}$ refer to gene number 4 for replication number 5 at time 2. Suppose that $\boldsymbol{Y_j^t}$ represents row vectors from the matrix $\boldsymbol{Y}$ and each row vector includes a sub-matrix $\boldsymbol{Y_{ij}}$ at discrete time t where $t = 1, 2, ..., T$. For example in Figure 1.4, $\boldsymbol{Y_{ij}^1}$ represents the R observation (patients) and p parameters (genes) at first time $(t = 1)$. Simply, $\boldsymbol{Y} = (Y^1, Y^2, \cdots, Y^T)$ which is $\boldsymbol{Y^1} = \boldsymbol{Y_j^1}$ where $j = 1, 2 \ldots, p$.

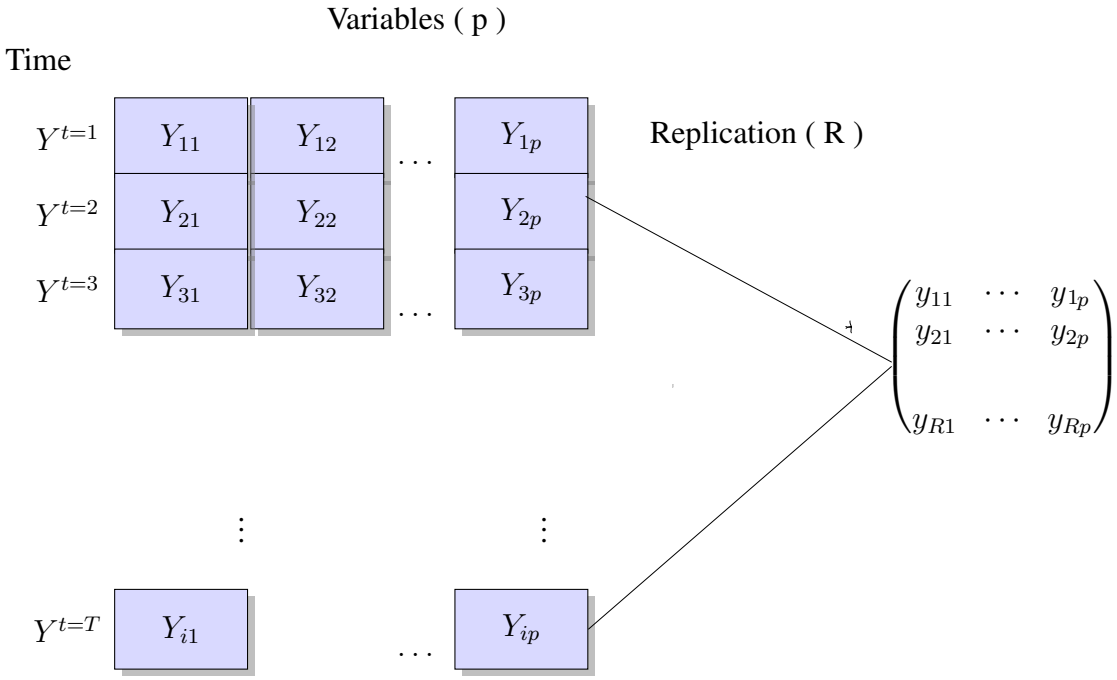


Figure 1.4: Illustration of the data for the dynamic Bayesian network model: each row vector in the matrix Y is represented as the as matrix Y_{ij} for every single time t on the left hand side.

The construction process of the DBN model is divided into two parts. Firstly, a time dependency is considered, i.e relations are invariant in time. Secondly, time-varying DBNs model, which we can compute the changing behaviour of the model for each time varied i.e both covariates and coefficients depends on time, for more detail see Kalli and Griffin (2014) for example. Our goal in the current study is the first kind, we assumed that a time slice is permitted, which the model is unchangeable over time points and the state of time t relies only on the previous time $t - 1$. Therefore, the gene regulations are modelled through this kind of modelling in Section 5.3.1.

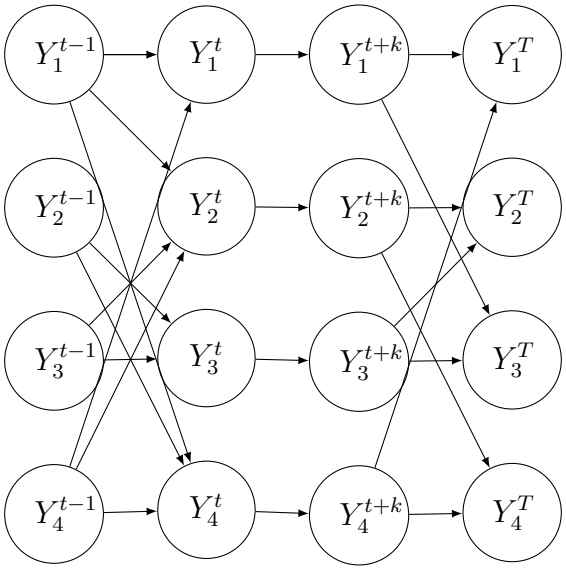


Figure 1.5: Simple example of a dynamic Bayesian network that includes four nodes, some edges (interaction between nodes) and different times slices, which displays graph have directed acyclic for most nodes. Where k represent any time point except the last time T. For example, the time t only depends on the previous the time $t-1$. This diagram recreated based on idea from Radhakrishnan et al. (2013).

The structure of the conditional probability distribution is represented by $P(Y_t|Y_{t-1})$ for $t = 2, 3..., T,$ which belongs to the first order Markov model. Therefore, the joint probability is factorised as

$$P(Y^1, Y^2, Y^3, ..., Y^T) = P(Y^1) \times P(Y^2|Y^1) \times P(Y^3|Y^2) \times ..... \times P(Y^T|Y^{T-1}). \quad (1.10)$$

where $j = 1, 2 \ldots, p$ is represents the number of variables, where the Equation 1.11 describes the product of conditional dependence for each variable at time t given the previous time:

$$P(Y^t|Y^{t-1}) = \prod_{j=1}^{p} P\Big(Y_j^t|Pa_j^{t-1}(G)\Big). \quad (1.11)$$

Figure 1.5 displays simple dynamic Bayesian network (regulation networks) consisting of four nodes (genes), all of them have the a feedback loops and interactions happen with each other over multiple time t, which every time slice is conditionally dependent on the earlier time.

Therefore, this aspect typically beneficial for building model based on time series gene expression data.

The conditional probability $P(Y_t|Y_{t-1})$ is also factorized for each node (gene) given its parent nodes $\left(Y_j^t|Pa_j^{t-1}(G)\right)$ as formulated in Equation 1.11 substituting in Equation 1.10, then the factorised joint probability distribution become,

$$P(Y^1, Y^2, Y^3, ..., Y^T) = P(Y^1)\prod_{t=2}^{T} P\Bigg(Y^t|Y^{t-1}\Bigg). \tag{1.12}$$

We note that a node at time t is conditionally dependent on the previous time $t-1$ for each variables. Then, the density function below describes the DBN model,

$$P(Y^1, Y^2, Y^3, ..., Y^T) = P(Y^1)\prod_{t=2}^{T}\Bigg\{\prod_{j=1}^{p} P\big(Y_j^t|Pa_j^{t-1}(G)\big)\Bigg\}. \tag{1.13}$$

$P\Bigg(Y_j^t|Pa_j^{t-1}(G)\Bigg)$ where $Pa_j^{t-1}(G)$ represents the state vector for the parent variables (genes) of the j^{th} gene at time $t-1$ and we do not have a parent of the initial node, that is, $Pa_j^0(G) = \phi$, where ϕ represents the empty set. Similarly to Bayesian networks, the dynamic Bayesian network model process has two stages. The first stage, is inference of the model, which includes selecting an appropriate statistical distribution for the JPD and then estimating parameters. The second stage is learning structures of the parameters, that is, selecting the best range of parameters that fit the model.

Chapter 2

Shrinkage in Linear Models

2.1 Introduction

In this chapter, we illustrate some approaches that help to solve the problems of over parametrised of linear models by shrinking the coefficients. Before looking at shrinkage approaches, we describe the statistical challenges of high-dimensional data. In the previous chapter, we briefly addressed the issue of high-dimensionality which arises from a dataset when the numbers of model parameters exceed by the number of sample observations, $p \gg n$. It has become a widespread problem due to technological advances in several fields of scientific research, especially in genomics, medicine, image processing and economics (Mallick and Yi, 2013). Note that for high-dimensional data, it is impossible to fit the regression coefficients of classical statistical models which are designed to deal with $p < n$. Consequently, building a model under sparsity to estimate the coefficients is a technique widely used by researchers (Tibshirani, 1996). The idea of sparsity assumptions in regression models is that few of the coefficients are different from zero and a significant amount of coefficients close to zero (Brimacombe, 2014). In order to overcome this problem, over the last few decades, statisticians have proposed various approaches to penalize model coefficients and to estimate the sparse models. We divide these methods into two main groups which are Bayesian shrinkage priors and non-Bayesian methods (frequentist methods).

Our objective in this chapter is to present some common shrinkage approaches either Bayesian shrinkage or and non-Bayesian penalty, that have been proposed until now and how shrinkage methods can overcome issues of high-dimensionality. We characterise the scale mixture of normal distributions as fundamental tools in creating Bayesian shrinkage prior distributions and also as an alternative to the Gaussian distribution for the error model. Therefore, we present two versions of the novel proposed Bayesian shrinkage methods in the next chapter. The outline of the current chapter is as follows: in Section 2.2, we review and state some penalty methods that have been designed for frequentist statistical model, that is, non-Bayesian methods. Then, the scale mixture normal distribution (SMN) is highlighted in Section 2.3. In Section 2.4 , the most common Bayesian shrinkage methods are demonstrated, which have scale mixture normal distribution structures. In the Section 2.5, we focus on some tools for choosing shrinkage prior value and evaluating variable selection that includes general information about Bayesian variable selection and how they participate to choose the tuning hyperparameter and then interpreting the model. Then, receiver operating characteristic curves analysis for assessing the performance of choosing hyperparameter for shrinkage priors, in the last subsection we highlight an overview of posterior predictive checking in order to check the estimation process.

2.2 Common Frequentist Regularization Methods

From a frequentist perspective, we can not use standard statistical methods $n > p$, such as ordinary least squares (OLS),

$$\text{argmin}_{\beta \in \mathbb{R}^p} ||y - X\beta||^2, \tag{2.1}$$

where X represents the matrix of independent variables, y is a vector of dependent variables and β is a vector of parameters.

In order to fit the coefficients in the regression model, Equation 2.1 has to be adjusted by adding a penalty term that plays an important role in fitting the parameters in the case of having large p and small n as follows:

$$\text{argmin}_{\beta \in \mathbb{R}^p} \Big(||y - X\beta||^2 + P(\beta) \Big) \tag{2.2}$$

where $P(\beta)$ refers to the specific penalty function used to overcome the high-dimensional problem.

Several penalized likelihood approaches have been developed in order to handle high dimensionality . At the same time, these methods have been extensively applied for selecting significant variables and estimating their impact on high dimensional problems (Fan and Tang, 2013). The LASSO (least absolute shrinkage and selection operator) has an L_1-penalty and is a popular approach to reach sparse estimates, which was introduced by Tibshirani (1996). This penalty makes the model sparse and controls the estimation process which estimates some coefficients as exactly equal to zero. In addition, the penalty function in Equation 2.2 is $P(\beta) = \gamma \sum_{i=1}^{p} |\beta_i|$, where $\gamma > 0$ is a tuning parameter and $i = 1, 2, \ldots, p$.

The SCAD penalty function (Smoothly Clipped Absolute Deviation) has been proposed by Fan and Li (2001), because the LASSO penalty method is inconsistent for variable selection, due to overestimate coefficients and having one tuning parameter. The tuning parameters in this penalty approach are computed by five fold cross-validation. The SCAD penalty has an oracle property and the mode of the posterior distribution is used to analyse the results. The SCAD penalty function is described by the equation bellow:

$$P_\gamma(\beta_i)_{\text{SCAD}} = \begin{cases} \gamma|\beta_i| & \text{if } |\beta_i| \leq \gamma, \\ \frac{|\beta_i|^2 - 2c\gamma|\beta_i| + \gamma^2}{2(c-1)} & \text{if } \gamma < |\beta_i| \leq c\gamma, \\ \frac{(c+1)^2}{2} & \text{if } |\beta_i| > c\gamma. \end{cases}$$

where $\gamma > 0$ and $c > 2$ are both adjustable parameters with $c = 3.7$ suggested in (Fan and Li, 2001). For example, Abegaz and Wit (2013) constructed the graphical model based on this method and used genetic data. Although the LASSO penalty method is attractive theoretically, it has some weaknesses that are addressed by Zou and Hastie (2005). The first weakness is the difficulty in selecting a the number of predictors which exceeds the sample size due to the nature of the convex optimization procedure.

According to Zou (2006), the LASSO penalty as a variable selection method might be unreliable in some situations. He proposed an alternative penalty function to the LASSO, which is called the adaptive Lasso. It is distinguished from the LASSO by having various weights for the coefficients which is overcome the issue of over shrinkage large coefficients and selecting unre-

lated coefficients. The penalty function in Equation 2.2 becomes $P(\beta) = \sum_{i=1}^{p} \gamma_i |\beta_i|$, where γ_i is a positive tuning parameter and $i = 1, 2, \ldots, p$. Tuning parameter selection is done by utilizing two-dimensional cross-validation. The LASSO has been discussed in the context of high-dimensional graph and asymptotic consistency results by Meinshausen and Bühlmann (2006). They noticed that the result of this method is inconsistent and the predictions. We display some other penalty functions in Table 2.1 that have been developed in order to shrink regression coefficients and others that can be found in the literature, due to our focus is on Bayesian shrinkage priors and to propose new version of them. We do not discuss them further due to our focus on Bayesian shrinkage.

Table 2.1: Some frequentist regularization functions and their tuning parameters.

Name of approaches	Penalty function	Tuning parameter
Ridge regression (Hoerl and Kennard, 1970)	$\gamma \sum_{i=1}^{p} \beta_i^2$	γ
Elastic net (Zou and Hastie, 2005)	$\gamma_1 \sum_{i=1}^{p} \lvert\beta_i\rvert + \gamma_2 \sum_{i=1}^{p} \beta_i^2$	γ_1, γ_2
Group lasso (Yuan and Lin, 2006)	$\gamma \sum_{k=1}^{K} \sqrt{\sum_{i=1}^{mk} \beta_{k_i}^2}$	γ, K number of groups (predictor covariates are divided into Kvarious groups) and mk sample size of groups

2.3 Scale Mixtures of Normal Distributions

In this section, we discuses scale mixtures of normal distributions (SMND). Our objective in this section is two-fold: firstly, to introduce some specific shrinkage prior distributions that are SMND and to explain how they participate in reducing high-dimensional data. We are going to describe some Bayesian shrinkage priors in the next section, and then our proposed Bayesian shrinkage prior is explained in Chapter 3. On the other hand, SMND are utilised as an alternative to the Gaussian assumption for building statistical models resulting in growing evidence real data applications that the normal distribution is not on appropriable choice for the error

model, especially in the presence of outliers values in data. Consider that scale mixtures of normal distributions provide a robust statistical model and they are elliptical symmetric distributions. We begin with an overview of scale mixtures of normal distributions that are used to build a statistical model. Andrews and Mallows (1974) defined scale mixtures of normal distributions for the first time as heavy-tails distributions, which play a crucial role in statistical models. Furthermore, Gonçalves et al. (2015) considered them one of the essential kinds of elliptical symmetric continuous distributions. Andrews and Mallows (1974) have defined the SMN distribution mathematically for a p-dimensional random vector as follows:

$$Y = \mu + k^{1/2}(\psi)W \tag{2.3}$$

where μ represents the location parameter for the random variable Y, and W has a normal distribution with precision equal to λ and mean equal to zero. Furthermore, $k(\psi)$ is a non- negative weight function, and ψ is its mixing non-negative random variable with probability density function $h(\psi|v)$, where v represents the location parameter or scale indexing for the distribution of ψ. The most common case that is used in the literature has $k(\psi) = \frac{1}{\psi\lambda^{-1}}$, which has a normal independent distribution (NI). This methods that $Y|\psi \sim N(Y|\mu, \psi^{-1})$ and Y has probability density function given by the formula below:

$$P(Y|\mu, \lambda, v, \psi)) = \int_0^\infty N\big(Y|\mu, \psi^{-1}\lambda^{-1}\big)h(\psi|v)d\psi. \tag{2.4}$$

There are various kinds of symmetrical classes of SMND that have a heavier tailed distribution than the normal distribution, which is considered as a suitable choice of the function $h(o|\nu)$ in Equation 2.3. It has been noticed that in the special case $Var(\psi) = 0$, SMND retrieves the usual normal distribution. In the following, we are going to present some different types of an important scale mixture normal distributions that have a heavy-tailed distribution for constructing a statistical model. We will use a SMND within linear regression the following chapters.

1 - The Variance Gamma Distribution

Madan and Seneta (1990) developed the Variance Gamma (VG) distribution for modelling for uncertainty in share market returns. The VG distribution has heavy-tails when contrasted with normal distribution. It has parameter $\nu > 0$, that is indicates to as the degrees of freedom and controls the shape of the VG distribution. The SMND density

representation of VG is similar to the Student-t distribution. The hierarchical structural of the VG distribution is expressed as follows:

$$Y|\psi \sim N\big(Y|\mu, \psi^{-1}\lambda^{-1}\big), \quad \psi|\nu \sim \text{IG}\bigg(\psi\bigg|\frac{\nu}{2}, \frac{\nu}{2}\bigg). \tag{2.5}$$

where ψ is positive and has the inverse gamma distribution (IG) with the rate and shape parameters both equal to $\frac{\nu}{2}$, and also $\nu > 0$. The IG distribution has the following form:

$$\text{IG}(X|a, b) = \frac{b^a}{\Gamma(a)} X^{-(a+1)} \exp\bigg(-\frac{b}{X}\bigg).$$

2 - The Student-t Distribution

The Student-t distribution is one of the appropriate choice for an alternative robust distribution instead of the normal distribution (West, 1987). The parameters of this distribution are the location μ, the scale λ^{-1} and the degrees of freedom ν. The probability density function of Student-t distribution as SMND can be expressed as in the following equation:

$$P(Y|\mu, \lambda, \psi) = \int_0^\infty N\big(Y|\mu, \psi^{-1}\lambda^{-1}\big) \quad \text{Ga}\bigg(\psi\bigg|\frac{\nu}{2}, \frac{\nu}{2}\bigg) d\psi. \tag{2.6}$$

Furthermore, this SMND can be expressed in terms of hierarchical distributions as in Equation (2.7):

$$Y|\psi \sim N\big(Y|\mu, \psi^{-1}\lambda^{-1}\big), \quad \psi \sim \text{Ga}\bigg(\psi\bigg|\frac{\nu}{2}, \frac{\nu}{2}\bigg), \tag{2.7}$$

where ψ has the gamma distribution with rate and shape both equal to $\frac{\nu}{2}$, and $\nu > 0$. This distribution is used widely as robust statistical model and was applied to gene regulation networks (Morrissey et al., 2010). We are going to utilise this distribution instead of Gaussian assumption of measurements error model in first part of Chapter 5, in order to build robust Bayesian regression model with measurements error model.

3 - The Slash Distribution

The slash distribution is one of the heavy-tailed distributions compared to the normal distribution that was proposed by Rogers and Tukey (1972). In a particular case, when $\nu \to \infty$, the normal distribution is obtained. The hierarchical form of the slash distribution is

expressed as:

$$Y|\psi \sim N\big(Y|\mu, \psi^{-1}\lambda^{-1}\big) \quad , \quad \psi \sim \text{Be}\big(\psi|\nu, 1\big). \tag{2.8}$$

where ψ has the beta distribution with the rate parameter equal to ν and shape parameter equal to 1.

In this section, we demonstrated different prior distributions that used instead of normal assumption in a regression model. These distributions belong to a family of distributions called scale mixture of normal distribution. Hence, we are going to describe some popular Bayesian shrinkage prior distribution namely, normal gamma (NG), Bayesian lasso, normal exponential gamma (NEG) and horseshoe prior distribution, among others in the next section. These proposed scale mixture normal distributions are used as Bayesian shrinkage prior distribution to obtain sparse coefficients in a statistical model.

2.4 Bayesian Shrinkage Approaches

In the context of Bayesian statistics, a number of shrinkage prior distributions have been suggested as an alternative to classical penalty methods in Section 2.2. From the Bayesian perspective, one shrinkage regularization method has a scale mixture of normal distribution in which the mean is equal to zero and the precision has a positive mixing distribution. These kinds of distributions are considered absolutely continuous priors (Mallick and Yi, 2013). The main property of SMND is to provide heavier tail behaviour that is beneficial to shrink those regression coefficients zero values. In this section, we are going to state some common Bayesian shrinkage methods in order to know how these kinds of shrinkage methods are developed. Then, we will apply and compare some of them to our proposed approaches in the following chapter.

The first Bayesian shrinkage prior utilised to solve the problem over parametrised regressions is the Bayesian LASSO which was proposed by Park and Casella (2008) and Hans (2009). The Bayesian LASSO prior is equivalent to a Laplace or the double exponential prior distribution; which is symmetrical and unimodal distribution that has a higher peak around the origin value compared to the Gaussian distribution. It is utilised to estimate regression coefficients based on

the posterior mode. The hierarchical structure of the Bayesian LASSO is as follows:

$$\begin{aligned} \beta_i|\psi^{-1}, \gamma_i^2 &\sim N(\beta_i|0, \psi^{-1}\gamma_i^2) \\ \gamma_i^2, |\tau &\sim Exp(\gamma_i^2|\tau), \end{aligned} \tag{2.9}$$

where β_i, indicates the i^{th} regression coefficients $(i = 1, 2, \ldots, p)$ with prior mean equal to zero and variance ψ^{-1} , γ_i^2 is i^{th} the hyperparameter of the scale mixture of Gaussian distributions and τ is tuning parameter of Bayesian lasso.

The horseshoe prior distribution is a type of scale normal mixture distribution proposed by Carvalho et al. (2010). They argued that the horseshoe prior has three important features which are sparsity, robustness, and tractability. It has heavy tails like the Cauchy distribution. Moreover, this shrinkage prior has both global and local shrinkage parameters which were not the case with previous shrinkage approaches. Empirical Bayes (Efron et al., 2001) EB is a method of statistical inference from the data observations to choose which the value of the hyperparameter prior that maximum marginal likelihood. It is utilised to benchmark the approach proposed in their paper. The horseshoe methods can also be formulated hierarchically as follows:

$$\beta_i|\psi_i, \theta \sim N(\beta_i|0, \psi_i^{-1}\theta), \psi_i^{-1/2} \sim C_+(0, 1), \tag{2.10}$$

where ψ^{-1} represents the variance, $C_+(0, 1)$ refers to the half-Cauchy distribution and θ represents the global shrinkage parameter. Even though this prior dose not have an exact analytical density function, it outperforms other prior distributions due to having a predictor-specific local shrinkage component and a global shrinkage component (Carvalho et al., 2009; Datta and Ghosh, 2013).

Griffin and Brown (2010) developed and generalised a kind of continuous prior called the double exponential distribution which is used in the Bayesian LASSO prior. To expand on this, they proposed a kind of modified Bayesian lasso and it has a particular form of normal scale mixture distribution with mean equal to zero. It is known as the normal-gamma prior distribution. They noticed that the double exponential prior distribution is inflexible and has negative effects on parameter estimation because it has a single hyperparameters formulation. This issue leads to shrinking some truly non-zero coefficients which are forced much closer to the value of zero. Indeed, the analysis process will have erroneous results and poor predictive performance. Grif-

fin and Brown (2010) did not use empirical Bayes to estimate the tuning parameters because the posterior distribution is highly multi-modal which make it very hard to apply the method. The hierarchical structure of the normal gamma prior distribution is as follows:

$$\begin{aligned} \beta_i &\sim N(\beta_i|0, \psi_i^{-1}) \\ \psi_i &\sim \text{Ga}\big(\psi_i\big|\omega, \frac{1}{2\gamma^2}\big), \end{aligned} \tag{2.11}$$

where ψ_i^{-1} is the variance of normal distribution. The variance of β_i is $\text{Var}(\beta_i|\omega, \gamma) = 2\omega\gamma^2$. The marginal density function $\pi(\beta_i)$ is given by:

$$\pi(\beta_i) = \frac{1}{\sqrt{\pi} 2^{\omega-\frac{1}{2}} \gamma^{\omega+\frac{1}{2}} \Gamma(\omega)} |\beta_i|^{\omega-1/2} K_{(\omega-1/2)}\left(\frac{|\beta_i|}{\omega}\right), \tag{2.12}$$

where $i = 1, 2, \ldots, p, K$ represents the modified the Bessel function of third Kind and γ, ω are tuning parameters.

After that, the normal exponential gamma (NEG) prior distribution was developed by Griffin and Brown (2011) as an alternative to the Bayesian lasso for regression problems. It has been used as a version of a shrinkage prior distribution in Bayesian regression especially in the case of estimating the regression coefficients of high-dimensional data. As well, it is a generalised lasso prior that has an extra layer; this is called a family hyper-lasso penalty functions (Griffin and Brown, 2011). Moreover, the reasons for using the NEG prior are the finite spike at zero and its heavy tails when the value of the shape parameter ω is small. We are going to explain this shrinkage prior in detail in the next chapter, so we can compare it to our proposed Bayesian shrinkage prior that belongs to a similar scale mixture of Gaussian distributions family. The density function of the NEG distribution is formulated as follows:

$$NEG(\beta_i|\gamma, \omega) = k \exp\left\{\frac{1}{4}\beta_i^2\gamma^2\right\} D_{-2\omega-1}(\gamma|\beta_i|), \tag{2.13}$$

where κ is a constant that satisfies $k = \frac{\omega\gamma}{\pi^{1/2}} 2^{\omega}\Gamma(\omega+1/2)$, $D_{-2\omega-1}(\gamma|\beta_i|)$ represents the parabolic cylinder function, and the parameters ω and γ control the shape of the tails and the scale, respectively. It includes different parameters for the scale distribution of the various regression coefficients.

Armagan et al. (2011) proposed for the first time a novel kind of zero mean and scale mixture of normal distribution based on the generalised three parameter beta distribution. This prior distribution is called the generalised beta mixture of Gaussian distribution, has heavy tails and has a complex form that makes this shrinkage prior more flexible so that it dose not over-shrinkage large coefficients. The hierarchical formation of it is given by:

$$\begin{aligned} \beta_i|\psi_i \sim N\big(\beta_i|0, \psi_i^{-1} - 1\big), \psi_i \sim \text{TPB}(\psi_i|a, b, \rho) \\ \text{TPB}(\psi_i|a, b, \rho) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \rho^b \psi_i^{b-1} (1-\psi_i)^{a-1} [1 + (\rho - 1)\psi_i]^{-(a+b)}, \end{aligned} \tag{2.14}$$

with $a, b, \rho > 0$, Armagan et al. (2011) suggested the better choice that $a \in (0, 1]$, $b \in (0, 1]$ and fixing ρ in the case when it is utilised on high-dimensional data. The horseshoe shrinkage prior is obtained when $a = b = \frac{1}{2}$ and $\rho = 1$.

Also, the double Pareto distribution is generalised as a hierarchical Bayesian shrinkage prior method by Armagan et al. (2013a). It is considered a type of scale mixture of normal distributions. Note that normal scale mixture distribution is a much more commonly used term, regards a recent advanced shrinkage approach. Furthermore, the density plot of the generalised double Pareto prior is similar to the density plot of the Laplace distribution at zero and it has tail behaviour similar to the Student-t distribution. They claim that it is not too hard to fit the coefficients in the statistical model and that inference can be done using a Gibbs sampler. The generalised double Pareto distribution (GDP) is described by the following formula:

$$f(\beta_i|\gamma, \alpha) = \frac{1}{2\gamma} \left(1 + \frac{|\beta_i|}{\gamma\alpha}\right)^{-(\alpha+1)}, \tag{2.15}$$

where β_i has the generalised double Pareto distribution, the prior for coefficients in the regression model and it has two parameters: $\gamma > 0$ is the scale parameter and $\alpha > 0$ is controlling the shape. In addition, the following shorthand notation is used,that is, $\beta_i \sim \text{GDP}(\gamma, \alpha)$.

The hierarchical structure of this prior distribution as a normal scale mixture distribution is given by Equation 2.16 and it has a similar hierarchical structure to the NEG prior distribution of Griffin and Brown (2011) which as a special case and leads to marginal distribution

where GDP($\gamma = \frac{\omega}{\alpha}, \alpha$).

$$\beta_i|\psi_i \sim N(\beta_i|0, \psi_i^{-1}), \psi_i^{-1} \sim \text{Exp}\left(\psi_i \middle| \frac{\zeta_i^2}{2}\right), \zeta_i \sim \text{Ga}(\zeta_i|\gamma, \omega). \tag{2.16}$$

Leng et al. (2014) proposed the Bayesian adaptive LASSO (BaLasso) methods, which generalised the Bayesian LASSO similar to Equation 2.13 where an EM-algorithm was used for inferring the parameters. Leng et al. (2014) used MCMC for updating the model parameters. Furthermore, they also take into account variable selection by using this prior distribution. The hierarchical construction of the BaLasso prior distribution is as follows:

$$\begin{aligned} \beta|\gamma^2 \sim N\big(\beta|0, D_\gamma\big), \gamma^2|\theta^2 \sim \prod_{i=1}^{p} \frac{\theta^2}{2} e^{-\frac{\theta^2\gamma^2}{2}} \\ \theta^2 \sim \prod_{i=1}^{p} (\theta_i^2)^{\omega-1} e^{-\delta\theta_i^2}, \end{aligned} \tag{2.17}$$

where D_γ is the diagonal matrix of $(\gamma_1^2, \gamma_2^2, \ldots, \gamma_p^2)$, θ^2 and γ^2 are $p \times 1$ vector of hyperparameters and both (ω, δ) represent the tuning parameters, where usually ω is fixed and δ is updated. Rajaratnam and Sparks (2015) have extended the Bayesian LASSO (Park and Casella, 2008) because they believed that convergence of the MCMC for the Bayesian LASSO is slow when dealing with a number of parameters that is larger than the sample size, that is, dealing with high-dimensional model. Consequently, they proposed an alternative method for reducing which is called the fast Bayesian LASSO. It is based on a blocked two-stages Gibbs sampler for updating the parameters of the Bayesian linear model, see the original paper for more details of this prior. Previously, we highlighted some popular penalty functions in order to create the model and obtain sparse estimation.

As we mentioned previously, the NEG prior distribution Griffin and Brown (2011) has been used to shrink coefficients of regression models. They claimed that their Bayesian prior includes non-convex penalized likelihoods and penalties. The NEG prior distribution includes two parameters for more flexibility in detecting significant variables in the model, and these parameters are λ which controls scale and ν which controls the shape, and heaviness. The EM-algorithm is used for estimating parameters. Consequently, Rockova et al. (2014) have proposed selecting variables in Bayesian linear regression by adding structural group information into the NEG prior (Griffin and Brown, 2011). These was achieved by allowing differential

penalization for each regression coefficient in the Bayesian LASSO. They considered an extra regression layer which relates to different penalty parameters to a group identification matrix. At the same time, they suggested that the hierarchical model can be simultaneously on two levels. Firstly, the regression layer for the continuous outcome with the predictors, and secondly, the layer for the penalty parameters with the grouping information. Furthermore, they mentioned that the smoothness was achieved at the penalty level instead of the regression coefficients. Furthermore, they noticed that using the EM-algorithm to compute their model so as to make it faster than MCMC methods, when dealing with high dimensionality.

In this thesis, we propose two novel Bayesian shrinkage prior distributions, which are the hierarchical Normal inverse Pareto (hNiP) and the Rescaled Beta hierarchical Normal inverse Pareto (ReB-hNiP), and they are considered to be a family of scale mixture of normal distribution. We will focus on these prior distributions in the following chapters. Moreover, we will use novel Bayesian shrinkage for modelling for different statistical models. To conclude our discussion about Bayesian shrinkage methods, that there are several options in the literature which will not be discussed in here.

2.5 Prior Parameter Selection

2.5.1 Bayesian Variable Selection

One of the interesting statistical issues that many statisticians have paid attention to is variable selection methods. A number of variable selection techniques have been studied from of the Bayesian perspective such as BMA- Bayesian model averaging, product space search, GVS - Gibbs variable selection (O'Hara et al., 2009), spike and slab method, among others. Frequently, for these kinds of situations, some adaptive variable selection methods are restricted in attention to choose variables that are the most important compared to others. Consequently, variable selection is considered a special case in selecting a statistical model according to (Mallick and Yi, 2013). The computational process is very time consuming of classic procedures of selecting the most important variables and, also having shortage for the unstable inherent (Breiman et al., 1996).

Spike and slab prior have become widespread in Bayesian variable selection approaches which are known as SSVS, stochastic search variable selection and was proposed by Mitchell and Beauchamp (1988). The spike-and-slab consists of two parts. The first part is the spike that focuses on its mass on zero, which permits the shrinkage of small effects towards the zero. Whilst the second part is the slab that allows mass spread widely around reasonable values for the coefficients in the regression model. In addition, the general form of the spike-and-slab prior is as follows:

$$\pi(\beta) \sim \theta\Psi_1(\beta) + (1-\theta)\Psi_0(\beta), \tag{2.18}$$

where the components of $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$ usually have Bernoulli distribution, that is, $\theta \in \{0, 1\}$ and $\Psi_0(\beta)$ represents the spike distribution for modelling and $\Psi_1(\beta)$ represents the slab distribution for modelling.

In light of the Bayesian analysis framework, MCMC is common technique that is utilised to fit model. Therefore, variable selection approaches can be applied in order to choose the most important variables. Recently, reducing high-dimensional data and selecting the best subset of variables has become easier compared with traditional methods due to advances in a number of Bayesian shrinkage prior distributions that were presented in the previous section. Besides, we can apply many criteria with shrinkage prior distributions to select appropriate prior values that lead to detecting the best choice of variables to include in a model. These criteria include the deviance information criteria (DIC), the Akaike information criterion (AIC) (AKAIKE, 1973), Bayesian information criterion (BIC) (Schwarz et al., 1978), extended Bayesian information criterion (EBIC) (Chen and Chen, 2012), DIC (Spiegelhalter et al., 2002) and WAIC (Vehtari and Gelman, 2014) among others. Selecting explanatory variables based on Bayesian shrinkage priors requires thresholding which is to distinguish which variables should be included in the model or not. This technique relies on posterior mean or median, and we will use the t-test and other tools for this purpose in the following chapters.

2.5.2 Hyperparameter Choice and Model Selection

Selecting tuning parameters in this thesis is based on two criteria which are the DIC and WAIC. Considering that the WAIC is a generalised type of DIC, but it is more convenient for the

Bayesian method to estimate the out-of-sample expectation, the computation begins with calculating the pointwise log likelihood then making an adjustment for the effective number of parameters to control for over-fitting. Consider that WAIC is flexible criteria to use in this thesis due to adopting respect to have the repeated measurement data in the model. The WAIC is given by:

$$\text{WAIC} = -2(\widehat{lpd} - \widehat{p}_{waic}). \tag{2.19}$$

Here, $\widehat{lpd}$ represents the pointwise log likelihood (log predictive density) and it is expressed as an approximation:

$$\widehat{lpd} = \sum_{j=1}^{N} \log\left(\frac{1}{K}\sum_{k=1}^{K} p(x_j|\Theta^k)\right), \tag{2.20}$$

while

$$\widehat{p}_{waic} = \sum_{j=1}^{N} \mathrm{V}_{k=1}^{K}\Big(\log p(x_j|\Theta^k)\Big), \tag{2.21}$$

where, Θ^k represents the k^th posterior draw of the parameter vector Θ, and $\widehat{p}_{waic}$ requires estimating the posterior the variance of the log predictive density for every single observation. Also, $\mathrm{V}_{k=1}^{K}\big(\log p(x_j|\Theta^k)\big)$ denotes the sample variance of log predictive densities, where k is number of MCMC iteration. We choose the minimum value of WAIC as the best model and the hyperparameters values in shrinkage prior that utilised for modelling they become the tuning parameters. We adopt WAIC criteria in the case when we handle measurement error model in Chapter 5. This have been done by adding one extra step before computing WAIC procedure, which is calculating the sum with respect to the repeated measurements R data because we have three-dimension matrix. For more details of measurements error model see Chapter 5.

The deviance information criterion, DIC (Spiegelhalter et al., 2002), derived form deviance $D(x, \Theta) = -2\log p(x|\Theta)$, and is given by the following equation:

$$\text{DIC} = \overline{D(x,\theta)} + P_D \tag{2.22}$$

where

$$P_D = \overline{D(x,\theta)} - D(x,\overline{\theta}), \tag{2.23}$$

Here $D(x,\overline{\theta})$ is the deviance at the posterior mean, while $\overline{D(x,\theta)}$ denotes the posterior mean deviance and P_D denotes the effective the number of parameters.

Moreover, we used another criteria, the extended BIC (Chen and Chen, 2012) that takes into consideration high-dimensionality $(n < p)$.

$$\text{EBIC}_\gamma(s) = -2l(\hat{\beta}_s) + v(s)\log(n) + 2v(s)\gamma\log(p). \tag{2.24}$$

where $l(\hat{\beta}_s)$ represents the log-likelihood function based on $\hat{\beta}_s$, the set of estimated coefficients identified as non-zero, and is given by

$$l(\hat{\beta}_s) = \left(\frac{-\hat{\lambda}}{2}\sum_{j=1}^{n}(Y_j - X_j\beta_s)^2 + \frac{N}{2}\log(\hat{\lambda})\right).$$

Here $\hat{\lambda}$ represents the estimated precision of the model, $v(s)$ refers to the number of non-zero coefficients that were selected and s is the set of those variables and coefficients that satisfy $|\frac{\hat{\beta}_i}{sd_{\hat{\beta}_i}}| > 1.9,6$ where $sd_{\hat{\beta}_i}$ represents the standard division. Although, $\gamma > 1 - \frac{1}{2k}$ where $0 < k < 1/3$ is a methods for choosing a consistent value that makes the false discovery rate (FDR) close to zero when the sample size is large, the advisable choice $\gamma = 0.5$ (Chen and Chen, 2012), because $\gamma = 0$ corresponds to classical BIC. We chose the model that has the minimum value of EBIC_γ and considered the prior values used to fit this model as tuning parameters.

2.5.3 Receiver Operating Characteristic Curves Analysis

The ROC curve used to visualize and measuring the numerical procedure for assessing the performance of binary classifier algorithms. It is widely used for comparing the performance of different classification in several fields (Fawcett, 2006), consider that it is trade-offs between false positives and true positives. The first step for creating the ROC curve to calculate true positive (TP), false positive (FP), true negative (TN) and false negative (FN). In our application, TP refers to the number of non-zero coefficients calculated by a method that are also non-zero in the model that generated the simulated data. The value FN represents the number of coefficients that are measured to be non-zero, but they are actually zero in the true model that generated simulated data and this case corresponds to the Type II error. Consider that FP is Type I error, because the value FP refers to the number of the coefficients that are computed to

be non-zero but, they are zero in the model that generated simulated data. Table 2.2 displays the relations between predictions (by specific model) and the observed values (simulated data). We aim to increase the values of sensitivity and specificity that leads to increase the efficiency of our model and the area under the curve AUC.

Table 2.2: Classification table for ROC curves, where positive is denoted by po and negative by ne.

		True-class	
		po	ne
Predicted-class	po	True Positives (TP)	False Positives (FP)
	ne	False Negatives (FN)	True Negatives (TN)

The second step is calculating two key statistics which are the true positive rate (TPR) and the false positive rate (FPR). The true positive rate (TPR) is also called the sensitivity and the false positive rate (FPR) is defined as 1-Specificity, where

$$\text{TPR (Sensitivity)} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}},$$

$$\text{Specificity} = \frac{\text{True Negatives (TN)}}{\text{True Negatives (TN)} + \text{False Positives (FP)}},$$

and

$$1 - \text{Specificity} = \text{FPR}$$

Then the ROC curve is produced by plotting the FPR on the x-axis against TPR or Sensitivity in y-axis. We will use such analyses in the following chapters. Furthermore, we are going to apply the measures in this section together with some criteria that were presented in the section 2.5.2, in order to evaluate the performance of our proposed hierarchical Bayesian shrinkage priors.

2.5.4 Posterior Predictive Checking

In this section, we give an overview of posterior predictive checking in order to detect any outliers between posited model and the data for more knowledge see the chapter six of Gelman et al. (2013). In practice, we will discuss how posterior predictive checking can be used. When

we fit a model in order to estimate parameters, we are usually interested in how well the model fits the data and in which aspects of the data generation process in nature are not captured by the model (Gelman and Shalizi, 2013). One way to measure the discrepancy between the data and the fitted model is to define test quantities. This is a scalar summary of the data and parameters which can be used as a standard if we would like to compare the data with predictive simulations. Test statistics can be generalized in order to allow dependence on the model parameters. By using this, we can directly summarize discrepancies between data and fitted model (Gelman and Shalizi, 2013). It has been proposed by Gelman et al. (2005) that checking models can be done using complete data. However, they also considered more general settings including latent variables (as missing data), non-ignorable missingness and ignorable missing data (Griffiths et al., 2010) .

The general posterior predictive distribution function represents as follows.

$$p(\widetilde{y} \,|\, y) = \int p(\widetilde{y} \,|\, \Theta)\, p(\Theta \,|\, y)\, d\Theta.$$

where y is vector of observed data with size $n \times 1$, Θ is set of unknown parameters in the model. While $\widetilde{y}$ is matrix of posterior predictive distribution after estimated parameters for each iteration and it represent by y^{rep}.

A test quantity is a function of the replicated data and original data with some additional parameters. The discrepancy in the between the test quantities can be evaluated using the replicated data and original data by calculating a $p-$value and watch for extreme tail area $p-$value. The selection of test quantity and appropriate posterior predictive distribution requires a consideration of the type of estimates required for the problem being considered. Extreme $p-$ value for the test quantities indicate areas of failure for the model which can be addressed by extending the model, or if it is ignored (Gelman et al., 2013). One of the popular $bayesplot$ package in R programming constructed by Gabry (2017), that included the coded of posterior predictive model checking which they based on the study for Gabry et al. (2017). Therefore, we are going to utilised this posterior predictive distribution in the following application chapters.

Chapter 3

A Novel Hierarchical Prior

3.1 Introduction

The idea behind using SMN as shrinkage priors is to allow for different amounts of mass concentration around the origin by varying the prior precision. Thus, the prior on the precision becomes the vehicle for controlling shrinkage. For instance, the Bayesian LASSO (Park and Casella, 2008) uses an exponential on the variance, thus allocating large prior probability to small (large) values of the prior variance (precision). The key intuition behind our approach is that values of the precision above certain threshold will make no practical difference to the amount of shrinkage achieved, so we would like to set an upper bound to the support of its prior.

One option would be simply to use a truncated exponential from above, but arguably a distribution with lighter tails would be more amenable for differential shrinkage. Thus, we propose to use an inverse Pareto distribution, in Equation 3.2, with upper bound (scale parameter) $b_0 > 0$ and shape parameter $\kappa > 0$. The shape parameter controls the amount of mass toward the upper bound of the distribution, with larger values assigning more mass at the upper end. In this thesis, we will fix b_0 and put a prior on κ. Given that $\kappa > 0$, a natural option amenable to Gibbs sampling is a Gamma distribution, the hNiP distribution.

Alternatively, if we wanted to have differential shrinkage, we would want a prior which assigns

high prior probability to small values of κ for those coefficients away from zero and large prior probability to large values of κ for coefficients close to zero. Our proposal is thus a Beta distribution with both parameters smaller than 1 (eg. $1/2, 1/2$), re-scaled to $(0, e_0)$; the ReB-hNip distribution.

In Section 3.2, we formalise these ideas and provide properties of these new distributions. Given that we will compare our method with some popular alternatives used in practice, we describe the NEG distribution of Griffin and Brown (2011) in Section 3.3, as it can be seen as the Bayesian version of the LASSO.

3.2 Hierarchical Structure of the Normal Inverse Pareto Distribution

As we previously described, there is a wide range of studies related to the topic of a scale mixture of Normal prior distributions; the main idea is to utilise the hierarchical structure of the prior distribution.

$$\pi(\beta) = \int N(\beta|0, \psi^{-1})\, \pi(\psi) d\psi, \tag{3.1}$$

where ψ represents the precision of the normal distribution, has a positive mixing distribution and is considered to be the second layer of hyperparameter of a shrinkage function. The fundamental idea behind using the SMN distribution is to have more mass of marginal distribution set around the origin, and also the tails of this hierarchical distribution are heavier than the tails of Student-t and normal distributions. Therefore, in this thesis, we assume that we are dealing with the prior distribution of the precision, $\pi(\psi)$, which is a mixing distribution and we want to choose a form for it that leads to $\pi(\beta)$ having high mass around zero and heavy tails, thus addressing over-shrinkage of large values of the regression coefficients.

We propose using the inverse Pareto distribution (Bernardo and Smith, 2000), which has been used in literature, under different names such as "power function" distribution as used by Lutful Kabir and Ahsanullah (1974)and Kifayat et al. (2012); and "inverse Pareto" distribution as

used by (Dallas, 1976; Pawlas and Szynal, 2000). The power function distribution is a flexible distribution which can utilised for constructing models for several types of data. Such data are related to the lifetime, reliability analysis, income, insurance and electric device for more information reader can see Shakeel et al. (2016) and Moothathu (1993). Other hand, inverse Pareto distribution has a heavy-tailed behaviour which is controlled by the shape parameter. Broadly speaking one could obtain the inverse Pareto distribution from Pareto distribution as a special case, these are one-to-one transformation. $X \sim Pa(X|a,b)$ then $\frac{1}{X} = y \sim IP(y|a,b)$, where Pa refers to Pareto distribution and IP is inverse Pareto distribution. Ni et al. (2015) used inverted Pareto as an absolutely continuous prior with gamma distribution to build nonlinear DAG model and fitting smoothing parameter. Furthermore, Morrissey et al. (2011) are used inverted Pareto as a prior distribution in the nonlinear sparse model.

The inverse Pareto distributions is follows,

$$\pi(X) = \text{iP}(X|\kappa, b_0) = \kappa\, b_0^{-\kappa}\, X^{\kappa-1}\ ,\ X \leq b_0, (\kappa, b_0 > 0), \tag{3.2}$$

where X is a random variable with shape parameter κ and scale parameter b_0. The mean of inverse Pareto distributions is $E(X) = \frac{\kappa\, b_0}{\kappa+1}$ and the variance is $Var(X) = \frac{\kappa b_0^2}{(\kappa+1)^2(\kappa+2)}$.

The main reason of using the NiP shrinkage prior distribution is that it has a special feature, the parameter space of b_0 controls shrinkage coefficients and also it bounds the precision from above, that is, upper bound). However, it has only one parameter that controls the different amount of mass onward upper bounds and lower-bounds, this is the fundamental feature for using it. Furthermore, it permits controlling the quantity of mass allocated toward the upper bound. The prior density is decreasing if $\kappa < 1$ which leads to the most of mass being close to zero. Figure 3.1 shows how different values of the shape parameter effect the shape of the distribution. In contrast, most of the mass is close to the upper bound value b_0 if $\kappa > 1$ due to the increasing prior density,

We find the marginal distribution of the regression coefficient β_i from the hierarchical structure of the normal scale mixture inverse Pareto distribution as follows:

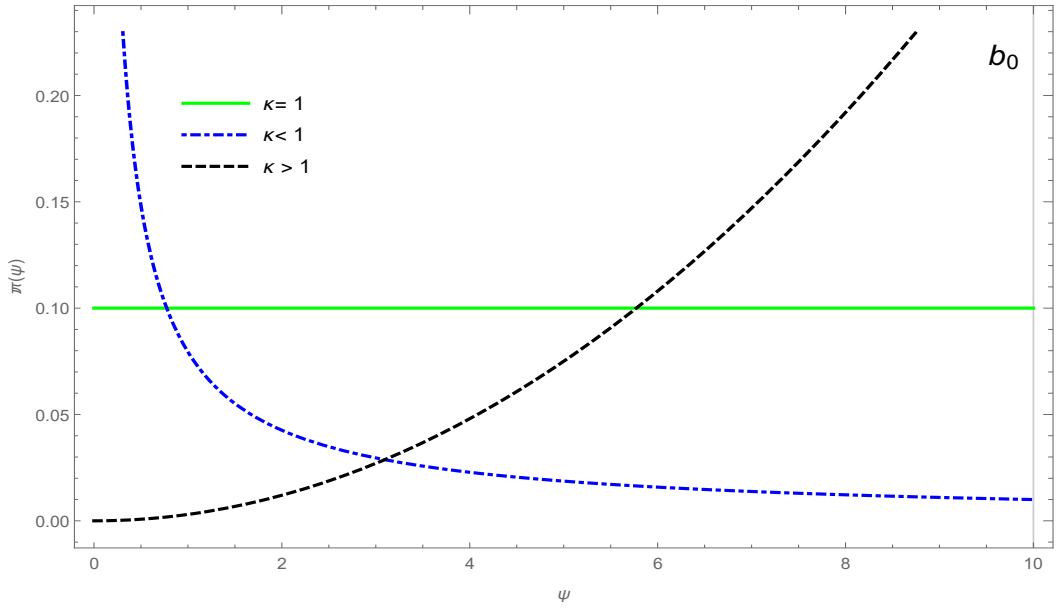
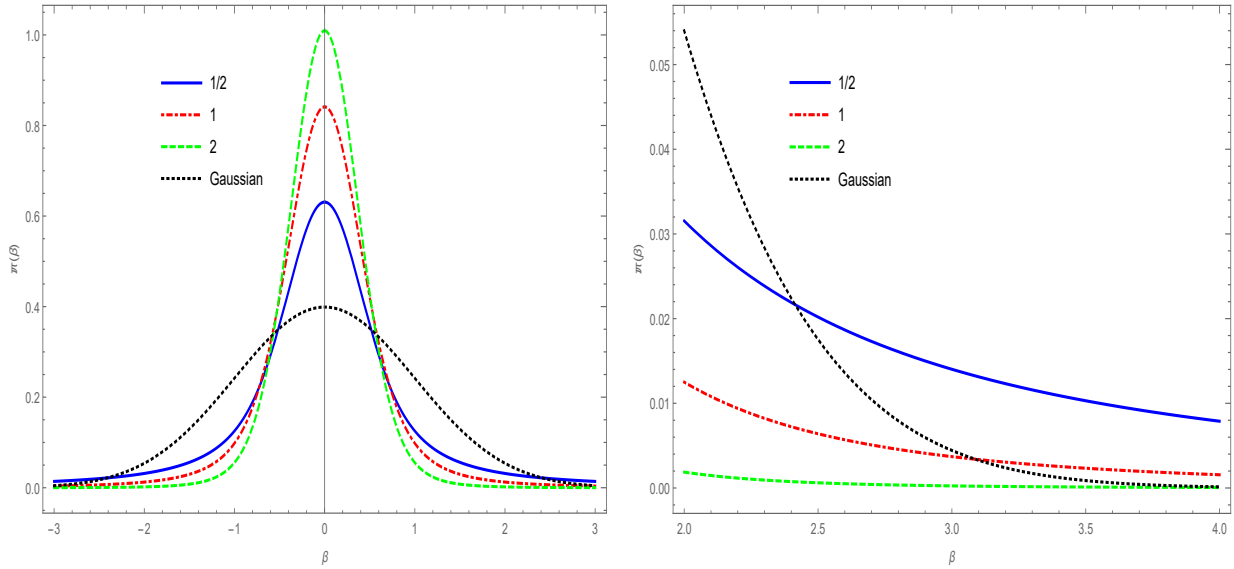


Figure 3.1: Density of the inverse Pareto distribution $\pi(\psi)$ with the values $b_0 = 10, \kappa = 1, \kappa > 1$ and $\kappa < 1$.

$$\begin{aligned} \text{NiP}(\beta_i|\kappa, b_0) &= \int_0^{b_0} N(\beta_i|0, \psi^{-1})\text{iP}(\psi|\kappa, b_0)d\psi = \int_0^{b_0} \frac{\psi^{\frac{1}{2}}\, e^{-\frac{1}{2}\psi\beta_i^2}}{\sqrt{2\pi}}\, \kappa\, b_0^{-\kappa}\psi^{\kappa-1}d\psi \\ &= \frac{2^\kappa \kappa\, b_0^{-\kappa}}{\sqrt{\pi}}|\beta_i|^{-(2\kappa+1)}\left(\Gamma\left[\kappa+\frac{1}{2}\right] - \Gamma\left[\kappa+\frac{1}{2}, \frac{b_0\beta_i^2}{2}\right]\right). \end{aligned} \tag{3.3}$$

It is clear that Equation 3.3 has an analytical closed form, but that is also includes the incomplete Gamma function $\Gamma(.,.)$ (Abramowitz et al., 1964). Figure 3.2 shows plots of the NiP distribution for several values of κ compared with Gaussian distribution for a fixed parameter b_0 and their tail behaviour in the right hand panel. It is clear that small values of κ lead to heavier tails. Figure 3.3 shows plots of the NiP distribution for different values of b_0 and fixed κ in order to demonstrate how the value of the upper bound effects the our proposed shrinkage prior. The Normal inverse Pareto (NiP) distribution has mean zero if $\kappa > \frac{1}{2}$ and precision $\frac{b_0(\kappa-1)}{\kappa}$ for $\kappa > 1$.

If we fix κ, then we fix the shrinkage, which is not advisable in practice. Thus we put a prior on κ and fix b_0. Instead of setting priors for both parameters of the inverse Pareto distribution,



(a) Density plots for NiP with $b_0 = 10$ and different values of κ.

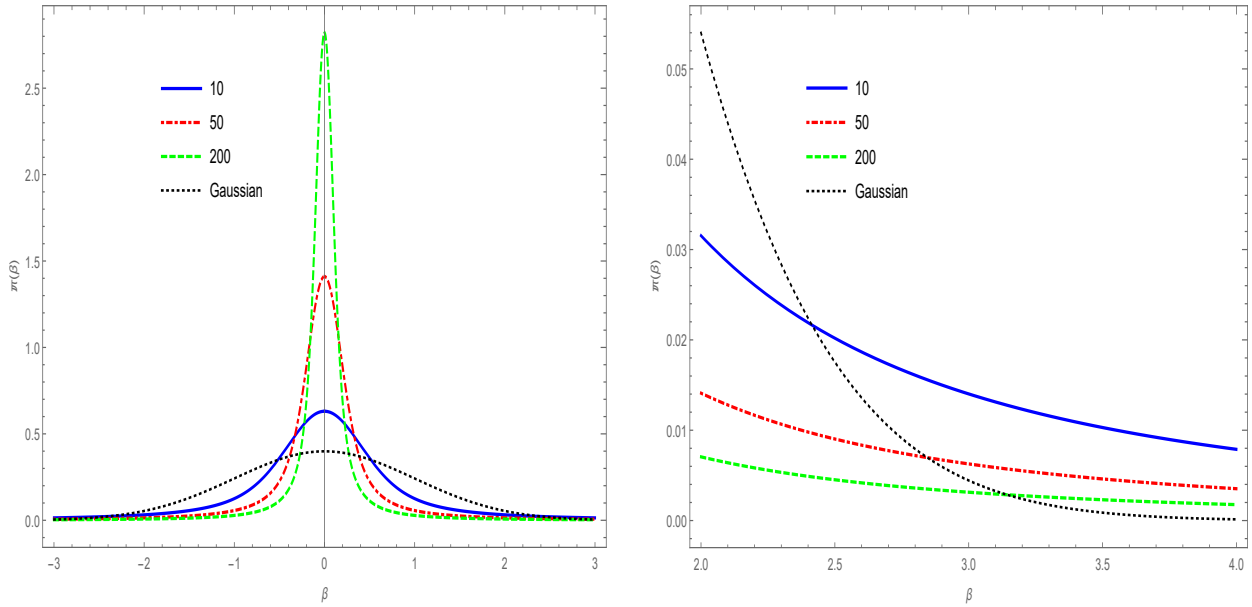
(b) The upper tail behaviour of NiP with $b_0 = 10$ and different values of κ.

Figure 3.2: Density plots for the Normal inverse Pareto (NiP) distribution for different values of κ and the black dotted line represents the Gaussian distribution

we will conduct a sensitivity analysis to discover which values of the threshold b_0 should be used. We propose two kinds of prior distribution for the hyperparameter κ, Gamma distribution and the Rescaled Beta to bound the shape parameter. In Section 3.2.1, we focus on the Gamma distribution as hyperprior for κ and then the Rescaled Beta distribution in Section 3.2.2.

3.2.1 Normal Inverse Pareto Distribution

Here, we consider the prior for the hyperparameter unbounded. The natural choice is a Gamma distribution, $Ga(c, d)$, where c represents the shape and d refers to the rate parameter. Then, we can obtain the hierarchical-NiP (hNiP) prior distribution as follows:



(a) Density plots for NiP with different values of b_0 and $\kappa = \frac{1}{2}$.

(b) The tail behaviour of NiP for different values of b_0 and $\kappa = \frac{1}{2}$.

Figure 3.3: Density plots for the Normal inverse Pareto (NiP) distribution for different values of b_0 and the black dotted line represents the Gaussian distribution.

$$\begin{aligned} \text{hNiP}(\beta|b_0, c, d) &= \int\int N(\beta|0, \psi^{-1})\ \text{iP}(\psi|\kappa, b_0)\ \text{Ga}(\kappa|c, d)\ d\kappa\ d\psi \\ &= \int_0^{b_0} \frac{\sqrt{\psi} c d^c e^{-\frac{\psi\beta^2}{2}} \left(\log\left(\frac{b_0}{\psi}\right) + d\right)^{-(c+1)}}{\sqrt{2\pi}} d\psi. \end{aligned} \tag{3.4}$$

Equation 3.4 dose not have an analytical closed form. It is still centered around the origin. The mean is follows,

$$\begin{aligned} E(\beta) &= E[E[E[\beta|\psi]]] \\ E(\beta) &= E[E[0]] = 0, \end{aligned} \tag{3.5}$$

because $\beta|\psi \sim N(\beta|0, \psi^{-1})$, therefore $E[\beta|\psi] = 0$, and the variance is

$$\begin{aligned} Var[\beta] &= E[Var[\beta|\psi]] + Var[E[\beta|\psi]] \\ &= E[E[Var[\beta|\psi]] + Var[E[\beta|\psi]]] \\ &= E[E[\frac{1}{\psi}]] + 0 \\ \text{where } E[\frac{1}{\psi}] &= \int_0^{b_0} \frac{1}{\psi} \kappa b_0^{-\kappa} \psi^{\kappa-1} d\psi = E[\frac{1}{b_0} \frac{\kappa}{\kappa - 1}] \text{ iff } \kappa > 1 \\ Var[\beta] &= \frac{1}{b_0} \frac{d^c}{\Gamma(c)} \int_0^{\infty} \frac{\kappa}{\kappa - 1} \kappa^{c-1} \exp^{-\kappa d} d\kappa, \end{aligned} \tag{3.6}$$

where $\psi \sim \text{iP}(\psi|\kappa, b_0)$ and $\kappa \sim Ga(\kappa|c, d)$, where the variance dose not exist, that is means the variance of β is infinite.

It may be interesting to consider the marginal mixing distribution of ψ, follows:

$$\begin{aligned} \pi(\psi|b_0, c, d) &= \int_0^{\infty} \text{iP}(\psi|\kappa, b_0)\ \text{Ga}(\kappa|c, d)\ d\kappa \\ &= \int_0^{\infty} \kappa b_0^{-\kappa} \psi^{\kappa-1}\ \frac{d^c}{\Gamma(c)} \kappa^{c-1} e^{-d\kappa} d\kappa \\ &= \frac{c}{d}\ \psi^{-1} \left(1 + \log\left(\frac{b_0}{\psi}\right)^{1/d}\right)^{-(c+1)} , \psi \leq b_0. \end{aligned} \tag{3.7}$$

It is clear that the parameter space of ψ is bounded from above. In the following formula, we present the moments of this mixing distribution for several orders, in particular,

$$E(\psi|b_0, c, d) = c\ e^d b_0 E_{c+1}(d) \qquad \text{and} \qquad E(\psi^2|b_0, c, d) = c\ e^{2d}\ b_0^2 E_{c+1}(2d). \tag{3.8}$$

The general form for m^{th} order moments is

$$E(\psi^m|b_0, c, d) = c\ e^{md} b_0^m E_{c+1}(md), \tag{3.9}$$

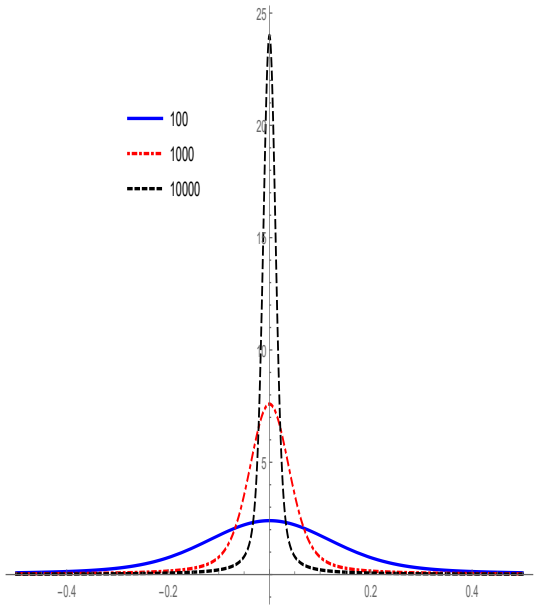
where

$$E_m(z) = \int_1^{\infty} t^{-m} e^{-zt} dt\ . \tag{3.10}$$

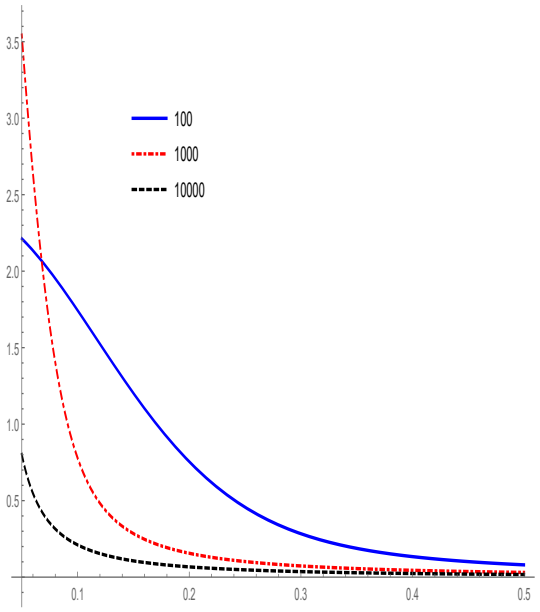
$E_m(z)$ represents the function called the exponential integral (Abramowitz et al., 1964). The hierarchical structure of the hNiP prior distribution is as follows:

$$\begin{aligned}\pi(\beta_i, \psi_i, \kappa_i |\, b_0, c, d) &= N(\beta_i|0, \psi_i^{-1})\ \text{IP}(\psi_i|\kappa_i, b_0)\ \text{Ga}(\kappa_i|c, d) \\ &= \frac{\psi_i^{\frac{1}{2}}\ e^{-\frac{1}{2}\psi_i\beta_i^2}}{\sqrt{2\pi}}\ \kappa_i b_0^{-\kappa_i}\psi_i^{\kappa_i-1}\frac{\kappa_i^{c-1}\ d^c\ e^{-\kappa_i d}}{\Gamma(c)} \\ &= \frac{b_0^{-\kappa_i}\ \kappa_i^c\ \psi_i^{\kappa_i-\frac{1}{2}}\ d^c\ e^{-\kappa_i d-\frac{\psi_i\beta_i^2}{2}}}{\sqrt{2\pi}\Gamma(c)}.\end{aligned} \tag{3.11}$$

Figures 3.4a and Figures 3.4b illustrate the density of the hierarchical normal inverse Pareto distribution and how its tail behaviour is changed for different values of the upper bound parameters b_0 with fixed c, d .



(a) Density plots for hNiP with various (b_0), the prior values of the hyperparameter κ are $c = 0.9$ and $d = 0.5$.



(b) The tail behaviour of the density plots for hNiP with several values of b_0 the prior values of the hyperparameter κ are $c = 0.9$ and $d = 0.5$.

Figure 3.4: Density plots for the Normal inverse Pareto (NiP) distribution for different values of b_0 and the black dotted line represents the Gaussian distribution with fixed c, d.

3.2.2 Rescaled Beta Hierarchical Normal Inverse Pareto Distribution

We can argue that large values of κ would lead to the same amount of shrinkage.One way of achieving is to differential shrinkage place a prior distribution on the hyperparameter κ of the following form:

$$\pi(\kappa) = \frac{e_0^{-1}}{B(a,b)} \left(\frac{\kappa}{e_0}\right)^{a-1} \left(1 - \frac{\kappa}{e_0}\right)^{b-1}, \tag{3.12}$$

where $a, b > 0$, when $a = b = \frac{1}{2}$, then the mean of the rescaled Beta distribution is $\frac{a\, e_0}{a+b} = \frac{1}{2}e_0$ and the variance is $\frac{abe_0^2}{(a+b)^2(a+b+1)} = \frac{1}{8}e_0^2$ see Appendix D for more information. This is not the only reason for bowing κ so we want to make the shrinkage adaptive so $a, b < 1$ as can seen in Figure 3.5 which prior of κ between 0 and e_0. Now the idea, which we would really like

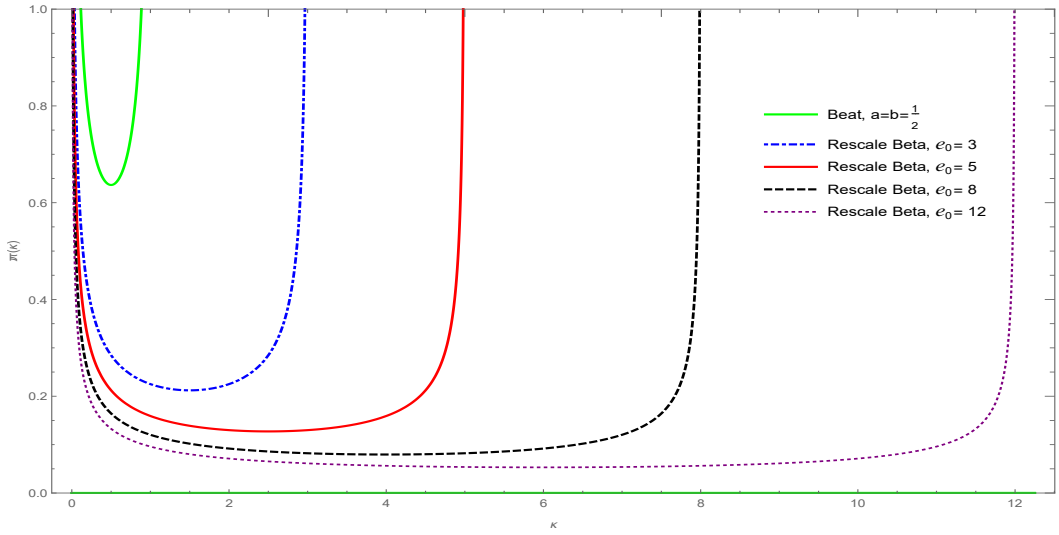


Figure 3.5: Density plots for the Rescaled Beta distribution $\pi(\kappa)$ with various values of the parameter e_0 and $a = b = \frac{1}{2}$.

to do, is to have a prior that will be flexible enough to put considerable of mass towards small value of the scale parameter κ, in order to not shrink some regression coefficients, and also put large of mass towards large values of this parameter. Therefore, we have Rescaled-Beta for this prior because we are trying to make a more flexible horseshoe shape. Equation 3.13 represent the structure of the Rescaled-Beta of the hierarchical normal inverse Pareto distribution (ReB-hNiP). This can be seen very clear that in Figure 3.5 where we compare with the standard horseshoe shape based on Beta distribution with $a = b = \frac{1}{2}$ with a range of values of Rescaled

Beta e_0. The bellow equation represents the Rescaled Beta hierarchical-NiP (ReB-hNiP) prior distribution:

$$\text{ReB-hNiP}(\beta|\kappa, b_0, a, b, e_0) = N(\beta|0, \psi^{-1}) \quad \text{iP}(\psi|\kappa, b_0) \quad \text{ReB}(\kappa|a, b, e_0). \tag{3.13}$$

$$\text{ReB-hNiP}(\beta|\kappa, b_0, a, b, e_0) = \frac{\psi^{\frac{1}{2}}\ e^{-\frac{1}{2}\psi\beta^2}}{\sqrt{2\pi}}\ \kappa b_0^{-\kappa}\psi^{\kappa-1}\frac{e_0^{-1}}{B(a,b)}\left(\frac{\kappa}{e_0}\right)^{a-1}\left(1-\frac{\kappa}{e_0}\right)^{b-1}. \tag{3.14}$$

The marginal distribution of (β) in 3.14 has no analytical expression, but the mean is still follows,

$$\begin{aligned} E(\beta) &= E[E[E[\beta|\psi]]] \\ E(\beta) &= E[E[0]] = 0, \end{aligned} \tag{3.15}$$

as we mentioned for hNiP prior, the $E[\beta|\psi] = 0$ for ReB-hNiP, because $\beta|\psi \sim N(\beta|0, \psi^{-1})$, and the variance is

$$\begin{aligned} Var[\beta] &= E[Var[\beta|\psi]] + Var[E[\beta|\psi]] \\ &= E[E[Var[\beta|\psi]] + Var[E[\beta|\psi]]] \\ &= E[E[\frac{1}{\psi}]] + 0 \\ \text{where}\ \ E[\frac{1}{\psi}] &= \int_0^{b_0}\frac{1}{\psi}\kappa b_0^{-\kappa}\psi^{\kappa-1}d\psi = E[\frac{1}{b_0}\frac{\kappa}{\kappa-1}]\ \ \text{iff}\ \ \kappa > 1 \\ Var[\beta] &= \frac{1}{b_0}\int_0^{e_0}\frac{\kappa}{\kappa-1}\frac{e_0^{-1}}{B(a,b)}\left(\frac{\kappa}{e_0}\right)^{a-1}\left(1-\frac{\kappa}{e_0}\right)^{b-1}d\kappa, \end{aligned} \tag{3.16}$$

$$Var[\beta] = \frac{\pi e^{-i\pi(a+1)}e_0^{-a}\left(\csc(\pi a)\left(\frac{e_0-1}{e_0}\right)^{b-1} - \frac{e_0^a(\cot(\pi a)+i)\Gamma(b)\,{}_2\tilde{F}_1\left(1,-a-b+1,1-a,\frac{1}{e_0}\right)}{\Gamma(a+b)}\right)}{b_0\ B(a,b)}\ \ \kappa > 1, e_0 > 1 \tag{3.17}$$

where ${}_2\tilde{F}_1[1, 1+a, 1+a+b, e_0]$ is the generalized Hypergeometric function (Gradshteyn and Ryzhik, 2014). Figure 3.6 shows the several plots of the rescaled Beta hierarchical Normal inverse Pareto density. Where ψ is the first layer of shrinkage prior that control amount of shrinkage when is large leads to shrink the coefficients and if the value of ψ close to zero, this is support to keep the coefficients in the model. κ is the second layer that control variabil-

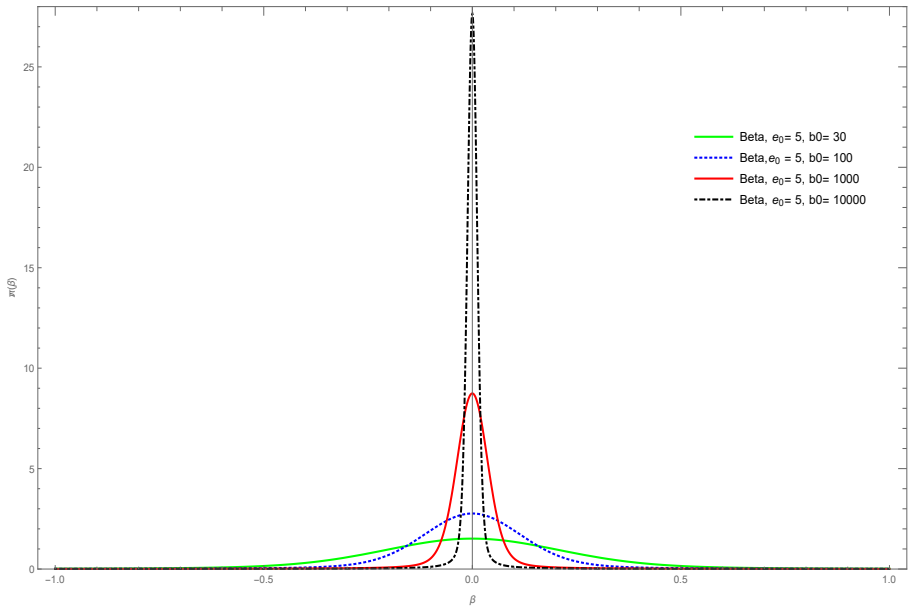


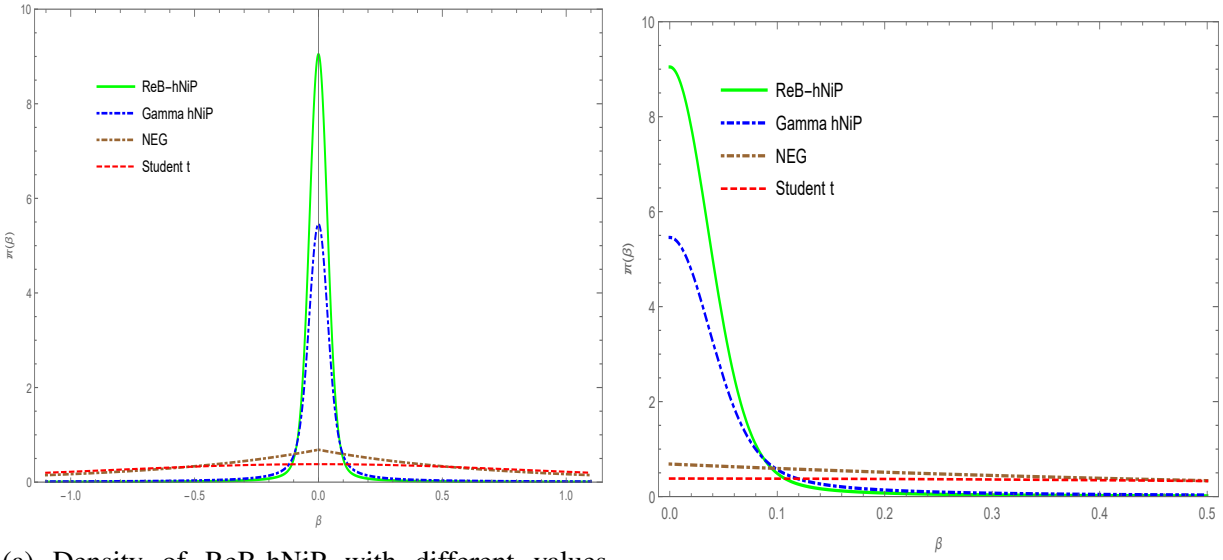
Figure 3.6: Density of ReB-hNiP with different values of b_0, $e_0 = 5$ where $a = b = \frac{1}{2}$.

ity of parameter ψ based on the value of rescaled beta parameters e_0. In addition, Figure 3.7 illustrates the marginal density functions of several Bayesian shrinkage prior including ReB-hNiP, hNiP, NEG and Student-t distribution, and their tail behaviour. It is clear that our proposed shrinkage prior has high of mass around zero and fatter tails compared with similar scale mixtures of Gaussian distributions. For illustration, we chose appropriate degree of freedom value $\nu = 5$, because growing the number of degrees of freedom leads to obtain normal distribution with mean is zero and variance is 1.

3.2.3 Posterior Consistency for Proposed Shrinkage Prior Distribution

Our particular focus is on shrinkage priors. We investigate the asymptotic behaviour of posterior distributions of regression coefficients in high-dimensional linear regression models as the number of dimensions grows with the observations. Particularly, in this section we focus on asymptotic posterior distribution as strongly consistent.

We consider dealing with the regression model such as $y_n = X_n\beta_0^n + \varepsilon_n$, where y_n is a vector of responses variable, X_n is the $n \times p_n$ is matrix of the independent variables, $\varepsilon_n \sim N(0; \lambda^{-1})$ with



(a) Density of ReB-hNiP with different values of b_0, $e_0 = 5$ Where $a = b = \frac{1}{2}$.

(b) The tail behaviour of several shrinkage prior.

Figure 3.7: Marginal density of different shrinkage priors including the hNiP with values $b_0 = 1000, c = 0.3$ and $d = 0.2$; represents by blue dashed dot line, and the solid green line refer to the ReB-hNiP with $b_0 = 1000, a = b = \frac{1}{2}$ and $e_0 = 5$; brown dotted dash line represents the brown dotted dash the NEG prior with $\omega = 4$ and $\gamma = \sqrt{4}$; and Student-t is presents by red dash line with degrees of freedom $\nu = 5$.

known the variance $\lambda^{-1} > 0$, and β_0^n is vector of the true coefficients that involve zero and non-zero values.

There are some sufficient conditions assumptions for posterior consistency, in order to show our shrinkage prior is posterior consistency as Armagan et al. (2013b) established in their article. We state the assumptions in the following points:

- $D1$. $p_n = O(n)$;
- $D2$. Let $\Delta_{n\ min}$ and $\Delta_{n\ max}$ are smallest and largest singular values of X_n respectively.
  Also, $0 < \Delta_{min} < \liminf_{n\to\infty} \frac{\Delta_{n\ min}}{\sqrt{n}} \leq \limsup_{n\to\infty} \frac{\Delta_{n\ max}}{\sqrt{n}} < \Delta_{max} < \infty$;
- $D3$. $\sup_{j=1,2,3,\ldots,p_n} |\beta_{ni}^0| < \infty$;
- $D4$. $q_n = O(n/\log n)$;

where $A_n = \{i : \beta_{ni}^0 \neq 0,\ i = 1,\ 2,\ \ldots, p_n\}$ and $|A_n| = q_n$ that symbolise the set of the number of zero and non-zero coefficients in β^0. Also $\rho \in (0,\ 2)$.

These assumptions are proposed by Armagan et al. (2013b), and also by utilizing Theorem 1 and the Equation A3 in mentioned article, that satisfy sufficient conditions to provide strong posterior consistency for a shrinkage prior in linear regression model. Where β_0^n represents both true zero and non-zero coefficients of regression vector. In order to deal with high-dimensional dataset, we let $p_n \to \infty$ and $n \to \infty$.

The main equation in Theorem 1 of Armagan et al. (2013b), is state as follow,

Theorem 3.2.1.

$$\pi_n\left(\beta_n :\ ||\beta_n - \beta_n^0|| > \frac{\Delta}{n^{\frac{\rho}{2}}}\right) > \exp(-M\ n), \tag{3.18}$$

for all

$$0 < \Delta < \frac{\epsilon^2\ \delta_{min}^2}{48\ \delta_{max}^2} \quad \textit{and} \quad 0 < M < \frac{\epsilon^2\ \delta_{min}^2\ \lambda}{32} - 3\ \Delta\ \delta_{max}^2\ \frac{\lambda}{2},$$

the proof Theorem 3.2.1 was provided and Equations A3 in original article of Armagan et al. (2013b) is follow;

$$\pi_n\left(\beta_n : ||\beta_n - \beta_n^0|| > \frac{\Delta}{n^{\frac{\rho}{2}}}\right) >= \prod_{i \in A_n}\left\{\pi_n\left(\beta_n : ||\beta_n - \beta_n^0|| > \frac{\Delta}{n^{\frac{\rho}{2}}}\right)\right\}\left\{1 - \frac{p_n\ n^\rho\ E\big(\sum\limits_{i \notin A_n} \beta_{nj}^2\big)}{(p_n - q_n)\Delta^2}\right\}, \tag{3.19}$$

Theorem 3.2.2. *The hierarchical normal inverse Pareto distribution hNiP satisfy strongly consistent posterior distribution under assumption $D1$ - $D4$ and the Equation 3.19, if $\psi_n = \frac{C^2}{\left(\log n\right)^2\ p_n\ n^\rho}$, for finite $C > 0,\ \kappa > 0$ and $0 < \psi < b_0$.*

The hierarchical normal inverse Pareto distribution hNiP is follows,

$$f(\beta_{ni}^0|\ b_{n0}, \psi,\ \kappa) = \frac{b_{n0}^{-\kappa_i}\ \kappa_i^c\ \psi^{\kappa_i - \frac{1}{2}}\ d^c\ e^{-\kappa_i d}\ e^{-\frac{\psi(\beta_{ni}^0)^2}{2}}}{\sqrt{2\pi}\Gamma(c)}.$$

To proof the Theorem 3.2.1, we compute the $E(\beta_{ni}^2)$ based on the variance is

$$\begin{aligned} Var[\beta] &= E[Var[\beta|\psi]] + Var[E[\beta|\psi]] = E[E[Var[\beta|\psi]] + Var[E[\beta|\psi]]] \\ &= E[E[Var[\beta|\psi]] + 0 \ \text{ where } \ E[Var[\beta|\psi]] = E(\beta^2) + [E[\beta|\psi], \\ &\qquad \text{therefore, } \ E(\beta^2) = E[Var[\beta|\psi]], \end{aligned} \tag{3.20}$$

therefore in this case, the variance of β is infinite.

$$\begin{aligned} &\pi_n\left(\beta_n:\ ||\beta_n - \beta_n^0|| > \frac{\Delta}{n^{\frac{\rho}{2}}}\right) \geq \left\{1 - \frac{p_n\, n^{\rho}\, E\big(\beta_{ni}^2\big)}{\Delta^2}\right\} \\ &\left(\frac{\Delta}{(\sqrt{p_n}\, n^{\rho/2})} \frac{b_{n0}^{-\kappa}\, \kappa^c\ d^c\ \exp\big\{-\kappa d\big\}}{\psi_n^{-(\kappa-\frac{1}{2})}\sqrt{2\pi}\Gamma(c)} \exp\left\{-\frac{\sup_{i\in A_n}(\beta_{ni}^0)^2 - \frac{\Delta^2}{(\sqrt{p_n}\ n^{\rho/2})}}{2\ \ \psi_n^{-1}}\right\}\right)^{q_n}. \end{aligned} \tag{3.21}$$

We putting the negative-logarithm for both sides of Equation (3.21), which became;

$$\begin{aligned} &-\log\ \pi_n\left(\beta_n:\ ||\beta_n - \beta_n^0|| > -\log\frac{\Delta}{n^{\frac{\rho}{2}}}\right) \leq -\log\left\{1 - \frac{p_n\, n^{\rho}\, E\big(\beta_{ni}^2\big)}{\Delta^2}\right\} \\ &-q_n \log\left(\frac{\Delta}{(\sqrt{p_n}\, n^{\rho/2})} \frac{b_{n0}^{-\kappa}\, \kappa^c\ d^c}{\big(\frac{C^2}{\big(\log n\big)^2\ p_n\ n^{\rho}}\big)^{-(\kappa-\frac{1}{2})}\sqrt{2\pi}\Gamma(c)}\right) + \kappa d\ \ + \left\{\frac{\sup_{i\in A_n}(\beta_{ni}^0)^2 + \frac{\Delta^2}{(\sqrt{p_n}\ n^{\rho/2})}}{2\ \ \big(\frac{C^2}{\big(\log n\big)^2\ p_n\ n^{\rho}}\big)^{-1}}\right\}, \end{aligned} \tag{3.22}$$

then,

$$\begin{aligned} &-\log\ \pi_n\left(\beta_n:\ ||\beta_n - \beta_n^0|| > -\log\frac{\Delta}{n^{\frac{\rho}{2}}}\right) \leq -\log\left\{1 - \frac{p_n\, n^{\rho}\, E\big(\beta_{ni}^2\big)}{\Delta^2}\right\} \\ &-q_n \log\left(\frac{b_{n0}^{-\kappa}\, \kappa^c\ d^c}{\big(\frac{C}{\log n}\big)^{-1}\ \left(\frac{C^2}{\big(\log n\big)^2\ p_n\ n^{\rho}}\right)^{-\kappa}\ \sqrt{2\pi}\Gamma(c)}\right) + \kappa d\ \ + \left\{\frac{\sup_{i\in A_n}(\beta_{ni}^0)^2 + \frac{\Delta^2}{(\sqrt{p_n}\ n^{\rho/2})}}{2\ \left(\frac{C^2}{\big(\log n\big)^2\ p_n\ n^{\rho}}\right)^{-1}}\right\} \end{aligned} \tag{3.23}$$

simplify-fractions

$$
\begin{aligned}
-\log\ \pi_n\left(\beta_n:\ ||\beta_n-\beta_n^0||>-\log\frac{\Delta}{n^{\frac{\rho}{2}}}\right)\leq-\log\left\{1-\frac{p_n\ n^\rho\ E\big(\beta_{ni}^2\big)}{\Delta^2}\right\}\\
-q_n\log(b_{n0}^{-\kappa}\ \kappa^c\ \ d^c)-q_n\log C+q_n\log\log n-2q_n\kappa\log C+2\kappa q_n\log\log n+\\
2q_n\kappa\ \log\ p_n\ n^\rho+\ q_n\kappa\ \log\left(\sqrt{2\pi}\Gamma(c)\right)+\kappa d\ \ +\frac{\Delta^2\sqrt{p_n}\ n^{\rho/2}}{2\ \ C^2}+\frac{C^2\ \sup_{i\in A_n}(\beta_{ni}^0)^2}{2\ \ \big(\log n\big)^2\ p_n\ n^\rho},
\end{aligned}
\tag{3.24}
$$

Based on assumptions $D1$ - $D4$, dominating term is the last one in Equation 3.24 with

$-\log\ \pi_n\left(\beta_n:\ ||\beta_n-\beta_n^0||>-\frac{\Delta}{n^{\frac{\rho}{2}}}<Mn\right)$ for all $M>0$. The result for all $M>0$ is being shown. Proof is completed.

On other hand, the posterior consistency studied by Song and Liang (2017), whom addressed the for shrinkage prior distributions that have polynomial tail. Specifically, they provided the bellow theorem, as general case of posterior consistency for any shrinkage prior's that satisfy. They stated that as follows,

The main context of (Theorem 3.1) of Song and Liang (2017), Assumed that conditions on $A(1)$ and $A(2)$:

The conditions $(A1)$ includes (1) All the covariates are uniformly bounded. For simplicity, we assume they are all bounded by 1. (2) We deal with high dimensional dataset $p_n\succeq n$. (3) There exist some integer $\bar{q}$ (depending on n and p_n) and fixed constant Ψ_0 such that $\bar{q}\succ s^*$ and $\Psi_m\in(X^TX)\geq n\Psi_0$ for any subset model $|\zeta_j|\leq\bar{q}$.

Where, p_n to indicate that the number of covariates can increase with the sample size n. We let s^* denote the size of the true model. We let $1(.)$ represents the indicator function. Ψ_{max} denote the largest and Ψ_{max} smallest eigenvalues of a square matrix. Also, ζ_j denote the size of the model. Both a and b are positive sequences, with $a\prec b$ means $\lim\frac{a}{b}=0$ and $a\asymp b$ that means $0<\liminf\ \frac{a}{b}\leq\limsup\ \frac{a}{b}<\infty$.

The second condition $(A2)$ involves (1) $s^*\ \log p_n\prec n$, where s is the size of the true model, and M represents some fixed constant. (2) $max\{|\beta_j^*\lambda|\}\ =\ \gamma_0E_n,$ for some fixed $\gamma_0\ \in$

$(0, 1)$, and E_n refer to non-decreasing with respect to n, also β_j^* indicated the true coefficients values and λ^* indicate the true value of precision parameter.

It hold for the linear regression models, and a polynomial-tailed prior is used. If $\log(E) = O(\log p_n)$, and the scaling parameter Ψ_n satisfies $\Psi_n \leq a_n p_n^{-(w+1)/(r-1)}$, $-\log \Psi_n = O(\log p_n)$ for some $w > 0$, then

- the posterior consistency Equation (3.25) based on (Song and Liang, 2017) which holds when $a_n \asymp \sqrt{s^* \log p_n / n}/p_n$;

$$\begin{aligned} P^*\left(\pi(||\beta - \beta^*|| \geq \frac{k_1\, \epsilon_n}{\lambda^*}|D_x^n) \geq e^{n\, k_2\, \epsilon_n}\right) \leq e^{-n\, k_3\, \epsilon_n}, and \\ P^*\left(\pi(||\beta - \beta^*||_1 \geq k_1\sqrt{s^*}\epsilon_n|D_x^n) \geq e^{n\, k_2\, \epsilon_n}\right) \leq e^{-n\, k_3\, \epsilon_n}, \end{aligned} \tag{3.25}$$

  where $k_1,\ k_2$ and k_3 are positive constants. $D_x^n = D_{x,y}$ represents the observed data.

- the model selection consistency Equation (3.26), satisfy when $a_n \prec \sqrt{\log p_n}/\sqrt{n}p_n$;

$$P^*\left\{\pi\big([\zeta(\boldsymbol{\beta}, \lambda^{-1}\big) = \zeta^*|D_x^n] > 1 - p_n^{-k_3}\right\} > 1 - p_n^{-k_4}, \tag{3.26}$$

  where k_3 and $k_4 < 2(w-1)$, which are positive constants. $min_{i\in\zeta^*}|\beta_i^*| \geq M_1\sqrt{s \log p_n/n}$ for sufficient large $M_1,\ s^* \log l_n \prec \log p_n$ and $u > 1$;

- the posterior approximation Equation (2.8) as mentioned in original article of the (Song and Liang, 2017), holds when $a_n \asymp \sqrt{s^* \log p_n/n}/p_n$; $min_{i\in\zeta^*}|\beta_i| \geq M_1\sqrt{s^* \log p_n/n}$ for sufficient large $M_1,\ s^* \log l_n \prec 1$ and $w > 1$;

$$\begin{aligned} \pi(\beta, \lambda^{-1}|D_x^n) \text{ converges in total variation to} \\ MVN\big(\beta_{\zeta^*}; \hat{\beta}_{\zeta^*}, \lambda^{-1}(X_{\zeta^*}X_{\zeta^*}^T)^{-1}\big) \prod_{i\notin\zeta^*} \pi(\beta|\lambda) Ga\left(\lambda, \frac{n - s^*}{2}, \frac{\lambda^{-1}(n - s^*)}{2}\right), \end{aligned} \tag{3.27}$$

where X_{ζ^*} represents sub design matrix based on to the model ζ parameters.

They noted that most commonly utilised polynomial decaying distributions satisfy as follows; $g(x) = C\ x^{-t}\ L(x)$, where $\lim_x L(x) = 1$ with the rate $|L(x) - 1| = O(x^{-t})$ where $t \geq$

0 and $L(x)$ stand for slowly varying function. We handle with MVN indicates multivariate normal distribution density with variance is λ^{-1} and Ga refers to the gamma distribution. Then, it is easy to see that if $min_{i\in\zeta^*}|\beta_i^*| > M_2\ \epsilon_n$ for some large M_2, $\Delta_n = O(\epsilon_n)$, then $s^*\ \log l_n \asymp\ s^*\ \epsilon_n/min_{i\in\zeta^*}|\beta_i^*|$. According to the previously explanation about the both proposed hNiP and ReBhNiP shrinkage prior distributions, which are the family scale mixture of normal distribution, and also of having polynomial decay tails distributions. For example, both proposed distribution have polynomial tails compare NEG Normal Exponential Gamma shrinkage prior Griffin and Brown (2007). Therefore, we can argue that both hNiP and ReBhNiP shrinkage prior satisfies the above conditions in Theorem (3.1) as stated in (Song and Liang, 2017) which is corresponding Equation (3.25) for posterior consistency distribution. The mixing distribution of scale parameter of hNiP shrinkage prior is $\psi^n|\kappa^n \sim iP(\psi^n|b_0,\kappa^n)Ga(\kappa^n|c,d)$, where $\Psi_n = \psi_n$ in the main theorem for posterior consistency. Also it true for ReBhNiP shrinkage prior distribution.

3.3 Normal Exponential Gamma Prior Distribution

The Normal Exponential Gamma prior (NEG) was proposed by Griffin and Brown (2007). It is used as a prior distribution (penalty method) to shrink regression coefficients in high-dimensional models by Griffin and Brown (2011, 2007), Hoggart et al. (2008), Rockova et al. (2014) and Li and Yao (2014). It is an advanced LASSO prior which is called a family of hyper-LASSO penalty methods. It has more mass towards zero and flatter tails compared to the double exponential distribution. The NEG distribution is one of popular scale mixture of normal prior which is derived from the generalised double exponential with hyper-prior of it is gamma distribution and some other Bayesian shrinkage prior have equivalent hierarchical structure as a special case. Consequently, for that reason. we would like to apply NEG prior in hierarchical structure using the MCMC algorithm. It can be described as a zero mean double exponential by Equation 3.28 which includes only one parameter that controls a scale mixture of the Normal distribution, as follows:

$$\text{DE}(x) = \int_0^\infty N\left(x|0,\phi\right) \text{Ga}\left(\phi\middle|1,\frac{\gamma^2}{2}\right) d\phi = \frac{\gamma}{2}\exp\{-\gamma|x|\}, \tag{3.28}$$

where the normal distribution $N(x|\mu, \sigma^2)$ has mean equal to zero and variance $\sigma^2 = \phi$. The Lasso penalty is obtained from the double exponential prior distribution with the mean zero. The NEG prior distribution can be represented as a mixture distribution as follows (Griffin and Brown, 2007; Hoggart et al., 2008):

$$\begin{aligned}\text{NEG}(\beta_i|\phi_i, \zeta_i, \omega, \gamma) &= N(\beta_i|0, \phi_i)\ \text{Ga}(\phi_i|1, \zeta_i)\ \text{Ga}\left(\zeta_i\big|\omega, \frac{1}{\gamma^2}\right). \\ \text{NEG}(\beta_i) &= \int_0^\infty \int_0^\infty \left(\frac{1}{\sqrt{2\pi}}\phi_i^{-\frac{1}{2}} \exp\left\{-\frac{\beta_i'\beta_i}{2\phi_i}\right\}\right)\left(\zeta_i \exp(-\phi_i\zeta_i)\right) \\ &\quad \left(\frac{(\frac{1}{\gamma^2})^\omega}{\Gamma(\omega)}\zeta_i^{\omega-1} \exp\left\{-\frac{\zeta_i}{\gamma^2}\right\}\right) d\phi_i d\zeta_i,\end{aligned} \tag{3.29}$$

$$\begin{aligned}\text{NEG}(\beta_i) &= \frac{\omega\gamma}{\sqrt{\pi}} 2^\omega \Gamma\left(\omega + \frac{1}{2}\right) \exp\left\{\frac{1}{4}\beta_i^2\gamma^2\right\} D_{-2(\omega+\frac{1}{2})}\big(\gamma|\beta_i|\big) \\ &= K \exp\left\{\frac{1}{4}\beta_i^2\gamma^2\right\} D_{-2\omega-1}(\gamma|\beta_i|),\end{aligned} \tag{3.30}$$

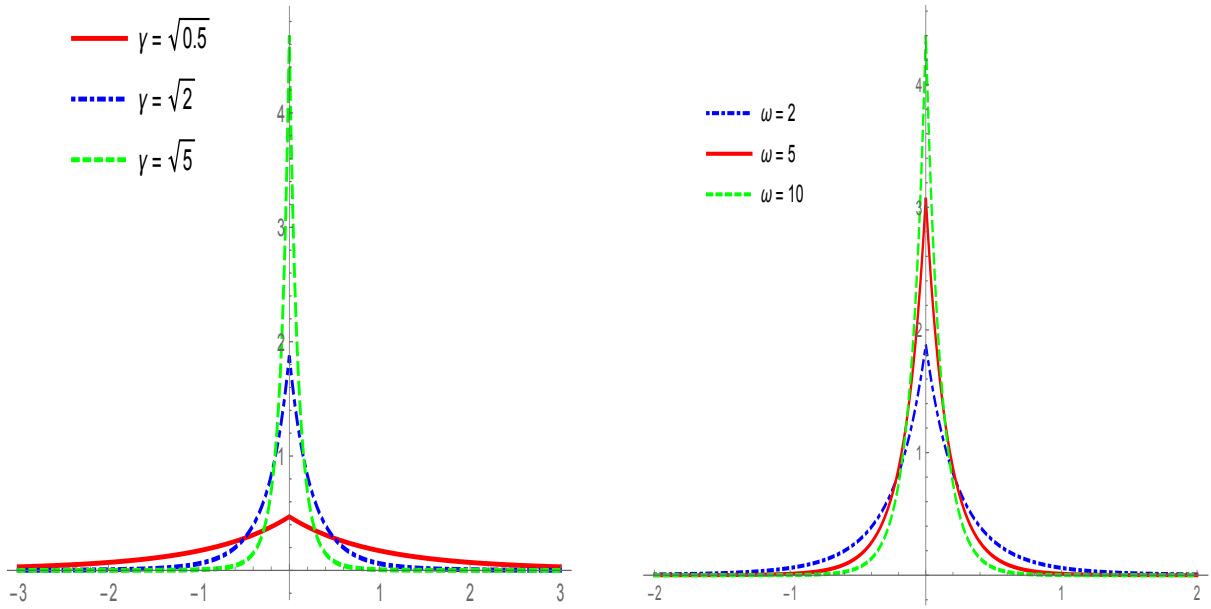
where K is constant and equal to $K = \frac{\omega\gamma}{\pi^{1/2}} 2^\omega \Gamma(\omega + 1/2)$ and the parameter ω controls the shape of the tails whilst γ control the scale. In addition, NEG has a sharp peak around zero with heavy tails. This distribution has mean equal to zero and $\text{Var}(\beta_i|\omega, \gamma) = \frac{1}{\gamma^2(\omega-1)}$ when $\omega > 1$. The highest point of the NEG at zero leads sparse solutions that are better for variable selection.

The structure of equation 3.30 has been widely used in the literature by researchers. We exploit the hierarchical structure of the NEG prior distribution to build a Gibbs sampler and we will compare with other shrinkage priors, so the hierarchical distribution is given as follows:

$$\begin{aligned}\pi(\beta_i|\phi_i, \zeta_i, \omega, \gamma) &= N(\beta_i|0, \phi_i)\ \ \text{Ga}(\phi_i\big|1, \zeta_i)\ \ \text{Ga}\left(\zeta_i|\omega, \frac{1}{\gamma^2}\right) \\ &= \frac{1}{\sqrt{2\pi}}\phi_i^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\frac{\beta_i'\beta_i}{\phi_i}\right\} \left(\zeta_i \exp\left\{-\phi_j\zeta_i\right\}\right) \frac{(\frac{1}{\gamma^2})^\omega}{\Gamma(\omega)}\zeta_i^{\omega-1} \exp\left\{-\frac{\zeta_i}{\gamma^2}\right\}.\end{aligned} \tag{3.31}$$

Figure 3.8 show plots of the marginal density function of the normal exponential gamma distribution for different values of the scale and shape parameters. Figure 3.8a displays three different graphs, where the shape is fixed $\omega = 2$ but the scale parameter differ when the scale is

equal to $\sqrt{5}$ which is represented by the green dashed line, the highest point of all three plots is reached. The second scale value is $\gamma = \sqrt{2}$ which has the second highest peak in the plot (the blue dashed-dotted line) then the red solid line has the flatted graph which belongs to the NEG distribution with scale $\gamma = \sqrt{0.5}$. In figure 3.8b, we have the fixed scale $\gamma = \sqrt{2}$ and have various values for the shape parameter ω. As we mentioned above, the shape parameter ω should larger than 1 because the variance dose not exist if ω is less than or equal to 1 (Griffin and Brown, 2007). The green dashed line where $\omega = 10$ has heavy tails. The blue dashed-dotted line is the density of the NEG prior distribution when $\omega = 2$. The red solid line has the second highest point at zero corresponding to $\omega = 5$. The reason to use several parameters values is to show how scale and shape parameters control tail behaviour of this distribution.



(a) Several values of γ and fixed $\omega = 2$.

(b) Several values of ω and fixed $\gamma = \sqrt{2}$.

Figure 3.8: Plots of the density function of the normal-exponential-gamma NEG distribution, the parameter ω control the shape of the distribution and the parameter γ controls the scale, where the highest point at zero aids sparse estimation

Chapter 4

Prior Calibration of the Bayesian Linear Regression Model

4.1 Introduction

The problem of a high-dimensional statistical model has been widely studied for many decades in modern datasets, both for linear and nonlinear models. The ordinary least squares (OLS) in Equation 2.1, is however infeasible to fit coefficients in the model. Therefore, numerous penalty approaches have been suggested to tackle this problem that can be found in the literature; see Mallick and Yi (2013). We explained some of them in Chapter 2. In high-dimensional statistical issues, perhaps the number of covariates exceeds the sample size $p > n$. In reality, only a few covariates have an impact on a model under investigation. To overcome issues of high-dimensionality, some penalty methods are suggested that shrink those regression coefficients that are not needed to interpret models based on the data (Brimacombe, 2014). The values of those unimportant parameters are close to zero, which produces the sparse coefficients of the regression model (Aijun et al., 2017). On this view, the shrinkage approaches developed and utilised in the past decades for both frequencies and Bayesian shrinkage, particularly based on the multivariate Bayesian linear regression model due to the standard statistical model. For examples, see Tibshirani (1996), Zou and Hastie (2005), Park and Casella (2008), Carvalho et al.

(2010), among others. To evaluate and select the best range values of hyperparameters, that is, tuning-parameter selecting for the shrinkage priors described in Chapter 3, we calibrate the proposed shrinkage priors based on the Bayesian linear regression model to reduce the dimension of datasets. Several shrinkage methods are used in order to compare the performance of each method based on different datasets and the signal-to-noise-ratio in stimulation datasets is also taken into consideration. In the following chapter, we will use some shrinkage methods highlighted in the Chapter 2, such as LASSO (Tibshirani, 1996), Bayesian LASSO (Park and Casella, 2008) and horseshoe (Carvalho et al., 2010) regularisation methods for the compression with the novel shrinkage prior hierarchical normal inverse Pareto (hNiP) and rescaled Beta hierarchical normal inverse Pareto (ReB-hNiP), which were explained in Section 3.2, and the hierarchical structure of the Normal exponential Gamma (NEG) prior in Equation 3.31. Bayesian posterior prediction can be done for typical datasets to evaluate and check that our models rely on novel shrinkage priors to know how the model working. Those shrinkage priors belong to the family of scale mixture of normal distribution and also considering is absolutely continues prior distributions. These are called a Bayesian regularisation priors distributions. Thus, implementing a Bayesian shrinkage prior leads to selecting non-zero variables due to having a heavy-tailed distribution by using some statistical tools such as the t-test or credible interval to estimate the regression parameters.

4.2 Bayesian Linear Regression Model Structure

In this section, we present the structure of the Bayesian linear regression model with some shrinkage prior distributions to shrink the high-dimensional regression coefficients in the model. In this case, we consider using the regression model. The general structure of the model is the following,

$$Y = X\boldsymbol{\beta} + \epsilon \tag{4.1}$$

where ϵ has a normal distribution, $\epsilon \sim N(\epsilon|0, \lambda^{-1})$ represents the error model $Var(\epsilon) = \frac{1}{\lambda}$, λ represents the precision of the model, $X \in \mathbb{R}^{n\times(p+1)}$ represents observed variables variables, that is, the design matrix, and $Y \in \mathbb{R}$ is the dependent variable. The sample size of $X = \{1, \boldsymbol{x_1}, \boldsymbol{x_2} \ldots, \boldsymbol{x_n}\}$, $\boldsymbol{x_i} = \{1, x_{i1}, x_{i2} \ldots, x_{ip}\}$, where $(i = 1, 2, \ldots, n), (j = 1, 2, \ldots, p + 1)$ and $Y = \{y_1, y_2 \ldots, y_n\}$. $\boldsymbol{\beta}$ is the $(p+1\times 1)$ coefficient vector that includes the intercept,

and n represents the number of observations.

The likelihood function is as follows:

$$L(\Theta; X, y) = (\lambda)^{n/2} \exp\left\{ -\frac{\lambda}{2}(\boldsymbol{y} - X\boldsymbol{\beta})' I(\boldsymbol{y} - X\boldsymbol{\beta}) \right\}. \tag{4.2}$$

where $\Theta = (\boldsymbol{\beta}, \lambda)$ and $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$, the prior distribution $\pi(\lambda)$ of error model is a gamma distribution with shape a_λ and a scale b_λ. In the following section, we use our proposed shrinkage prior; hNiP is a scale mixture of normal with a hierarchical normal inverse Pareto distribution with gamma distribution, and ReB-hNiP is a rescaled beta hierarchical normal inverse Pareto distribution as a prior of the coefficients $\pi(\beta)$. The full conditional distribution can be established for each named priors. On the other hand, the full conditional distribution of the hierarchical structure of the normal exponential gamma NEG prior of Griffin and Brown (2011) can be found in Appendix E. Finally, we have displayed the pseudo-code for the MCMC algorithm of the following sections for the current chapter in Appendix C.1.

4.2.1 hNiP Distribution as a Prior of β

In this section we construct the hierarchical high-dimensional regression model. We assume that the regression parameter β has an hNiP shrinkage prior distribution, which is explained in the previous chapter, especially in Equation 3.11, to shrink the coefficient and then obtain a sparse model. The likelihood, prior and posterior distribution for the novel shrinkage regularisation method of the hNiP prior distribution is separately addressed in the following section.

4.2.1.1 The Prior and Posterior Distributions

One of the most important parts is addressing the high-dimensional Bayesian regression model which employs a specific prior distribution for regression coefficients. Thus, according to the model construction of the current study, our proposed Bayesian shrinkage distribution plays an important role in shrinking those regression values close to zero. Therefore, we set possible prior distributions for each parameter that participate in building the model. We use our

proposed hNiP prior distribution in this section. The prior distributions are as follows:

$$\pi(\Theta^*) = \frac{b_\lambda^{a_\lambda}\lambda^{a_\lambda-1}\exp\{-b_\lambda\lambda\}}{\Gamma(a_\lambda)}\prod_{j=1}^{p+1}\frac{\psi_j^{\frac{1}{2}}\exp\left(-\frac{1}{2}\psi_j\beta_j^2\right)}{\sqrt{2\pi}}\kappa_j b_0^{-\kappa_j}\psi_j^{\kappa_j-1}\frac{\kappa_j^{c-1}d^c\exp(-\kappa_j d)}{\Gamma(c)}. \quad (4.3)$$

We let Θ^* be the parameters space; $\Theta^* = \{\beta_j, \lambda, \kappa_j, \psi_j\}$. The prior distribution of the coefficient parameter β_j is the hNiP prior distribution, as mentioned in the previous section. The gamma distribution is the prior for the error model with the shape parameter $a_\lambda = 2$ and the scale is $b_\lambda = 0.6$, which leads to a relatively flat prior distribution. The second layer of the hNiP prior distribution is $\pi(\psi_j)$, which has two parameters: the b_0 scale parameter and the shape κ, which controls the control amount of shrinkage. This has a third-layer prior, which is the gamma distribution $\text{Ga}(c, d)$ with the expected value $E(\kappa_j) = \frac{3}{2}$ because if $\kappa < 1$ prior of values ψ close to zero and leads to coefficients are non-zero, vice verses if κ much larger than 1 causes to shrink all coefficients. A variance of κ is finite and have differ values based on the value of both hyperparameters c and d. Consider that both c and d are hyperparameters and should be selected sensitively.

The posterior distribution for the Bayesian linear regression model with hNiP shrinkage priors is given by the following:

$$\pi(\Theta^*|\boldsymbol{y}, X) \propto \pi(\Theta^*)L(\boldsymbol{y}|X, \Theta^*),$$

$$\begin{aligned}\pi(\Theta^*|\boldsymbol{y}, X) = (\lambda)^{n/2}\exp\left\{-\frac{\lambda}{2}(\boldsymbol{y}-X\boldsymbol{\beta})'I(\boldsymbol{y}-X\boldsymbol{\beta})\right\}\frac{b_\lambda^{a_\lambda}\lambda^{a_\lambda-1}\exp\{-b_\lambda\lambda\}}{\Gamma(a_\lambda)} \\ \prod_{j=1}^{p+1}\left(\frac{\psi_j^{\frac{1}{2}}\exp\left(-\frac{1}{2}\psi_j\beta_j^2\right)}{\sqrt{2\pi}}\kappa_j b_0^{-\kappa_j}\psi_j^{\kappa_j-1}\frac{\kappa_j^{c-1}d^c\exp(-\kappa_j d)}{\Gamma(c)}\right).\end{aligned} \quad (4.4)$$

4.2.1.2 The Full Conditional Posterior Distributions

To quantify and update the posterior distributions, the full conditional distribution is found for each parameter. Straightforwardly, the marginal posterior distributions for each parameter are computed by the Gibbs sampler (see Appendix A.1 for information about MCMC and the Gibbs sampler) because it is difficult to update parameters analytically. We have divided the

parameters into two parts to avoid repetition. Thus, the first part includes those parameters that have the same structure for all types of shrinkage priors. The second part is related to the regression parameters and their specific shrinkage prior characterisations.

The full conditional posterior distributions are presented in the following parts:

First part: The full conditional posterior distribution precision of the model λ is given by the following:

$$\lambda \sim \text{Ga}\left(\lambda \middle| a_\lambda + \frac{n}{2}, b_\lambda + \frac{1}{2}(\boldsymbol{y} - X\boldsymbol{\beta})' I(\boldsymbol{y} - X\boldsymbol{\beta})\right). \tag{4.5}$$

The precision of the model λ follows the gamma distribution and is updated by the Gibbs sampler.

Second part: In this part, the full conditional distribution is demonstrated by the parameters that belong to the hNiP prior distribution and updating the coefficients according to those parameters.

The shape parameters κ is the second level of the hierarchical normal inverse Pareto distribution and one of the hyperparameters of the prior β_j. The full conditional posterior distribution of this parameter κ is as follows:

$$\kappa_j \sim \text{Ga}\left(\kappa_j \middle| c + 1, \log\left(\frac{b_0}{\psi_j}\right) + d\right). \tag{4.6}$$

The parameter ψ_j is a scale mixture normal distribution for the first level of the hNiP prior distribution, and ψ is the vector $(p+1) \times 1$ of the positive values. If the values of $\psi_j \to 0$ the coefficient β_j are included in the model, consider that a coefficient is non-zero. On the contrary, if it is true that $\psi_j \to \infty$ a coefficients shrinks close to zero and is excluded from the model. The full conditional posterior distribution of the scale mixture normal distribution ψ_j is given by the following:

$$\psi_j \sim \text{TrGa}\left(\psi_j | b_0, \kappa_j + \frac{1}{2}, \frac{\beta_j^2}{2}\right), \tag{4.7}$$

where the truncated gamma (TrGa) distribution is bounded from above; for more information, see Appendix D. We let $\text{TrGa}(x|b_0, a, b) = \text{Ga}(x|a, b) I_{x<b_0}$, where $I_{x<b_0}$ represents the condition of the upper bound.

The full conditional posterior distribution of the coefficient parameter $\hat{\boldsymbol{\beta}}$ is updated based on the multivariate normal distribution and intercept included in these vector values as follows:

$$\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}|\mu_\beta, \Omega_\beta^{-1}\right), \tag{4.8}$$

where $\mu_{\boldsymbol{\beta}} = \lambda\Omega_{\boldsymbol{\beta}}^{-1}X'\boldsymbol{y}$, $\Omega_{\boldsymbol{\beta}} = (\lambda X'X + \Sigma_\psi)$ and Σ_ψ is a matrix by $(p+1)\times(p+1)$. The diagonal of this matrix is equal to $\{\psi_1, \psi_2 \ldots, \psi_{p+1}\}$.

4.2.2 Rescaled Beta hNiP as Prior of β

The model structure in this section is same as the hierarchical, high-dimensional regression model in Equation 4.2. However, the only difference in this section is using the ReB-hNiP prior distribution for the regression parameters. Therefore, the prior and posterior distributions rely on the ReB-hNiP prior distribution highlighted separately in the following steps. We are not repeating the full conditional distribution for the precision parameter model λ, the coefficient $\hat{\boldsymbol{\beta}}$ and the parameter ψ_j because they correspond to the Equations 4.5, 4.7 and 4.8, respectively. Θ is the set of parameters $\Theta^* = \{\beta_j, \lambda, \kappa_j, \psi_j\}$.

Similar to the previous section, the Bayesian shrinkage prior distribution ReB-hNiP is applied in this section to construct the Bayesian sparse linear regression model. We display only the prior distribution and possible prior distributions for all the parameters which contributed for creating Bayesian linear regression model. Consequently, the prior distributions of the current model are the following:

$$\begin{aligned}\pi(\beta, \lambda, \kappa, \psi) = &\frac{b_\lambda^{a_\lambda}\lambda^{a_\lambda-1}\exp\{-b_\lambda\lambda\}}{\Gamma(a_\lambda)} \\ &\prod_{j=1}^{p+1}\left(\frac{\psi_j^{\frac{1}{2}}\exp\left(-\frac{1}{2}\psi_j\beta_j^2\right)}{\sqrt{2\pi}}\kappa_j b_0^{-\kappa_j}\psi_j^{\kappa_j-1}\frac{e_0^{-1}}{B(a,b)}\left(\frac{\kappa_j}{e_0}\right)^{a-1}\left(1-\frac{\kappa_j}{e_0}\right)^{b-1}\right).\end{aligned} \tag{4.9}$$

Our proposed ReB-hNiP shrinkage method is used as the prior distribution of the coefficient parameters β_i, to overcome the problem of high-dimensionality. The posterior distribution of Bayesian linear regression, for which the ReB-hNiP shrinkage prior is utilised for modelling in

the currant section, has the following formula:

$$\pi(\Theta^*|\boldsymbol{y}, X) \propto \pi(\Theta^*)L(\boldsymbol{y}|X, \Theta^*)$$

$$\begin{aligned}\pi(\Theta^*|\boldsymbol{y}, X) = (\lambda)^{n/2} \exp\left\{-\frac{\lambda}{2}(\boldsymbol{y} - X\boldsymbol{\beta})' I (\boldsymbol{y} - X\boldsymbol{\beta})\right\} \frac{b_\lambda^{a_\lambda} \lambda^{a_\lambda - 1} \exp\{-b_\lambda \lambda\}}{\Gamma(a_\lambda)} \\ \prod_{j=1}^{p+1} \left(\frac{\psi_j^{\frac{1}{2}} \exp\left(-\frac{1}{2}\psi_j \beta_j^2\right)}{\sqrt{2\pi}} \kappa_j b_0^{-\kappa_j} \psi_j^{\kappa_j - 1} \frac{e_0^{-1}}{B(a,b)} \left(\frac{\kappa_j}{e_0}\right)^{a-1} \left(1 - \frac{\kappa_j}{e_0}\right)^{b-1} \right).\end{aligned} \tag{4.10}$$

The full conditional posterior distribution are presented only for those parameters which have different formulas than previous hNiP prior distributions. Most parameters in the model structure are updated directly based on the Gibbs sampler, as for the hNiP model.

The full conditional posterior distribution of the hyperparameter κ_j is the following:

$$\pi(\kappa_j) \propto b_0^{\kappa_i} \psi_j^{\kappa_i - 1} \frac{e_0^{-1}}{B(a,b)} \left(\frac{\kappa_j}{e_0}\right)^{a-1} \left(1 - \frac{\kappa_j}{e_0}\right)^{b-1}, \tag{4.11}$$

where $j = 1, 2, \ldots, (p+1)$, the values of κ_j should satisfy the condition $0 < \kappa_j \leq e_0$ and $a = b = 1/2$. The rescaled parameter e_0 addressed later. The formula in 4.11 dose not have a typical distribution. Therefore, the Metropolis Hastings algorithm is used to update posterior distribution. We provide more details about the pseudo-code in Appendix C.1. The full conditional posterior distribution of the coefficients parameters $\boldsymbol{\beta}$ follows Equation 4.8.

4.2.3 The NEG Prior

Similar to the formulation of the hNiP and ReB-hNiP prior, we have presented the prior and full conditional distributions only for parameters which are different from previous formulas based on the hierarchical structure of the NEG prior in Equation 3.31. Furthermore, the prior and posterior distributions are presented in Appendix E; the main difference prior distributions are related to the NEG prior distribution without repeating the full posterior parameters and formulas which have similar formats to priors that have been mentioned before under the same Bayesian linear model.

4.3 Simulated Data

In this section, we present the procedure of simulating the data. The purpose of the simulated data was to assess the efficiency of our discussed novel shrinkage priors, to sensitively select and calibrate the values of the hyperparameters based on Bayesian linear regression models, and, subsequently, to compare our proposed shrinkage priors with some shrinkage methods described in the first section of the current chapter. In the comparison process, we will utilise the best range values of the selected hyperparameters. The best way to calibrate is by using simulated data. Thus, we simulate different datasets under multivariate linear regression models. The general formulation of the linear regression is $Y = \beta X + \epsilon$, where the independent variables is $X \in \mathbb{R}^{n \times p+1}$, that is, the design matrix, each vector of X being generated from a normal distribution with a mean of 2 and variance of 25, due to values of observations differ from zero and with large variability. The first column of the matrix of random variables is a vector of one. The β coefficients are a sparse vectors of true values which include various values to address our efficient shrinkage priors. The error of the regression model has a Gaussian distribution with $\epsilon \sim N(.|0, \lambda^{-1})$, where λ is the precision of the model. We set different precision values for all datasets scenarios in which $\lambda = (0.02, 2, 2000)$, that is, we take into consideration the signal-to-noise-ratio to analyse the efficiency of the hNiP and ReB-hNiP prior distributions. The simulated datasets have been divided into four different main scenarios, each of them consists of one or more different datasets. Each case scenario is utilised for specific purposes in the following sections. For example, choosing the best range of optimal hyperparameters values are chosen for novel Bayesian shrinkage. Therefore, we compare some Bayesian and non-Bayesian shrinkage with the posterior prediction distribution in all cases, in which β_1 is the intercept.

- Scenario A: We let $p < n$. The sample size is $n = 80$, the number of independent variables is $p = 35$, 26 out of 36 coefficients are exactly equal to zero, and the rest are non-zero.

- Scenario B: We let $p > n$.

  There are three datasets that are different only in terms of the precisions of their signal-to-noise-ratio; $\lambda = (0.02, 2, 2000)$. To analyse the results and calibrate hyperparameters correctly, we symbolise the simulation samples for every scenario by adding sample code (I , II, III) based on the respective precision of the model.

- Scenario B1: $n = 35, p = 45$ and the non-zero coefficients are 17 out of 46 , where the first coefficient (β_1) is the intercept.

  $\beta_{(1,\ 4,\ 9,\ 15,\ 16,\ 21,\ 22,\ 23,\ 25,\ 34,\ 37,\ 38,\ 39,\ 42,\ 44,\ 45,\ 46)} = (40.25, -2.827, 1.547, 0.984, -1.482,$
  $-0.143, 2.831, 3.101, -7.985, -0.633, -47.359, 2.069, 0.847, -5.339, 3.858,$
  $2.493, 11.889).$

  The scenario B1 is divided into (B1I, B1II, B1III) where the precision of the mode are $\lambda = (0.02, 2, 2000)$, respectively.

- Scenario B2: $n = 35, p = 100$ and the non-zero coefficients are 16 out of 101 for (β).

  $\beta_{(\ 1,\ 8,\ 9,\ 14,\ 19,\ 24,\ 25,\ 39,\ 48,\ 54,\ 76,\ 85,\ 87,\ 91,\ 98,\ 101)} = (-2.360, 1.584, -4.189, 3.671, -1.882,$
  $-5.665, -126.665, 16.481, 37.604, 0.062, -218.994, -6.303, 5.718, -23.554,$
  $6.965, -21.616).$

  The datasets in this scenario are B2=(B2I, B2II, B2III), where the precisions of the mode are $\lambda = (0.02, 2, 2000)$, respectively.

- Scenario B3: $(n = 50, p = 250)$ and the non-zero coefficients are 14 out of 251.

  $\beta_{(1,\ 33,\ 37,\ 52,\ 59,\ 62,\ 63,\ 161,\ 180,\ 193,\ 195,\ 222,\ 247,\ 251)} = (40.25, -5.415, -1.995, -1.741,$
  $52.031, -1.199, 6.079, 0.393, -0.087, 3.025, 2.99, 3.978, -6.859, 0.369).$

  The datasets in this scenario are B3=(B3I, B3II, B3III), where the precisions of the mode are $\lambda = (0.02, 2, 2000)$, respectively.

- Scenario C: $n = 200, p = 1000$ and the non-zero coefficients are 11 out of 1001.

  $\beta_{(1,\ 71,\ 212,\ 326,\ 333,\ 434,\ 584,\ 681,\ 784,\ 847,\ 1001)} = (-2.36, 1.359, 1.469, 0.3031,$
  $0.999, -0.8125, 2.336, -116.491, 8.678, 13.734, -0.851).$

  The datasets in this scenario are B4= (B4I, B4II, B4III), where the precisions of the model are $\lambda = (0.02, 2, 2000)$, respectively.

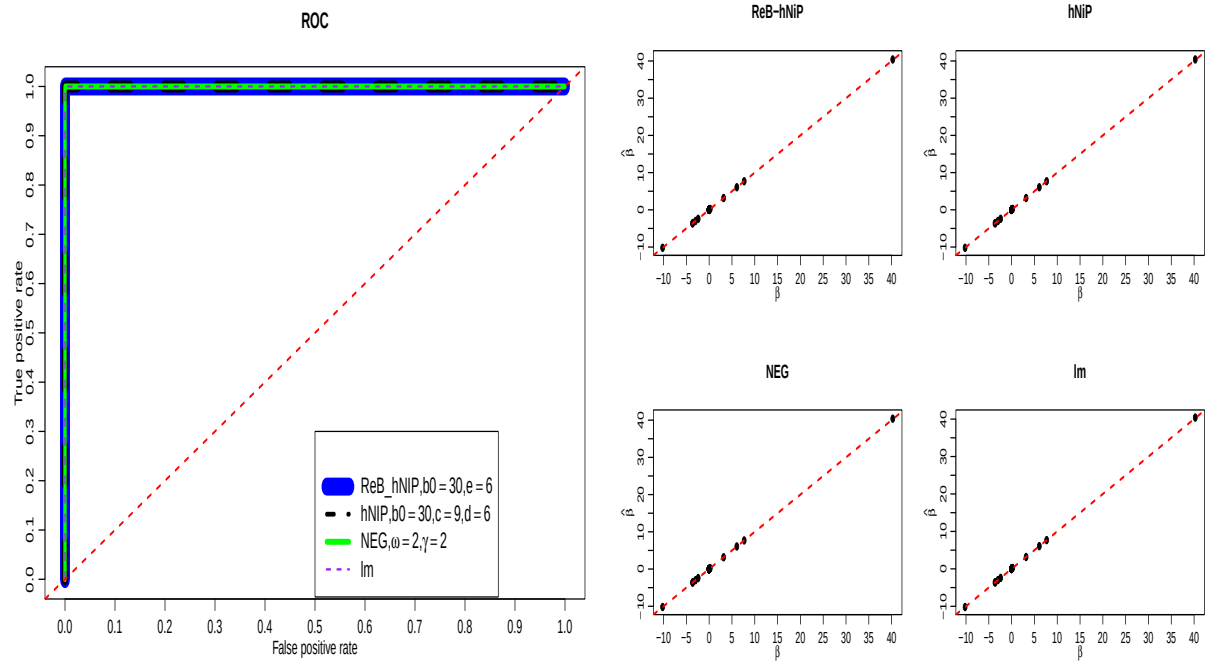
- Scenario D: $n = 200, p = 5000$ and the number of non-zero coefficients are 5 out of 5001, $\beta_{(1,\ 2,\ 3,\ 50,\ 500)} = (10, 5, 5, -3, -3)$.

  The datasets in this scenario D=(DI, DII, DIII), where the precisions of the model are $\lambda = (0.02, 2, 2000)$, respectively.

4.4 Applying Shrinkage priors to Non-high Dimensional Data

To address a novel statistical method, it is necessary to check the performance and traces its plot convergence under a statistical model. Thus, in this section, the proposed priors discussed in Section 3.2 are implemented with a standard Bayesian linear regression model, $n > p$. Also, we address the behaviour of the hyperparameters of the proposed shrinkage for both hNiP and ReB-hNiP in terms of prediction consistency visually. For that reason, a pre-request test was done. We applied the dataset that was described as "Scenario A" in Section 4.3, a non-high dimensional dataset. We run an MCMC algorithm for the suggested shrinkage model based on diverse shrinkage priors, therefore are compared with: NEG with $\omega = 2$ and $\gamma_\phi = \sqrt{2}$. Both proposed priors are also compared with a multivariate linear model by using lm R code. As we discussed in Section 3.2, we concentrate on the mean of hyperparameter κ around $\frac{3}{2}$ which is not allowed for very large variability with $c = 9$ and $d = 6$ the upper bounds of that is $b_0 = 30$. In the hNiP prior distribution, the shape parameter controls the shrinkage amount of coefficients regression. The ReB-hNiP is the third shrinkage prior that is used to estimate regression model. In these cases, we have to choose only the prior value b_0, e_0 because our aim is to obtain the density distribution plot of the hyperparameter, similar to a horseshoe curve, a feature we obtained where $a = b = \frac{1}{2}$ in the beta distribution. Furthermore, we based it on $e_0 = 6$ and $b_0 = 30$. Finally, the prior distribution for the error model has the gamma distribution $a_\lambda = 2, b_\lambda = 0.6$.

The MCMC algorithm was run for scenario A, the dataset being $n = 80, p = 35$ is based on all the methods mentioned above. The number of iterations was $M = 30000$ and we kept 75% in a total of iteration. We noticed that, posterior means of the MCMC outcome for all different models were the same as by their true values. In order to check the performance of models, we have compared all four methods by ROC curves as can be seen in Figure 4.1a. It is clear that all outcomes for all models had similar performance. Because all the curves tended to cluster in the close upper-left corner of graph, that is, as we explained in Section 2.5.3. All regression coefficients in the models are fitted exactly and are same as to true value. The scatter plots for the true coefficients parameter against of the estimated posterior based on on 3-time standard deviation (3sd) error bar values for each models are show separately in Figure 4.1b. These graphs shows all the coefficients estimated with tiny variability.

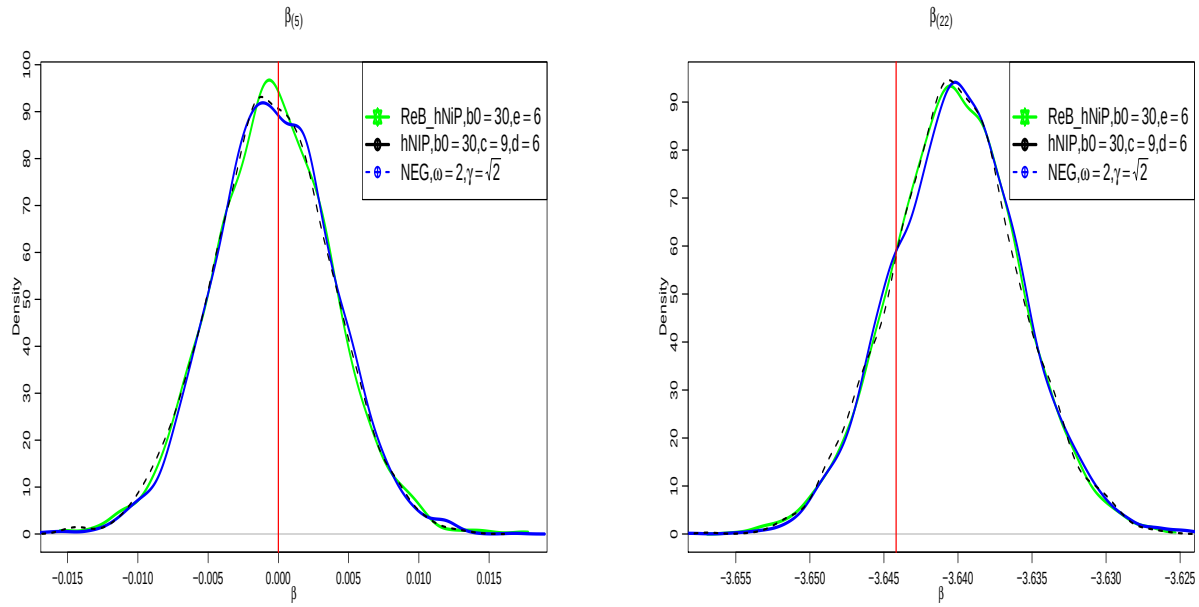


(a) Receiver operating characteristic (ROC) curves (b) Scatter plots

Figure 4.1: Estimated regression coefficients. Panel (a) shows the ROC curves and scatter plots for all the shrinkage priors; the green line represents for the NEG with $\omega = 2$ and $\gamma = \sqrt{2}$, and the thick bold blue line is refers to ReB-hNiP $b_0 = 30, e_0 = 6$. The black, dotted, dashed line represents to the hNiP $b_0 = 30, c = 9$ and $d = 6$, and multivariate linear model lm is described by dashed purple line. Panel (b) shows scatter plots for all the methods described in the ROC panel. The sample size is $n = 80$ and $p = 35$.

Subsequently, we plotted the posterior densities as examples of both non-zero and zero regression coefficients, as can be seen in Figure 4.2. Furthermore, we have established traces for all resulying relying on the shrinkage priors in Figure B.6 to reveal stability chains of the posterior distribution for each shrinkage priors convergent to the true value. Figure 4.3 displays the the plots of the posterior density distributions of two different coefficients involving zero and non-zero for all cases. Both panels indicate that the posterior predictive density distributions are well-fitted.

To conclude, in scenario A, to check our proposed Bayesian shrinkage model construction in this section, we used posterior predictive checking, which was discussed in the Section 2.5.4. Consequently, posterior predictive distribution plots for both hNiP and ReB-hNiP are displayed

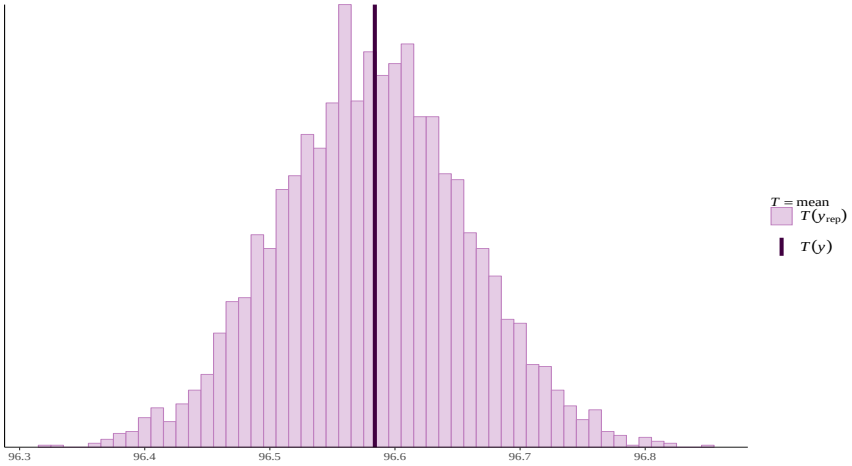


(a) Zero coefficient parameter β_5

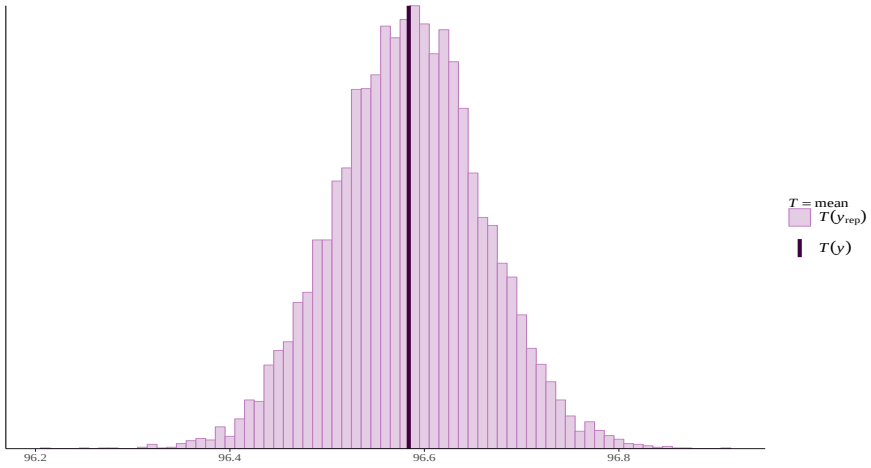
(b) Non-zero coefficient parameter β_{22}

Figure 4.2: Posterior density curves for regression coefficients (β_5, β_{22}) for zero and non-zero regression coefficients, respectively. The green lines refers to the posterior density of the ReB-hNiP model, the black lines represents to the hNiP model and dashed blue lines is refers to the NEG model. Red vertical line is the true value of the coefficient (β_5, β_{22}). The sample size of $n = 80$ and the number of coefficients was $p = 35$.

in Figure 4.3, which shows significant results when we compared them with test statistics by using the mean of the dependent variable y compared to the mean of the posterior predictive $\hat{y}$. By using R Package *bayesplot* (Gabry, 2017). Finally, there is sufficient data to supports prior distributions to estimate the coefficients regression model.



(a) hNiP



(b) ReB-hNiP

Figure 4.3: Posterior predictive distribution checking plots for the Bayesian regression model based on ReB-hNiP and hNiP shrinkage priors. The black vertical line represents the test statistics and is the mean of the dependent variable of the true value, and the histogram represents the mean posterior predictive replication data $\hat{y}$.

4.5 Hyperparameter Calibration of High-dimensional Data

In this section, we address the procedure of selecting hyperparameters values for our priors. Before explaining our methodology, we will recap some information. An important task for any Bayesian statistical models is variables selection. In traditional methods, this is done by adding a penalty term to the likelihood function, minimise it and estimate the coefficients. We have addressed some frequentist methods in Section 2.2 and a penalty function that includes at least one or more tuning parameters that control the amount of shrinkage. The tuning parameter can be fixed by several methods, such as 10-fold cross-validation, Bayesian information criteria (BIC) (Schwarz et al., 1978), among others. However, only employing penalty methods is not enough to select variables; for that purpose, researchers used several approaches, such as the spike and slabs prior in Equation 2.18 and some other strategies that can be found in Section 2.5.1. Recently, shrinking the high-dimensional parameters of coefficients and selecting the variables is a method that has been widely utilised compared to more traditional methods. Therefore, the number of Bayesian shrinkage prior distributions were used that have a scale mixture of Gaussian distribution and are also called "Bayesian regularisation methods". The hyperparameters for Bayesian shrinkage priors are fixed in various ways, such as using gamma distribution with appropriate scale and shape parameters or an empirical Bayes after finding the marginal likelihood estimation of the hyperparameters; see Bayesian LASSO form Park and Casella (2008) as an example. This kind of prior consideration, that is, absolute continuous distribution and, at the same time, shrinking the coefficients results in load mass around zero. Also, heavy-tailed are beneficial to keep the coefficients different from zero. Consequently, selecting variables is done by using simple tools, such as a t-test.

The main objective in this section is to calibrate the hyperparameters for both hNiP and ReBhNiP priors. Furthermore, we try setting guidelines for researchers, for implementing both shrinkage priors for specific modelling. We do the same procedure for the NEG prior distribution, if possible. We consider the high-dimensional datasets $p > n$ for which we use the simulated data Scenario B1 from Section 4.3, especially for those cases scenario where the precision of model $\lambda = 2$, because it is our goal to investigate the performance of the novel shrinkage priors and select the optimal value range of the hyperparameters. Consequently, we will generalise the results for other sample sizes.

To assess the sensitivity of the shrinkage priors, we run an MCMC algorithm for the Bayesian linear regression model depending on several candidates values for each hNiP, ReB-hNiP and NEG prior. For example, the candidate values that were used for the hNiP prior start with $c = 0.3, 0.4, 0.5, \ldots, 9$, where $d = 0.2, 0.3, \ldots, 6$ with our assumption that the expected value of the gamma distribution was $E(\kappa) = \frac{c}{d} = 1.5$, and for ReB-hNiP we used different values of rescaled beta $e_0 = 2, 3, 4, \ldots, 12$. The upper-bound parameters of the inverse Pareto distribution is the second level, such as $b_0 = (30, 50, 100, 250, 500, 1000, 5000, 10000)$ is used of several datasets. The posterior distribution is conducted numerically with variability values of prior distributions for each dataset (B1II, B2II, B3II). To check the coefficients, which are zero or not, we test each posterior coefficient relying on the t-test $= \frac{|\hat{\beta}_j|}{sd(\hat{\beta}_j)}$ with probability level 0.05, that is, thresholding by an approach 95% credible interval. Hence, in this chapter, for inference, the posterior mean is considered. WIAC and DIC criteria, as described in Section 2.5.2 was used for model selecting and detecting the optimal model. Some measures were used to check the accuracy of selecting either the same non-zero or zero regression parameters that include TP, FP, FN, and TN, which are described in Section 2.5.3. The MCMC ran $M = 30000$ burning 50% in total. The results for some datasets are shown graphically in the following section to analyse them.

The results of the MCMC algorithm are presented graphically and summarised in the tables for some of the datasets mentioned above. The first datasets is B1II, the sample size for which is $n = 35$ and $p = 45$; see Section 4.3 for more details about it, where the numbers of non-zero regression coefficients have an intercept equal to 17. The results for both the hNiP and ReB-hNiP shrinkage models are summarised in Table 4.1 based on the current datasets. Thus, the number of estimated non-zero coefficients are selected by thresholds at a 1.96 credible interval, which is denoted in Table 4.1 by No.$\hat{\beta}$. FN (false negative) represents the number of estimated non-zero coefficients, but they are zero coefficients in the true datasets opposite each upper bound in each table row.

The number of estimated non-zero coefficients approaches the number of true non-zero coefficients in the simulated data as long as the value of the upper-bounds parameter b_0 increases and the values of WAIC and DIC decrease. The optimal range of hyperparameter values that can be used for models in this dataset is denoted by e_0^*, c^* which leads to good-fit sparse coefficients. To explain the figures in Table 4.1, consider an example: the upper bound $b_0 = 5000$, the

Table 4.1: Hyperparameter prior calibration based on WAIC and DIC criteria for both ReB-hNiP and hNiP shrinkage priors distribution using several hyperparameter values. The used dataset has asize of $n = 35$ and $p = 45,$ and the number of non-zero regression coefficients in true value is 17. No.$\hat{\beta}$ represents the number of estimated non-zero regression coefficients, and both e_0^*, c^* describe the bounds on the hyperparameters for hNiP and ReB-hNiP, respectively. $e_{0(\text{min})}$ illustrates the rescaled beta parameter value of the ReB-hNiP prior and has the minimum WAIC, $c_{(\text{min})}$ represents the hyperparameter value of the hNiP prior, which has the minimum WAIC. FN refers to false negative values.

		ReB-hNiP						**hNiP**					
Dataset	b_0	**No.**$\hat{\beta}$	(e_0^*)	**FN**	$(e_{0_{min}})$	**WAIC**	**DIC**	**No.**$\hat{\beta}$	c^*	**FN**	$c_{(\text{min})}$	**WAIC**	**DIC**
	30	15	3-12	0	6	95.53	109.37	15	≤ 1.8	0	9.6	92.58	106.25
	100	16	2-12	0	9	90.45	104.49	16	≤ 2.8	0	0.9	92.28	106.51
$n = 35,\ p = 45$	250	17	2-12	0	12	90.54	110.27	16	≤ 3.3	0	2.4	91.82	106.53
	500	17	2-12	0	7	88.09	110.27	17	≤ 3.0	0	1.8	89.19	103.17
	1000	17	2-12	0	5	85.08	102.27	17	≤ 2.4	0	1.4	86.77	100.73
	2500	17	2-12	0	8	79.86	99.18	17	≤ 2.1	0	1.2	80.46	94.24
	5000	17	2 - 7	0	6	78.47	93.46	17	≤ 1.5	0	1.5	78.5	91.68
	10000	17	2 - 7	0	2	80.32	92.23	17	≤ 1.3	0	0.3	78.9	91.02

number $\hat{\beta} = 17$ and FN equal to zero for both shrinkage prior models hNiP and ReB-hNiP. The optimal bounds of the hyperparameter e_0 is restricted between two and seven; however, using any value outside this bound leads to over-shrinkage, which is also true for the hNiP prior the hyperparameter c bound between 0.3 and 1.5, similar to ReB-hNiP utilising any values greater than $c > 1.5$ that cause the over-shrinkage model. e_{0min} represents the optimal value of the ReB-hNiP model that has the minimum WAIC and DIC exactly fitting the model based on $b_0 = 5000$. Furthermore, c_{min} is the optimal value of hNiP and has a minimum WAIC and DIC compared to other values of b_0.

Concluding this example, the values in column e_0^*, c^* represents bounds of the hyperparameters in this restriction range. If we use values out of these restriction ranges, the number of coefficients aggressively shrink to zeros. On the other hand, regarding the sensitivity analysis of the hyperparameter (tuning parameters) selection for the Bayesian regression model depending on the NEG prior distribution. We only shows some results graphically for the NEG model, because it is difficult to summarise the MCMC outcomes in a table, such as hNiP and ReB-hNiP priors, due to unstable number of non-zero regression coefficients from several models. Figure 4.4 shows the relationship between parameter bounds and the hyperparameters $c, d,$ based on thehNiP prior distribution. There are three lines from the above panel in Figure 4.4. The

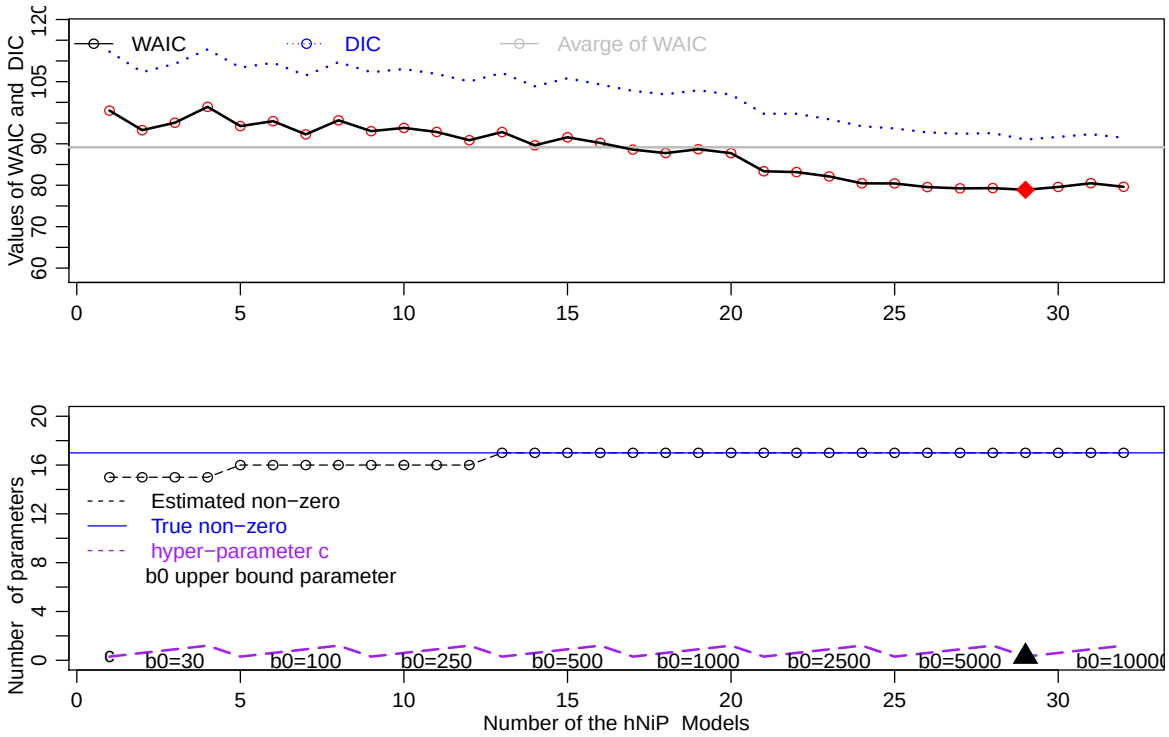


Figure 4.4: The relation between upper bounds, hyperparameters and both criteria (WAIC and DIC) using the hNiP shrinkage prior and the best number of candidate values of the hyperparameters. In the top panel, the a blue dotted line represents the DIC, and the black line with red circle illustrates the WAIC values for the number of models. The grey solid lines refers to the average values of the WAIC. In the bottom panel, the blue solid line indicates the number of true non-zero coefficients. The black dashed lines with black circles represents the number of estimated regression coefficients. The purple dashed line denotes the used candidate values for the parameter upper bound b_0 and the best range of the hyperparameters $c = \{0.3, \ldots, 1.5\}$. The dataset $(n = 35,\ p = 45)$ is used.

solid grey line indicates the average of WAIC, the blue, dotted line illustrates the DIC criteria and the solid, black line with red circles represents the WAIC criteria. Both DIC and WAIC decline as long as the value of upper-bound b_0 increases the limited range of c, d. We believe that $c \leq 1.5$ is the best range overall for the upper-bound parameter b_0 because it has the minimum values of the WAIC, which is explained by the solid red dot opposite the black triangle in the bottom panel. The large upper-bound parameter b_0 with small prior values c, d has minimum values for both WAIC and DIC. Furthermore, the smallest values of b_0 shrink those coefficient values, which are very small in the dataset toward the zeros; particularly, the coefficients that are between 0 and 1, become zero. Therefore, the results lead us to use the large b_0 with

enough small values of both c, d, because they are more flexible to fit the regression parameters due to the large variance of the second layer of the hierarchical hyperparameter for the hNiP prior. It helps to shrink small values of the regression coefficients in the model, and this explanation is true for the ReB-hNiP prior with hyperparameter bounds rescaled Beta e_0.

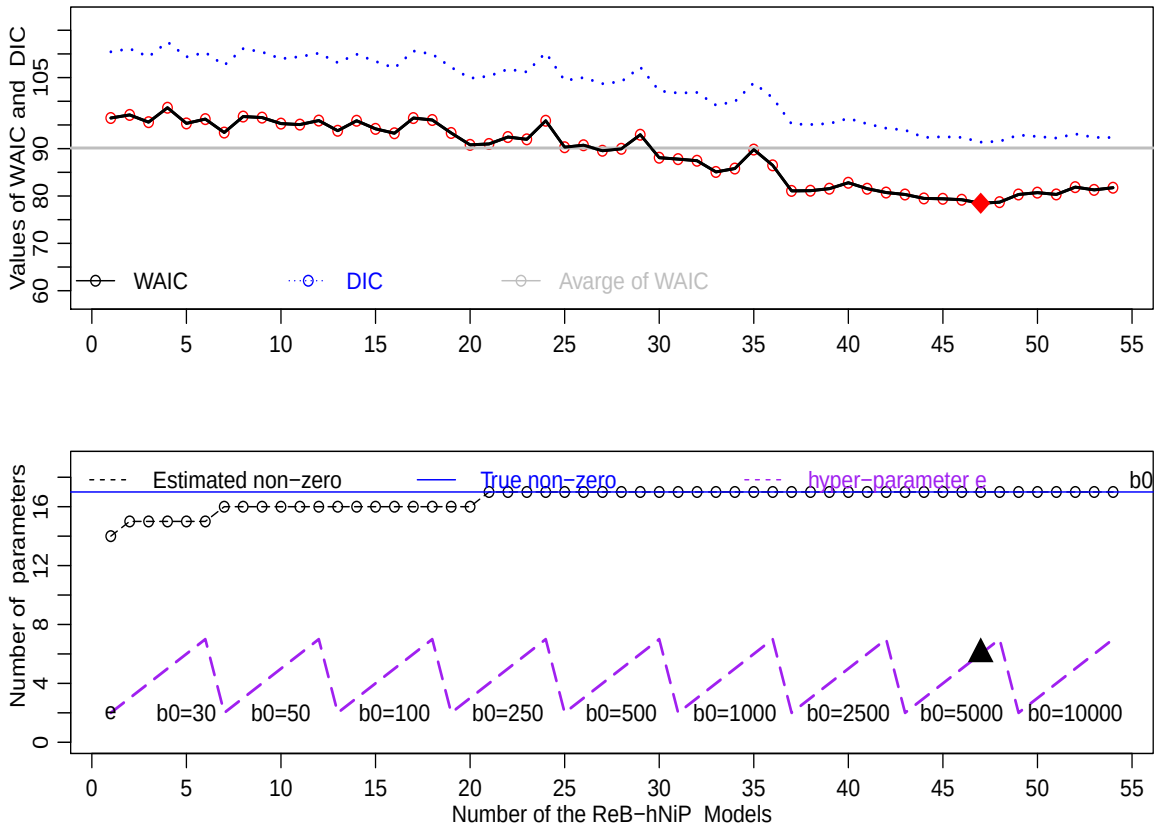


Figure 4.5: The relation between the upper bounds, hyperparameters and both criteria (WAIC and DIC) by using the ReB-hNiP shrinkage prior and the best number of candidate values of hyperparameters. In the top panel, the blue dotted line represents the DIC, and the black line with red circle illustrates the WAIC values for the number of the models. The solid grey line refers to the average of the WAIC. In the bottom panel, the blue solid line indicates the number of true non-zero coefficients. The black-dashed line with black circles represents the number of estimated regression coefficients. The purple dashed line denotes the used candidate value's upper bounds b_0 and the best range of hyperparameters e_0. The dataset had a size of $n = 35$ and $p = 45$ was used.

On the other hand, the hyperparameter sensitivity analysis and selection by using ReB-hNiP shrinkage and the NEG prior distribution were addressed by utilising WAIC and DIC as we did

with the hNiP model. Consequently, we have displayed the relationship between the parameter bounds and the hyperparameter e_0 in Figure 4.5 for the ReB-hNiP prior distribution. The results show similar changes compared to the hNiP plots. However, we have put restrictions on the range of the hyperparameter of the rescaled beta e_0, and its interval of restriction of parameter e_0 becomes smaller as long as the parameter values of the upper bound b_0 become larger.

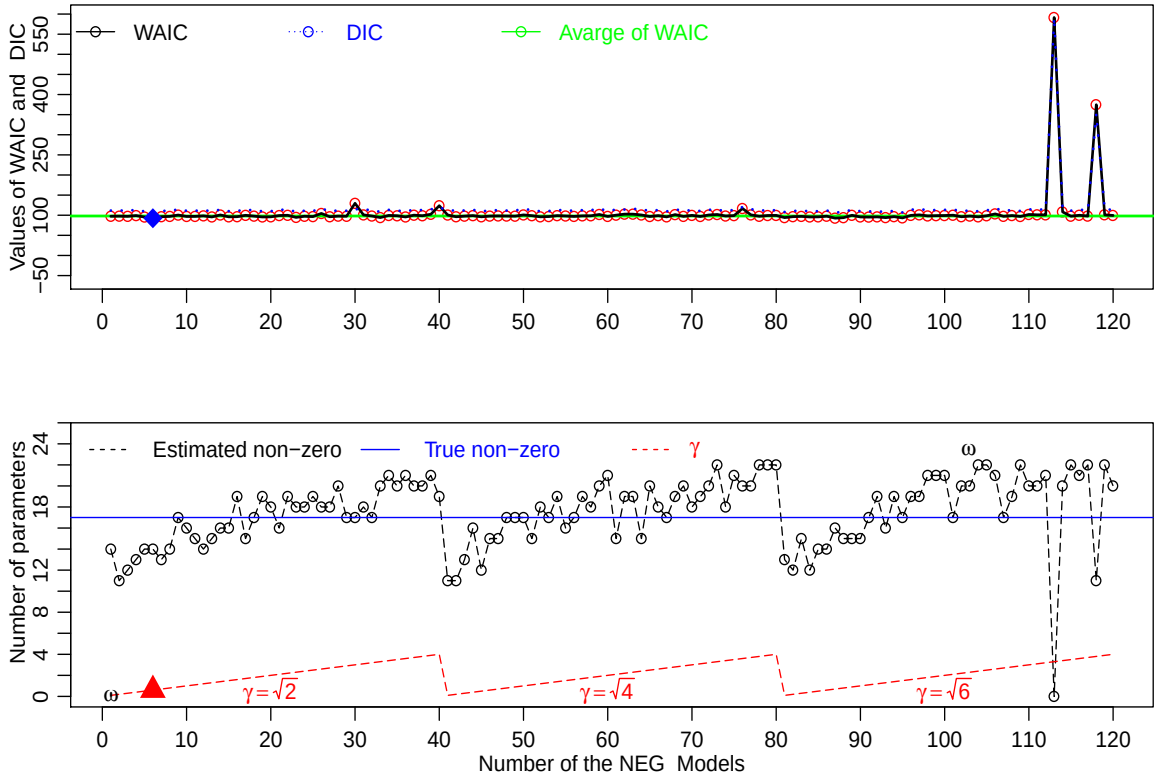


Figure 4.6: The relation between upper bounds, hyperparameters and both criteria (WAIC and DIC) by using the hierarchical NEG shrinkage structure prior and the best candidate values of the hyperparameters ω, γ. In the top panel, the blue dotted line represents the DIC and the black line with red circle illustrates the WAIC values for the number of the models. The solid green line refers to the average value of the WAIC. In the bottom panel, the blue solid indicates the number of true non-zero regression coefficients. The black dashed line with black circle represents the number of estimated regression coefficients. The red dashed line denotes the used candidate values of the hyperparameter for the best range. The dataset had a size of $n = 35$ and $p = 45$ was used.

As a consequence, we will not only use our priors in the next section to show the performance

of our proposed shrinkage prior, but we will also utilise the regularisation methods highlighted in Section 4.1. Regarding the tuning-parameters selection for the NEG shrinkage prior, WAIC and DIC were implemented, as with the ReB-hNiP and hNiP priors. However, the model based on the NEG shrinkage prior does not provide suitable results, which is clear from Figure 4.6. The number of estimated non-zero coefficients for most cases did not correspond to the same coefficients in the estimated data, and for the candidates values, such as c, d and e_0, were used to fit the model, the coefficients were over-estimated.

As we mentioned above, the plots were shown only for the bounded range of the hyperparameters. Thus, if we utilise values larger than these ranges, most coefficients aggressively shrink close to zero, which leads to increased WIAC and mean square error (MSE). Some figures are displayed in the Appendix B which show the plots for all ranges of the hyperparameters that were addressed with a sensitivity analysis and hyperparameter selection for the proposed Bayesian shrinkage prior that was used for the current section. The graphs clearly provide an overview of how the number of estimated non-zero coefficients changes WAIC and DIC by the both hyperparameters for the hNiP, ReB-hNiP and NEG shrinkage priors, for example Figure B.3.

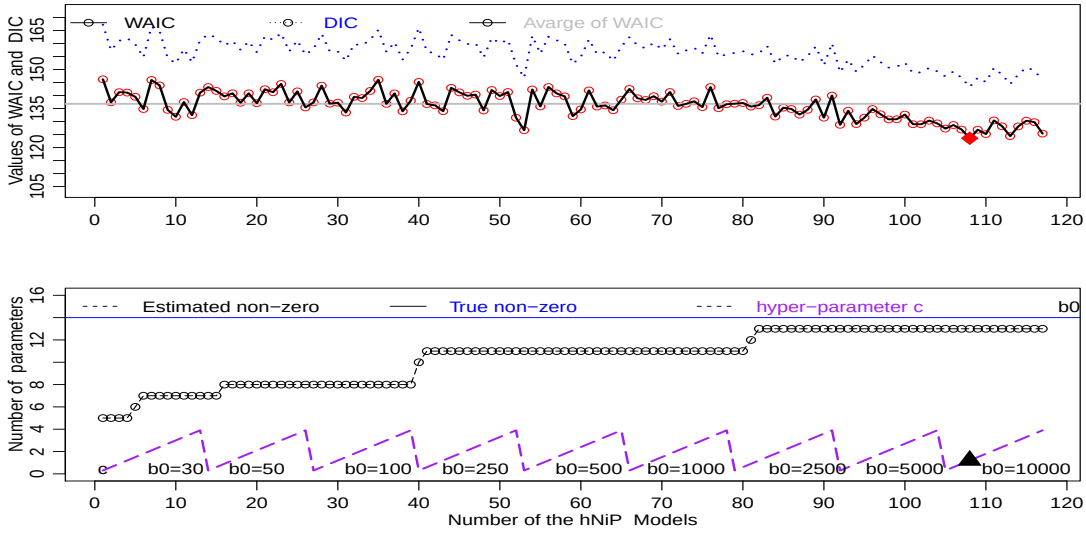
Table 4.2: Hyperparameter prior calibration based on WAIC and DIC criteria for both ReB-hNiP and hNiP shrinkage prior distribution when using different hyperparameter values. The sample size of the dataset is scenario B4, that is, $n = 50$ and $p = 250$. The number of true non-zero coefficients is 14. No.$\hat{\beta}$ represents the number of estimated non-zero regression coefficients ${e_0}^*, c^*$ describes the bounds on hyperparameters, $e_{0(\text{min})}$ illustrates the rescaled beta value parameter of the ReB-hNiP prior that has a minimum WAIC, and $c_{(\text{min})}$ represents the hyperparameter value of the hNiP prior, which has the minimum WAIC. FN refers to false negative values.

		ReB-hNiP						**hNiP**					
Dataset	b_0	**No.**$\hat{\beta}$	(e_0^*)	**FN**	(e_{0min})	**WAIC**	**DIC**	**No.**$\hat{\beta}$	c^*	**FN**	$c_{\textbf{(min)}}$	**WAIC**	**DIC**
$n = 50,\ p = 250$	30	7	≤ 12	0	8	126.53	146.51	7	1.5-10	0	3	131.84	152.33
	100	10	≤ 12	0	9	131.96	150.01	8	≤ 10	0	9	131.41	151.23
	250	11	≤ 12	0	6	130.45	150.5	11	≤ 10	0	3.9	131.49	151.94
	500	11	≤ 12	0	9	132.65	152.96	11	≤ 10	0	0.3	126.67	146.85
	1000	11	≤ 12	0	4	132.74	152.37	11	≤ 10	0	6	132.71	152.45
	2500	13	≤ 12	0	8	130.52	150.83	13	1.2-6.3	0	4.8	130.85	150.77
	5000	13	≤ 12	0	6	125.55	145.45	13	≤ 4.5	0	0.3	128.75	148.96
	10000	13	≤ 12	0	8	121.85	141.84	13	≤ 4.2	0	1.2	123.58	143.3

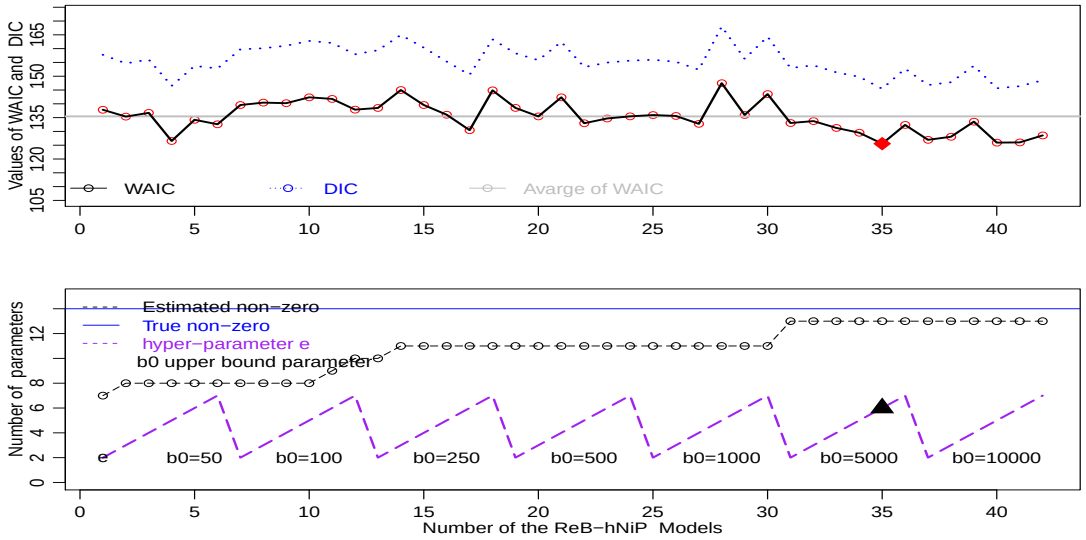
To evaluate and select the tuning parameters of the shrinkage prior distributions mentioned

previously and then make a decision about how both proposed Bayesian shrinkage priors are implanted, we used the different simulated datasets which were described previously. As in the first dataset case, we had some results. For example, Table 4.2 consists of minimum values for WAIC and DIC with restriction ranges for both e_0^*, c^*, which are hyperparameters for both the ReB-hNiP and hNiP and the dataset with $n = 50$ and $p = 250$, which refers to scenario B3II. Meanwhile, the plots in Figures 4.7 correspond to the same data in the last table. Likewise, the quantities of WAIC and DIC declined as the values of the upper bounds increased. The bound of hyperparameter c also decreased, similar to the previous example. Most of the coefficients were significantly convergent to their true value; only one coefficient was not because its true value is very small, that is, $\beta_i = -0.086$, to investigate and fit this value, we try to use larger value of the upper bonds b_0 later in this chapter. As we noticed from Figures 4.7, both priors, the minimum values of the WAIC and DIC, were located between the intervals of values that are very close to each other. Choosing any values in this range should thus produce similar results. For this reason, we applied the simulation datasets (B1II, B2II, B3II), which are explained in Section 4.3, as datasets in the case $p > n$. Figures 4.8 and 4.7 displayed the behaviour of the hyperparameter calibration and the selection of the best range of candidate values based on the shrinkage prior that studied in this section to guide future research. Figure B.4 displayed the behaviour of the hyperparameter calibration and the selection of the all range of the candidate values based on NEG shrinkage prior for sample size $n = 35$ and $p = 100$.

We apply both proposed shrinkage methods under this suggestion in the following sections to fit the high-dimensional datasets and compare them with some Bayesian and non-Bayesian shrinkage methods, as stated in the first section of current chapter. The proposed methods using regression models can be checked by posterior predictive distribution to check how our methods were satisfactory modelled.

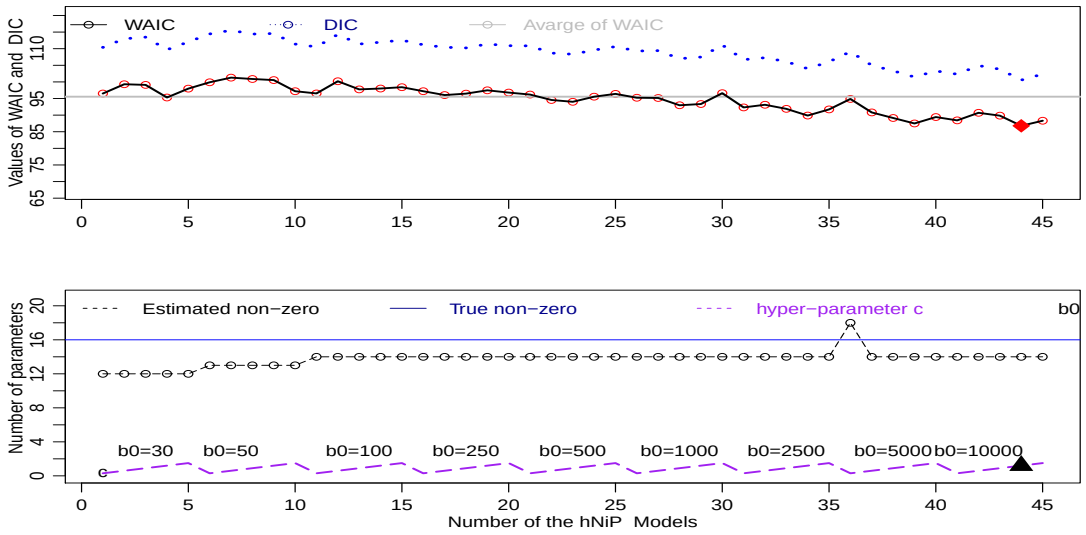


(a) hNiP

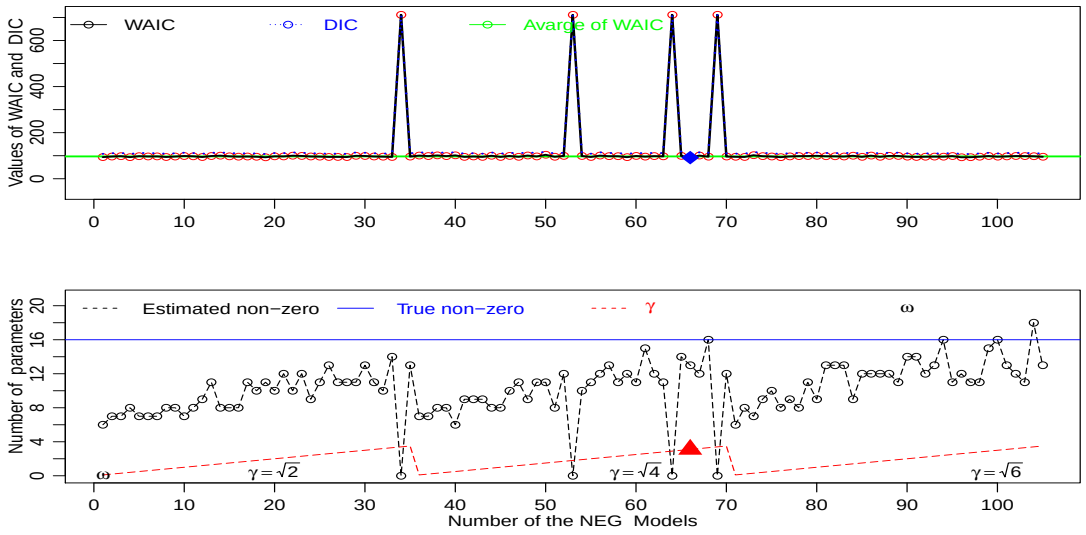


(b) ReB-hNiP prior

Figure 4.7: Panels (a) hNiP and (b) ReB-hNiP shows the relation of the hyperparameter calibration between upper bounds and second-layer hyperparameters using (WAIC and DIC) criteria for the best range candidate values. Panels (a) and (b) are illustrated as follows: The blue dotted line represents the DIC, and the black line with red circle illustrates the WAIC values for the number of the models. The grey solid lines refers to the average value of the WAIC. In the bottom panel, the blue solid line indicates the number of true non-zero coefficients. The black dashed lines with black circle represents the number of estimated regression coefficients. The purple dashed line denotes the candidate values used as the upper bounds b_0 and the best range of hyperparameters either for hNiP c or ReB-hNiP e_0. The dataset had a size of $n = 50$ and $p = 250$ was used.



(a) hNiP



(b) NEG prior

Figure 4.8: Graphs (a) hNiP and (b) NEG shows the relation of the hyperparameter calibration between upper bounds and second-layer hyperparameters using (WAIC and DIC) criteria for the best range candidate values. Graphs (a) and (b) are illustrate as follows: The blue dotted line represents the DIC, and the black line with red circle illustrates the WAIC values for the number of the models. The grey solid line refers to the average value of the WAIC. In the bottom panel, the blue solid line indicates the number of true non-zero coefficients. The black dashed line with black circle represents the number of estimated regression coefficients. Purple dashed line denotes the several values of used candidate upper bounds b_0 and the best range of hyperparameters either for hNiP c or NEG shrinkage prior with ω and γ. The dataset had a size of $n = 35$ and $p = 100$ was used.

4.6 Numerical Comparison of Some Regularization Methods

In this section, we compare and evaluate the performance of our proposed shrinkage prior distribution with the same family of scale mixtures of normal distributions, including the Bayesian LASSO (Park and Casella, 2008) and horseshoe (Carvalho et al., 2010) and also non-Bayesian regularization methods, such as SCAD (Fan and Li, 2001) and LASSO (Tibshirani, 1996). The simulated datasets for scenarios B and C are utilised and explained in Section 4.3. The findings are conducted using R packages, both *ncvreg* (Breheny et al., 2018) and *lars* (Hastie et al., 2007) for SCAD and LASSO, respectively, and the tuning parameters were selected by 10-fold cross-validation. The *monomvn* package (Gramacy et al., 2019) in R language was used for the horseshoe shrinkage prior and the Bayesian LASSO (Blasso) assuming an error model have gamma prior of shape 2 and scale 0.6, which we used in our proposed model as flat prior. We sett the *monomvn* package code (RJ= off, Reversible Jump) to compute the posterior distribution as an original paper for both the Bayesian LASSO and horseshoe. Therefore, we select the same variables as the procedure modelled in the previous section. Thus, we set the gamma distribution as a prior of the hyperparameter, as utilised in both the Bayesian LASSO and horseshoe with the default values of shape 10 and scale 5 suggested by the $monomvn$ package.

To run the MCMC algorithm faster for both the Bayesian LASSO and horseshoe, we have to set the length of thin equal as 5. Two different hyperparameter values of our proposed shrinkage priors are used based on the discussed of the results in the previous section to establish and assess the performance of the proposed Bayesian shrinkage prior, especially in high-dimensional datasets. The hyperparameter values of hNiP are $c = 0.9, d = 0.6$, which provide the large variability of κ and the value of the hyperparameter $e_0 = 5$ with two different values of upper bound $b_0 = (10000, 100000)$ for the proposed shrinkage prior distributions. We ran the MCMC algorithm with $M = 20000$ iterations and kept half the total iteration for the shrinkage methods belonging to the family of scale mixtures of normal distributions by applying them on the datasets described in scenario B. The iterations were $M = 10000$ for the dataset in scenario C due to time consumption and limited time-run of the model in the system of the iceberg (high performance computing cluster networks) that I used to run my model at University of Sheffield.

We present our overall outcomes for three different datasets with the purpose of the evaluation and performance of all the Bayesian and non-Bayesian shrinkage priors mentioned above that

were used throughout this section. We have presented the average results of the 100 posterior replication samplers for all the dataset cases and for all the shrinkage methods in Section 4.3. The results are displayed for each dataset separately in the following tables based on the posterior mean distribution. For example, Table 4.3 shows the figures of some shrinkage priors, to

Table 4.3: Comparison and performance between some Bayesian shrinkage priors and frequentist methods, in which the sample size is $n = 35, p = 45$. λ indicates the values of the error model. The number of true non-zero coefficients is 17. Where FN is false negative, FP false postie, AUC is area under curves and MSE is mean square error.

	λ	**No.**$\hat{\beta}$	**Sensitivity**	**Specificity**	**MSE**	**FN**	**FP**	**AUC**
hNiP	0.02	16	100%	96.67%	0.023	0	1	97.83%
	2	17	100%	100%	0.00008	0	0	100%
	2000	17	100%	100%	0.00005	0	0	100%
ReB-hNiP	0.02	15.83	100%	95.52%	0.8701	0	1.5	96.46%
	2	16.92	100%	99.78%	0.0617	0	0.08	99.83%
	2000	16.57	98.82%	98.82%	0.0200	0.45	0.75	98.98%
Bayesian LASSO	0.02	17	100%	69.00%	610.239	0	13	71.70%
	2	17	100%	69.00%	631.826	0	13	71.70%
	2000	17	100%	69.00%	635.704	0	13	71.70%
Horseshoe	0.02	17	98.71%	96.68%	0.0899	0.22	1	97.34%
	2	17	100%	100%	0.0047	0	0	100%
	2000	17	100%	100%	0.0001	0	0	100%
LASSO	0.02	46	38.46%	66.80%	25491.57	20.46	4.23	46.33%
	2	46	41.52%	75.00%	25270.26	19.58	3.14	50.61%
	2000	46	35.06%	58.10%	25276.26	20.6	5.95	42.28%
SCAD	0.02	46	100%	80.56%	25143.77	0	7	84.78%
	2	46	100%	80.56%	25440.94	0	7	84.78%
	2000	46	100%	80.56%	25434.95	0	7	84.78%

examine the performance of the variables selection for the shrinkage priors explored in former sections for the dataset in Scenario B1 with error noise cases $(\lambda = 0.02, 2, 2000)$. The summarised results represent the average sensitivity, specificity, No.$\hat{\beta}$ (the number of estimated non-zeros), false negatives (FN), false posties (FP), areas under curves (AUC), all of which were explained in Section 2.5.3, as well as the mean square error (MSE) for the dataset described in B1, which had a sample size of $n = 35$ and $p = 45$.

As we can see from the figures in the Table 4.3, our proposed shrinkage prior distribution outperformed all other shrinkage methods. For instance, the estimated regression coefficients are exactly the same as the coefficients of the true values for both the dataset cases in which

the precision of model was $\lambda = 2$ and 2000, and both the FP and the false negatives (FN) were zero. In the noise dataset, when $\lambda = 0.02$, the average of sensitivity is 100% and the specificity is 96.67% for hNiP methods, which are better than horseshoe methods because the sensitivity and specificity of the later methods are 98.71% and 96.68% respectively. Despite the second proposed ReB-hNiP shrinkage prior outperforming the Bayesian LASSO (Park and Casella, 2008), SCAD (Fan and Li, 2001) and LASSO (Tibshirani, 1996) methods, but it is not satisfactory compared ton either the hNiP nor horseshoe method based on the data scenario B1.

Table 4.4: Comparison and performance between some Bayesian shrinkage and frequentist methods, in which the sample size is $n = 50$, $p = 250$, the number of iterations is $M = 20000$ and the number of true non-zero coefficients is 14. FN is false negative, FP is false positive, AUC is area under curves and MSE is mean square error. For both hNiP and ReB-hNiP, the upper bounds is $b_0 = 10^4$, and for both hNiP* ReB-hNiP*, that is, proposed shrinkage priors with the upper bounds which is $b_0 = 10^6$. λ indicates the values of the error model.

Models	λ	**No.**$\hat{\beta}$	**Sensitivity**	**Specificity**	**MSE**	**FN**	**FP**	**AUC**
hNiP	0.02	11.07	100%	98.78%	0.0008	0	2	98.83%
	2	13	100%	99.60%	0.0003	0	1	99.60%
	2000	13	100%	99.58%	0.0001	0	1	99.60%
ReB-hNiP	0.02	11.41	100%	98.90%	0.00014	0	2.65	98.90%
	2	13	100%	99.58%	0.00040	0	1	99.60%
	2000	13	100%	99.58%	0.00016	0	1	99.60%
Bayesian LASSO	0.02	1	100%	94.80%	113.953	0	13	94.80%
	2	1	100%	94.80%	115.981	0	13	94.80%
	2000	1	100%	94.80%	113.336	0	13	94.80%
Horseshoe	0.02	11	100%	98.80%	0.0027	0	3	98.80%
	2	14	100%	100%	0.0195	0	0	100%
	2000	14	100%	100%	0.0002	0	0	100%
LASSO	0.02	48.36	94.08%	94.08%	3216.49	46.36	12	76.75%
	2	46.86	94.58%	94.58%	3223.79	43.92	11.06	78.10%
	2000	44.66	94.61%	94.61%	3226.28	41.78	11.12	78.92%
SCAD	0.02	9	4.14%	94.08%	3216.49	3	12	76.75%
	2	9	66.67%	96.69%	31138.45	3	8	95.62%
	2000	9	66.67%	96.69%	31128.40	3	8	95.62%
hNiP*	0.02	12.31	100%	99.3%	0.00009	0	1.69	99.33%
	2	14	100%	100%	0.00002	0	0	100%
	2000	14	100%	100%	0.00003	0	0	100%
ReB-hNiP*	0.02	13.03	99.50%	99.35%	0.00044	0.07	1.6	99.35%
	2	14	100%	100 %	0.00007	0	0	100%
	2000	1	100%	99.8%	0.0003	0	0.2	99.8%

Moreover, Table 4.4 displays the figures of some shrinkage priors after increased the number of the coefficients parameters, which shows the performance of the proposed shrinkage priors for variables selection similar to the former sections for the dataset in the previous Scenario.

Above all, the Bayesian linear regression model is computed relying on the hNiP prior distribution, which is faster than the models based on the ReB-hNiP prior distribution. The MCMC algorithm was run for a different number of iterations and kept 50% of all iterations for Regression models that rely on hNiP, ReB-hNiP, horseshoe and the Bayesian LASSO. All models are applied on a personal computer (PC) with of 8GB (RAM) and an i7 double (CPU) and 2.5 GHz. We reported the time computation for all three dataset scenarios which implemented for different shrinkage regularisation methods. Table 4.5 shows the time computing for each model. As can be seen there, hNiP is the fastest model among ReB-hNiP, horseshoe and the Bayesian LASSO. For example, a run utilising the dataset in scenario C, that is, $n = 200, p = 1000$ is took 8 hours and 53 minutes, whereas the horseshoe method took 17 hours and 45 minutes and 27 seconds with thin every 5 iterations.

Table 4.5: Comparison of the time computation of hNiP, ReB-hNiP, horseshoe and Bayesian LASSO shrinkage priors under a linear regression model for several datasets. The number of iterations is $M = 20000$ for both datasets $(n = 35, p = 45)$ and $(n = 50, p = 250)$. $M = 10000$. for dataset $(n = 200, p = 1000)$. hh represents the hours, mm represents the minutes and ss refers to seconds.

	$n = 35, p = 45$	$n = 50, p = 250$	$n = 200, p = 1000$
Models	hh:mm:ss	hh:mm:ss	hh:mm:ss
hNiP	00:01:05	00:21:38	08:53:00
ReB-hNiP	00:02:19	00:28:48	09:55:00
horseshoe	00:06:00	00:34:00	18:40:00
Bayesian LASSO	00:03:25	00:32:00	17:45:27

Table 4.6: Comparison and performance between some Bayesian shrinkage and frequentist methods, in which $n = 200, p = 1000,$ the number of iterations is $M = 10000$ and the number of true non-zero coefficients is 11. FN is false negative, FP is false positive, AUC is area under curves and MSE is mean square error.

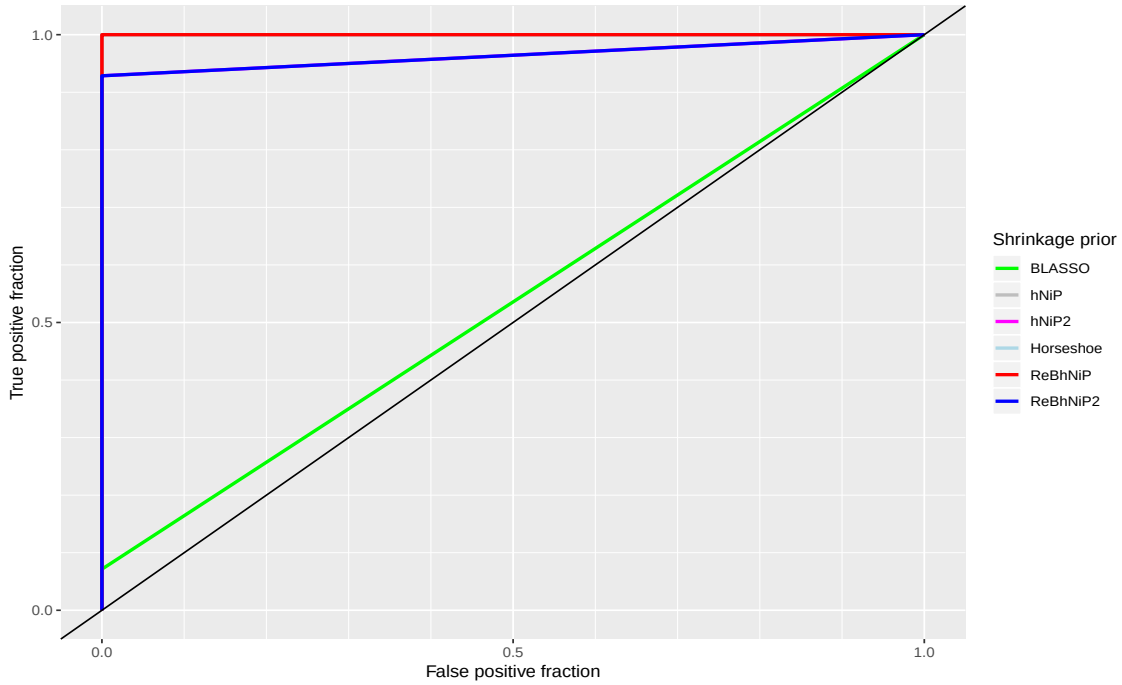
	λ	**No.**$\hat{\beta}$	**Sensitivity**	**Specificity**	**MSE**	**FN**	**FP**	**AUC**
hNiP	0.02	10	100%	99.90%	0.0014	0	1	99.90%
	2	10	100%	99.90%	0.0017	0	1	99.90%
	2000	10	100%	99.90%	0.0026	0	1	99.90%
ReB-hNiP	0.02	10	100%	99.90%	0.0042	0	1	99.90%
	2	10	100%	99.90%	0.0033	0	1	99.90%
	2000	10	100%	99.90%	0.0025	0	1	99.90%
Bayesian LASSO	0.02	3	100%	99.20%	15.4141	0	8	99.20%
	2	3	100%	99.20%	16.3453	0	8	99.20%
	2000	3	100%	99.20%	16.0345	0	8	99.20%
Horeshoe	0.02	14.81	74.42%	100%	0.2929	3.81	0	99.62%
	2	13.67	82.14%	100%	0.4522	2.66	0	99.73%
	2000	13.62	82.41%	100%	0.6012	2.62	0	99.74%
LASSO	0.02	11	0.85%	98.87%	23637.29	116	10	87.41%
	2	11	1.05%	98.89%	23589.65	188	9	80.32%
	2000	11	0.00%	98.81%	23588.01	78	11	91.11%
SCAD	0.02	11	100%	99.30%	51563.721	0	7	99.30%
	2	11	100%	99.30%	51620.649	0	7	99.30%
	2000	11	100%	99.30%	51610.471	0	7	99.30%
hNiP*	0.02	13.17	83%	99.9%	0.0024	2	0.3	99.6%
	2	11	100%	100%	0.0003	0	0	100%
	2000	11	100%	100%	0.00002	0	0	100%
ReB-hNiP*	0.02	15.02	75.57%	99.80%	0.0021	3.5	1	99.75%
	2	11	100%	100%	0.00009	0	0	100%
	2000	11	100%	100%	0.00014	0	0	100%

For visual compression and representation, we establish the results produced by the ROC curves in Section 2.5.3, Figure 4.9 displays an example of the ROC curves of coefficients for the Bayesian shrinkage prior. The ROC curves concentrate in the upper-left corner means that the the most shrinkage varaible selection had larger AUC values.

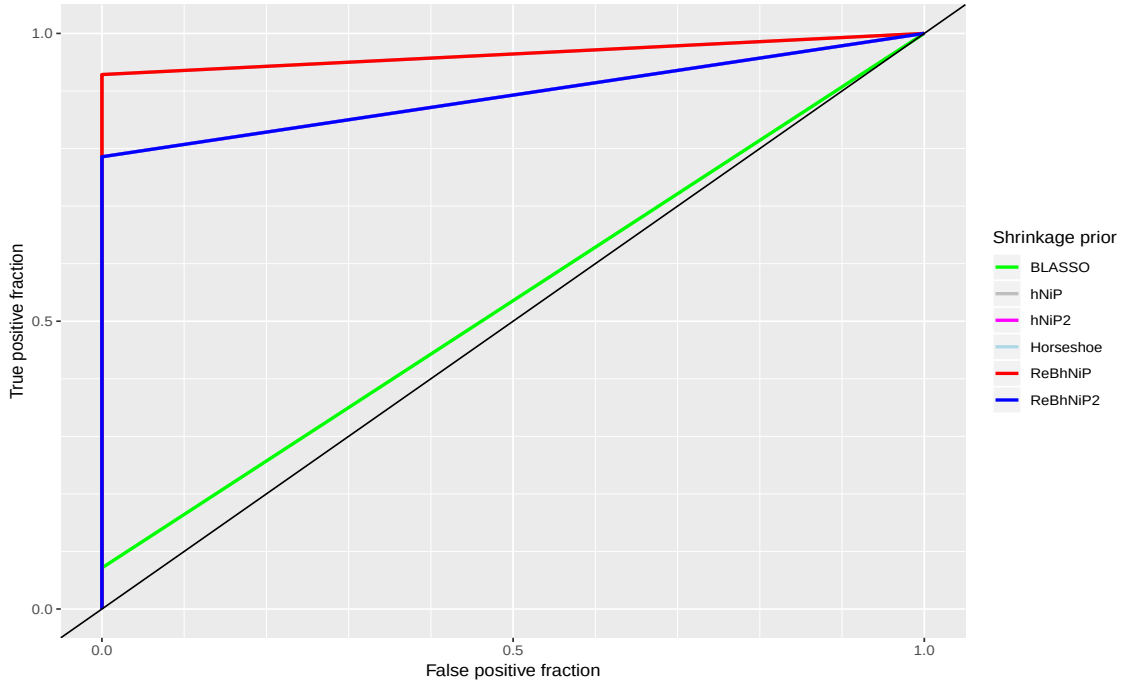
Figure 4.10b, three lines represent the output of the regression model: the red line for ReB-hNiP2 with $b_0 = 10^6$, the blue lines for ReB-hNiP with $b_0 = 10^4$ and the green line for the results of the Bayesian LASSO. Nevertheless, some other lines are hiding due to identical performances and the plots corresponding to the mentioned lines, which effectively hides them.

For example, red lines for ReB-hNiP2 corresponds to pink line for hNiP. Also,the blue line hides both the hNiP with $b_0 = 10^4$ and horseshoe. Despite of some shrinkage methods have corresponding ROC curves, they could have different MSEs, such as the MSE=0.00214 for ReB-hNiP with $b_0 = 10^4$ corresponding to the horseshoe with MSE= $0.2929,$ as can be seen in the ROC curve in Figure 4.10.

We randomly chose some MCMC samplers from the results for the 100 replications in the previous section, for both datasets $(n = 50, p = 250$ and $n = 200, p = 1000)$ and also for both the hNiP and ReB-hNiP shrinkage models. We used the mean as a statistics test, to compare the predictive prediction distribution a mean for the observed data y. Figure 4.11 shows the graphical predictive distribution of sample size $n = 50$ and $p = 250,$ which is clearly illustrates the mean (black line) of the observed data for the proposed scale mixture of the normal distributions hNiP and ReB-hNiP. Similar to the previous example, Figure 4.12 shows similar performances with of sample size $n = 200$ and $p = 1000$ in terms of the posterior prediction.

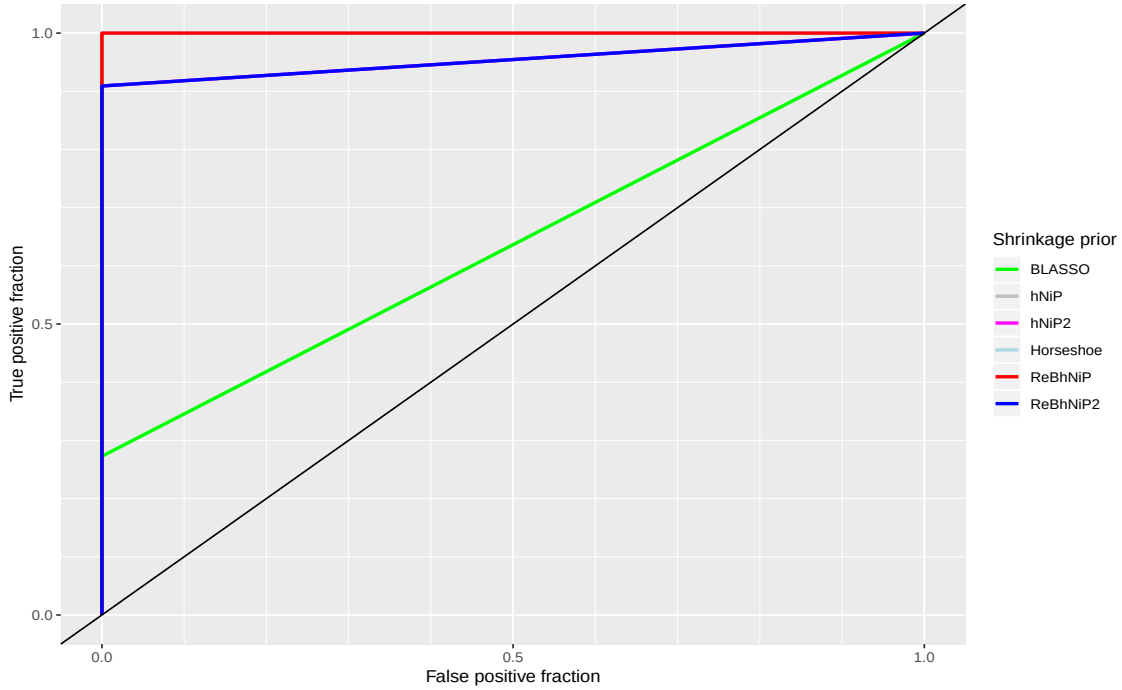


(a) $\lambda = 2000$

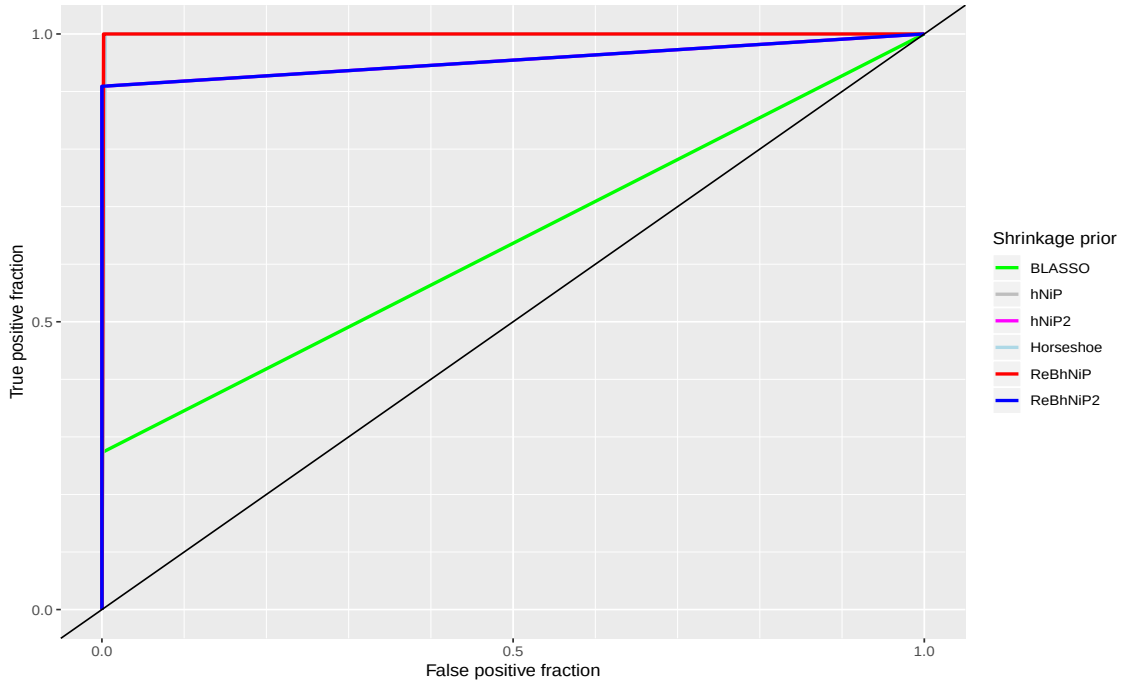


(b) $\lambda = 0.02$

Figure 4.9: ROC curves of some Bayesian shrinkage priors, including the Bayesian LASSO, the horseshoe, hNiP$(b_0 = 10000, c = 0.9)$, and ReB-hNiP$(b_0 = 10000, e_0 = 6)$, hNiP2$(b_0 = 100000, c = 0.9)$, ReB-hNiP2$(b_0 = 100000, e_0 = 6)$ for both cases where $\lambda = 0.02$ and $2000, n = 50$ and $p = 250$.

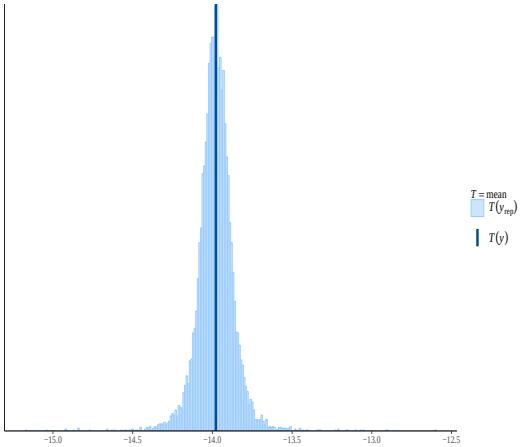


(a) $\lambda = 2000$

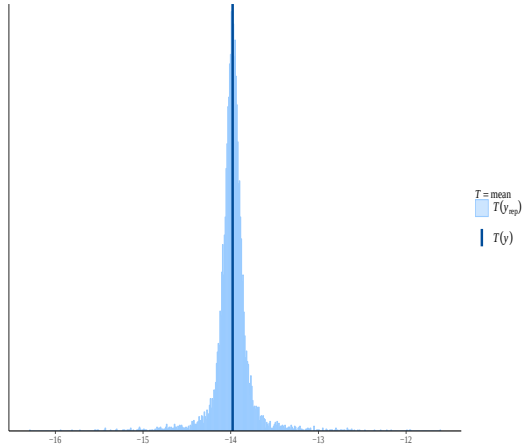


(b) $\lambda = 0.02$

Figure 4.10: ROC curves of some Bayesian shrinkage priors, including the Bayesian LASSO, the horseshoe, the hNiP$(b_0 = 10000, c = 0.9)$, the ReB-hNiP$(b_0 = 10000, e_0 = 6)$, hNiP2$(b_0 = 100000, c = 0.9)$, and ReB-hNiP2$(b_0 = 100000, e_0 = 6)$ for both cases where $\lambda = 0.02$ and 2000, $n = 200$ and $p = 1000$.

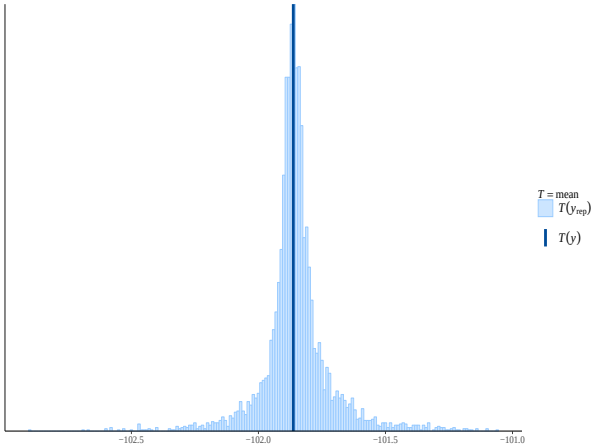


(a) Predictive distribution model based on ReB-hNiP shrinkage, where sample size $n = 50$ and $p = 250$.

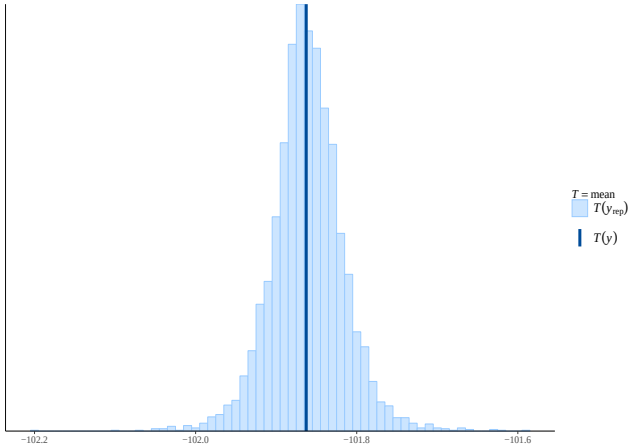


(b) Predictive distribution model based on ReB-hNiP shrinkage, where sample size $n = 50$ and $p = 250$.

Figure 4.11: Model checking using the predictive distribution model based on ReB-hNiP and hNiP shrinkage prior distribution, the dataset precision of model being $\lambda = 0.02$ and for both cases sample size $n = 50$ and $p = 250$.



(a) Predictive distribution model based ReB-hNiP shrinkage, where p sample size $n = 200$ and $p = 1000$



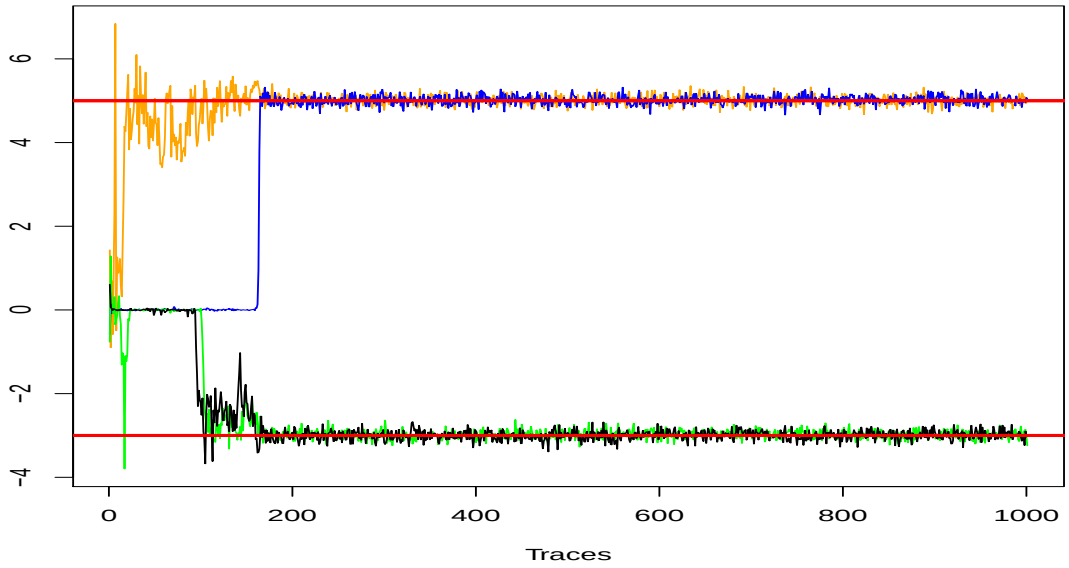
(b) Predictive distribution model based on hNiP shrinkage, where sample size $n = 200$ and $p = 1000$.

Figure 4.12: Model checking using the predictive distribution model based on ReB-hNiP and hNiP shrinkage prior distribution, the dataset precision of model being $\lambda = 0.02$ and in both cases the sample size being $n = 200$ and $p = 1000$.

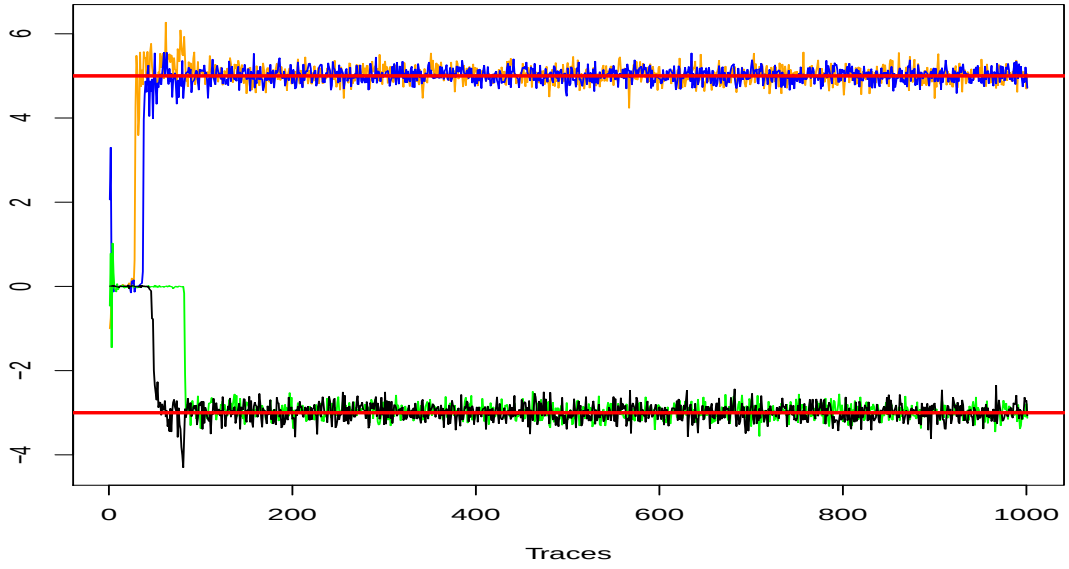
In the case of dealing with high-dimensional data $p >> n$, we applied a regression model based on the proposed Shrinkage prior method explained in Section 4.3, particularly in scenario D, in which $n = 200$ and $p = 5000$. Consider that the D datasets are high-dimensional and sparse because only five values of the true coefficients differ from zero. We ran the MCMC algorithm only for $M = 1000$ iterations due to time limitations our high-performance computing system (iceberg) which is not allowed to run the MCMC algorithm for more than 168 hours. Our MCMC algorithm for both shrinkage methods requires much more time to reach $M = 10000$ iterations. With only $M = 1000$ iteration for all the datasets in scenario D for the hNiP and ReB-hNiP shrinkage methods, the computing time was 113 hours and 43 minutes for the hNiP model and 120 hours and 10 minutes for the ReB-hNiP model. Even though, these result are not reliable due to the Short runs. But show that all the cases satisfactorily estimated four out of five non-zero regression coefficients; only the intercept parameter shrunk to zero under the hNiP hyperparameters $b_0 = 10000, c = 0.9$, and $d = 0.6$, the ReB-hNiP upper-bound parameter $b_0 = 10000$, and a rescaled Beta parameter $e_0 = 6$. Table 4.7 shows the good performance of both the proposed Bayesian shrinkage priors for the high-dimensional datasets. The values of the WAIC criteria cluster together for each shrinkage prior method with the same dataset. As can be seen in Figures 4.13 the traces of two sample hNiP and ReB-hNiP were stable and convergent to the true values after a few hundred iterations that support our variable selection. We also used an appropriate values for both shrinkage priors.

Table 4.7: Comparison of Bayesian regression models based on proposed shrinkage priors, including hNiP and ReB-hNiP for different single-noise datasets when $p >> n$. The figures show the number of estimated non-zero coefficients as No.$\hat{\beta}$, false positive as FP, false negative as FN and WAIC criteria. The hyperparameters of hNiP is $b_0 = 10000,\ c = 0.9, d = 0.6$. The ReB-hNiP upper-bounds parameter is $b_0 = 10000$ and the rescaled beta parameter is $e_0 = 6$. The number of non-zero coefficients that included the intercept parameter are non-zero coefficients in 5 datasets of sample size $n = 200$ and $p = 5000$.

		hNiP				**ReB-hNiP**			
	Error model λ	**non-zero $\hat{\beta}$**	**FP**	**FN**	**WAIC**	**non-zero $\hat{\beta}$**	**FP**	**FN**	**WAIC**
	0.02	4	1	0	541.95	4	1	0	426.33
n=200, p=5000	2	4	1	0	651.43	4	1	0	507.25
	2000	4	1	0	478.23	4	1	0	436.16



(a) Traces plots for the ReB-hNiP model, where the red line represents the true value.



(b) Trace plots for the hNiP model for non-zero coefficients, where the red line represents the true value.

Figure 4.13: Traces plots for the linear regression model depending on shrinkage priors; the hyperparameters values are $b_0 = 1000, e_0 = 6$ for ReB-hNiP in panel (a) and $b_0 = 10000, c = 0.9$ and $d = 0.6$ for hNiP in panel (b). The red line represents the true values some non-zero coefficient β, where β_2 and $\beta_3 = -3$ indicates by black and green colour respectively, and β_{50} and$\beta_{500} = 5$ represents by blue and orange colour respectively. The sample size is $n = 200$ and $p = 5000$. The number of iterations is $M = 1000$.

4.7 Summary of Chapter

This chapter is based on utilising the proposed Bayesian shrinkage prior distributions hNiP and ReB-hNiP with a Bayesian linear regression model. The calibration of the optimal hyperparameter values of proposed methods was addressed. These were compared with some Bayesian shrinkage and non-Bayesian methods methods to examine their performances in terms of variable selection and MCMC algorithm run time. We concluded that the best ranges of the hyperparameters (tuning parameters) that can be used for both our proposed Bayesian shrinkage prior distributions (hNiP and ReB-hNiP) are as follows; first, for the upper-bound parameter b_0 for both shrinkage priors in the case of the high-dimensional data, we suggest using $b_0 = (2500, \ldots, 100000)$, that is, larger than 2500 up to 100000 for the cases in which $p \gg n$, because it controls the scale of the inverse Pareto distribution as the scale mixture of the normal distribution, which is flexible for shrinking small regression coefficients due to heavy tails. Second, the advisable values of the hNiP shrinkage prior distribution for the hyperparameter is $0.3 \leq c \leq 1.5$ to choose the value of d taking into consideration the assumption that $E(\kappa) = 1.5$. Third, regrading the ReB-hNiP tuning parameters, the results suggest using the values of the hyperparameter rescaled Beta $e_0 = \{4, 5, 6\}$. Lastly, the results show that hierarchical structure of the NEG prior provide some useful suggested values of hyperparameters similar to both others shrinkage prior. We will discover more about of parameter estimation using NEG methods in the future.

On the other hand, the Bayesian linear regression model based on the hNiP prior distribution is faster than the models based on the ReB-hNiP prior distribution, horseshoe, and Bayesian LASSO. Also, the variable selection rules for the hNiP prior distributions satisfactorily outperform the other named shrinkage methods. The MCMC algorithm, which was run for several iterations, kept 50% of them. In the next chapter, we address hyperparameter calibration based on a linear model with a measurement error model similar to the procedures of this chapter.

Chapter 5

High-dimensional Measurement Error with Linear Model via Bayesian Shrinkage Prior

5.1 Overview of Measurement Error Model

In various fields of science, statisticians face typical issues in terms of analysing data. In term of the effect and presence the issue of Measurement Error (ME), data analysis has been known and researched for some decades. Measurement error is also called error-in-variable (Buonaccorsi, 2010; Fuller, 1987). ME occurs in either the covariate variable, response variable or both of them. Ignoring such type of errors in replication, usually leads to biased estimates of coefficient parameters in regression model, that is, the observation in single or more variables is not measured accurately by utilising one of the measures, tools, inaccurate sampling, or imperfect observations (Buonaccorsi, 2010). Furthermore, simple linear regression is infeasible to fit parameters in the model, also confidence intervals and parameters estimation might suffer from biases error, in the case if ME is disregarded. Thus, it might cause the loss of power for variables selection or relation through variables and the substantial features of the data might be hidden. In the case of having replicated data, it is difficult to ignore ME factor. This is be-

cause ignoring it will lead to an extremely misled understanding of the amount of link between response and independent variables (Gilks, 1999). On the other hand, parameter estimation based on the simple linear regression is not realistic for interpretation of the model. The structure of regression model with measurement error model becomes complex and it has different formulation compared with the standard linear model. Several studies cover the matter of measurement error models (Buonaccorsi, 2010; Fuller, 1987; Gilks, 1999). Recently, Morrissey et al. (2010) studied the interaction network between genes. They built a model under Bayesian perspective with measurement error. This model depends on spike and slab prior distribution for coefficient parameters, and measurement error is represented by Student-t distribution. In my experience, for the first time Lachos et al. (2011) used the scale mixture of Normal SMN distribution as robustness distribution instead of the Gaussian assumption in the case dealing with measurement error models; because they proposed that both random errors and covariates follow SMN distribution for the bivariate linear model. They considered local influence analysis for measurement error models. Simultaneously, an expectation maximization (EM) algorithm is used to update sampler of the parameters. Furthermore, a bivariate linear regression model with measurement error was examined under heavy tails distribution by Lin and Cao (2013). They established that the novel structure of a repeated measurement error model has unobserved variables. The noise of the model follows SMN distribution. So, the results of this model provided an attractive robust distribution when compared with usual Normal distribution. An alternative distribution to the Normal distribution was used in the multivariate measurement error model by Cao et al. (2015). They consider that SMN distribution is one of the heavy-tails distributions and the model includes SMN distribution that covers different kinds of data, such as unpair or/and unequal replicated data. Here, the SMN distribution is an adaptable choice and creates an interesting robust distribution instead of the usual normal assumptions. This model deals with the maximum likelihood estimation (MLE) for estimating parameters and it relies on expectation-maximization (EM algorithm) for sampling.

In this chapter, we assume dealing with unobserved variables, and our model structure depends on the simple linear regression and measurement error. In order to compute observed variables (true variables), which includes unobserved plus an additive error, the structure of the linear regression model with measurement error can be identified as,

$$Y_i = \alpha + \boldsymbol{\beta}\boldsymbol{X}_i + \epsilon_i. \tag{5.1}$$

Also, the measurement error equation is described as follows,

$$z_{ijr} = X_{ij} + \eta_{ijr}. \tag{5.2}$$

Suppose that $\boldsymbol{X}_i = X_{1i}, X_{2i}, \ldots, X_{ni}$ is a vector of unobserved variable with size $1 \times p$, where $(i = 1, 2, \ldots, p), (j = 1, 2, \ldots n)$ and z_{ijr} represents single observed variables from matrix with size $n \times p \times R$ where $r = 1, 2, \ldots, R$, and Y_i is dependent variable with size $1 \times n$, where n is a sample size, p represents the number of parameters and R refers to the number of replications. $\boldsymbol{\beta}$ is the vector of coefficient parameters of size $1 \times p$ and α is an intercept of the model. All parameters are assumed to be independent, with ϵ_i and η_i both are independent and ϵ_i represents the error term of the model which has normal distribution with mean equals to zero and precision equals to λ. η_i is measurement error and has normal distribution with mean equals to zero and precision equals to τ, and z_{ir} has normal distribution with mean equal to X_i and precision τ. In the next section, we will describe our ME models, particularly the models that are related to overcoming measurement error problem in the Bayesian framework. The prior for the unknown parameters are implemented on dynamic Bayesian networks with measurement error based on the autoregressive model AR(1). For more information bout autoregressive model, the reader is referred to Radhakrishnan et al. (2013).

5.2 High-dimensional Measurement Error with Multivariate Linear Regression Model via Bayesian Shrinkage Prior

The effect and presence of matter in data analysis for Measurement Error have been recognized and investigated in several directions for many decades. So far, we assumed in the earlier works that the explanatory variables are completely measured, that is, covariates are free from measurement error. In fact, the issue of measurement error can include in covariates with at least a small amount value that happens in several research fields. As can be noticed that in the previous section, we highlighted the general form and some literatures of measurement error. In the following section, we put emphasis on some other studies and give the details of the structure for the high-dimensions Bayesian linear regression in term of covariates as measurement error model. In light of the standard regression model, without correction from measurement error,

the parameters estimation will be from biased (Carroll et al., 2006). Therefore, a replicated measurement is used in the explanatory variables in order to correct the model and estimate the measurement error through regression model (Sørensen et al., 2014).

In recent years, researchers have tended to utilise a penalty function to penalize the likelihood in regression models to correct the measurement error model. This type of correction approaches have been studied by Liang and Li (2009); Ma and Li (2010); Rosenbaum et al. (2010); Rudelson and Zhou (2015); Rudelson et al. (2017); Sørensen et al. (2014, 2015); Xu and You (2007). Both Rosenbaum et al. (2010) and Sørensen et al. (2014) have used lasso and Dantzig selector (DS) methods for correcting the measurement error in the linear regression model, in which assumption of sparse model was taken into consideration. While, the smoothly clipped absolute deviation (SCAD) penalty function is studied in solving this model problem by Xu and You (2007). The corrected version of the penalty functions include; lasso, SCAD and DS. They investigated and modified in to remove the effect of the measurement error in the regression model and selecting the parameters were studied by some authors such as Liang and Li (2009); Ma and Li (2010); Sørensen et al. (2015).

To our knowledge of correcting the issue of measurement error in a regression model based on Bayesian shrinkage prior, distribution or no has one used this kind of Bayesian regularization prior distribution. Accordingly, we will utilise our proposed Bayesian shrinkage prior distribution to correct and select the coefficients in Bayesian regression with measurement error model which they involve; hNiP, ReB-hNiP prior distribution and NEG prior distribution. All these priors were studied in the last section for fitting the coefficients of Bayesian linear regression model. Hence, a Bayesian perspective is used to estimate unobserved variables and unknown parameters. Also, a repeated measurement is taken into consideration. Likewise, to the last section 4.5, for reducing the high dimensionality in the Bayesian regression model, we utilise same Bayesian shrinkage prior to that presented in the Chapter 3.

5.2.1 The Model Formulation

Multivariate Bayesian regression with measurement error model was described in the previous section of this chapter. In the following sections, we illustrate the formulation of this

model. We have only one formula of likelihood function. However, utilising different prior distributions lead to obtain various posterior distributions. We consider Equation 5.1 for the linear regression model and the Equation 5.2 for the measurement error model. ϵ_i has normal distribution, $\epsilon_i \sim N(\epsilon_i|0, \lambda^{-1})$ represents random error. It has $\text{Var}(\epsilon) = \frac{1}{\lambda}$ and unobserved data denoted by $\boldsymbol{x_i}$ and depended variable $y_i \in \Re$, where the size of unobserved is $X = \{\boldsymbol{x_1}, \boldsymbol{x_2} \ldots, \boldsymbol{x_n}\}$, that is, $X \in \mathcal{M}_{n\times p}$, $\boldsymbol{x_i} = \{x_{i1}, x_{i2} \ldots, x_{ip}\}$ where $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, p$ and $y_i = \{y_1, y_2 \ldots, y_n\}$. Also, $\boldsymbol{\beta}$ is $(p \times 1)$ coefficient vector and n represents the number of observation. In here, we assume that η_{ijr} has Student-t distribution in order to make a robust model, $\eta \sim \text{St}(\eta|0, \nu, \tau)$ where ν is degree of freedom, τ represents the scale and $i = 1, 2, \ldots, n$, $j = 1, 2, 3, \ldots p$. Where z_{ijr} represents the observed value for n observation, p variables and $r = 1, 2, \ldots, R$ is the number of replications. Both λ and τ are independent.

In Bayesian framework, prior distribution plays a crucial role for building the model. Particularly, mixing model such as high dimensional multivariate regression with measurements error model require a set of prior distributions for fitting regression coefficients. Firstly, we formulate the likelihood function as follows,

$$\begin{aligned} L(\Theta; X, z) = (\lambda)^{n/2} \exp\left\{-\frac{\lambda}{2}(\boldsymbol{y} - \mathcal{X}\mathcal{B})' I (\boldsymbol{y} - \mathcal{X}\mathcal{B})\right\} \\ \prod_{j=1}^{p}\prod_{i=1}^{n}\prod_{r=1}^{R} \tau^{\frac{1}{2}} \delta_{ijr}^{\frac{1}{2}} \exp\left\{-\frac{\tau\delta_{ijr}}{2}(z_{ijr} - x_{ij})^2\right\} \left(\frac{\frac{\nu}{2}^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})}(\delta_{ijr})^{\frac{\nu}{2}-1} \exp\left\{-\frac{\nu}{2}\delta_{ijr}\right\}\right), \end{aligned} \tag{5.3}$$

where Θ is a set of unknown parameters $\Theta = (\mathcal{B}, \tau, \lambda, \phi, \zeta, \nu, \delta)$, $\mathcal{X} = (1, X)$ is the matrix of unobserved data, $\mathcal{B} = (\alpha, \boldsymbol{\beta})$ and $(\alpha, \boldsymbol{\beta}) \in \mathbb{R}^{p+1}$ and δ is scale of prior degree parameter ν. We have described above and considered that measurement error model has robust modelling follows has Student-t distribution. We use hierarchical structure of Student-t distribution as explained in Section 2.3. In the following sections, posterior distribution is described based on shrinkage prior distributions in order to overcome the issue of high dimensional data and measurements error in covariates variables.

5.2.2 The Model with hNiP Prior Distributions

5.2.2.1 The hNiP Prior Distributions for β_i

In this part of our study, we utilize a prior distribution as convenient flat distribution for both error model and measurement error model which is gamma distribution. However, the main difference in our model is using of shrinkage prior distributions for regression coefficients. The prior distribution of the parameters is as follows,

$$\pi(\alpha,\beta,\lambda,\tau,\psi,\kappa,\nu) = \frac{d_\tau^{c_\tau}}{\Gamma(c_\tau)}\tau^{c_\tau-1}\exp\{-d_\tau\tau\}\left(\frac{b_\nu^{a_\nu}}{\Gamma(a_\nu)}\nu^{a_\nu}\exp\{-\nu b_\nu\}\right)$$
$$\frac{b_\lambda^{a_\lambda}}{\Gamma(a_\lambda)}\lambda^{a_\lambda-1}\exp\{-b_\lambda\lambda\}\prod_{j=1}^{p+1}\frac{\psi_j^{\frac{1}{2}}\exp\left(-\frac{1}{2}\psi_j\beta_j^2\right)}{\sqrt{2\pi}}\kappa_j b^{-\kappa_j}\psi_j^{\kappa_j-1}\frac{\kappa_j^{c-1}d^c\exp(-\kappa_j d)}{\Gamma(c)}. \quad (5.4)$$

The prior distribution for coefficient β_i is scale mixture of normal distribution based on hNiP prior as defined in Section 3.2.1. We will discus this in the later section in details. a_ν, c_τ and a_λ represent the shape parameters of gamma distribution and b_ν, b_λ, d_τ represents the scale parameters for gamma distribution as prior distribution for all $\nu, \lambda, \tau,$ respectively.

$\pi(\Theta|y,z) \propto \pi(\Theta)\pi(y|X,\Theta),$

The posterior distribution under hNiP methods can be described as follows:

$$\pi(\Theta|y,z) = \lambda^{n/2}\exp\left\{-\frac{\lambda}{2}(\boldsymbol{y}-\mathcal{X}\mathcal{B})'I(\boldsymbol{y}-\mathcal{X}\mathcal{B})\right\}\prod_{j=1}^{p}\prod_{i=1}^{n}\prod_{r=1}^{R}\tau^{\frac{1}{2}}\delta_{ijr}^{\frac{1}{2}}\exp\left\{-\frac{\tau\delta_{ijr}}{2}(z_{ijr}-x_{ij})^2\right\}$$
$$\left(\frac{\frac{\nu}{2}^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})}(\delta_{ijr})^{\frac{\nu}{2}-1}\exp\left\{-\frac{\nu}{2}\delta_{ijr}\right\}\right)\frac{d_\tau^{c_\tau}}{\Gamma(c_\tau)}\tau^{c_\tau-1}\exp\{-d_\tau\tau\}\left(\frac{b_\nu^{a_\nu}}{\Gamma(a_\nu)}\nu^{a_\nu}\exp\{-\nu b_\nu\}\right)$$
$$\frac{b_\lambda^{a_\lambda}}{\Gamma(a_\lambda)}\lambda^{a_\lambda-1}\exp\{-b_\lambda\lambda\}\prod_{j=1}^{p+1}\frac{\psi_j^{\frac{1}{2}}\exp\left(-\frac{1}{2}\psi_j\beta_j^2\right)}{\sqrt{2\pi}}\kappa_j b^{-\kappa_j}\psi_j^{\kappa_j-1}\frac{\kappa_j^{c-1}d^c\exp(-\kappa_j d)}{\Gamma(c)}. \quad (5.5)$$

The full conditional distribution of unobserved variables x_i is presented by Equation 5.6 and we update the unobserved covariates based on it. The full conditional distribution of unobserved variable is given in Equation 5.6 and it is general formulation of any types of prior which will

be applied in the Bayesian regression with measurements error model, so the full conditional distributions is

$$\boldsymbol{x_i} \sim N\left(\boldsymbol{x_i}\middle|(\lambda\boldsymbol{\beta}\boldsymbol{\beta}' + \tau(\sum_{r=1}^{R}\delta_{ijr})I_p)^{-1}\big(\lambda\boldsymbol{\beta}(y_i-\alpha) + \tau(\sum_{r=1}^{R}\delta_{ijr}z_{ijr})\big), (\lambda\boldsymbol{\beta}\boldsymbol{\beta}' + \tau(\sum_{r=1}^{R}\delta_{ijr})I_p)\right). \tag{5.6}$$

The Gibbs sampler will be used to update the parameters for this hNiP modelling. The parameter for second layers has Truncated Gamma TrGa distribution which is bounded from above. The parameters ψ and κ are hyperparameters prior of the regression coefficients β_j in the case relying on hNiP distribution. Where $j = 1, 2 \ldots p+1$.

The full conditional posterior distribution of scale of normal distribution is ψ_j, which is given by,

$$\psi_j \sim \text{TrGa}\left(\psi_j\middle|b_0, \kappa_j + \frac{1}{2}, \frac{\beta_j^2}{2}\right), \tag{5.7}$$

and, the full conditional posterior distribution of the third layer of hNiP prior κ is,

$$\kappa_j \sim \text{Ga}\left(\kappa_j\middle|c+1, \log\left(\frac{b_0}{\psi_j}\right) + d\right). \tag{5.8}$$

The full conditional posterior distribution for regression coefficients $\mathcal{B}$ is as follows,

$$\mathcal{B} \sim N\left(\mathcal{B}\middle|\mu_\beta, \Omega_\beta^{-1}\right), \tag{5.9}$$

where $\mu_\beta = \lambda\Omega_\beta^{-1}\mathcal{X}'\boldsymbol{y}$, $\Omega_\beta = (\lambda\mathcal{X}'\mathcal{X} + \Sigma_\psi)$ and Σ_ψ is a matrix of size $(p+1)\times(p+1)$ and the diagonal of this matrix is equal to $\{\psi_1, \psi_2 \ldots, \psi_{p+1}\}$ and $\alpha \in \mathcal{B}$.

The posterior distribution of δ_{ijr} is given by,

$$\delta_{ijr} \sim \text{Ga}\left(\delta_{ijr}\middle|\frac{\nu+1}{2}, \frac{\nu + \tau(z_{ijr} - x_{ij})^2}{2}\right). \tag{5.10}$$

The full conditional distribution of measurements error model τ follows the Gamma distribution follows,

$$\tau \sim \text{Ga}\left(\tau\middle|c + \frac{RNp}{2}, d + \frac{1}{2}\Big(\sum_{r=1}^{R}\sum_{j=1}^{p}\sum_{i=1}^{n}\delta_{ijr}(z_{ijr} - x_{ij})^2\Big)\right). \tag{5.11}$$

We use Metropolis-Hastings algorithm (Appendix A.1.2) for updating parameter degree of freedom ν and the full conditional distribution is

$$\pi(\nu|rest) \propto \left(\frac{\frac{\nu}{2}^{\frac{Rnp\nu}{2}}}{(\Gamma(\frac{\nu}{2}))^{Rpn}} (\prod_{i=1}^{n}\prod_{j=1}^{p}\prod_{r=1}^{R}(\delta_{ijr})^{\frac{\nu}{2}-1}) \exp\left\{ -\frac{\nu}{2}\sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{r=1}^{R}\delta_{ijr} \right\} \right) \left(\frac{b_\nu^{a_\nu}}{\Gamma(a_\nu)}\nu^{a_\nu}\exp(-\nu b_\nu) \right). \tag{5.12}$$

The full conditional posterior distribution of error model λ is

$$\lambda \sim \text{Ga}\left(\lambda \middle| a_\lambda + \frac{n}{2}, b_\lambda + \frac{1}{2}(\boldsymbol{y} - \mathcal{X}'\mathcal{B})'I(\boldsymbol{y} - \mathcal{X}'\mathcal{B}) \right), \tag{5.13}$$

where all parameters are described previously.

5.2.3 The Model with ReB-hNiP Prior Distributions

In this section we utilize Rescaled Beta of hierarchical scale mixture of normal distribution ReB-hNiP as discussed in section 3.2.2.

5.2.3.1 The Prior Distributions (ReB-hNiP Prior for β_i)

The prior distributions for the unknown parameters $\Theta = \{\alpha, \beta_i, \lambda, \tau, \psi_i, \kappa_i, \nu\}$ are defined as follows,

$$\pi(\Theta) = \frac{d_\tau^{c_\tau}}{\Gamma(c_\tau)}\tau^{c_\tau-1}\exp\{-d_\tau\tau\}\left(\frac{b_\nu^{a_\nu}}{\Gamma(a_\nu)}\nu^{a_\nu}\exp\{-\nu b_\nu\}\right)\frac{b_\lambda^{a_\lambda}}{\Gamma(a_\lambda)}\lambda^{a_\lambda-1}\exp\{-b_\lambda\lambda\}$$
$$\prod_{j=1}^{p+1}\frac{\psi_j^{\frac{1}{2}}\exp\left(-\frac{1}{2}\psi_j\beta_j^2\right)}{\sqrt{2\pi}}\kappa_j b_0^{-\kappa_j}\psi_j^{\kappa_j-1}\frac{e_0^{-1}}{B(a,b)}\left(\frac{\kappa_j}{e_0}\right)^{a-1}\left(1-\frac{\kappa_j}{e_0}\right)^{b-1}. \tag{5.14}$$

The posterior distribution on unknown parameters $\pi(\Theta|y,z)$ using Bayesian framework is defined as follows,

$$\pi(\Theta|y,z) \propto \pi(\Theta)\pi(y|X,\Theta). \tag{5.15}$$

where $\pi(\Theta)$ is the prior distribution and $\pi(y|X,\Theta)$ is likelihood.

Hence, combining the ReB-hNiP shrinkage prior $\pi(\Theta)$ and the likelihood $\pi(y|X,\Theta)$, we can obtain the posterior distribution of unknown parameters as follows,

$$\pi(\Theta|y,z) = \lambda^{n/2}\exp\left\{-\frac{\lambda}{2}(\boldsymbol{y}-\mathcal{X}\mathcal{B})'I(\boldsymbol{y}-\mathcal{X}\mathcal{B})\right\}\prod_{j=1}^{p}\prod_{i=1}^{n}\prod_{r=1}^{R}\tau^{\frac{1}{2}}\delta_{ijr}^{\frac{1}{2}}\exp\left\{-\frac{\tau\delta_{ijr}}{2}(z_{ijr}-x_{ij})^2\right\}$$
$$\left(\frac{\frac{\nu}{2}^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})}(\delta_{ijr})^{\frac{\nu}{2}-1}\exp\left\{-\frac{\nu}{2}\delta_{ijr}\right\}\right)\frac{d_\tau^{c_\tau}}{\Gamma(c_\tau)}\tau^{c_\tau-1}\exp\{-d_\tau\tau\}\left(\frac{b_\nu^{a_\nu}}{\Gamma(a_\nu)}\nu^{a_\nu}\exp\{-\nu b_\nu\}\right)$$
$$\frac{b_\lambda^{a_\lambda}}{\Gamma(a_\lambda)}\lambda^{a_\lambda-1}\exp\{-b_\lambda\lambda\}\prod_{j=1}^{p+1}\frac{\psi_j^{\frac{1}{2}}\exp\left(-\frac{1}{2}\psi_j\beta_j^2\right)}{\sqrt{2\pi}}\kappa_j b_0^{-\kappa_j}\psi_j^{\kappa_j-1}\frac{e_0^{-1}}{B(a,b)}\left(\frac{\kappa_j}{e_0}\right)^{a-1}\left(1-\frac{\kappa_j}{e_0}\right)^{b-1}. \tag{5.16}$$

The full conditional distribution of unobserved variables x_i updated based on Equation 5.6.

In order to compute the posterior distribution, we find the full conditional distribution for unknown parameters which they presented in the following equations.

The parameters ψ and κ are hyperparameters prior of the β_j when using ReB-hNiP distribution. The full conditional posterior distribution for ψ_j is given by:

$$\psi_j \sim \text{TrGa}\left(\psi_j|b_0,\kappa_j+\frac{1}{2},\frac{\beta_j^2}{2}\right). \tag{5.17}$$

The full conditional posterior distribution of hyperparameter of the coefficients κ_j is

$$\pi(\kappa_j) \propto b_0^{\kappa_i}\psi_j^{\kappa_i-1}\frac{e_0^{-1}}{B(a,b)}\left(\frac{\kappa_j}{e_0}\right)^{a-1}\left(1-\frac{\kappa_j}{e_0}\right)^{b-1}, \tag{5.18}$$

where $j=1,2,\ldots,p+1$, and the values of κ_j should satisfy the condition $0<\kappa_j\leq e_0$ and $a=b=\frac{1}{2}$, where e_0 is recaled parameters for ReB-hNiP shrinkage prior distribution. This is implemented utilizing Metropolis Hasting algorithm as detailed in Appendix A.1.2.

The following parameters have the same posterior distribution, therefore we only represent them mathematically in the following equation. The full conditional posterior distribution of regression model $\mathcal{B}$ parameter is corresponding on Equation 5.9, with respect to ReB-hNiP

model for updating the Σ_ψ is a matrix of size $(p+1)\times(p+1)$, and the diagonal of this matrix is equal to $\{\psi_1, \psi_2 \ldots, \psi_{p+1}\}$.

$$\delta_{ijr} \sim \text{Ga}\left(\delta_{ijr}\middle|\frac{\nu+1}{2}, \frac{\nu+\tau(z_{ijr}-x_{ij})^2}{2}\right). \tag{5.19}$$

$$\tau \sim \text{Ga}\left(\tau\middle|c+\frac{RNp}{2}, d+\frac{1}{2}\sum_{r=1}^{R}\sum_{j=1}^{p}\sum_{i=1}^{n}\delta_{ijr}(z_{ijr}-x_{ij})^2\right). \tag{5.20}$$

Metropolis-Hastings algorithm (Appendix A.1.2) is used for updating ν and the full conditional distribution is described in the Equation 5.21.

$$\begin{aligned}\pi(\nu|rest) \propto \frac{\frac{\nu}{2}^{\frac{Rnp\nu}{2}}}{(\Gamma(\frac{\nu}{2}))^{Rpn}}(\prod_{i=1}^{n}\prod_{j=1}^{p}\prod_{r=1}^{R}(\delta_{ijr})^{\frac{\nu}{2}-1})\exp\left\{-\frac{\nu}{2}\sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{r=1}^{R}\delta_{ijr}\right\} \\ \left(\frac{b_\nu^{a_\nu}}{\Gamma(a_\nu)}\nu^{a_\nu}\exp(-\nu b_\nu)\right).\end{aligned} \tag{5.21}$$

5.2.4 The NEG prior distributions

In this section, we establish the Bayesian model construction, prior distribution and the full conditional distribution based on NEG prior distribution as formulated in Equation 3.30.

5.2.4.1 The Prior Distributions of NEG Prior for β_i

The prior distributions for all parameters are as follows:

$$\begin{aligned}\pi(\phi,\beta,\lambda,\tau,\nu) = \frac{d_\tau^{c_\tau}}{\Gamma(c_\tau)}\tau^{c_\tau-1}\exp\{-d_\tau\tau\}\left(\frac{b_\nu^{a_\nu}}{\Gamma(a_\nu)}\nu^{a_\nu}\exp\{-\nu b_\nu\}\right)\frac{b_\lambda^{a_\lambda}}{\Gamma(a_\lambda)}\lambda^{a_\lambda-1}\exp\{-b_\lambda\lambda\} \\ \prod_{j=1}^{p+1}\frac{1}{\phi_j^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\mathcal{B}_j'\frac{1}{\phi_j}\mathcal{B}_j\right\}\prod_{j=1}^{p+1}(\zeta_j\exp(-\phi_j\zeta_j))\prod_{j=1}^{p+1}\frac{(\frac{1}{\gamma^2})^\omega}{\Gamma(\omega)}\zeta_j^{\omega-1}\exp\left\{-\frac{\zeta_j}{\gamma^2}\right\},\end{aligned} \tag{5.22}$$

where Σ_Φ is matrix of size $(p+1)\times(p+1)$ and the diagonal of this matrix is equal to $\{\frac{1}{\phi_1}, \frac{1}{\phi_2}\ldots, \frac{1}{\phi_{p+1}}\}$. The prior distribution of coefficient parameters β_i is Normal Exponential Gamma. Therefore

the usual formula of prior distribution is $\pi(\Theta|y,z) \propto \pi(\Theta)\pi(Y|X,\Theta)$.

The posterior distribution of the Bayesian regression with measurement error model is as follows:

$$\begin{aligned}
&\pi(\Theta|y,X) = \\
&\lambda^{n/2}\exp\left\{-\frac{\lambda}{2}(\boldsymbol{y}-\mathcal{X}\mathcal{B})'I(\boldsymbol{y}-\mathcal{X}\mathcal{B})\right\}\prod_{j=1}^{p}\prod_{i=1}^{n}\prod_{r=1}^{R}\tau^{\frac{1}{2}}\delta_{ijr}^{\frac{1}{2}}\exp\left\{-\frac{\tau\delta_{ijr}}{2}(\mathrm{z}_{ijr}-x_{ij})^2\right\} \\
&\left(\frac{\frac{\nu}{2}^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})}(\delta_{ijr})^{\frac{\nu}{2}-1}\exp\left\{-\frac{\nu}{2}\delta_{ijr}\right\}\right)\frac{d_\tau^{c_\tau}}{\Gamma(c_\tau)}\tau^{c_\tau-1}\exp\{-d_\tau\tau\}\left(\frac{b_\nu^{a_\nu}}{\Gamma(a_\nu)}\nu^{a_\nu}\exp\{-\nu b_\nu\}\right) \\
&\frac{b_\lambda^{a_\lambda}}{\Gamma(a_\lambda)}\lambda^{a_\lambda-1}\exp\{-b_\lambda\lambda\}\prod_{j=1}^{p+1}\frac{1}{\phi_j^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\mathcal{B}_j'\frac{1}{\phi_j}\mathcal{B}_j\right\}\prod_{j=1}^{p+1}\zeta_j\exp(-\phi_j\zeta_j)\prod_{j=1}^{p+1}\frac{(\frac{1}{\gamma^2})^\omega}{\Gamma(\omega)}\zeta_j^{\omega-1}\exp\left\{-\frac{\zeta_j}{\gamma^2}\right\},
\end{aligned} \tag{5.23}$$

where the prior distribution of degree of freedom (ν) is follows Gamma distribution. We update the unobserved covariates x_i based on Equation 5.6.

NEG prior distribution have two order hierarchical hyperparameters. The ζ is the second order, which should be updated before the first layer of scale mixture parameter. Then, the full conditional distribution is as follows:

$$\zeta_j \sim \text{Ga}\left(\zeta_j\middle|\omega+1,(\phi_j+\frac{1}{\gamma^2})\right). \tag{5.24}$$

The first order hierarchical hyperparameters is ϕ which cannot be represented by usual probability distribution function. Therefore, Metropolis-Hastings algorithm (Appendix A.1.2) is used for updating ϕ_j.

The proposal distribution is Gamma with shape equal to $(1+\phi_j \times pb_\phi)$ and rate equal to (pb_ϕ) . We are fixing the mode at the current value and use pb_ϕ to control acceptance rate.

$$\pi(\phi_j|rest) \propto \phi_j^{-\frac{1}{2}}\exp\left\{-\left(\frac{\beta_j'\beta_j}{2\phi_j}+\zeta_j\phi_j\right)\right\}. \tag{5.25}$$

We are updating a vector of ϕ_j where $(j=1,2,\ldots,p+1.$

$$\prod_{j=1}^{p+1} \pi(\phi_j|rest) \propto \prod_{j=1}^{p+1} \phi_j^{-\frac{1}{2}} \exp\left\{-\left(\frac{\beta_j'\beta_j}{2\phi_j} + \zeta_j\phi_j\right)\right\}. \tag{5.26}$$

Therefore, the full conditional distribution of the regression coefficients $\mathcal{B}$ based on the NEG prior updated by Equation 5.9. The only deference here are using $\mu_\beta = \lambda\Omega_\beta^{-1}\mathcal{X}'\boldsymbol{y}$, $\Omega_\beta = (\lambda\mathcal{X}'\mathcal{X} + \Sigma_\Phi)$ and Σ_Φ is matrix by $(p+1)\times(p+1)$ which diagonal of this matrix is equal to $\{\frac{1}{\phi_1}, \frac{1}{\phi_2} \ldots, \frac{1}{\phi_{p+1}}\}$. The full conditional distributions for τ, λ, δ and ν have similar formulation as in Equations 5.11, 5.13, 5.10 and 5.12 respectively. Thus, we will not repeat the full conditional distribution for those parameters named above.

5.2.5 Synthetic Data

In order to examine our model numerically, some dataset sample are generated. The simulated data of the multivariate Bayesian regression with measurement error model is as follows:

$$Y = \alpha + X\boldsymbol{\beta} + \epsilon, \tag{5.27}$$

where random error $\epsilon \sim N(.|0, \lambda^{-1})$ and $Var(\epsilon) = \frac{1}{\lambda}$, where unobserved data is $\boldsymbol{x_i}$ and $Y_i \in \Re$, and the size of $\boldsymbol{x_i} = \{x_{i1}, x_{i2} \ldots, x_{ip}\}$ where $i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, p$. α is intercept parameter and $\boldsymbol{\beta}$ is $(1 \times p)$ coefficient matrix and p represents the number of parameters. Measurement error model part is represented as follows:

$$Z_{ir} = X_i + \eta_{ir}, \tag{5.28}$$

where $\eta \sim \text{St}(\eta\big|0, \tau, \nu)$ and ν represents the degree of freedom, τ refers to scale parameter and Z represents the size $n \times p \times R$ of observed variables and $\alpha = \beta_1$ for all datasets.

We simulated the first dataset **(D1)** under the below values for all parameters and coefficients parameters the values are as follow:

$\alpha = -5,\quad \beta_3 = 0.25, \beta_6 = -0.5, \beta_8 = 1, \beta_{16} = -3, \beta_{20} = 0.75, \beta_{24} = 0.5, \beta_{27} = -1.5, \beta_{32} = 1.4, \beta_{34} = 1.25, \beta_{39} = -2, \beta_{40} = 1.5$ and the rest values of $\beta_j = 0$.

Thus, the number of coefficients parameters $p = 40$, sample size $n = 30$, the replication for each samples $R = 25,\ \lambda = 10$ represents the precision of noise of the model, $\tau = 15$ is precision of measurement error.

The second dataset **(D2)** is simulated where the values of the non-zero coefficients parameters are as follows:

$(\alpha = 10.25, \beta_6 = -0.636, \beta_7 = 1.61, \beta_{12} = -17.74, \beta_{14} = 10.80, \beta_{16} = 9.45, \beta_{29} = -3.406, \beta_{32} = 7.305, \beta_{33} = -1.97, \beta_{40} = 1.89, \beta_{45} = -1.97, \beta_{47} = 3.74, \beta_{49} = 6.56, \beta_{57} = -5.42, \beta_{59} = 2.497, \beta_{65} = 1.60, \beta_{81} = 0.409.$

The number of coefficients parameters $p = 80$, sample size $n = 50$, the replication for each sample $R = 25,\ \lambda = 10$ represents the precision of noise of the model, $\tau = 15$ represents the precision of measurement error.

In order to show the effect and performance of our proposed shrinkage prior distribution, we also generated two different sample size datasets with three different signal-noise error for each of them. We describe these data as follows:

The dataset (D3): In this dataset the sample size is $n = 30$ and $p = 40$, where $\alpha = 40.25$, $\beta_9 = 1.460, \beta_{14} = 0.252, \beta_{36} = -4.819, \beta_{41} = -7.603$ and the design matrix X was simulated under normal distribution mean of 2 and the variance of 25. Based on these information we also simulated three different datasets D3I, D3II and D3III where the error mode and measurements error model are $(\lambda = 10, \tau = 15), (\lambda = 0.1, \tau = 15)$ and $(\lambda = 0.1, \tau = 1.5)$ respectively.

The dataset (D4): the sample size is $n = 50$ and $p = 80$. We also generated three different datasets based on coefficients regression of $\alpha = 40.25,\ \ \beta_8 = -3.625, \beta_{48} = 0.612, \beta_{52} = 0.879, \beta_{54} = 1.042, \beta_{67} = -2.424, \beta_{73} = 4.618, \beta_{74} = 1.115, \beta_{81} = 1.975$ and the design matrix X was simulated under normal distribution mean $= 2$ and the variance $= 25$. Based on these information we have simulated three different dataset D4I, D4II and D4III where the error mode and measurements error model are $(\lambda = 10, \tau = 15), (\lambda = 0.1, \tau = 15)$ and $(\lambda = 0.1, \tau = 1.5)$ respectively.

5.2.6 Hyperparameters Selection on Linear Regression with Measurements Error Model

This section highlights the main results that obtained from simulated data as described in the former section. The simulated data is used in high dimensional multivariate Bayesian regression with measurement error model. In this part of current chapter, we utilised WAIC in Equation 2.19 after it has modified, to detect the best tuning parameters or the best bounds range for hNiP, ReB-hNiP and NEG Bayesian shrinkage priors that stated in Chapter 3. As we mentioned previously, we have two different sample size for simulated datasets that implemented i.e dataset D1 and D2 (see Section 5.2.5), with several candidate hyperparameter values for all shrinkage priors hNiP, ReB-hNiP and NEG, in order to choose the minimum values of WAIC based on high dimensional multivariate Bayesian regression with measurement error model. In this section, we run model for $M = 30000$ iterations and kept 50% of iterations, where the prior values of the parameter the error model is the same as stated in the previous chapter for both λ is $a_\lambda = 2, b_\lambda = 0.62$ and for measurement error τ is $a_\tau = 2, b_\tau = 0.62$ to have a flat prior distribution. Table 5.1 shows the summarized figures of tuning hyperparameters selection

Table 5.1: Hyperparameters selection in the Bayesian regression with Measurements error model based on both ReB-hNiP and hNiP shrinkage prior distributions with different prior values, for dataset where n = $30, p$ = $40, R$ = 25 and the number non-zero coefficients is equal 9. No.$\hat{\beta}$ represents the number of the estimated non-zero coefficients and e_0^*, c^* presents the range bounds on hyperparameters which every upper bounded b_0 has typical limitation . $e_{0\text{ min}}$ illustrates the value of rescaled beta for minimum value of WAIC for ReB-hNiP model. c_{min} represent the value of hyperparameter for hNiP model has minimum value of the WAIC criteria.

	hNiP					**ReB-hNiP**			
	b_0	**No.**$\hat{\beta}$	c^*	**WAIC**	c_{min}	e_0^*	**No.**$\hat{\beta}$	**WAIC**	$e_{0\text{ min}}$
$n = 30, p = 40, R = 25$	100	7	$0.1 \leq 15$	-145.3	6.8	2-15	7	-144.27	14
	500	8	$0.1 \leq 15$	-148.43	2.8	2-13	8	-146.81	10
	1000	8	$4 < c^* \leq 15$	-147.56	10.2	2-15	8	-145.91	13
	5000	9	$0.1 \leq 2.5$	-152.38	5.8	2-15	9	-152.89	11
	10000	9	$0.1 \leq 2.5$	-153.23	8.4	2-15	9	-154.26	12

based on the minimum values of WAIC for both hNiP and ReB-hNiP Bayesian shrinkage prior to the first dataset D1. The number of estimated non-zero regression coefficients represents by No. $\hat{\beta}$. For each upper bounds parameter b_0, we have a range of values for the optimal hy-

perparameter for both hNiP and ReB-hNiP shrinkage prior distribution that leads to obtaining the best set of variables selection that has minimum values of WAIC criteria. e_0^*, c^* presents the range bounds on hyperparameters for hNiP and ReB-hNiP respectively, with specific upper bounded b_0 . Each hyperparameter has a typical limitation which leads to the best performance for each bound to fit the coefficients.

Figure 5.1 shows the trend of WAIC values; in the top panel, the best bounds hyperparameter values c, d and of upper bounds $b_0,$ parallel with bottom panel so as to standardize hyperparameters values of hNiP for satisfactory variables selection. Results of Table 5.1 and Figure 5.1 can help us to make decision about variables selection. We have achieved the number of non-zero coefficients in the simulation data which are exactly the same as estimated non-zero regression coefficients depending on the range limits hyperparameters. Also, we have got approximately lower values of the WAIC compared with other values of hyperparameters, when the upper bounds of prior distribution is large enough, such as $b_0 = 5000$ or 10000 and bounds parameter for hNiP shrinkage prior distribution between $c = 0.1$ and 2.5. However, it is clear from the plots that when the value of upper bounds b_0 is lower than 5000 it leads to under estimated.

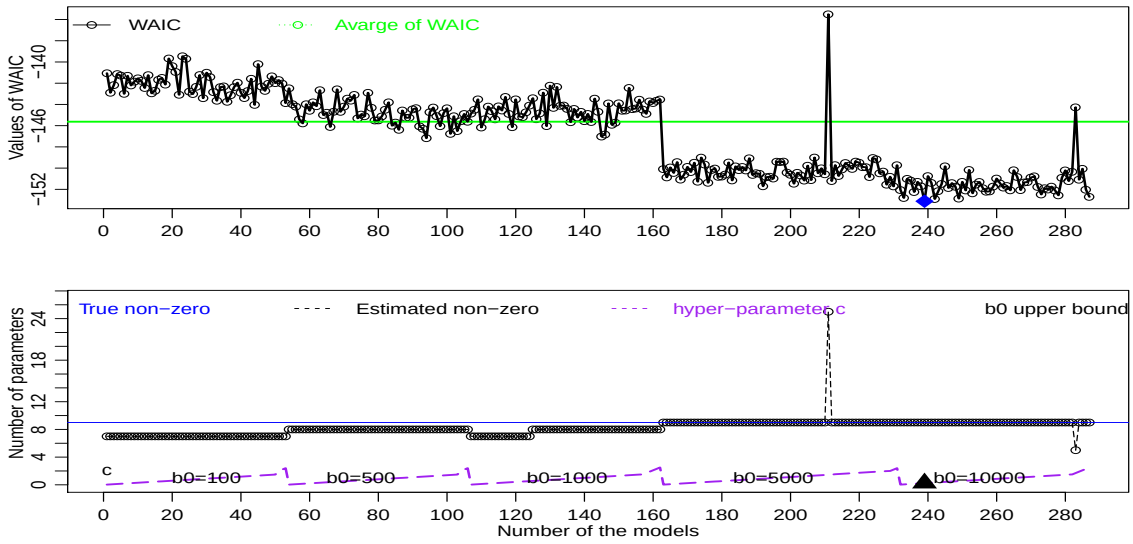


Figure 5.1: Hyperparameter parameters selection for the hNiP prior distribution, where top panel shows the WAIC values represented by black lines with black-circle and green-solid-line refers to average of WAIC. The bottom panel displays three lines, where blue solid line indicates the number of non-zero coefficients in the data D1. Black-dash-dot represents the number of estimated coefficients and the purple dash line represents different values of upper bounds b_0 and parallel with hyperparameters c, d and different values of upper bounds parameters b_0.

As we can notice, the trend of WAIC has slightly fluctuated and declined with increased upper bounds b_0, then suddenly fell-down with large values of upper bounds, and the plots of the number estimated parameters increased as much as b_0 rise up. In this example, the hyperparameters is bounded from $0.1 \leq c \leq 2.5$, because choosing any value is out of this range, will lead to coefficients aggressively shrink toward zeros, especially when b_0 = 10000.

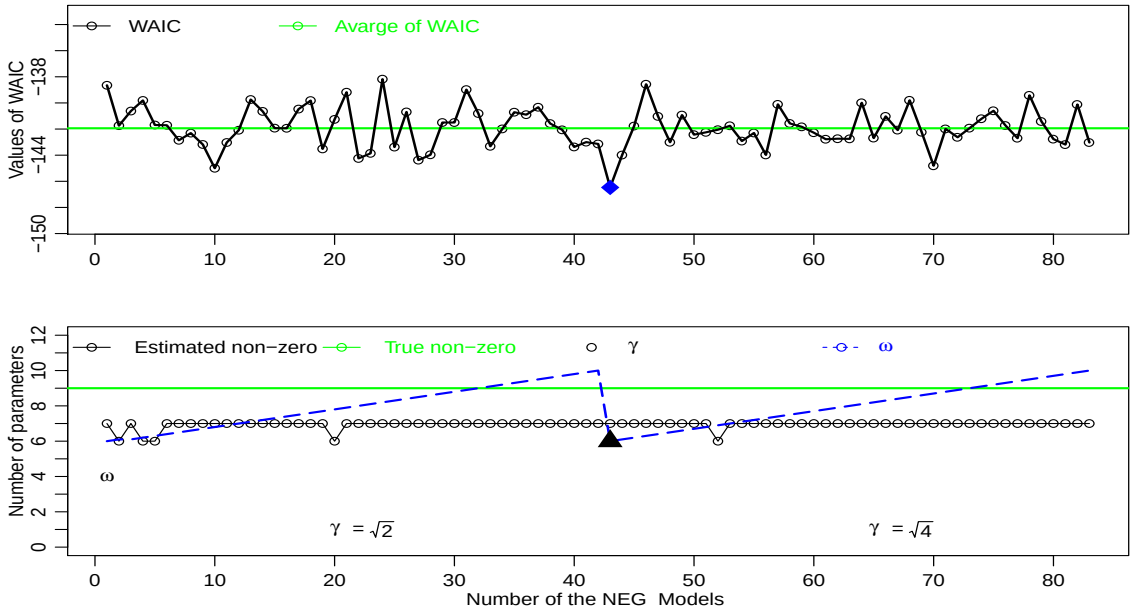


Figure 5.2: Hyperparameters selection for NEG shrinkage prior and the best range of WAIC criteria. The the panel shows two different lines, where WAIC is represented by black line with black circle and green solid line refers to the average of WAIC. The bottom panel displays three lines, where green solid line indicates the number of true non-zero coefficients in the simulation data, where $n = 30$ and $p = 40$. Black-dash-dot line represents the number of estimated coefficients and the blue dash line represents different values of upper bounds γ and parallel with hyperparameters ω.

On other hands, the results of the model based on NEG prior is shown in the Figure 5.2 which the plots in the top panel is fluctuated for the WAIC values with different values for the both hyperparameters ω, γ. Therefore, as can be seen in Figure 5.2 the line of estimated non-zero regression represented by black-dash dot line are always under the green solid line which refers to the number of non-zero coefficients in the simulation dataset. This results indicates that fitting model by using NEG shrinkage prior is under estimated for several hyperparameter values for both ω, γ. Furthermore, in general, the variable selection for some used values in this chapter based on regression model with robust measurement error model are more stable, that is, the

number of non-zero coefficients are close to each other for several hyperparameters, compared with the results of regression model based on NEG prior distribution in the previous chapter.

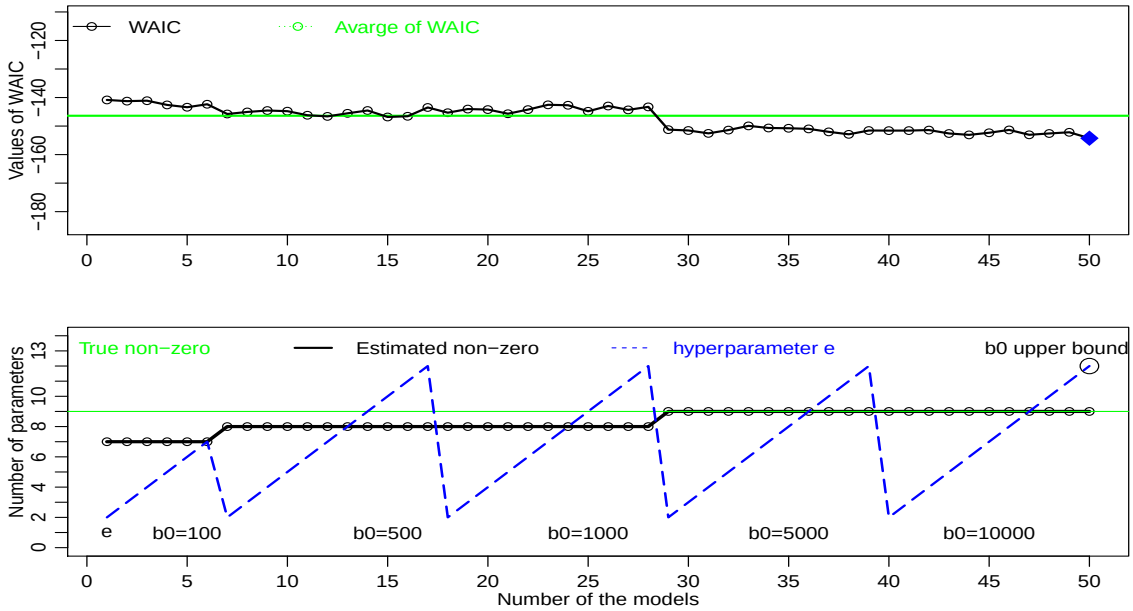


Figure 5.3: Hyperparameters parameters selection for ReB-hNiP; where the top panel shows two different lines where WAIC is represented by black lines with black-circle and green solid line refers to average of WAIC. The bottom panel displays three lines, where green solid line indicates the number of non-zero coefficients in the simulation data, where $n = 30, p = 40$. Black-dash-dot represents the number of estimated coefficients and the blue dash-line represents different values of upper bounds b_0 and parallel with hyperparameters e_0.

Table 5.2 includes summarized results of the minimum values of WAIC and the number of estimated parameters for both hNiP and ReB-hNiP shrinkage prior for dataset D2, sample of $n = 50, p = 80$. We discussed in the case D1, the value of WAIC decreases while the value of upper bounds b_0 increases in restriction bounds for both hyperparameters including rescaled beta e_0 and gamma prior c, d as second layer of scale mixture normal distribution. To support our hyperparameters calibration based on this case example, we show some results in the following plots. Figures 5.4, 5.5 and 5.6 shows the plots of the best range values of hyperparameters for all Bayesian shrinkage priors used in this section which have minimum values of WAIC. Both panels in each figure show the relation between changing the value of the upper bounds of hNiP shrinkage prior, hyperparameters and the best range of WAIC values criteria.

Table 5.2: Summarized Hyperparameters selection in the Bayesian regression with Measurements error model based on both ReB-hNiP and hNiP shrinkage prior distributions with different prior values for dataset, where $n = 50, p = 80, R = 25$ and the number non-zero coefficients is equal 17. No.$\hat{\beta}$ represents the number of the estimated non-zero coefficients and e_0^*, c^* presents the range bounds on hyperparameters which every upper bounded b_0 has typical limitation. $e_{0\,\text{min}}$ illustrates the value of rescaled beta for minimum value of WAIC for ReB-hNiP model. $c_{\,\text{min}}$ represent the value of hyperparameter for hNiP model has minimum value of the WAIC criteria.

		hNiP				**ReB-hNiP**			
	b_0	**No.**$\hat{\beta}$	c^*	**WAIC**	c_{min}	e_0^*	**No.**$\hat{\beta}$	**WAIC**	$e_{0\textbf{(min)}}$
$n = 50, p = 80$	100	16	1.5	-313.40	0.45	2 - 15	16	-312.03	4
	500	17	1.5	-316.69	1.32	3 - 15	17	-314.74	8
	1000	17	1.5	-317.89	0.72	4 - 13	17	-321.94	7
	5000	17	1.5	-326.03	0.78	5 - 13	17	-326.87	3
	10000	17	1.5	-325.20	1.08	6 - 12	17	-323.81	4

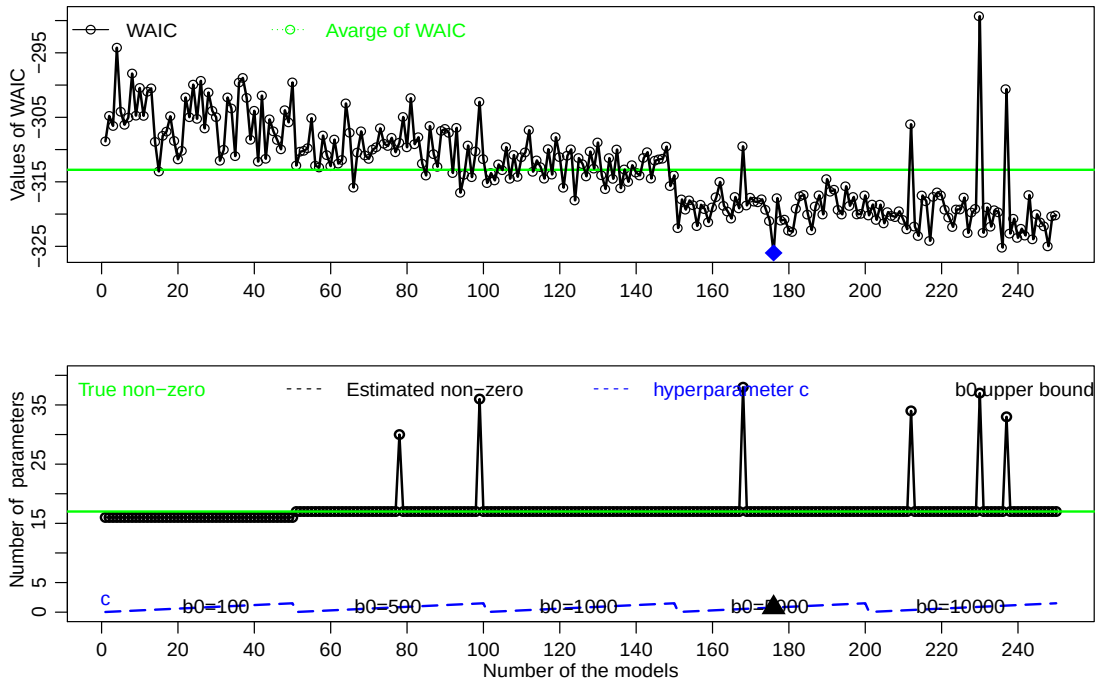


Figure 5.4: Hyperparameters parameters selection for hNiP; where the top panel shows two different plots where WAIC is represented by black lines with black-circle line and green solid line refers to average of WAIC. The bottom line displays three lines, where blue solid line indicates the number of non-zero coefficients in the simulation data, where $n = 50, p = 80$. Black-dash-dot represents the number of estimated coefficients and the blue dash line represents different values of upper bounds b_0 and parallel with hyperparameters c, d.

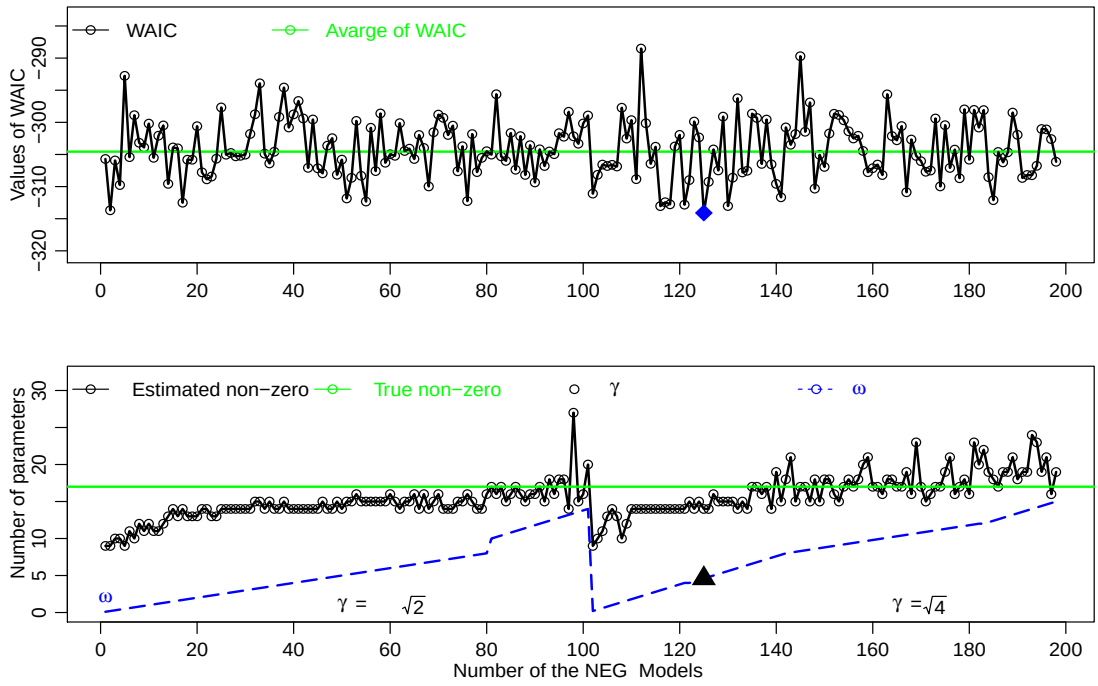


Figure 5.5: Hyperparameters parameters selection for NEG; the top panel shows two different lines, where WAIC is represented by black lines with black-circle and green solid line refers to average of WAIC. The bottom panel displays three lines, where green solid line indicates the number of non-zero coefficients in the simulation data where $n = 50, p = 80, R = 25$. Black-dash-dot represents the number of estimated coefficients and the blue dash line represent different values of upper bounds γ and parallel with hyperparameters ω.

A summary of first part of this chapter, the hyperparameters selection based on Bayesian regression model with measurements error model which are similar to the previous procedure in the Chapter 4. Consequently, we can use the bonds range of the hyperparameters c for hNiP where $c \in [0.1, 1.5]$ as the best closed range interval values with respect of the excepted value of second layer of shrinkage prior $E(\kappa) = \frac{c}{d} = 1.5$, and the upper bounds parameters $b_0 \in (5000, \ldots, 10000)$. Also, the best range value of ReB-hNiP shrinkage prior distribution is as follows; hyperparameter $e_0 \in [4, 7]$ which represents closed set, and upper bounds parameters $b_0 \in (5000, \ldots, 10000)$, that is, choosing any values from 5000 till 10000. Although of the number of non-zero coefficients are slightly close together for some candidate value for NEG shrinkage prior, we still have to address this method in terms of the stability and convergences for very close values to zero. Therefore, we will not use it for comparison in the following section.

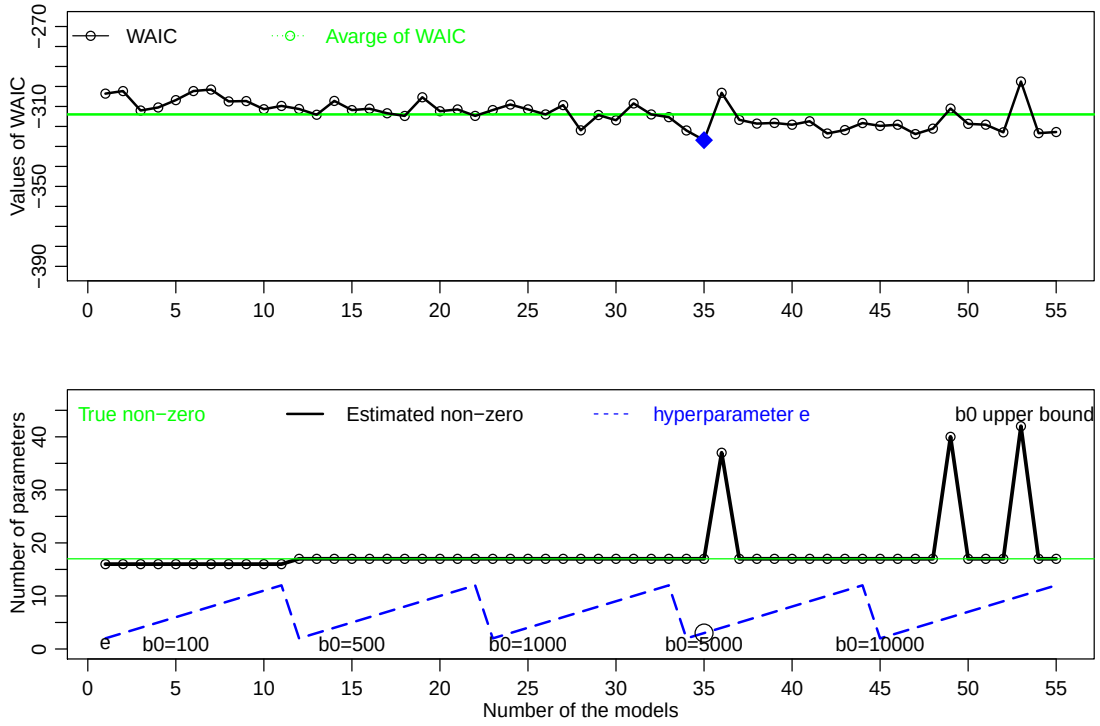


Figure 5.6: Hyperparameters parameters selection for ReB-hNiP; the top panel shows two different lines, where WAIC is represented by black lines with black-circle and green solid line refers to average of WAIC. The bellow panel displays three lines where green solid line indicates the number of non-zero coefficients in the simulation data where $n = 50, p = 80, R = 25$. Black-dash-dot represents the number of estimated coefficients and the blue dash line represent different values of upper bounds b_0 and parallel with hyperparameters e. In bottom panel, black triangle presents the prior values of γ which has minimum WAIC.

5.2.7 Numerical Results of Simulated Data

In this section, we will analyse the MCMC results for 100 replication posterior distributions for two different sample sizes that includes all scenarios for datasets D3 and D4, which we explained in Section 5.2.5. The linear regression model with measurements error model based on our proposed shrinkage priors are implemented. The MCMC algorithm have been run for $M = 20000$ iterations and kept 40% of iterations. The prior values for λ and τ are similar for both cases, in which the values for error of model λ are $a_\lambda = 2, b_\lambda = 0.62$ and for measurement error τ are $a_\tau = 2, b_\tau = 0.62$. The parameter of the degree of freedom ν belongs to the Student-t distribution, which is used as scale mixture Normal distribution for measure-

ment error. The values of hyperparameters for all hNiP, ReB-hNiP priors are chosen depending on the appropriate values in the best bounds range that we have obtained in the previous section. Furthermore, the prior values of hyperparameters κ of $c = 0.33$ and $d = 0.22$ for all case scenarios in the dataset D3 where $n = 30, p = 40$ and for the sample size for dataset D4 is $n = 50$ and $p = 80$. Also, $b_0 = 10000$ is the upper bounds parameter for hNiP and ReB-hNip shrinkage prior distribution, and the values of rescaled beta parameter is $e_0 = 7$. Posterior mean values based on Bayesian shrinkage priors are computed and compared by using average of 100 replication measures including sensitivity, specificity and MSE which are displayed in Table 5.3.

Table 5.3: Results average of 100 replication MCMC sampler for some proposed Bayesian shrinkage hNiP andReB-hNiP. The measures and performance used for be comparison between them are Sensitivity, Specificity and MSE. Several cases scenario based on signal to noise error are considered for both datasets D3 and D4. The number of iterations was equal to $M = 20000$. λ represents the error model and τ is measurement error model.

Dataset		**hNiP**			**ReBhNiP**		
	$\lambda,\ \tau$	**Sensitivity**	**Specificity**	**MSE**	**Sensitivity**	**Specificity**	**MSE**
$n = 30, p = 40$	$\lambda = 10, \tau = 15$	100%	99.90%	0.0158	100%	100%	0.0065
	$\lambda = 0.1, \tau = 15$	100%	100%	2.0273	100%	100%	3.181
	$\lambda = 0.1, \tau = 1.5$	100%	100%	0.0300	100%	100%	0.0588
$n = 50, p = 80$	$\lambda = 10, \tau = 15$	100%	100%	0.0019	100%	99.94%	0.0024
	$\lambda = 0.1, \tau = 15$	100%	100%	0.150	100%	100%	0.790
	$\lambda = 0.1, \tau = 1.5$	100%	100%	0.0255	97.57%	99.69%	.0644

The MCMC results based on the hNiP shrinkage prior show that hNiP is outperforming compared with similar proposed methods of the ReB-hNip for both datasets and cases scenario of the different noise model. We compared models through sensitivity, specificity and measures of mean square error as shown in Table 5.3 by using posterior mean values. For instance, the average of sensitivity for the dataset D3, where $n = 30, p = 40$ are equal to 100% with for cases scenarios for both λ and τ. In this situation, if sensitivity and specificity for typical datasets are identical, we compare MSE of the them to investigate the performances, either ReB-hNiP or hNiP prior distribution is used. For example, the MSE for dataset $(n = 30$ and $p = 40)$ where error mode $\lambda = 0.1$ and measurement error $\tau = 1.5$ is equal to 0.03 for hNiP model and it is smaller than the MSE for the same dataset based on ReB-hNiP equal to 0.0588. This indicated

that model based on hNiP give better results than ReB-hNiP. Variables selection through both methods satisfactory, because both sensitivity, specificity are approximately close to 100% and is telling us that non-zero regression coefficient and zero coefficients estimated exactly in this section.

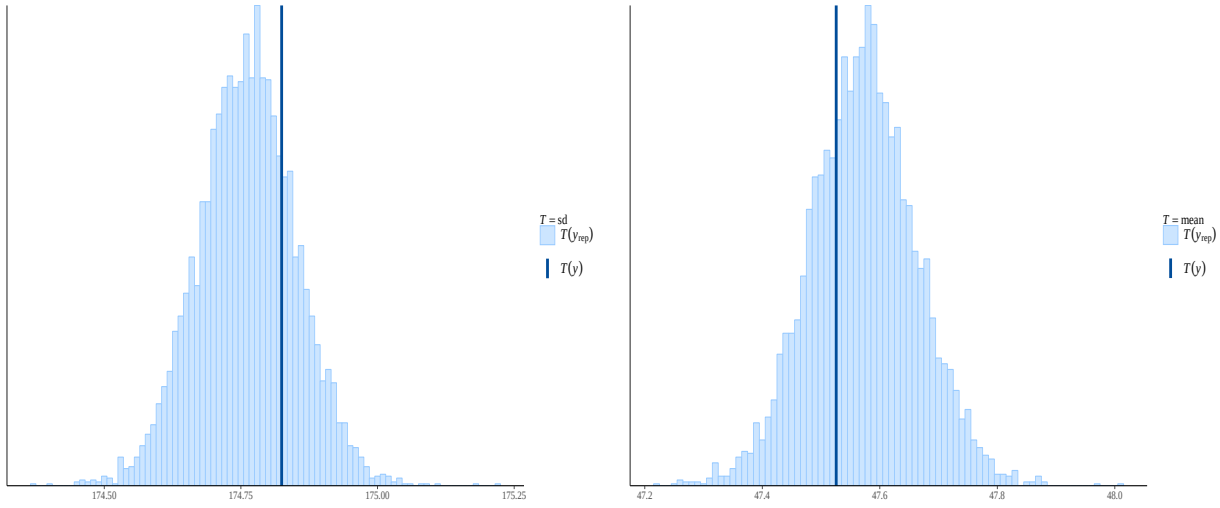
On the other hand, we can notice that when there is large difference between error model and measurements error model, for example, when $\lambda = 0.1$ and $\tau = 15$, the variables are selected completely but the average of the MSE are larger than MSE for different datasets with closer value for both, the error model and measurements error model. For example, the values of MSE is equal to 0.0019 for dataset with $\lambda = 10$ and $\tau = 15$, but MSE is equal to 0.15 for similar datasets with $\lambda = 0.1$ anf $\tau = 15$ in same sample size. This indicated that MSE for noise datasets is larger than other datasets, and it takes longer time to convergence to the true values. Therefore, the MCMC outcomes for the measurement error model has better results based on stander linear regression model, which we implemented in the previous chapter.

Table 5.4: Comparing computation time of Bayesian regression with measurements error model based on both ReB-hNiP and hNiP shrinkage prior distribution for different datasets. The number of iteration of all models is equal to $M = 20000$.

	Models	
Datasets	**hNiP**	**ReB-hNiP**
$n = 30, p = 40$	7 Minutes	12 Minutes
$n = 50, p = 80$	49 Minutes	50 Minutes

Regarding to the time of computation for measurements error with linear regression model, hNiP prior is faster than ReB-hNiP prior. Table 5.4 displays the times of computing every models for different datasets and Bayesian shrinkage priors, where the number of iteration is $M = 20000$ and we kept 50% of M, for all models in PC of 8 GB RAM and processor of i7 Double CPU 2.5 GHz. As we have discussed, the procedure of selecting hyperparameters values for each proposed Bayesian shrinkage prior distributions were done by WAIC criteria. Moreover, we applied MCMC algorithm based on the appropriate values in the best bounds range (see Section 5.2.6) to know the performance of each utilised shrinkage prior distributions in the last section.

Now, we are going to demonstrate the results in several ways for a particular dataset that utilised previously, in order to check our modelling through graphical posterior predictive distribution. We will check, to know how the chain of the parameters are converged to true values. Consequently, we have chosen one sample of MCMC results from replication sampler for datasets and case scenarios when using Bayesian linear regression with measurement error model, particularly, the datasets D4 and all cases scenarios.

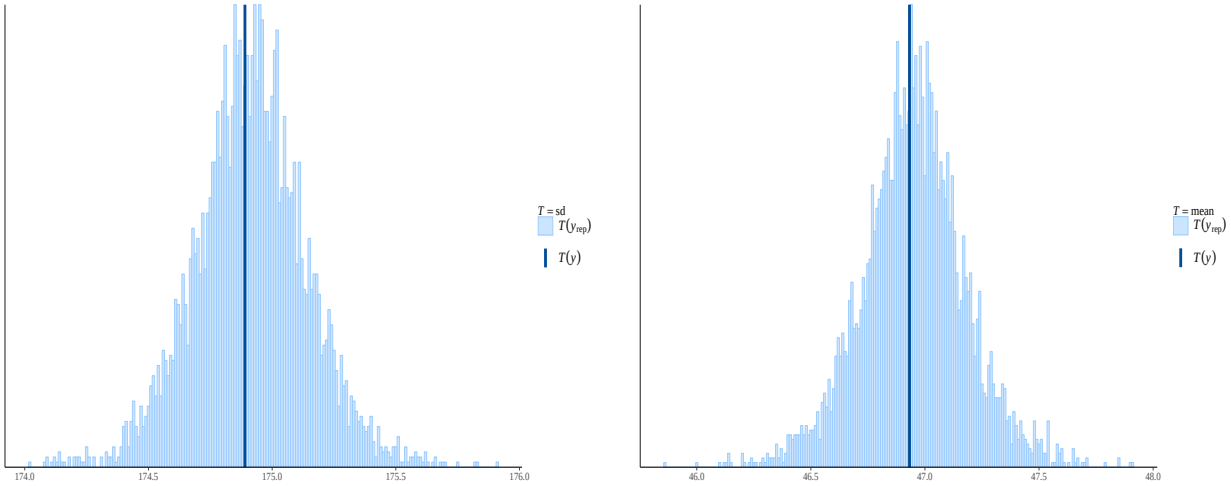


(a) sd (b) mean

Figure 5.7: Posterior predictive distributions based on the hNiP shrinkage prior distribution, for the dataset when $\lambda = 10, \tau = 15$ and sample size $n = 50$ and $p = 80$. Black vertical lines is a test statistics in panel (a) is standard division sd, and in panel (b) is mean. The brighter region on the panel is a histogram of test statistics for posterior replication.

Figure 5.7, 5.8 and 5.9 shows histograms of the posterior predictive distributions based on the procedure that described by Gelamn et al (2013). These graphs is produced using the *bayesplot* package Gabry (2017), in R programming. The black vertical lines represent the test statistics for true dependent variables, and we used both mean and standard division as test statistics . Figures 5.7 shows the graphical posterior predictive distribution for MCMC results of hNiP model for D4 dataset and for case scenario $\lambda = 10$ and $\tau = 15$. Figures 5.8 and 5.9 are conducted for dataset explained in D4 for cases scenarios when $\lambda = 0.1$ and $\tau = 15$ for both utilised hNiP and ReB-hNiP shrinkage prior respectively. We see that, based on the results of theses the graphs that Bayesian regression with measurement error model using our proposed shrinkage prior distribution are satisfactory as the black vertical lines in all plots are moderated

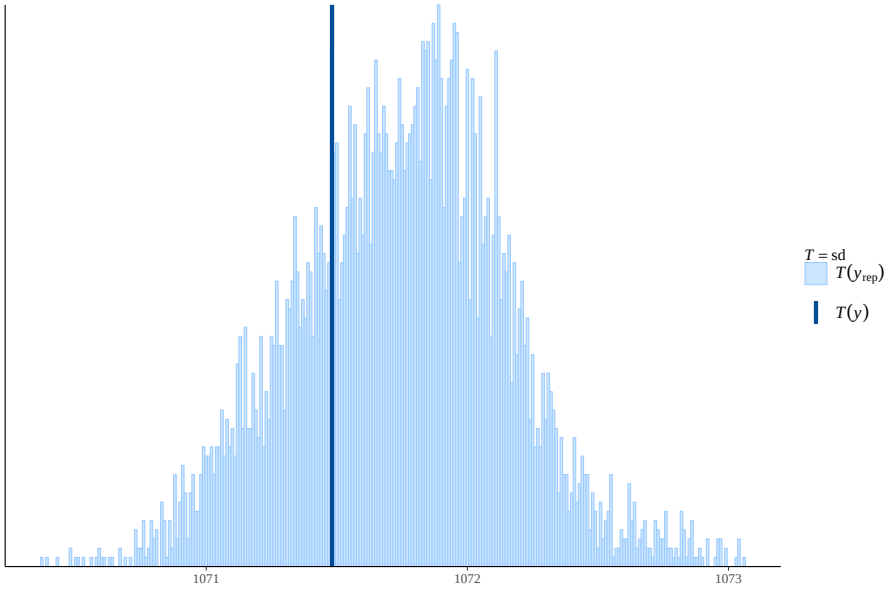
in the middle of the histograms of the posterior predictive distribution.



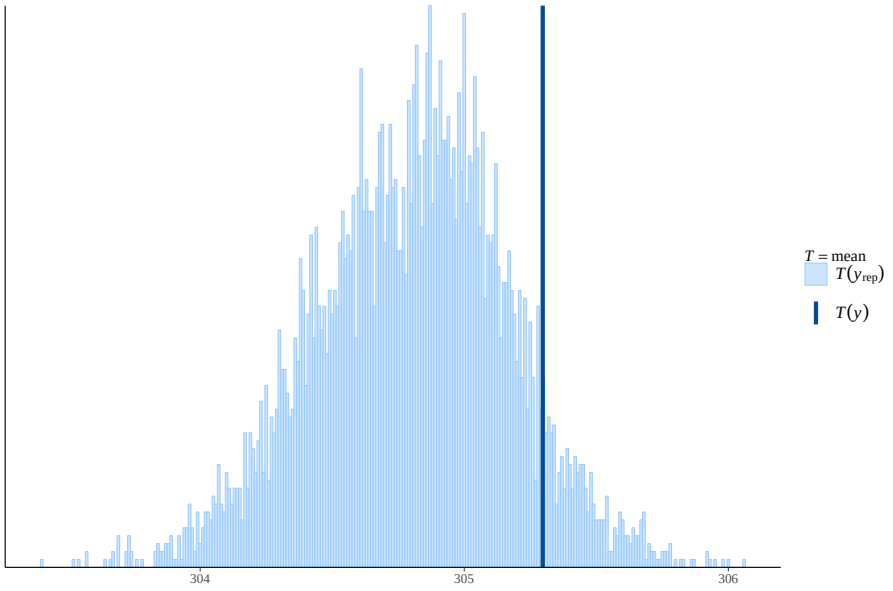
(a) sd (b) mean

Figure 5.8: Posterior predictive distributions based on the ReB-hNiP shrinkage prior distribution, for the dataset when $\lambda = 0.1, \tau = 15$ and sample size $n = 50$ and $p = 80$. Black vertical lines is a test statistics in panel (a) is standard division sd, and in panel (b) is mean. The brighter region on the panel is a histogram of test statistics for posterior replication.

To investigate the stability and convergences of the trace plots of non-zero regression coefficients, The traces are presented in the Figures 5.10 which it is clearly show the coefficients are convergent for sample size $n = 50, p = 80$ and $R = 25$. Where $\lambda = 0.01$ and $\tau = 15$.



(a) sd



(b) mean

Figure 5.9: Posterior predictive distributions based on the hNiP shrinkage prior distribution, for the dataset D4 when $\lambda = 0.1, \tau = 15,$ where sample size $n = 50$ and $p = 80$. Black vertical lines is a test statistics in panel (a) is standard division sd, and in panel (b) is mean. The brighter region on the panel is a histogram of test statistics for posterior replication.

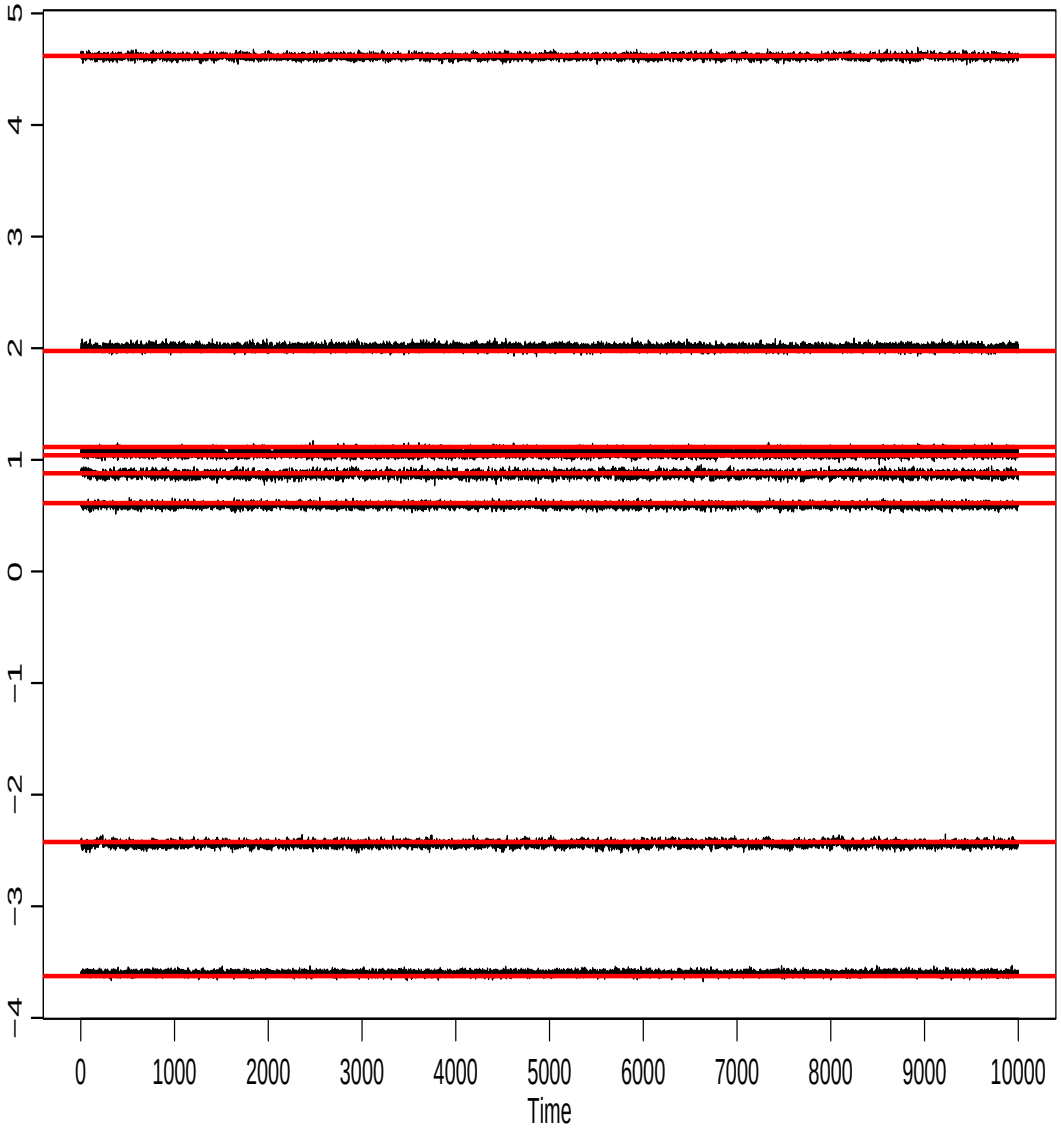


Figure 5.10: Trace plots for non-zero coefficients β for regression model with measurements error based on ReB-hNiP shrinkage priors, where solid-red lines refers to true values of coefficients, black line represents the traces of estimated regression coefficients. The D4 datasets with sample size is $n = 50,\ p = 80,\ R = 25$, and error mode $\lambda = 0.1$ and measurements error $\tau = 15$.

5.3 Dynamic Bayesian Network with Measurement Error Model

5.3.1 Introduction

Probabilistic Graphical Models (GMs) is a type of statistical methodology which has been widely used over the past few decades. With this kind of methodology, the estimation of gene regulatory networks (GRNs) is done by graphical models. Thus, GRNs play a crucial role in different cellular and biological systems, which is beneficial for discovering and treating diseases. Additionally, a graphical model studies the relationship and interactions between covariates; typically this relationship is described by the probability between variables and is presented in the graph. Common dataset that are used in this kind of modelling are gene expression data, because this methodology has developed significantly to be able to investigate diseases. Gene expression data are divided into two main types that are utilised for estimation and learning the GRN structure. The first one is time series data where data (gene expression levels) are measured at different time points sequentially, and the second one is steady-space that measures data at steady-space levels for several samples (Bar-Joseph, 2004; Fan et al., 2017). Furthermore, the network model can be studied by time series data (time crosses data) due to the information of dynamic features.

Constructing the model of genetic regulatory networks from data, particularly genetic data, is one of the main methods for inferring the interaction uncertainties between genes. It is known that a general autoregression method for building gene networks is impractical because of the naturally time-consuming inference of the model and also for obtaining results of a model under study. Therefore, researchers have developed various computational methods for modelling by different techniques, as we illustrated in the previous chapter. Consider that the dynamic Bayesian network DBN is one of the most common statistical models to identify the advanced interactions between genes, which are described by different kinds of genetic network models. Thus, in this part of the chapter, we deal with a high-dimensional dynamic Bayesian network model relying on an AR(1) structure and a measurement error model. In the usual DBN model, it is assumed that variables are measured exactly, where it considers the data as observed without error. However, in the situation of having measurement error within covariates, true variables do not exist. In this case, we take into consideration that the variables are unobserved. We focus on

a measurement error model which has replicated data and involves additional errors. Ignoring measurement errors in modelling leads to misleading amounts of the link between response and independent variables (Gilks, 1999).

Several wide-ranging studies cover the material about measurement error models that is presented in the first section of this chapter and also in the literature of studies related to sparse dynamic Bayesian network models. Furthermore, a number of studies for inferring the structure of a DAG as explained in section 1.5.1 have been demonstrated, therefore we will deal with high-dimensional data to learn the structure of the dynamic model in this thesis. Moreover, a number of inference methods have been suggested to reduce and estimate the topology of a DAG. For example, Murphy (2002) applied some Bayesian structure learning techniques for DBNs; the BNT (Bayes Net Toolbox) is a package in Matlab used for this purpose. Meinshausen and Bühlmann (2006) first applied LASSO to network inference that used cross-validation for selection variables.

Opgen-Rhein and Strimmer (2007) proposed James Stein shrinkage prior, which is another technique used in shrinking the coefficients of the autoregressive AR(1) model, . They showed that this method outperformed many classical methods. The network construction is fixed by constraining the edges in order to reduce the number of parameters. In addition, the partial correlations are used in a selection of subsequent models. The methods in this model outperformed other methods in the case of having a high dimension and a small sample size. G1DBN is another method for learning DBNs that depends on first-order conditional dependencies. This was proposed by Lèbre (2009), who tested partial regression coefficients to reduce variables for the first step. Therefore, this test was based on the t-test to obtain final networks and estimated regression coefficients of the controlled network. Shojaie and Michailidis (2010) showed that the LASSO is the only reliable variable selection technique under strict conditions. The adaptive lasso can consistently estimate the true parameters under standard regularity assumptions. For more details about assumptions, see Section 4 of Shojaie and Michailidis (2010). They considered these two penalties in order to discover a sparse estimate of the adjacency matrix, that is, covariance matrix of a DAG, particularly variables that have the same inherent natural ordering. Morrissey et al. (2011) used a semi-parametric spline model. It allowed for non-linear connections to simultaneously network edge-selections reached by utilizing Gibbs variable selection. They proposed a fully Bayesian set up for dealing with a smoothness parameter for

spline regression. Fu and Zhou (2013) improved a LASSO-penalized likelihood method of structure learning of Gaussian Bayesian networks (GBNs) and a block-wise coordinate descent algorithm is taken into account in this model. They considered it as an issue in the general situation, where the ordering of the variables is unknown.

From previous problems that were mentioned, another issue has arisen which is dealing with a dataset that are suffering from high dimensionality, gene expression level as an example. In addition, with such an issue, we cannot base it on the usual standard linear regression model because the number of parameters G is much larger than the number of the observations n. In the time series data $G = p \times p$ and $n = T \times p$ where T refers to time and p represents the number of genes. The challenging task with this kind of problem is how to learn the model structure under the assumption of sparsity, that is, establishing the graph network by a few edges which represent the relationship between nodes (variables /genes).

To overcome this issue in this study, we have utilised our proposed Bayesian shrinkage prior distribution which includes hNiP and ReB-hNiP, which are discussed in chapter 3, for estimating sparse coefficients. These prior distributions are heavy-tailed distribution and shrink the regression coefficients close to zero. Thereby, our model construction is similar to Gaussian model of Morrissey et al. (2010), but it differs through the methods of shrinkage. Furthermore, two new kinds of Bayesian shrinkage prior distribution are used in order to learn about unknown parameters, especially the coefficients matrix in our study. In the following sections, we present the model framework, full conditional distributions, simulated data, and results of simulated data. In section 5.3.10 we compare our proposed model with Morrissey et al. (2010).

5.3.2 Model Setup

In this section, we explain the framework of dynamic Bayesian networks with measurement error model as we highlighted in section 1.5.2. A Bayesian approach has been used to estimate unobserved variables and parameters. We will introduce our model in Equation 5.30 which follows the autoregressive model, and Equation 5.31 represents measurement error model. DBN model construction can be shown in terms of a directed acyclic graph, which has the joint

probability as follows:

$$P(Y^t|Y^{t-1}) = \prod_{j=1}^{p} P\left(Y_j^t|Pa_j^{t-1}(G)\right). \tag{5.29}$$

Where DBN are discussed in details in Section 1.5.2. The general form of the model can be described as follows:

$$Y^t = A + BY^{t-1} + \epsilon^t, \tag{5.30}$$

where $\epsilon^t \sim N(\epsilon|0, \lambda^{-1})$ represents the random error and the unobserved data is denoted by $Y^t \in \mathbb{R}^p$, where the size of Y^t is $(p \times 1)$, $t = 1, 2, 3, \ldots, T$ is an unobserved variable and $Y_i^t = \{Y_1^t, Y_2^t, \ldots, Y_p^t\}$ is a single value. The dimension of the coefficient matrix B is $G = p \times p$ which represents the interaction between variables (genes) and each vector of B denoted by β_i where zero denote absence of a relation between variables and non-zeros indicate a relation between variables i and j, where $i = 1, 2, \ldots, p$ and $j = 1, 2, \ldots, p$. Also, note that p represents the number of variables (genes) and A is an intercept parameter that is a vector of size $p \times 1$, where $A = \{\alpha_1, \alpha_2, \ldots, \alpha_p\}$. The measurement error model is presented as follows:

$$X_{ir}^t = Y_i^t + \eta_{ir}^t, \tag{5.31}$$

where the measurement error model η_{ir}^t has zero mean normal distribution and variance τ^{-1}. The likelihood function for both observed and unobserved variables is as follows:

$$\begin{aligned} L(\Theta) = \prod_{t=2}^{T} \lambda^{p/2} \exp\left\{\frac{-\lambda}{2}(Y^t - A - BY^{t-1})' I (Y^t - A - BY^{t-1})\right\} \\ \prod_{i=1}^{p}\prod_{t=1}^{T}\prod_{r=1}^{R} \tau^{\frac{1}{2}} \exp\left\{\frac{-\tau}{2}(X_{ir}^t - Y_i^t)^2\right\}, \end{aligned} \tag{5.32}$$

where Θ represents the set of unknown parameters in the model.

5.3.3 The Prior and Posterior Distributions for ReB-hNiP Prior for β_i

In this section, we establish the prior distributions for each parameter in our model, and then full conditional distributions are presented. For the first case, we use ReB-hNiP as the prior distribution of coefficients β_{ij} as we discussed in Chapter 3. We have used a Gamma dis-

tribution as the prior for both the precision and the measurement error with shapes a_λ, a_τ and scales b_λ, b_τ, respectively. The intercept α_i has a flexible flat prior distribution, where P_α and A^* are constant prior values for the intercept. R refers to the number of replications, T time points and p represents the number of covariates. Also e_0 is hyperparameter the rescaled Beta for κ and b_0 refers to the upper bounds hyperparameter. The prior distribution represents as follows:

$$\begin{aligned}\pi(\alpha,\beta,\lambda,\tau,\psi) = {} & \frac{b_\tau^{a_\tau}}{\Gamma(a_\tau)}\tau^{a_\tau-1}\exp\{-b_\tau\tau\}\,\frac{d_\lambda^{c_\lambda}}{\Gamma(c_\lambda)}\lambda^{c_\lambda-1}\exp\{-d_\lambda\lambda\} \\ & \prod_{i=1}^{p}\prod_{j=1}^{p}\frac{\psi_{ij}^{\frac{1}{2}}\exp\left(-\frac{1}{2}\psi_{ij}\beta_{ij}^2\right)}{\sqrt{2\pi}}\kappa_{ij}b_0^{-\kappa_{ij}}\psi_{ij}^{\kappa_{ij}-1}\frac{e_0^{-1}}{B(a,b)}\left(\frac{\kappa_{ij}}{e_0}\right)^{a-1}\left(1-\frac{\kappa_{ij}}{e_0}\right)^{b-1} \\ & \prod_{i=1}^{p}P_{\alpha_i}^{\frac{1}{2}}\exp\left\{\frac{P_\alpha(\alpha_i-A^*)}{2}\right\}\frac{q_\alpha^{1/2}}{\sqrt{2\pi}}\exp\left\{\frac{-q_\alpha A^*}{2}\right\}.\end{aligned} \tag{5.33}$$

Hence, the posterior distribution given by:

$$\begin{aligned}P(\Theta|Data) = {} & \prod_{t=2}^{T}\lambda^{p/2}\exp\left\{\frac{-\lambda}{2}(Y^t-A-BY^{t-1})'I(Y^t-A-BY^{t-1})\right\} \\ & \prod_{i=1}^{p}\prod_{t=1}^{T}\prod_{r=1}^{R}\tau^{\frac{1}{2}}\exp\left\{\frac{-\tau}{2}(X_{ir}^t-Y_i^t)^2\right\}\frac{b_\tau^{a_\tau}}{\Gamma(a_\tau)}\tau^{a_\tau-1}\exp\{-b_\tau\tau\}\,\frac{d_\lambda^{c_\lambda}}{\Gamma(c_\lambda)}\lambda^{c_\lambda-1}\exp\{-d_\lambda\lambda\} \\ & \prod_{i=1}^{p}\prod_{j=1}^{p}\frac{\psi_j^{\frac{1}{2}}\exp\left(-\frac{1}{2}\psi_{ij}\beta_{ij}^2\right)}{\sqrt{2\pi}}\kappa_{ij}b_0^{-\kappa_{ij}}\psi_{ij}^{\kappa_{ij}-1}\frac{e_0^{-1}}{B(a,b)}\left(\frac{\kappa_{ij}}{e_0}\right)^{a-1}\left(1-\frac{\kappa_{ij}}{e_0}\right)^{b-1} \\ & \prod_{i=1}^{p}P_{\alpha_i}^{\frac{1}{2}}\exp\left\{\frac{P_\alpha(\alpha_i-A^*)}{2}\right\}\frac{q_\alpha^{1/2}}{\sqrt{2\pi}}\exp\left\{\frac{-q_\alpha A^*}{2}\right\}.\end{aligned}$$

5.3.4 The Full Conditional Distributions of ReB-hNiP Prior for β_i

In the following equations, we present the full conditional distributions for all unknown parameters. It is clear that through finding the full conditional distributions analytically, most of the parameters have specific conjugate statistical distribution. Therefore, the Gibbs sampler is used to update those parameters that have a typical probability distribution. However, only the hyperparameter κ has not been described by a specific probability distribution. Consequently,

we used the Metropolis-Hastings algorithm for updating it.

We update the unobserved variables Y for $t = 2, \ldots, T-1$ as follows:

$$Y_i^t \sim N\left(Y_i^t \middle| \frac{\lambda\left(\alpha_i + \sum_{j=1}^{p} \beta_{ij} Y_j^{t-1} + \beta_{ii}\left(Y_i^{t+1} - \alpha_i - \sum_{i \neq j}^{p} \beta_{ij} Y_j^t\right)\right) + \tau R \bar{X}_i^t}{\lambda(1+\beta_{ii}^2) + \tau R}, \lambda(1+\beta_{ii}^2) + \tau R\right), \tag{5.34}$$

where $i = 1, 2, 3 \ldots p$ and $t = 2, \ldots, T-1$.

Regarding to updating the unobserved variables Y^T, we rely on the following formula:

$$Y_i^{t=T} \sim N\left(Y_i^{t=T} \middle| \frac{\lambda\left(\alpha_i + \sum_{i \neq j}^{p} \beta_{ij} Y_i^{t=T-1} + \beta_{ii} Y_i^{t=T-1}\right) + \tau R \bar{X}^{t=T}}{\lambda + R\tau}, (\lambda + R\tau)\right). \tag{5.35}$$

We supposed that $Y^1 \sim N(Y|0,1)$. The intercept parameter of the model is represented by α_i and is updated as follows:

$$\alpha_i \sim N\left(\alpha_i \middle| \frac{\lambda\left(\sum_{t=2}^{T} Y_i^t - \sum_{t=2}^{T} \sum_{j=1}^{p} \beta_{ij} Y_j^{t-1}\right) + A^*}{(T\lambda + p_\alpha)}, T\lambda + \sigma_\alpha\right), \tag{5.36}$$

where,

$$A^* \sim N\left(A^* \middle| \frac{pP_\alpha \bar{\alpha}}{(q_a + pP_\alpha)}, (q_a + pP_\alpha)\right). \tag{5.37}$$

The precision λ of the model can be computed using;

$$\lambda \sim \text{Ga}\left(\lambda \middle| c_\lambda + \frac{pT}{2}, d_\lambda + \frac{1}{2}\sum_{t=2}^{T} (Y^t - A - BY^{t-1})' I (Y^t - A - BY^{t-1})\right). \tag{5.38}$$

The precision of the measurement error model τ is updated using:

$$\tau \sim \text{Ga}\left(\tau \middle| a_\tau + \frac{RT(G-1)}{2}, b_\tau + \frac{1}{2}\sum_{i=1}^{p}\sum_{t=1}^{T} (X_i^t - Y_i^t)^2\right), \tag{5.39}$$

where X_i^t is a vector of size $R \times 1$.

The parameters ψ and κ are hyperparameters of the prior for β_{ij} in the case of using ReBeta-hNIP distribution. The full conditional distribution for ψ_{ij} is given by:

$$\psi_{ij} \sim \text{TrGa}\left(\psi_{ij}\middle| b_0, \kappa_{ij} + \frac{1}{2}, \frac{\beta_{ij}^2}{2}\right), \tag{5.40}$$

where TrGa stands for the Truncated Gamma distribution that is bounded from above. The full conditional distribution of the hyperparameters κ_{ij} is as follow:

$$\pi(\kappa_{ij}) \propto b_0^{\kappa_{ij}} \psi_{ij}^{\kappa_i j-1} \frac{e_0^{-1}}{B(a,b)} \left(\frac{\kappa_{ij}}{e_0}\right)^{a-1} \left(1 - \frac{\kappa_{ij}}{e_0}\right)^{b-1}. \tag{5.41}$$

which is not a standrad probability distribution. For that reason, Metropolis-Hastings is used for updating κ_{ij}. Note that the values of κ_{ij} should satisfy the condition $0 < \kappa_{ij} \leq e_0$.

The full conditional distribution for $(\boldsymbol{\beta_i})$ is as follows:

$$\boldsymbol{\beta_i} \sim N\left(\boldsymbol{\beta_i}\middle| \mu_{\boldsymbol{\beta_i}}, \Omega_{\boldsymbol{\beta_i}}^{-1}\right), \tag{5.42}$$

where $\boldsymbol{\beta_i}$ represents a single vector in B. $\mu_{\boldsymbol{\beta_i}} = \lambda\, \Omega_{\boldsymbol{\beta_i}}^{-1}\left(\sum_{t=2}^{T} Y^{t-1}(Y_i^t - \alpha_i)\right)$, $\Omega_{\boldsymbol{\beta_i}} = \left(\lambda\big(\sum_{t=2}^{T}(Y^{t-1}Y^{t-1'})+\Sigma_{\psi_i}\right)$ and Σ_{ψ_i} is a $p \times p$ matrix with diagonal equal to $\psi_i = \{\psi_1, \psi_2 \ldots, \psi_p\}$.

5.3.5 The Prior and Posterior Distributions of hNiP Prior for β_i

The prior distributions are the same as those used in section 5.3.3, except using the hNiP prior for coefficients as follows:

$$\pi(\alpha,\beta,\lambda,\tau,\psi) = \frac{b_\tau^{a_\tau}}{\Gamma(a_\tau)}\tau^{a_\tau-1}\exp\{-b_\tau\tau\}\,\frac{d_\lambda^{c_\lambda}}{\Gamma(c_\lambda)}\lambda^{c_\lambda-1}\exp\{-d_\lambda\lambda\}$$
$$\prod_{i=1}^{p} P_{\alpha_i}^{\frac{1}{2}}\exp\left\{\frac{P_\alpha(\alpha_i-A^*)}{2}\right\}\frac{q_\alpha^{1/2}}{\sqrt{2\pi}}\exp\left\{\frac{-q_\alpha A^*}{2}\right\} \tag{5.43}$$
$$\prod_{i=1}^{p}\prod_{j=1}^{p}\frac{\psi_{ij}^{\frac{1}{2}}\exp\left(-\frac{1}{2}\psi_{ij}\beta_{ij}^2\right)}{\sqrt{2\pi}}\kappa_{ij}b^{-\kappa_{ij}}\psi_{ij}^{\kappa_{ij}-1}\frac{\kappa_{ij}^{c-1}d^c\exp(-\kappa_{ij}d)}{\Gamma(c)}.$$

The posterior distributions given by:

$$P(\Theta|Data) = \prod_{t=2}^{T}\lambda^{p/2}\exp\left\{\frac{-\lambda}{2}(Y^t-A-BY^{t-1})'I(Y^t-A-BY^{t-1})\right\}$$
$$\prod_{i=1}^{p}\prod_{t=1}^{T}\prod_{r=1}^{R}\tau^{\frac{1}{2}}\exp\left\{\frac{-\tau}{2}(X_{ir}^t-Y_i^t)^2\right\}\frac{b_\tau^{a_\tau}}{\Gamma(a_\tau)}\tau^{a_\tau-1}\exp\{-b_\tau\tau\}\,\frac{d_\lambda^{c_\lambda}}{\Gamma(c_\lambda)}\lambda^{c_\lambda-1}\exp\{-d_\lambda\lambda\}$$
$$\prod_{i=1}^{p}\prod_{j=1}^{p}\frac{\psi_{ij}^{\frac{1}{2}}\exp\left(-\frac{1}{2}\psi_{ij}\beta_{ij}^2\right)}{\sqrt{2\pi}}\kappa_{ij}b^{-\kappa_{ij}}\psi_{ij}^{\kappa_{ij}-1}\frac{\kappa_{ij}^{c-1}d^c\exp(-\kappa_{ij}d)}{\Gamma(c)}$$
$$\prod_{i=1}^{p}P_{\alpha_i}^{\frac{1}{2}}\exp\left\{\frac{P_\alpha(\alpha_i-A^*)}{2}\right\}\frac{q_\alpha^{1/2}}{\sqrt{2\pi}}\exp\left\{\frac{-q_\alpha A^*}{2}\right\}. \tag{5.44}$$

5.3.6 The Full Conditional Distributions for hNiP Prior β_i

The full conditional distributions for all unknown parameters are presented in the following points. Analytically, the full conditional distributions for all unknown parameters have specific probability distributions. Thus, Gibbs sampler is used to conduct MCMC for all parameters.

Updating unobserved variables for $t = 2,\ldots,T-1$ is as follows:

$$Y_i^t \sim N\left(Y_i^t\middle|\frac{\lambda\left(\alpha_i+\sum_{j=1}^{p}\beta_{ij}Y_j^{t-1}+\beta_{ii}\left(Y_i^{t+1}-\alpha_i-\sum_{i\neq j}^{p}\beta_{ij}Y_j^t\right)\right)+\tau R\bar{X}_i^t}{\lambda(1+\beta_{ii}^2)+\tau R},\lambda(1+\beta_{ii}^2)+\tau R\right), \tag{5.45}$$

where $i = 1,2,3,\ldots,p$ and $t = 2,\ldots,T-1$.

Regarding computation for both cases of $t = 1$ and $t = T + 1$, we use the following formulas:

$$Y_i^{t=T} \sim N\left(Y_i^{t=T}\middle|\frac{\lambda\left(\alpha_i + \sum_{i\neq j}^{p}\beta_{ij}Y_i^{t=T-1} + \beta_{ii}Y_i^{t=T-1}\right) + \tau R\bar{X}^{t=T}}{(\lambda + R\tau)}, \lambda + R\tau\right), \tag{5.46}$$

where,

$$\alpha_i \sim N\left(\alpha_i\middle|\frac{\lambda\left(\sum_{t=2}^{T}Y_i^t - \sum_{t=2}^{T}\sum_{j=1}^{p}\beta_{ij}Y_j^{t-1}\right) + A^*}{(T\lambda + p_\alpha)}, T\lambda + \sigma_\alpha\right), \tag{5.47}$$

and

$$A^* \sim N\left(A^*\middle|\frac{pP_\alpha\bar{\alpha}}{(q_a + pP_\alpha)}, (q_a + pP_\alpha)\right). \tag{5.48}$$

$$\lambda \sim \text{Ga}\left(\lambda\middle|c_\lambda + \frac{pT}{2}, d_\lambda + \frac{1}{2}\sum_{t=2}^{T}(Y^t - A - BY^{t-1})'I(Y^t - A - BY^{t-1})\right), \tag{5.49}$$

where,

$$\tau \sim \text{Ga}\left(\tau\middle|a_\tau + \frac{RT(G-1)}{2}, b_\tau + \frac{1}{2}\sum_{i=1}^{p}\sum_{t=1}^{T}(X_i^t - Y_i^t)^2\right), \tag{5.50}$$

and X_i^t is a vector of size $R \times 1$. Also, the full conditional distribution of the hyperparameters κ_{ij} is defined as follows:

$$\kappa_{ij} \sim \text{Ga}\left(\kappa_{ij}\middle|c + 1, \log\left(\frac{b_0}{\psi_{ij}}\right) + d\right), \tag{5.51}$$

where $i, j = 1, 2 \ldots, p$.

We have explained in the previous chapter $\text{TrGa}(x|b_0, a, b)$ which is the Truncated Gamma distribution and it is bounded from above. The parameters ψ and κ are hyperparameters of the prior for β_{ij} in the caseof using the ReB-hNiP distribution. The full conditional distribution for ψ_{ij} is given by:

$$\psi_{ij} \sim \text{TrGa}\left(\psi_{ij}\middle|b_0, \kappa_{ij} + \frac{1}{2}, \frac{\beta_{ij}^2}{2}\right). \tag{5.52}$$

The full conditional distribution for $\boldsymbol{\beta_i}$ is defined as follows:

$$\boldsymbol{\beta_i} \sim N\left(\boldsymbol{\beta_i}\middle|\mu_{\boldsymbol{\beta_i}}, \Omega_{\boldsymbol{\beta_i}}^{-1}\right), \tag{5.53}$$

where $\boldsymbol{\beta_i}$ represents a single vector in matrix B, $\mu_{\boldsymbol{\beta_i}} = \lambda\,\Omega_{\boldsymbol{\beta_i}}^{-1}\left(\sum_{t=2}^{T} Y^{t-1}(Y_i^t - \alpha_i)\right)$, $\Omega_{\boldsymbol{\beta_i}} = \lambda\left(\sum_{t=2}^{T}(Y^{t-1}Y^{t-1'}\right) + \Sigma_{\psi_i}$ and Σ_{ψ_i} is a $p \times p$ matrix with diagonal equal to $\psi_i = \{\psi_1, \psi_2 \ldots, \psi_p\}$.

The pseudo-code for both dynamic Bayesian networks with measurements error models based on hNiP and ReB-hNiP are presented in Appendix C.3.

5.3.7 Simulated Data

In order to check the performance of shrinkage prior distributions and to evaluate our model, we generated different data samples. The synthetic data of the dynamic Bayesian network with measurement error depends on both the autoregressive model and measurement error model as follows:

$$Y_i^t = \alpha_i + BY_i^{t-1} + \epsilon^t, \tag{5.54}$$

and,

$$X_{ir}^t = Y_i^t + \eta_{ir}^t, \tag{5.55}$$

where $\eta_{pr}^t \sim N(\eta|0, \tau^{-1})$, $i = 1, 2, \ldots, p$, random error $\epsilon^t \sim N(\epsilon|0, \lambda^{-1})$ and unobserved variables $Y^t \in \mathbb{R}^p$. The size of Y^t is $p \times 1$, $t = 1, 2, \ldots, T$ is unobserved variables and $Y_i^t = \{Y_1^t, Y_2^t, \ldots, Y_p^t\}$ is a single value, where Y_i^t has mean of zero and variance of 25. Also, p refers to the number of parameters (genes), R is the number of replications (patients) and T represents the number of times where data has been collected. Also, α_i is the intersection value and B is $p \times p$ coefficient matrix. We have simulated data under two cases, Case A1 and Case A2. The structures for both cases are described in the following points:

- **Case A1:**

  In this case, we suppose that we have enough information to infer parameters in the model, which means we do not have the problem of high-dimensionality. We assumed that the size of the stimulated dataset is $p = 10, R = 21$ and $T = 50$. The initial values for the precision of the AR(1) model and the precision of the measurement error are $\lambda = 10$ and $\tau = 15$ respectively. The diagonal of the coefficients matrix B

is equal to 0.3, while only a few non-diagonal elements of B are non zero which are: $\beta_{2,9} = -0.5, \beta_{3,7} = -0.7, \beta_{4,3} = -0.5, \beta_{4,6} = 0.4, \beta_{5,3} = 0.6, \beta_{7,8} = 0.64, \beta_{10,6} = 0.25$ and $\beta_{10,9} = 0.7$.

- **Case A2:**

  In this case, we simulate a dataset with the problem of high dimensionality. We assume that the size of this sample is $p = 40, R = 21$ and $T = 20$. The initial values for the precision of the AR(1) model and the precision of measurement error are equal to $\lambda = 10$ and $\tau = 15$ respectively. The diagonal of coefficients matrix B is equal to 0.3, while only a few non-diagonal elements of B are non zero which are: $\beta_{2,9} = -0.5, \beta_{3,7} = -0.7, \beta_{4,3} = -0.5, \beta_{4,6} = 0.4, \beta_{5,3} = 0.6, \beta_{7,8} = 0.64, \beta_{10,6} = 0.25$ and $\beta_{10,9} = 0.7$.

5.3.8 Results From Simulated Data

In this section, we compare the results of selecting of tuning parameters based on our proposed Bayesian shrinkage prior distributions, when applied to dynamic Bayesian networks with a measurements error model. As we explained earlier when discussing the results of tuning parameters selection in the previous studies, the WAIC criteria had significantly better results compared to the DIC and EBIC$_\gamma$ criteria. Therefore, again we utilised the WAIC criteria, in order to detect the best bounds range of upper bounds parameters and hyperparameters of both hNiP and ReB-hNiP prior distributions. Simulated data presented in the Section 5.3.7 are the same as that currently used.

Selecting variables is an importation part of this section, so we used three different credible intervals (CI.s) which are $90\%, 95\%$ and 99% CI.s to identify whether coefficients are different from zero or not,that is, identifying the non-zero coefficients. Consequently, a link (edge) between two variables (nodes) is absent if the values of the coefficients are equal to zero. On the contrary, if the values of the coefficients are not equal to zero, a link between two nodes exists, meaning that one variable regulates another one either positively or negatively. Selecting the minimum value of WAIC criteria is not enough to detect the best hyperparameters prior values of any prior distribution, especially for our Bayesian shrinkage prior distribution, in order to know how corresponding the estimated non-zero coefficients to non-zero coefficients based on

simulated data. For that reason, we compute true positive (TP) and false negative (FN) values as detailed in Section 2.5.3 for each model and WAIC values. In addition, we aim to decrease the value the false negative FN to be close to zero, because it is representing the second type error, which is influential in such current kind of modelling, related to medical or genetic fields. We try to avoid the selecting variables that are not participating for design modelling via simulated data. We have simulated two examples to understand the procedure of selecting the hyperparameters and comparing shrinkage prior distributions. The first example is Case A1; where we consider $n > p$, that is, it is not a high-dimensional data problem), and there is much information to provide estimation and fit the model based on dynamic networks with both proposed Bayesian shrinkage prior distributions.

We now summarize the results of the gene regulatory network model with hNiP prior distribution. It is clear that in this shrinkage prior there are two parameters that control the amount of shrinkage which are the upper bound b_0 and the hyperparameter of the gamma prior c. For selecting these two tuning parameters, we analysed the MCMC results based on a sequence of values of c, which starts from 0.1 to 6 by 0.1 steps for several values of b_0. Moreover, we present the results of WAIC, TP, FN and total non-zeros by different graphs in Figure 5.11. The trend of WAIC declined as much as the values of b_0 increased and that lead to shrinking most of the coefficients closest to zero. The values of WAIC for the first half of the models are quite similar as the upper bounds value for this range being close together where $b_0 \leq 100$, which leads to shrinking less coefficients close to zero. On the contrary, the value of WAIC gradually declined as a result of increasing the value of b_0 and also the numbers of non zero coefficients are reduced too.

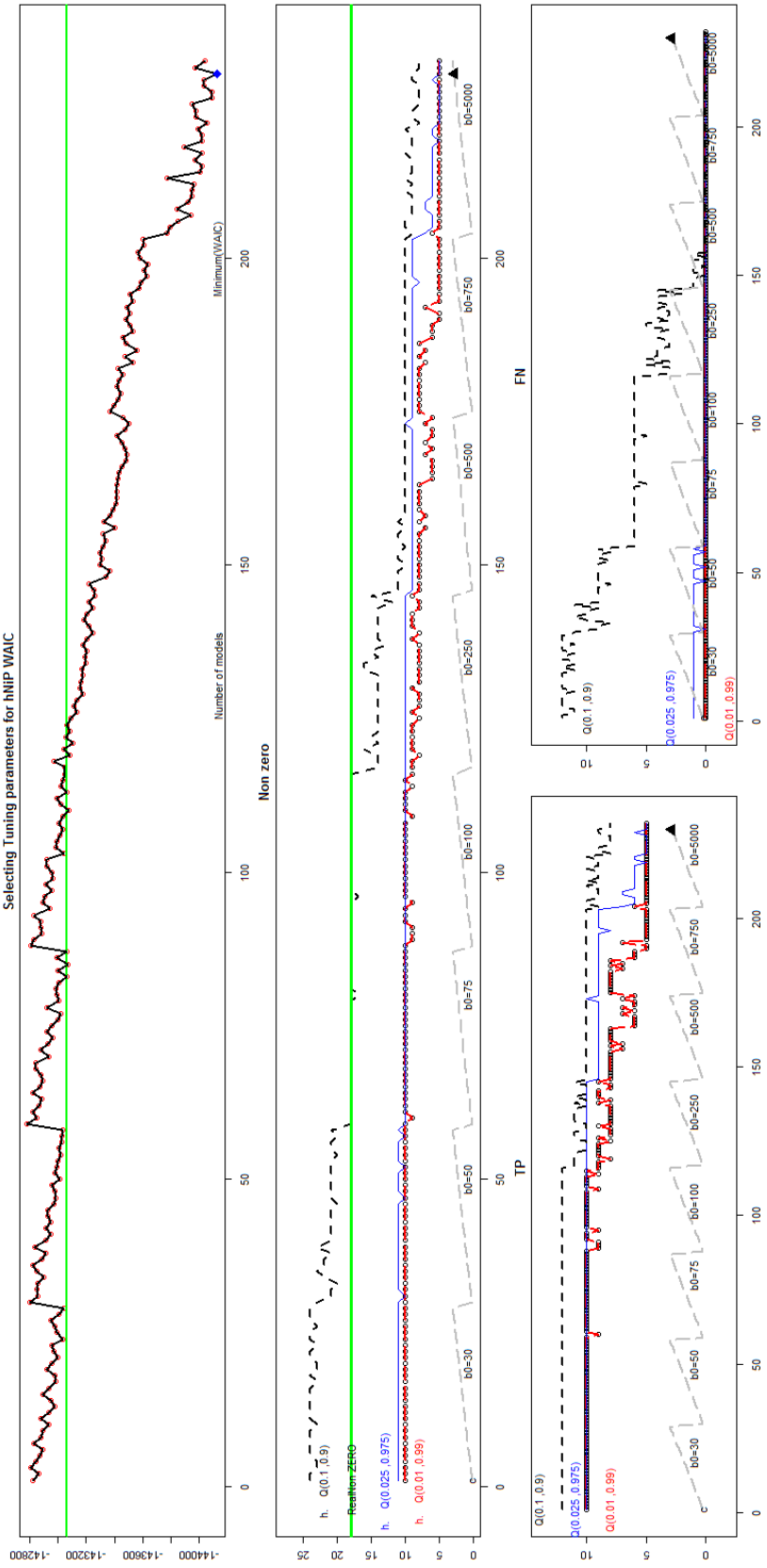


Figure 5.11: All graphs present the tuning parameters selection for WAIC for the dynamic Bayesian network with measurements error model and hNiP for shrinkage. In the top panel, the black line with red circle represents the WAIC values for the number of models and the solid green line is the average of it. In the middle graph, there are five different lines. The green-solid line refers to the number of non-zero coefficients in the simulated data, the number of estimated non-zero coefficients based on a credible interval are represented by the black-dashed line $(90\%$ CI), blue-solid line $(95\%$ CI) and black-dot-red line 99% CI). The gray dashed-line indicates the different values of the upper bounds b_0 of the hNiP prior with different values of the hyperparameters c on it. The right bottom graph shows the number of the non-zeros based on TP values. The left bottom graph shows the number of non-zeros based on FN values. These graphs are for sample Case A1 where $T = 50, R = 21, p = 10$ and the hNiP is used.

In the beginning of this section, we highlighted a selection of variables based on three different credible intervals (CI), $90\%, 95\%$ and 99%. This is to evaluate the performance of the current modelling because it is not easy to decide which parameters are equal to zero or not. Consequently, we plot MCMC results for TP, FN and total non-zero coefficients for all different values that are used here, as can be seen in the middle and bottom graphs Figure 5.11. Our goal is to diagnose whether the non-zero coefficients are estimated correctly, that is, they are the same as the non-zeros in simulated data, or they are false negative coefficients. Furthermore, there are four different lines in the middle graphs which include: a green-solid line referring to the number of non-zero coefficients in simulated data, the number of estimated non-zero coefficients is based on black-dash lines 90% CI, blue-solid line 95% CI, and black-dot-red line 99% CI. The grey dash-line indicates the different values of upper bounds b_0 of hNiP prior with different values of the hyperparameter c. Also, the bottom right graph shows the number of non-zeros based on TP values. The bottom left graph shows the number of non-zeros based on FN values. Although there are an estimated number of non-zeroes belonging to TP, it is not corresponding to the non-zeroes in the simulated data, especially as most of the diagonal elements of coefficients matrix shrink close to zero. The values of WAIC have similar trends for a range of upper bounds $b_0^* = (30, 50, 75, 100)$ and the values of WAIC slightly declined when the value of hyperparameters c and d increases for every b_0^*. The number of non-zero coefficients are the same for this range of upper bounds. Nevertheless, increasing the value of b_0 has significantly affected the shrinkage of the coefficients.

As we can see from Figure 5.11 the value of WAIC decrease as long as the value of b_0 rises, especially when $b_0 > 100$, and also the number of estimated non-zeros decline for both TP and FN. As a result, WAIC achieves its minimum value when $b_0 = 5000$ in our experiments. However, increase of the value of b_0 leads to shrinking the most coefficients and the amount of TP is less than the number of non-zero in the simulation data even though FN become zero. Although WAIC is declining with increasing values of b_0, we prefer using the tuning parameters of upper bounds around $b_0 = 75, 100$ in this example, owing to the fact that we have a significant number of TP and zero FN in this range of b_0 when using 95% and 99% CI.

In order to evaluate the capability of both shrinkage prior distributions for used fitting and shrinking coefficients in this particular model, we obtained the results of several models based on different values of the upper bound parameter b_0 and rescaled hyperparameter e for the

ReB-hNiP prior distribution. Through the aforementioned procedure, and similar to the hNiP shrinkage prior, tuning parameters are selected. Furthermore, we chose the hyperparameters such that $a = b = \frac{1}{2}$ which leads to a shape like horseshoe. In this case, selecting tuning parameters is challenging because there is no clear trend in the value of WAIC. Thus, for evaluating and detecting the best range of hyper-prior values for ReB-hNiP prior distribution, TP, FN and WAIC are computed. The minimum values of WAIC is -145780 whereas the number of TP= $6, 7$ and FN= $0, 2$ based on 95% and 99% credible intervals, respectively. As we highlighted before, the trend of WAIC is not stable, while the number of FN is close to zero as long as the value of b_0 increases and the number of TP decreases and stable around 4 using the credible intervals that mentioned previously. We observe that the best range of values for this shrinkage prior is $b_0 \geq 250$, especially for upper bounds between $b_0 = 250$ and $b_0 = 1000$ because in this range of values the number of TP are around 5 and 8 and the FN approach zero as we can see in Figure 5.12.

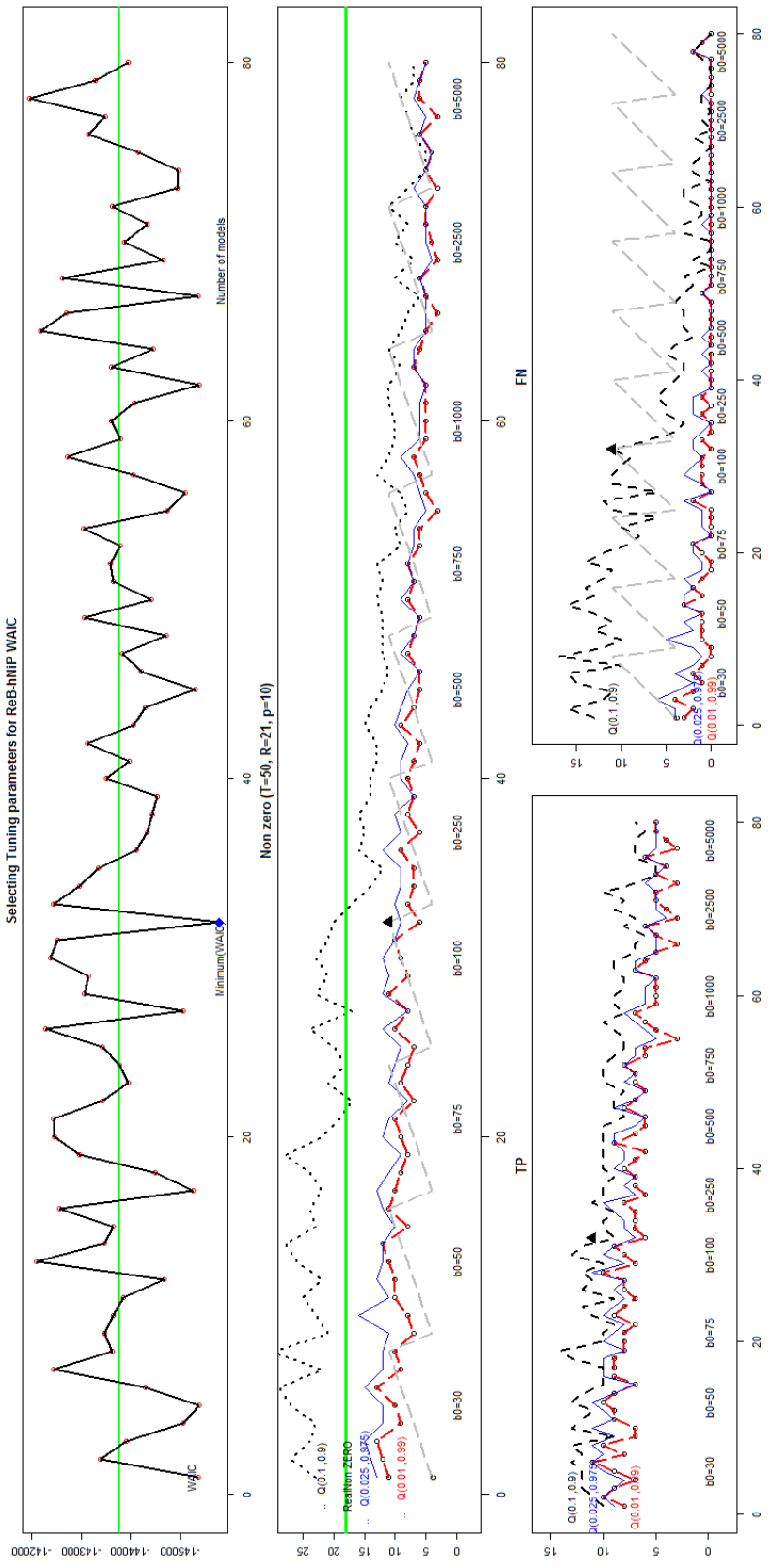


Figure 5.12: All graphs present the tuning parameters selection for WAIC for the dynamic Bayesian network with measurements error model and ReB-hNiP for shrinkage. In the top panel, the black line with red circle represents the WAIC values for the number of models and the solid green line is the average of it. In the middle graph, there are five different lines. The green-solid line refers to the number of non-zero coefficients in the simulated data, the number of estimated non-zero coefficients based on a credible interval are represented by black-dashed line $(90\%\text{CI})$, blue-solid line $(95\%\text{CI})$, and black-dot- red line $(99\%\text{CI})$. The grey dashed-line indicates the different values of upper bound, b_0 of the ReB-hNiP prior with different values of the hyperparameters of rescaled beta e on it. The right bottom graph shows the number of non-zeros based on TP values. The left bottom graph shows the number of non-zeros based on FN values. These are for case A1 where $T = 50, R = 21, p = 10$ and ReB-hNiP prior is used.

As we have noticed, both (ReB-hNiP and hNiP) proposed Bayesian shrinkage prior distributions have unsatisfactory results over all prior values that we used for both of them, therefore we did not estimate exactly the same or very close to the number of non-zero coefficients. Also the diagonal of it is aggressively shrunk to zero. In spite of these unsatisfactory results, hNiP out-performance ReB-hNiP based on the simulated dataset in case A1 where we have enough information to fit the model. In the next case, we focus on a high-dimensional problem.

Our aim in this example is to investigate the estimation of the coefficients matrix with different prior values and also select appropriate tuning parameter values for both proposed shrinkage distributions for a dynamic Bayesian network with measurement error, in case of having high-dimensional data. In recent years a number of shrinkage prior distribution have been used in order to shrink the high-dimensional regression coefficient matrix and to find interactions between genes. Also, the popular problem in such studies are having huge number of covariates (genes) and a few observations, and also a few time points. In light of this idea, we simulated a high-dimensional dataset as described in Case A2, and then we have run our MCMC algorithm and analysed the results. The number of non-zero coefficients in the simulated data is equal to 48 including the diagonal of the coefficients matrix. The MCMC algorithm is run for $10,000$ iterations with a burn-in of 25% of sample. WAIC, TP and FN are computed for evaluating and comparing the results of both Bayesian shrinkage prior distributions (hNiP, ReB-hNiP). We make a summary of the results of Case A2 similar to the former case based on several credible intervals, $90\%, 95\%$ and 99% CI, as presented for hNiP in Figure 5.13 and for ReB-hNiP in Figure 5.14. The number of TP is stable around 1 based on 95% and 99% CI. and dose not exceed 2 depend on 90% CI, in spite of the decreasing trend in WAIC and increasing the value of upper bounds parameter b_0 for several hyperparameter prior values in either the hNiP or the ReB-hNiP shrinkage prior distributions.

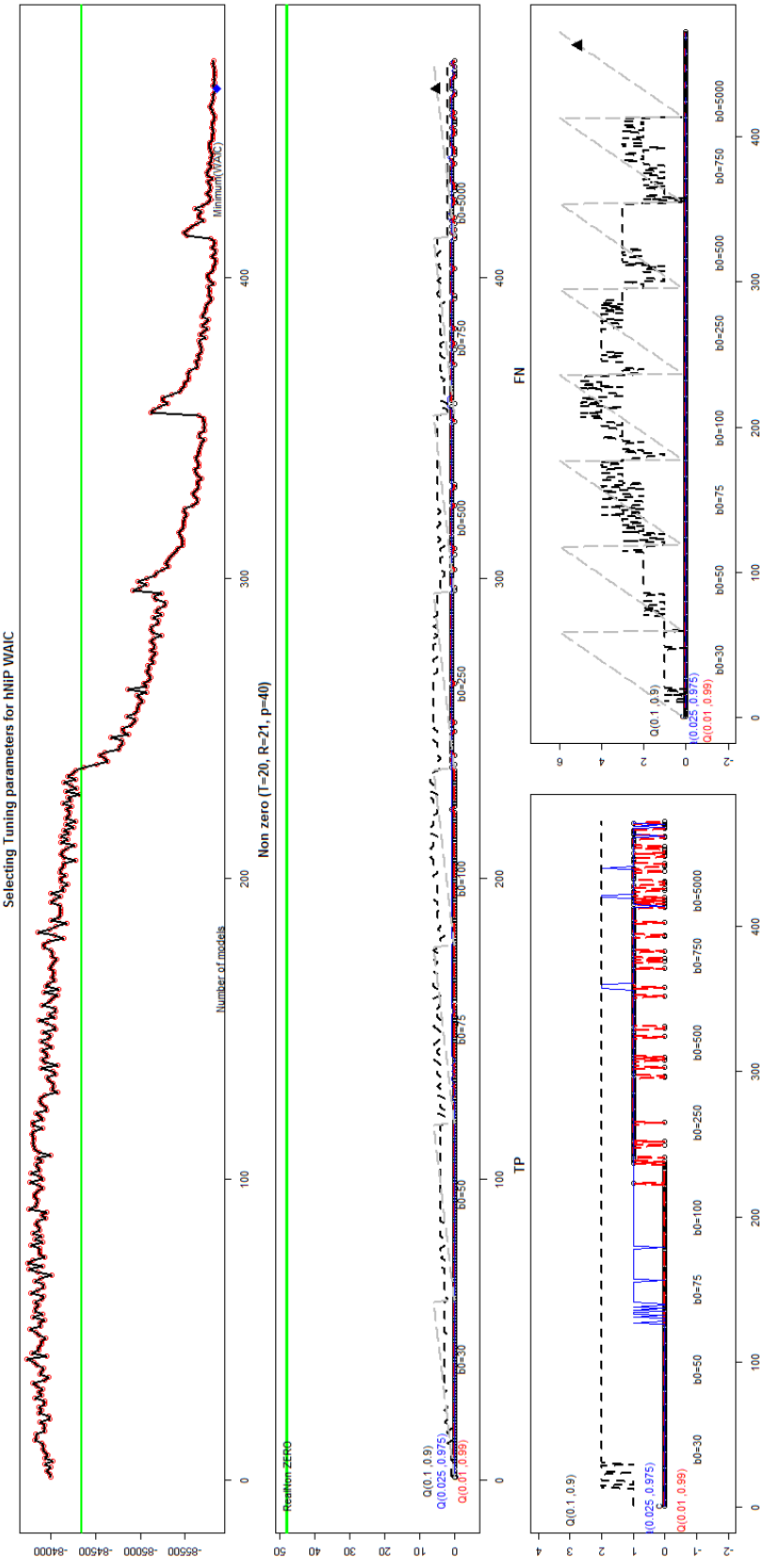


Figure 5.13: All graphs present the tuning parameters selection for WAIC for the dynamic Bayesian network with measurements error model and hNiP for shrinkage. In the top panel, the black line with red circle represents the WAIC values for the number of models and the solid green line is the average of it. In the middle graph, there are five different lines. The green-solid line refers to the number of non-zero coefficients in the simulated data, the number of estimated non-zero coefficients based on a credible interval are represented by black-dashed line $(90\%\text{CI})$, blue-solid line $(95\%\text{CI})$, and black-dot- red line $(99\%\text{CI})$. The grey dashed-line indicates the different values of upper bound, b_0 of the prior with different values of the hyperparameters of gamma prior as second layer e on it. The right bottom graph shows the number of non-zeros based on TP values. The left bottom graph shows the number the non-zeros based on FN values. These results are for case A2 where $(T = 20, R = 21, p = 40)$ when hNiP prior is used.

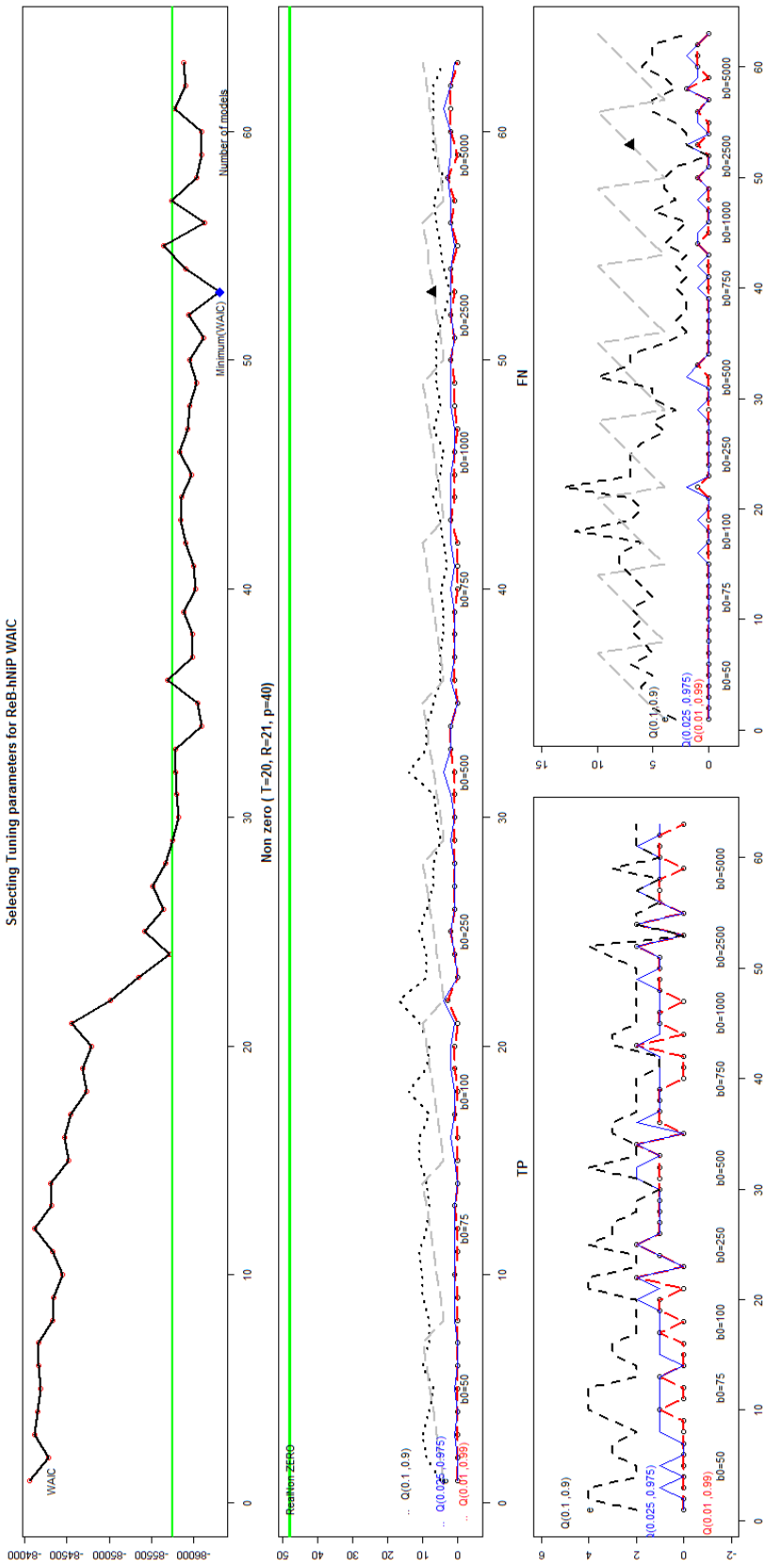


Figure 5.14: All graphs present the tuning parameters selection for WAIC for the dynamic Bayesian network with measurements error model and ReB-hNiP for shrinkage. In the top panel, the black line with red circle represents the WAIC values for the number of models and the solid green line is the average of it. In the middle graph, there are five different lines. The green-solid line refers to the number of non-zero coefficients in the simulated data, the number of estimated non-zero coefficients based on a credible interval are represented by black-dashed line $(90\%\text{CI})$, blue-solid line $(95\%\text{CI})$, and black-dot- red line $(99\%\text{CI})$. The grey dashed-line indicates the different values of upper bound, b_0 of the ReB-hNiP prior with different values of the hyperparameters of rescaled beta e on it. The right bottom graph shows the number of non-zeros based on TP values. The left bottom graph shows the number the non-zeros based on FN values. These results are for Case A2 where $T = 50, R = 21, p = 10$ and when ReB-hNiP prior is used.

Regarding the results that have been displayed in Figures 5.13 and 5.14 for both priors mentioned, it is difficult to select appropriate values significantly that lead to shrinking the coefficients matrix toward the actual value of the simulated data. We can conclude that hNiP and ReB-hNiP prior distributions aggressively shrink the most coefficients toward zero when we deal with dynamic Bayesian networks with measurement error model. Therefore, to further study our modelling approach, we have simulated a new dataset that is created similarly of example Case A1 but with the values of the diagonal coefficients increased from 0.3 to different values around 0.8. The results show that the number of non-zero coefficients with hNiP prior declines as the value of upper bound b_0 increases, therefore the number of FN decreases, especially when the value of $b_0 > 100$ which leads to FN approaching zero, whereas TP falls toward 5. Furthermore, variables are selected based on credible intervals of size 95% and 99% CI. Nevertheless, the values of WAIC are unstable as long as the value of b_0 are changed. Besides, the value of the precision parameter of AR(1) model, λ, and the precision of the measurement error model, τ, do not converge to their true values, and also the trends of both of them are unstable over all models that were computed based on several prior values for the hNiP prior.

Moreover, these outcomes can be interpreted in the way that increasing the values of the diagonal matrix (self-regulation values are large) is unsatisfactory for fitting the coefficients in the proposed model, using either prior hNiP or ReB-hNiP. We showed the results for this examples in Figures 5.15 and 5.16 in which plots represents the WAIC, TP, FN and total non-zero coefficients. The computation time to collect the MCMC results for dynamic Bayesian network with measurement error model is shown in Table 5.5 for both shrinkage prior distributions used. Consequently, we notice that our model is faster if the prior distribution based on hNiP for both cases of high-dimensional and non-high-dimensional datasets.

Table 5.5: Comparing computation time of the dynamic Bayesian network with measurement error model based on both ReB-hNiP and hNiP shrinkage prior distributions for different datasets. The number of iterations $M = 10000$.

Dataset	**Dynamic Bayesian Network with Measurement Error Model**	
	hNiP	**ReB-hNiP**
$T = 50, p = 10, R = 21$	4 Minutes, 15 seconds	7 Minutes, 11 seconds
$T = 20, p = 40, R = 21$	10 Minutes, 20 seconds	20 Minutes, 40 seconds

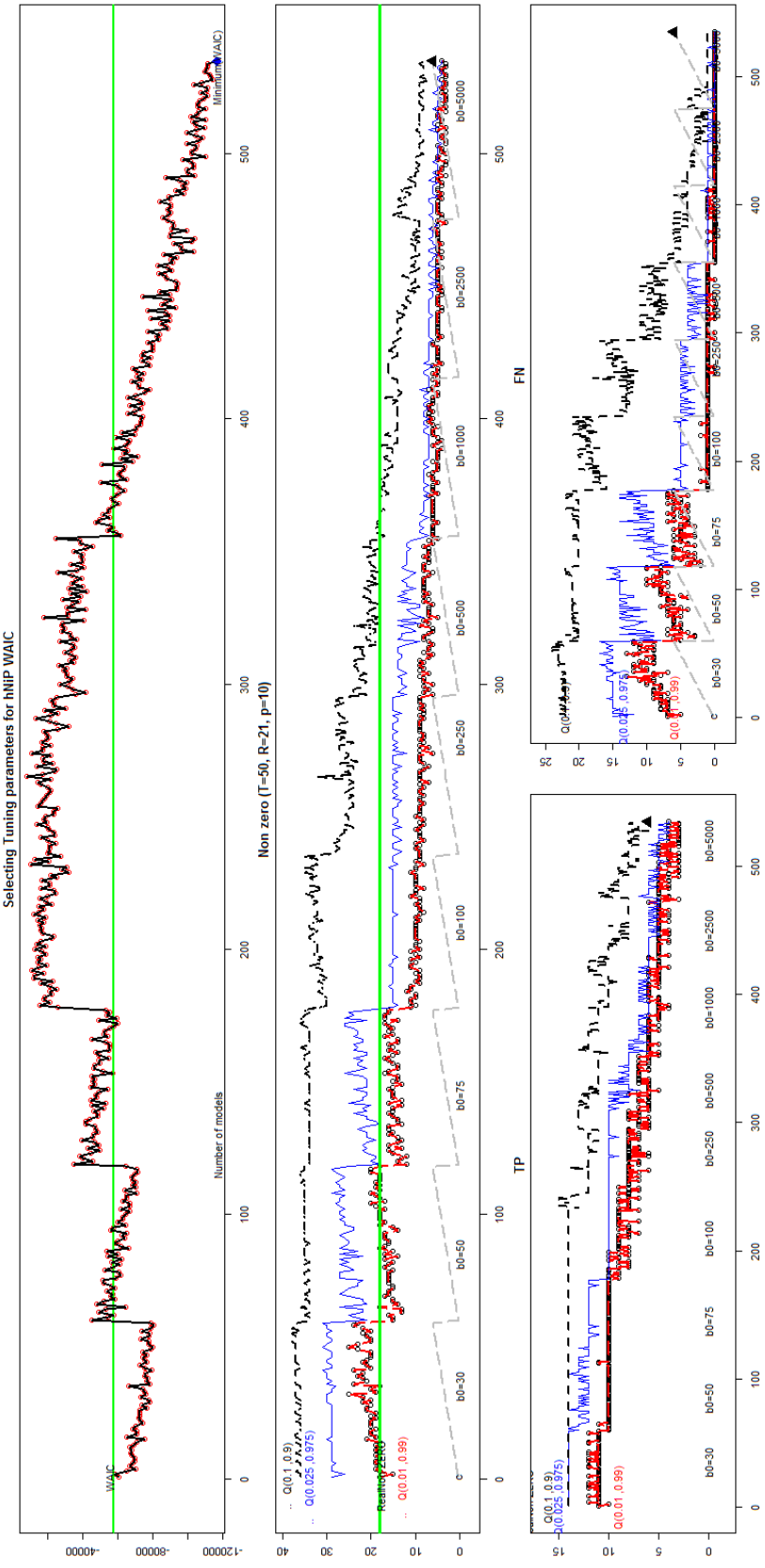


Figure 5.15: All graphs present the tuning parameters selection for WAIC for the dynamic Bayesian network with measurements error model and hNiP for shrinkage. In the top panel, the black line with red circle represents the WAIC values for the number of models and the solid green line is the average of it. In the middle graph, there are five different lines. The green-solid line refers to the number of non-zero coefficients in the simulated data, the number of estimated non-zero coefficients based on a credible interval are represented by black-dashed line $(90\%\text{CI})$, blue-solid line $(95\%\text{CI})$, and black-dot- red line $(99\%\text{CI})$. The grey dashed-line indicates the different values of upper bound, b_0 of the prior with different values of the hyperparameters of rgamma e on it. The right bottom graph shows the number of non-zeros based on TP values. The left bottom graph shows the number the non-zeros based on FN values. Theses results are for Case A3 where $(T = 50, R = 21, p = 10)$ and when hNiP with larger and different values of diagonal coefficients matrix is used.

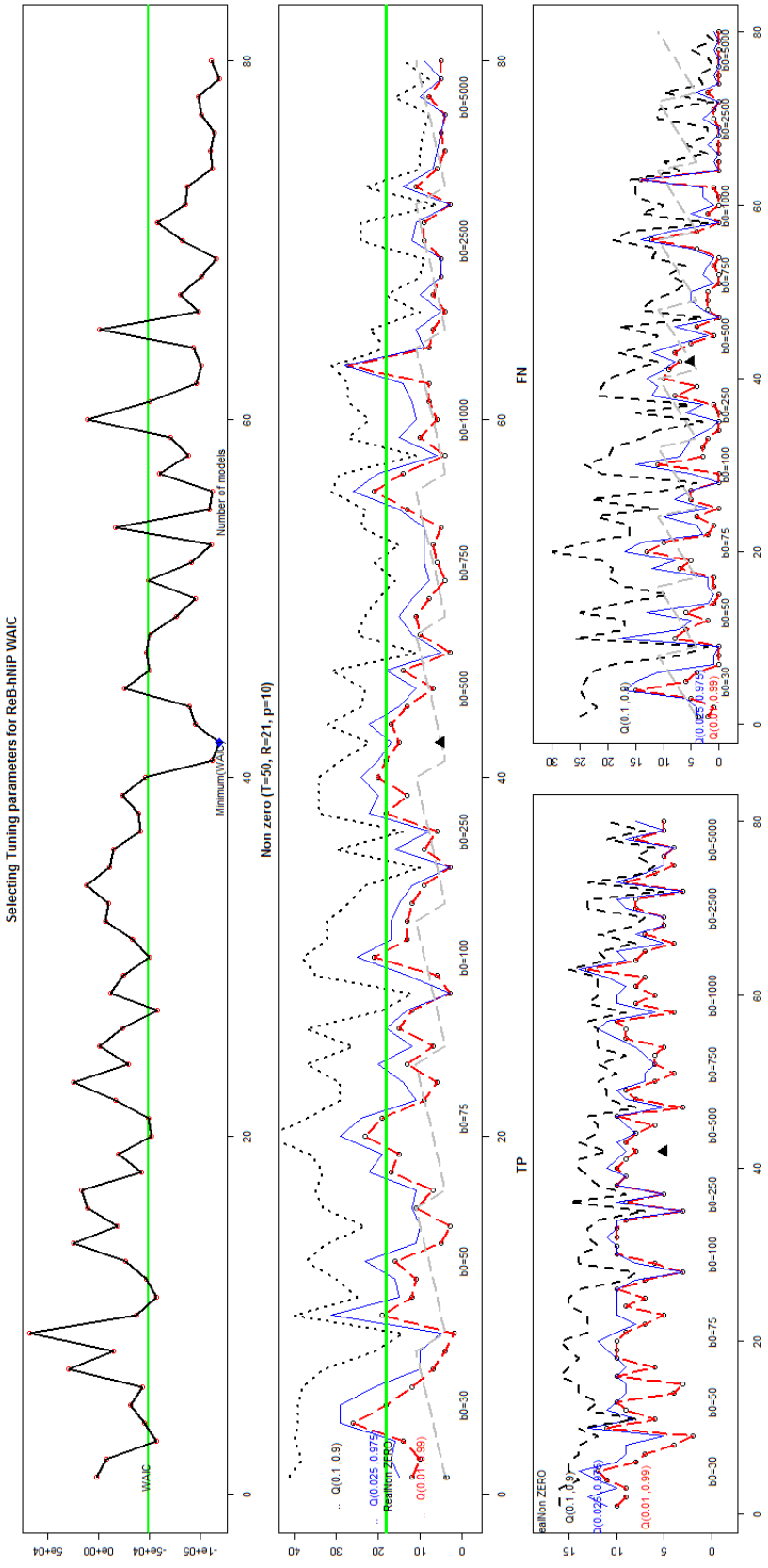


Figure 5.16: All graphs present the tuning parameters selection for WAIC for the dynamic Bayesian network with measurements error model and ReB-hNiP for shrinkage. In the top panel, the black line with red circle represents the WAIC values for the number of models and the solid green line is the average of it. In the middle graph, there are five different lines. The green-solid line refers to the number of non-zero coefficients in the simulated data, the number of estimated non-zero coefficients based on a credible interval are represented by black-dashed line $(90\%\text{CI})$, blue-solid line $(95\%\text{CI})$, and black-dot- red line $(99\%\text{CI})$. The grey dashed-line indicates the different values of upper bound, b_0 of the prior with different values of the hyperparameters of rescaled beta e on it. The right bottom graph shows the number of non-zeros based on TP values. The left bottom graph shows the number the non-zeros based on FN values. These results are for Case A3 where $T = 50, R = 21, p = 10$ and when ReB-hNiP with larger and different values of diagonal coefficients matrix is used.

5.3.9 Alternative Link Selection

In this section, the dynamic Bayesian network with measurements error model based on hNiP prior distribution is applied to the simulated dataset that is described in Case A1, in order to investigate which of the coefficients are convergent and have a significant t-test. The outcomes are obtained for twenty MCMC chains, where the number of iterations $M = 40000$ and 25% of iterations are discarded as burn in. In addition, the prior values were chose based on the results of the last section, so we used $b_0 = 750, c = 3$ and $d = 2$. The selection of variables is c using both the t-test and quantiles $Q_{0.10}$ and $Q_{0.90}$. Consequently, the findings show that 10% of the higher t-test values of coefficients are the same as the non-zero coefficients identified by the quantiles and, indeed are non-zero in the simulated data. We have ranked the coefficients based on their t-test values from highest to lowest in Table 5.6. The figure with the most significant coefficient has higher value t-test depending on posterior distribution values. The results in the last column of the table indicate the frequencies of each coefficient regarding the overall rank of the MCMC chain.

Table 5.6: Frequency and rank of the 10% of the t-test for higher coefficients. Dynamic Bayesian network with measurement error model / hNiP prior distribution are used for 20 MCMC chains.

Rank	**Coefficients**	**Order of coefficients matrix**	**Frequencies**
1	$\beta_{5,3}$	25	20
2	$\beta_{3,7}$	63	11
3	$\beta_{4,3}$	24	11
4	$\beta_{10,9}$	90	20
5	$\beta_{2,2}$	12	20
6	$\beta_{4,6}$	54	16
7	$\beta_{8,7}$	68	15
8	$\beta_{2,9}$	82	19
9	$\beta_{3,3}$	23	20
10	$\beta_{8,8}$	78	20
11	$\beta_{4,3}$	43	19
12	$\beta_{2,3}$	22	19
13	$\beta_{8,4}$	38	19
14	$\beta_{7,3}$	27	10
15	$\beta_{6,7}$	66	7
16	$\beta_{9,9}$	89	8
17	$\beta_{9,9}$	89	7
18	$\beta_{10,5}$	50	14
19	$\beta_{5,5}$	45	13
20	$\beta_{4,4}$	34	15

For example, coefficient $\beta_{5,3}$ has a the highest significant t-test which takes the first rank for all MCMC chains, while coefficient $\beta_{3,7}$ has second rank with frequency 11 and 9 times in the third rank. We noticed that all coefficients in the table have the same frequencies over all results based on 10% coefficients. Moreover, both the precision of error model and precision parameters of the AR(1) are convergent and close to their true values where the posterior median are approximate by $\tau = 14.85$ and $\lambda = 8.75$, respectively. Finally, we plot the relationship coefficients in the network in Figure 5.6.Although, some of the coefficients non-zero in the simulated data, they are estimated as zero due to shrunk to zero such as for $\beta_{5,5}, \beta_{4,4}$ and $\beta_{9,9}$.

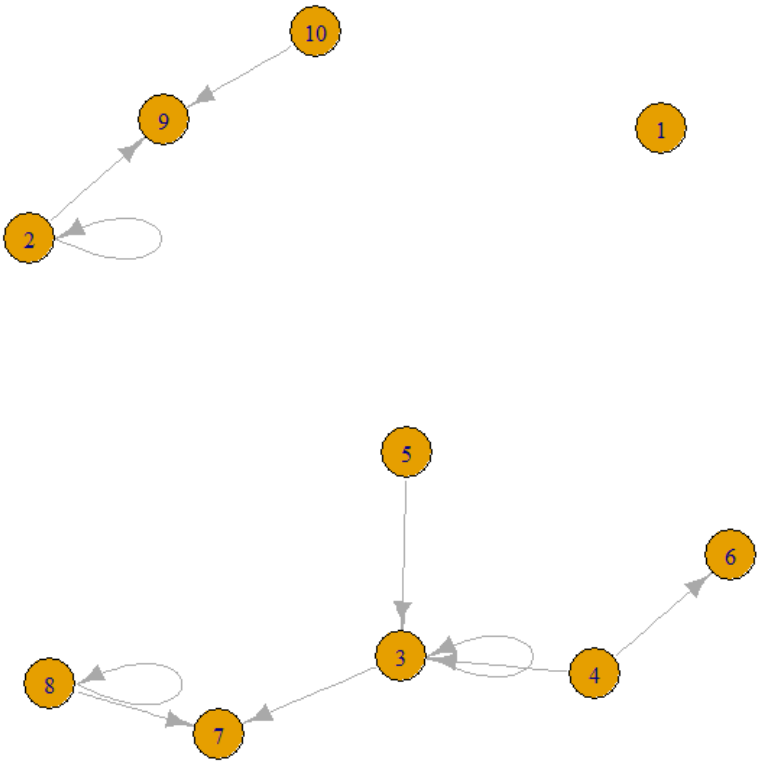


Figure 5.17: Graphical network for all coefficients is the dynamic Bayesian network with measurements error model / hNiP is shrinkage, where the circle represents the node and edges is value of coefficients, by using sample Case A1.

5.3.10 Comparison Example

We compared our dynamic Bayesian network with the measurements error model, to the dynamic Bayesian networks in case of linear interactions between genes, where used by Morrissey et al. (2010). Moreover, we have applied Arabidopsis thaliana's circadian clock dataset that is used in $GRENITS$ package for Morrissey (2012). Arabidopsis thaliana circadian clock data (Locke et al., 2006) "Arabidopsis thaliana, was first proposed to comprise a feedback loop in which two partially redundant genes, Late Elongated Hypocotyl (LHY) and Circadian Clock Associated 1 (CCA1), is the expression of their activator, Timing Of Cab Expression 1 (TOC1)." Data to consist of four replication and time represents hour, and then the log was calculated, where dataset consist of $T = 50, R = 4$ and $p = 5$ and it is two dimensional matrixes. Therefore, we converted to three dimensions array in order to utilise it in our proposed model. The MCMC have been run for 10000 iterations and 2500 iterations are burned for both hNiP and RiB-hNiP shrinkage prior distribution. The appropriate prior values used to fit the dynamic Bayesian network with measurement error model based on both priors are $b_0 = 75$ and the gamma hyperparameter for hNiP are $c = 3$ and $d = 2$, and the rescaled Beta hyperparameter for ReB-hNiP is $e_0 = 6$.

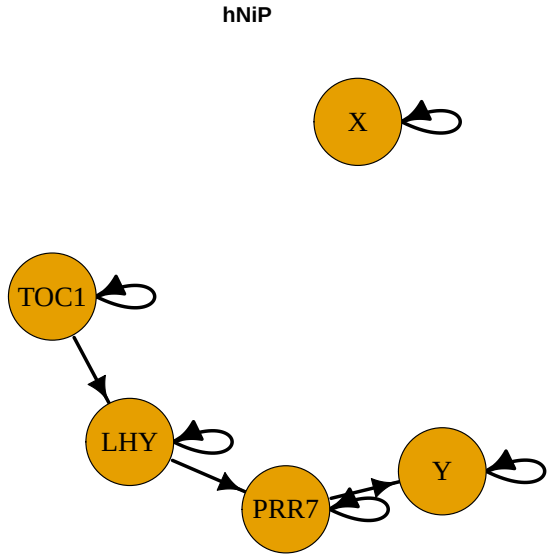
The MCMC results of coefficients for our proposed linear dynamic models for both hNiP and ReB-hNiP prior distributions have been analysed using 90% CI, which is presented in the networks of Figure 5.18. Results show that there are four nodes and three links which refer to the regulation of the genes in our models based on a t-test on the posterior median. The positive values in the coefficients matrix indicates the activation of gene - another gene, where we can see LHY activates to PRR7 and PRR7 activates to Y, where the value of the coefficients are 0.22 and 0.66, respectively, for the ReB-hNiP prior and are equal to 0.22 and 0.4 for the hNiP prior. However, the negative values of a gene prevent another gene, as TOC1 prevents LHY and the value of this coefficient is equal to -0.47 for the model that is based on the ReB-hNiP prior and -0.44 for hNiP prior. Regarding the self-regulation of genes in the dataset, we estimated all coefficients depending on the dynamic models where only one self-regulation variable Y is not detected for the model based on ReB-hNiP.

On the other hand, the results based on the GRENITS package indicate that there are more links (edges) between variables (nodes) as three links have been detected in this model, which are

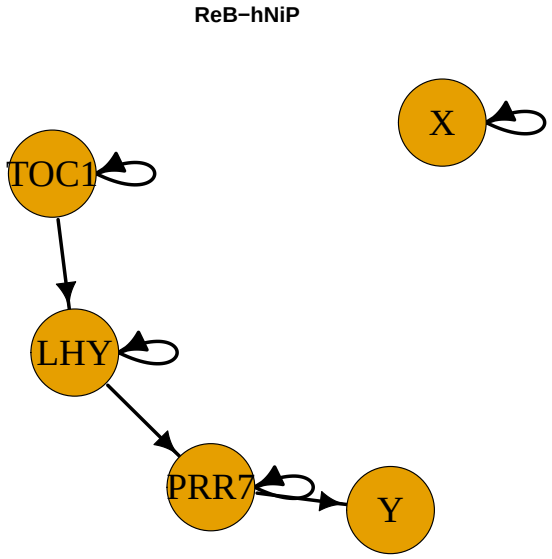
Table 5.7: frequency and rank of the 50% t -test of higher coefficients. Dynamic Bayesian network with measurement error model based on both ReB-hNiP and hNiP prior distribution MCMC chain run for 10000 iterations when "Athaliana ODE 4NoiseReps" dataset is used.

	Dynamic Bayesian Network with Measurement Error Model / (ReB-hNiP and hNiP prior)					
Rank	**Coefficients**	**t-test hNiP**	**Orderβ in matrix**	**Coefficients**	**t-test ReB-hNiP**	**Orderβ**
1	$\beta_{X,X}$	5.975	13	$\beta_{X,X}$	5.178	13
2	$\beta_{TOC1,TOC1}$	3.069	7	$\beta_{TOC1,TOC1}$	2.826	7
3	$\beta_{PRR7,PRR7}$	2.478	25	$\beta_{PRR7,PRR7}$	2.492	25
4	$\beta_{PRR7,Y}$	2.302	24	$\beta_{Y,PRR7}$	2.398	24
5	$\beta_{LHY,LHY}$	2.242	1	$\beta_{LHY,LHY}$	2.368	1
6	$\beta_{TOC1,LHY}$	2.123	6	$\beta_{LHY,TOC1}$	1.781	6
7	$\beta_{LHY,PRR7}$	1.878	5	$\beta_{PRR7,LHY}$	1.599	5
8	$\beta_{Y,Y}$	1.759	19	$\beta_{TOC1,LHY}$	1.363	2
9	$\beta_{LHY,TOC1}$	1.688	2	$\beta_{Y,Y}$	1.223	19
10	$\beta_{Y,LHY}$	1.493	12	$\beta_{LHY,Y}$	1.153	16
11	$\beta_{X,Y}$	1.277	3	$\beta_{Y,X}$	1.119	14
12	$\beta_{X,PRR7}$	1.249	15	$\beta_{PRR7,X}$	1.106	15
13	$\beta_{LHY,X}$	1.185	16	$\beta_{X,LHY}$	0.962	3

not found in our proposed model. These three coefficients are represented as TOC1 $\rightarrow$ PRR7, TOC1 $\rightarrow$ X and X $\rightarrow$ TOC1. Therefore, as we highlighted in the results,the tuning parameters selection, the dynamic Bayesian network with measurement error model did not select the coefficients precisely. Thus, we can summarise and compare our results with the linear Gaussian model. Furthermore, we conclude that our proposed models are less efficient than dynamic Bayesian network model with assumption of linearity of Morrissey et al. (2010), because some of the coefficients which refer the edges between variables that have been hidden and they did not select some important links. Figure 5.19 shows the plot by heat map of the posterior mean of the coefficients matrix of all models that we obtained from MCMC results for based on the above mentioned data. We can conclude that from the Figure 5.19 the results for both proposed shrinkage priors are quite similar , but there is some coefficients are detected by Morrissey et al. (2010) model which are not detected by our proposed shrinkage methods.

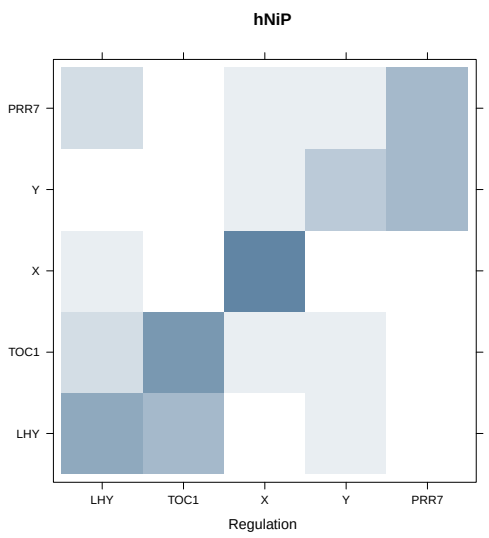


(a)

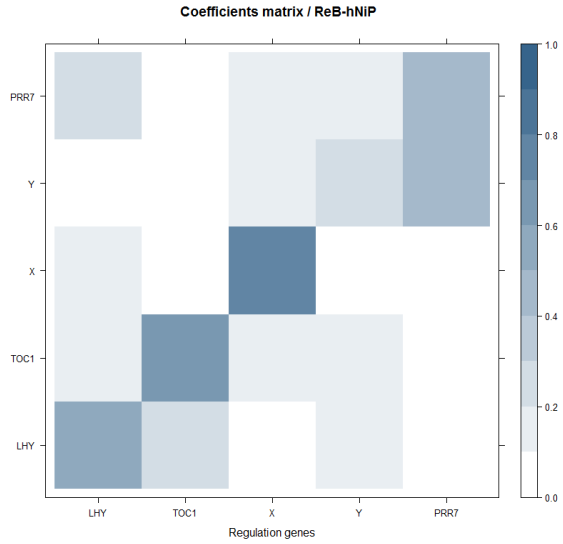


(b)

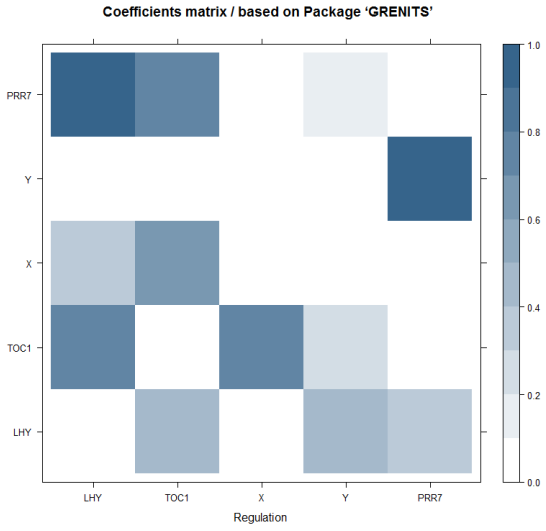
Figure 5.18: Dynamic Bayesian network with measurement error model based on both hNiP and ReB-hNiP prior applied to a simulated dataset (Athaliana ODE 4NoiseReps). Panel (a) is networks based on hNiP and Panel (b) is networks based on ReB-hNiP. Orange Circles are represents nodes (genes) and arrows indicates the interactions between genes. The sample size is $T = 50, p = 5, R = 4$.



(a) hNiP shrinkage prior



(b) ReB-hNiP shrinkage prio



(c) GRENITS package

Figure 5.19: Comparing results between the model described in the GRENITS package as the standard model and our dynamic Bayesian network with measurement error model based on the hNiP prior when applied to the simulated dataset (Athaliana ODE 4NoiseReps). Plot (c) represents results based on the dynamic Bayesian network with linear Gaussian interaction of Morrissey et al. (2010). Plot (a) and plot (b) refers to our dynamic Bayesian network with hNiP prior and ReB-hNiP prior respectively. The dark colors indicate the interaction between two genes with high probability. The sample size of $T = 50, p = 5, R = 4$.

5.4 Summary of Chapter

In this chapter, we implemented our proposed Bayesian shrinkage prior distributions on both the Bayesian linear regression with measurement error model and the dynamic Bayesian network with measurement error model. to concluded, in the first part of this chapter, both proposed Bayesian shrinkage prior distributions were satisfactorily utilised to fit the Bayesian linear regression with measurement error model, therefore tuning parameters are selected in order to shrink the parameters toward the origin values depending on the adopted WAIC in cases of having measurement error. In addition, we have summarised an MCMC chain for a hundred samples of the posterior distribution to present the performance and compare the prior distributions that are used as part of the modelling. In the second part, we have utilised both hNiP and ReB-hNiP as prior distribution of coefficients in Dynamic Bayesian network with measurements error model. This model is based on AR(1) model due-to having times series data. Also, WAIC is applied to select the best range of hyperparameter values. The results were unsatisfactory to detect the best range values of tuning parameters and sufficient number of non-zeros corresponding to the simulated data. This is because most of the autoregressive coefficients aggressively shrunk to zeros in case of non-high dimensional data, and all coefficients aggressively become zero when we have high dimensional data. In the next chapter, we will implemented dynamic Bayesian network on real gene expression data relies on some hyperparameter values for both proposed shrinkage priors.

Chapter 6

An Application of Bayesian Dynamic Network to Genetic Data

6.1 Introduction

Over the last decades, numerous methods have been advanced that facilitated the investigation of the interactions among covariates (genes). Particularly, the Gene Regulatory Network (GRN) model that has been explained in Section 5.1, is one of the popular approachs that deals with genetic data. Therefore, Dynamic Bayesian Networks (DBN) is a type of GRN model. We consider a special structure of modelling which is combines DBN with measurement error model, that was highlighted in Section 5.3.1, in conjunction with our proposed shrinkage priors (explained in Chapter 3) which are used to control the high-dimensionality and leads to the estimation of a sparse coefficients matrix. In this chapter, the methodology mentioned previously is applied to a real gene expression cardiovascular dataset. Thus, this real dataset presents the gene expression levels for patients that are suffering from cardiovascular diseases over several time points. We explain the source of the dataset in the following sections. In Section 1.1, we gave a brief introduction to both cardiovascular diseases CVD and genetics. Hence, in Section 6.3, we explain the source of our real data and how we chose a sub-sample of those genes that related to one kind of CVD. Section 6.2, we review a few researches that applied statistical

methodologies to CVD data. Our goal is twofold; firstly, we select those genes that have impact on CVD, that is, detect some gene that regulates others genes based on time-crosses data. Secondly, we aim to understand the ability of the dynamic Bayesian networks with measurement error model to fit this type of data.

6.2 CVD and Statistical Methods

In this section, we give a brief survey of some studies that are related to cardiovascular diseases using statistical methodologies. The datasets used in the following studies including genetic and non-genetic data. Five different statistical approaches were utilised for distinguishing the set of genes that might cause human heart failure (Huang et al., 2005). These approaches are nearest shrunken, penalized partial least squares, random forest, LASSO and partial least squares. The authors argued that it is difficult to distinguish different factors that cause human heart failure in a single dataset of gene expression. These methods were used on two different real genetic data sources when collecting tissues from a number of patients. The first tissue samples were collected from the University of Minnesota / USA during the placement of a left ventricular assist device (LVAD). Approximately $22,000$ genes were measured from 30 patients. With type of human genome the Affymetrix (U133A) chip was used for analyzing data. The second genetic tissue data was collected from the Harvard medical school and the Affymetrix (HG U-133 plus 2) chips were utilised to analyze data which involved $5,4678$ probe set where $22,277$ probe sets were used in this study for 36 patients under CVD. Huang et al. (2005) reported that all approaches named above yield similar performance results for each simulation dataset used. However, they have obtained two different results for each real dataset; the list of genes obtained from the first dataset is not the same genes as the second one.

Wilson et al. (1998) used logistic regression and Cox-proportional methods to build a model for detecting the correlation between cholesterol, blood pressure and coronary heart disease (CHD). In addition, the investigation of the variables had been considered for both categorical and continuous variables. The variables studied in this article include: LDL-cholesterol, age, NCEP cholesterol, diabetes, blood pressure and smoking. The proportional hazards regressions model is used for analysing data. The results found that every factor has an impact on CHD differently

for both genders when categorical variables or continuous variables are used. For example, 28% of men and 29% of women who had CHD were diagnosed with blood pressure which is correlated with CHD risk factors. Ghosh and Marco (1999) studied probability Bayesian Networks based on the epidemiology data for patients with Heart disease. 17 different variables have been used for heart disease, 13 variables represent predecessors and the other four variables represent children. They claimed that some of the parent variables did not have any direct effect on heart disease. Nevertheless, they have an influence on the particular nodes for their parents. For instance, heart disease is affected indirectly by some risk factors even if they have high-blood-pressure. Therefore, the main drawbacks for this model, is flat they did not take into account some important factors, such as environment, age, genetic, and some others. Marshall et al. (2010) applied DBNs to predict the survival of patients who suffered from a type of CVDs called Ischaemic Heart Disease. They used a package called $''deal''$, which is built under R software. The data is not genetic, but there are metabolic syndrome risk factors. Five variables found from fourteen were highlighted in networks that appear to affect survival including: smoking, diastolic blood pressure, and BMI both at 5 year follow-up, and "inchihd10", indicating whether a patient dies with IHD at 10 years. Furthermore, they noticed that the result of systolic blood pressure was conditionally independent with survival given diastolic blood pressure.

6.3 Sources of Data

In this section, we provide with a general explanation of the type and source of our real dataset. Different types of data play an important role in statistical analysis, as well as, the kind of methodology applied, especially in the medical filed. One of the most interesting types of medical data which researchers tend to use in this field of statistical analysis is genetic data, especially gene expression levels. This is because this information can help detecting those genes cause diseases which leads to the production of better the treatments. Gene expression is the transcription levels of the DNA which is measured by microarrays. It can be obtained from the level of protein sequences while protein synthesis occurs particularly when the mRNA translates to protein (Bani, 2009). Repeated measurements are taken in some studies by researchers. Thus, the repeated measurements generate data when the measurements are taken from each of exper-

imental unit at each one of several time point. We obtained a sample of microarray data (gene expression levels) from the royal Hallamshire Hospital, Sheffield, UK. The data was taken from the blood of patients who had pain in their chest when they attended hospital. The patients with cardiac symptoms had been diagnosed were selected and then permission was obtained from the patients by signing a written consent form. The form requested the patients to take a blood sample for scientific studies. The samples were taken on five occasions $(1^{st}$ day, 3^{rd} day, 7^{th} day, 30^{th} day and 90^{th}day$)$. On the first event $(1^{st}$ day$)$, the blood was taken from 33 patients and the summary of the other days data are presented in Table 6.1. Pearson et al. (2009) have developed an uncertainty propagation method and created an R package which is called PUMA. They take into consideration the speed and scope in terms of high implementation compared to the previous uncertainty approaches for analysing Affymetrix GeneChip data. The real data were processed by a researcher whose applied the PUMA package (Propagating Uncertainty in Microarray Analysis). Pearson et al. (2009) argued that point estimation is mostly used to analyses microarray data (gene expression) and uncertainty in such estimation methods is ignored when analysing data.

T = Day	R**=Number of patient**	**Probe-sets**
1^{st} day	33	31343
3^{rd} day	25	30415
7^{th} day	32	30006
30^{th} day	32	30888
90^{th} day	31	29801

Table 6.1: A summary of the gene expression levels for 33 patients who are suffering from CVD. The number of replicates, the number of probe sets and **T** the time sampling.

As we highlighted, the data are collected on five separate occasions with a number of patients so that we have repeated measurement data. There are two main issues that have an impact on the dataset. Firstly, some patients were unable to attend some of the scheduled appointments leading to missing data. This was considered a major issue, particularly from the second occasion till the last one. Consequently, this issue resulted in missing values for patients at different times (patients are considered the replication in our study). Secondly, the number of probe sets was mismatched to the genechip at more than one occasion, and this makes the dataset comprise missing values in terms of the covariates (probe-sets). This issue results in measuring the Affymetrix arrays of the probe-sets during the data collection. In order to deal

with the problem of missing values, we try to clean the data by removing all missing replication (which in this dataset refers to patients) and probe-sets. After that, we combined the data over all times (where gene expression levels are collected at 5 different times as we explained above) because we handle missing values in this thesis. The missing values make the computation of any model expensive and very difficult, particularly with dynamic networks and high dimensional models. For that reason, we decided to clean our dataset via two main stages from missing values. In the first stage, we removed the missing values in terms of the replication, which resulted in removing 12 of the 33 patients, that is, columns of the dataset. In the second stage, we removed the rows of the dataset (probe-sets) that involve missing values, and then we combined all data files in terms of time. Therefore, as a result we have 24958 probe-sets out of 36085 excluding any missing values for the 21 patients (replication) and 5 different times. The data still consists of a large amount of variables and it would result in high computation in terms of length, feasibility and cost of implementing the model. Therefore, in the next section, we explain selecting a subsample from the complete dataset.

6.3.1 Sample of Real CVD Data

The computation of applying the dynamic Bayesian network model to a high-dimensional dataset takes a long time, particularly, when dealing with a very large dataset for different time-cross points. The reason for choosing a sub-sample is because it is time consuming. If we did not chose a sub sample, it takes a very long time to compute a model and address DBN structures based on those genes that are related to this kind of disease. Then we will apply it to all genes in the dataset if we obtained satisfactory results. Consequently, a subsample of probe-sets has been chosen from our dataset to detect some genes that have an impact on CVDs . We followed on the literature that studies such genes, for example Morris et al. (2016), Seo et al. (2006) and Shiffman and Porter (2000) in doing that. We have chosen sample probe-sets from annotation of the HG-U133plus2 array. The probe-sets have been selected from the annotation file, especially using the column called Gene Ontology Biological Process which includes information about most probe-sets and indicates probe-sets associated to the words heart and cardiovascular. Then, a sample of $1,193$ probe-sets was selected in annotation bank file. In the first stage, we kept only three main columns in order to investigate which probe-sets belong to heart and CVD, which includes: Probe-set," Gene Symbol, and Gene Ontology Biological

Process then we removed Gene Ontology Biological Process"column. In the second main crucial stage, we combined the sample annotation with our dataset of real gene expression level which mentioned previously so as to reveal how many probe-sets match in our dataset with those probe-sets id that chosen in the first stage. In this stage, we selected 758 probe-sets of real gene expression data for all times either some of probe-sets values missing or not. The number of probe-sets are different from one time to another which is clear from Table 6.2. For example, in the second time slice dataset we only have 645 out of 758 probe-sets. After that, we selected the maximum value for each unique gene for each patient ID overall times, because for some genes we have more than one probe-sets. For instance, the data collected at fifth time of collecting data, we have only 319 out of 370 gene (see Table 6.2). Overall number of the genes recognised are equal to 370 that includes missing values for one time or more (as described in our dataset in the previous section), and also the number of genes are different for each times due to some gene mismeasuring in one or more time points. In the final stage, the number of genes chosen as sub-sample of CVD gene expressions equal to 274, after we have combined all 5 data file and removed the missing values (NA) for all observation over the time.

Table 6.2: Summary of the procedure selecting sub-sample of CVD gene expression level from real dataset,where T_1, T_2, T_3, T_4 and T_5, represents the dataset in five different time collected.

	Notes	T_1	T_2	T_3	T_4	T_5
First stage:	CVD probe-sets from Annotation HU G-133-PULES 2					
Second stage	Selecting probe-set for every time	657	645	623	658	629
	Genes	370	370	370	370	370
Third stage	Selecting genes (includes missing values)	323	323	312	327	319
Forth stage	CVD gene expression levels (without missing values)	274	274	274	274	274

6.4 Application on Dynamic Bayesian Networks Model

We have implemented our dynamic Bayesian network with measurement error models in Section 5.3.2 on the sub-sample of real gene expression data as presented in Section 6.3.1. The coefficient matrix has relied on both class hNiP and ReB-hNiP shrinkage prior distributions, which the all posterior distributions are mathematically described in Section 5.3.4 and 5.3.6. As

a reminder, our sub-sample real dataset refers to gene expression levels for 21 replication (R) data (CVD patients), 274 genes (covariates) that collected over five different time points $T = 1, 2, 3, 4, 5$. Four MCMC chains run for 5000 iterations and we burn-in 1250 iterations, for each proposed shrinkage prior distributions, two set of appropriate prior values are used in order to fit the models. The prior values for both parameters precision AR(1) model and precision measurements error models are $a_\tau, c_\lambda = 2$ and $b_\tau, d_\lambda = 0.62$ respectively. Computing MCMC chain for model based on hNiP prior takes 35 hours and 44 minutes and to model based ReB-hNiP takes 44 hours and 7 minutes.

The procedure of variable selection are done by differing zero and non-zero coefficients values. The quantile 0.1 and quantile 0.90 of posterior values are computed for each coefficients. To handle this situation, we are setting the hyperparameter values of ReB-hNiP as follows: $b_0 = 750, e_0 = 6$ for the first choice; and $b_0 = 250, e_0 = 6$ for the second choice. We have chosen an appropriate hyperparameter values of hNiP based on results of simulated data, which in the first case is $b_0 = 750, c = 6$ and $d = 4$ and in the second case is $b_0 = 750, c = 3$ and $d = 2$. Our goal is to fit the coefficients in DBNs with measurement error model by utilising different prior values of coefficients in analysing the results for the DBNs model via both shrinkage prior distributions. Likewise, all MCMC chain are run, therefore posterior distribution based on the results for CVD sub-sample are conducted. The results for all coefficients in DBNs models were aggressively shrunk toward zero for all different hyperparameters used. Furthermore, there were no any coefficients different from zero for all models that have been implemented.

On the other hand, we used an alternative approach to analyse the results and variable selection which based on computing the t-test for every single coefficients of posterior distribution using both prior distributions (hNiP and ReBhNiP). Table 6.3 present 25 high values of the t-test, for the estimated coefficients after ranking and ordering each $\beta_{i,j}$ in coefficients matrix B in Table 6.3. It can be noticed that the maximum values for some t-test are less than 0.25 and others are very close to zeros which indicates that there are no significant results, that is, there is no links between the nodes. Consequently, both techniques for analysing MCMC chain have similar results because all coefficients are aggressively shrunk toward zeros.

Table 6.3: Top 25 gene interactions by forming t-test, with 95% equally for posterior interval, mean and median. For dynamic Bayesian network with measurement error model based on both ReB-hNiP and hNiP prior distributions. MCMC chain for 5000 iteration. The quantiles for both models are presented under hNiP with $b_0 = 750, c = 3$ and $d = 2$ and ReB-hNiP $b_0 = 250, e_0 = 6$.

	ReB-hNiP prior			**Quantiles**			**hNiP prior**			**Quantiles**		
Rank	β_{ij}	Mean	t-**test**	$q_{0.025}$	$q_{0.5}$	$q_{0.975}$	β_{ij}	Mean	t-**test**	$q_{0.025}$	$q_{0.5}$	$q_{0.975}$
1	$\beta_{235,131}$	0.181	0.212	-0.317	0.006	3.04	$\beta_{136,67}$	0.015	0.125	-0.117	0.005	0.18
2	$\beta_{264,183}$	0.184	0.206	-0.201	0.006	3.30	$\beta_{159,24}$	0.006	0.120	-0.100	0.005	0.13
3	$\beta_{96,131}$	0.187	0.197	-0.215	0.005	3.08	$\beta_{163,36}$	0.013	0.119	-0.122	0.006	0.16
4	$\beta_{36,247}$	0.257	0.196	-0.392	0.004	4.78	$\beta_{47,36}$	0.005	0.118	-0.113	0.006	0.17
5	$\beta_{172,131}$	0.178	0.189	-0.252	0.006	3.01	$\beta_{159,159}$	0.006	0.117	-0.119	0.006	0.18
6	$\beta_{136,183}$	0.167	0.188	-0.209	0.003	3.12	$\beta_{70,24}$	0.005	0.116	-0.123	0.006	0.17
7	$\beta_{268,242}$	-0.198	0.187	-3.494	-0.007	0.24	$\beta_{159,47}$	0.011	0.116	-0.112	0.005	0.17
8	$\beta_{159,23}$	0.158	0.185	-0.269	0.005	2.95	$\beta_{206,264}$	0.012	0.114	-0.112	0.005	0.15
9	$\beta_{172,94}$	0.179	0.184	-0.290	0.002	3.24	$\beta_{240,67}$	0.005	0.113	-0.114	0.006	0.16
10	$\beta_{248,80}$	0.165	0.184	-0.191	0.005	2.98	$\beta_{98,20}$	0.019	0.113	-0.108	0.005	0.19
11	$\beta_{205,139}$	0.179	0.183	-0.215	0.003	3.29	$\beta_{229,67}$	0.011	0.113	-0.113	0.005	0.16
12	$\beta_{211,133}$	-0.166	0.183	-3.028	-0.006	0.20	$\beta_{98,53}$	0.011	0.111	-0.116	0.006	0.16
13	$\beta_{98,177}$	0.275	0.182	-0.249	0.003	5.45	$\beta_{39,199}$	0.015	0.111	-0.131	0.005	0.18
14	$\beta_{209,146}$	0.167	0.182	-0.322	0.004	3.12	$\beta_{167,112}$	0.003	0.111	-0.115	0.005	0.15
15	$\beta_{98,131}$	0.155	0.182	-0.192	0.005	2.95	$\beta_{215,234}$	0.014	0.111	-0.116	0.005	0.17
16	$\beta_{167,139}$	0.183	0.181	-0.464	0.004	3.37	$\beta_{149,136}$	0.007	0.110	-0.115	0.005	0.16
17	$\beta_{98,247}$	0.193	0.181	-0.245	0.005	3.57	$\beta_{175,98}$	0.015	0.110	-0.119	0.005	0.16
18	$\beta_{163,195}$	-0.110	0.177	-1.911	-0.005	0.17	$\beta_{181,224}$	0.012	0.110	-0.119	0.006	0.16
19	$\beta_{256,247}$	0.166	0.177	-0.259	0.003	2.96	$\beta_{268,39}$	0.011	0.110	-0.123	0.004	0.20
20	$\beta_{211,23}$	0.139	0.177	-0.220	0.004	2.80	$\beta_{47,159}$	0.013	0.109	-0.124	0.005	0.17
21	$\beta_{274,192}$	0.144	0.176	-0.173	0.005	2.61	$\beta_{24,152}$	0.002	0.109	-0.116	0.004	0.17
22	$\beta_{70,23}$	0.123	0.175	-0.195	0.007	2.19	$\beta_{202,171}$	0.008	0.109	-0.121	0.005	0.16
23	$\beta_{67,183}$	0.149	0.175	-0.216	0.005	2.52	$\beta_{255,58}$	0.009	0.109	-0.122	0.005	0.16
24	$\beta_{211,131}$	0.152	0.175	-0.283	0.003	2.98	$\beta_{255,235}$	0.007	0.109	-0.116	0.004	0.17
25	$\beta_{107,247}$	0.179	0.174	-0.396	0.006	3.24	$\beta_{50,58}$	0.010	0.108	-0.120	0.004	0.16

Consequently, both proposed shrinkage methods are not convenient to fit the links (coefficients) in our constricting dynamic Bayesian network, particularly for the mixing model which is built under the autoregressive model and measurement error. Figures 6.1 and 6.2 display the graphs of posterior densities and traces of precision error term of model λ and precision measurements error τ. Note that posterior distribution seem skew. We cannot show any graphical network due to suggested procedures providing no links between the nodes (variables).

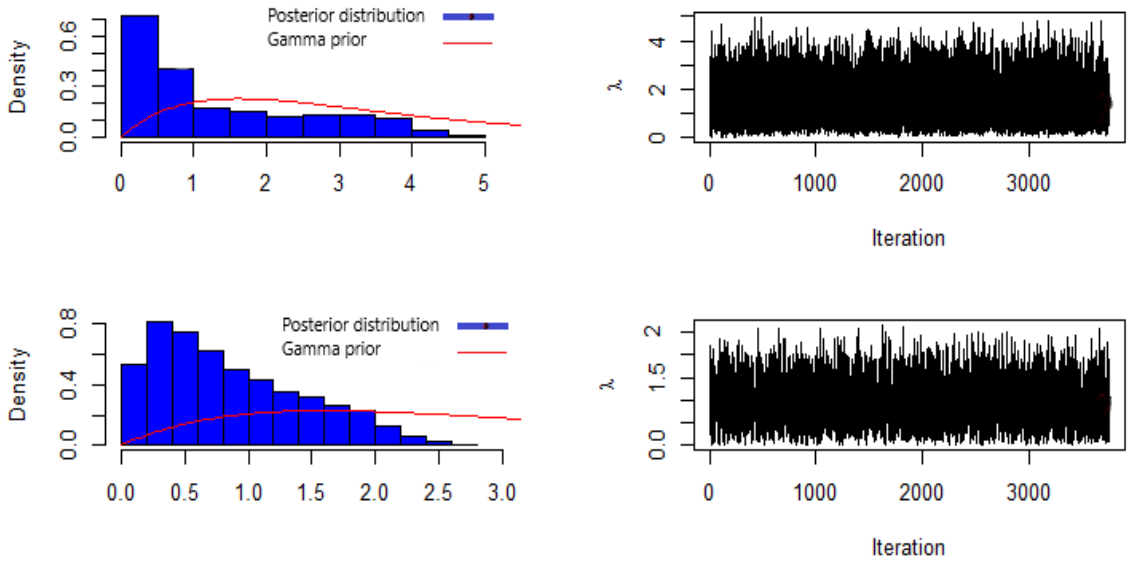


Figure 6.1: Traces and density plot for precision λ of error term in dynamic Bayesian networks. The blue histogram refers to the posterior distribution of its error model based on DBNs. Red lines represents the gamma prior of error model. The above graphs refers to the chain with hNiP hyperparameters values is $b_0 = 750, c = 6$ and $d = 4$. The bottom graphs indicates the results based on hyperparameter values of $b_0 = 750, c = 3$ and $d = 2$.

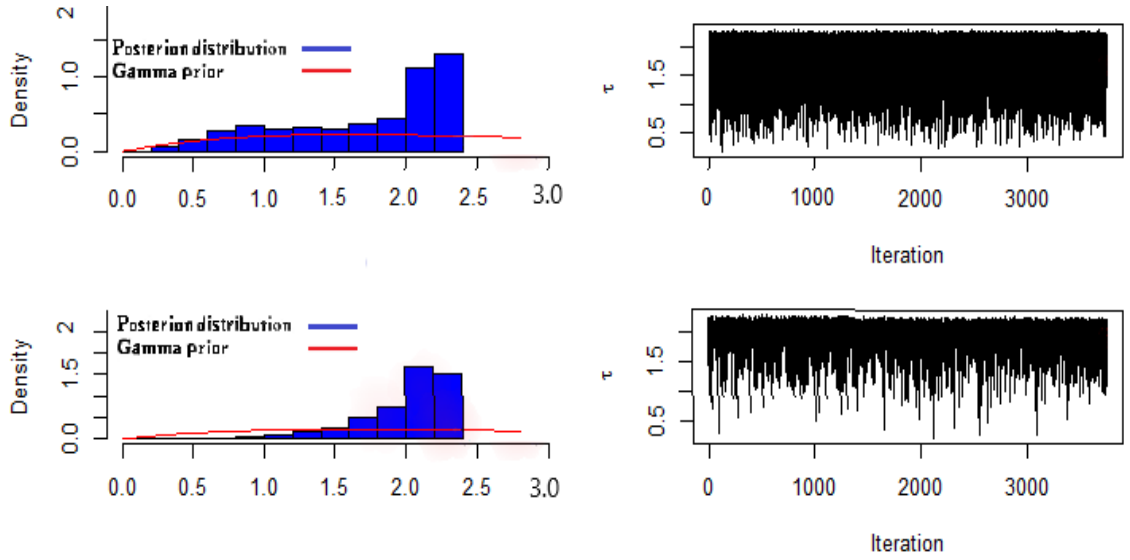


Figure 6.2: Traces and density plot for precision τ of measurements error in dynamic Bayesian networks model. The blue histogram refers to the posterior distribution of the measurement error model based on DBNS, red lines represent the gamma prior of its measurement error. The above graphs refers to the chain with hNiP hyperparameters values is $b_0 = 750, c = 6$ and $d = 4$. The bottom graphs indicates the results based on hyperparameter values of $b_0 = 750, c = 3$ and $d = 2$.

6.5 Summary of Chapter

In this chapter, we have applied our dynamic Bayesian network with measurement error models using a sample of gene expression level dataset taken form cardiovascular diseases patients. We can conclude that both proposed shrinkage priors are not useful to create DBNs under our structure model in this thesis. One might address this kind of modelling with an alternative algorithm instead of MCMC, because the results show that for all coefficients shrunk to zero or modifying the proposed Bayesian shrinkage priors to avoid shrinkage aggressively all coefficients. Therefore, this kind of modelling is no helpful for shrinkage parameters in the high-dimensional dataset.

Chapter 7

Conclusion and Future work

In this chapter, we summarise the main results obtained from our investigations in this thesis and display our contributions and some considerations for future work to tackle the problems which we have faced in our study.

7.1 Conclusions

The high-dimensional data becomes one of the biggest challenging choice for investigation several kinds of statistical modelling. This is because the number of explanatory variables succeed the number of observations. In order to overcome a high-dimensional issues, statisticians proposed a large number of regularization methods regarding penalizing likelihoods function, with a view to reducing the dimensions and gaining a sparse estimation of model parameters. From a Bayesian perspective, shrinkage prior is a Bayesian regularization method based on scale mixture Normal distribution, which is proposed to shrink the coefficients in a model toward the origin (West, 1987). Consequently, our contributions in this domain of the statistical issue, was to propose two new versions of Bayesian shrinkage prior distributions which belong to the family of scale mixture Normal distribution. Considering that they are absolutely continuous prior distributions, we have aimed to work on different kinds of probability distribution as mixing distribution of the scale Gaussian distribution and also consuming the time

when using MCMC sampling algorithm. In Chapter 3, we presented both hierarchical Normal inverse Pareto prior distribution (hNiP) and rescaled beta hierarchical Normal inverse Pareto prior distribution (ReB-hNiP).

Bayesian shrinkage prior distribution were applied, not only as a penalization function but also as Bayesian variables selection with extra tools to evaluate whether the coefficients differ from zero or not. Besides, we have explained the producer on how hNiP affects prior distribution on coefficients. Moreover, one of the most challenges in regularization method is selecting the tuning parameters for hNiP and ReB-hNiP priors, therefore, we used some statistical calibrations to overcome this challenges by thresholds of the regression coefficients. Furthermore, several types of creations were studied including WAIC, DIC and EBIC. However, only WIAC criterion was applied because it has minimum values compared with other mentioned creations. We evaluated the performance of both shrinkage priors comparing with some Bayesian shrinkage prior distributions, including the Bayesian LASSO (Park and Casella, 2008), the horseshoe prior distribution (Carvalho et al., 2009). We also used some non-Bayesian methods including; LASSO and SCAD based on 100 replications of posterior distributions. Also, we calibrated Normal exponential Gamma distributions (NEG) in Chapter 4 in the same way as the hNiP and the ReB-hNiP shrinkage priors. In the Chapter 4, we utilised the multivariate Bayesian linear regression model as simple and standard model to select the best range of the hyperparameters values. The best range values of tuning parameters were selected for all proposed regularization methods using the creations mentioned above.

As we discussed in previous chapters, variables selection was done by shrinkage priors with t-test as an extra tool or credible interval due to evaluating whether the posterior mean of the regression coefficients values have significantly approximately close to true values or not. The results showed that both of the proposed priors outperformed compared to other Bayesian and non-Bayesian shrinkage methods which we have used in different case scenarios taking into consideration the signal to noise dataset. In addition, results show that values of WAIC creations are smaller than DIC and EBIC_γ which means that WAIC creation was better to use. Nevertheless, the performance of the hNiP is satisfactory compared with the ReB-hNiP in terms of time computation of running every single MCMC chains and also use sensitivity and specificity, overall 100 posterior distribution replication samplers.

On the other hand, our contribution in Chapter 5 was to use the proposed Bayesian shrinkage prior distributions with more complicated model structure. We separated that into two kinds of modelling: measurement error with multivariate linear regression model and dynamic Bayesian networks with measurement error. It is clear that the main part for constricting these two kinds of modelling is an additive error equation which is measurements error (ME) model. Measurements error occurs typically in the explanatory variables and the observation of a single or more variables that is not measured accurately by a tool. Ignoring such types of measurement errors in replication and covariates variables leads to insufficient coefficients estimation. Also, it might cause the loss of power for detecting the explanatory variables. The first part of Chapter 5 was focused on measurement error with multivariate linear regression model. Similar to Chapter 4, hyperparameters values were selected based on different cases scenarios of dataset, based on proposed Bayesian shrinkage prior distributions and NEG shrinkage priors. In addition, the precision measurement error model we proposed as a scale mixture of normal distribution which is Student-t distribution as an alternative for Gaussian assumption. Moreover, all hyperparameters prior values were selected based on WIAC after thresholding. Consequently, replication of MCMC samplers have been obtained for both the hNiP and the ReB-hNiP models, and the performance of both shrinkage prior distributions showed satisfactory selected variables. Therefore, the outcomes of the model depended on the hNiP and the ReB-hNiP that have approximately same performance, but the MCMC chain has converged faster for the hNiP model compared to the ReB-hNiP. The prediction of the model was checked based on the posterior predictive distribution through $bayeplot$ an package in R language, which based on the paper of Gelman and Shalizi (2013). Thus, the results show that the both linear regression model with/ without measurement error model are working well.

In the second part, we presented the dynamic Bayesian networks with measurement error model. Particularly, we used our proposed shrinkage priors to create DBN model based on linear Gaussian autoregressive model AR(1) and measurement error model. This type of modelling is important in the case of studying gene regulations networks, because the coefficient matrix of AR(1) represents the interaction between one gene to other genes over several time points, especially the high-dimensional data dominated by gene expression levels in this kind of modelling. Another benefit of investigating this kind of model by detecting suitable treatments of diverse diseases by selecting the most variables (genes) to other genes. Typically, results of such models are represented by graphical models. Therefore, the structure of this complex model was de-

pending on the hNiP and the ReB-hNiP prior distributions assuming that both error terms have Gaussian distributions. Likewise, Chapter 5, the adopted WAIC criterion was used in order to select the best range of hyperparameters values for both proposed Bayesian shrinkage priors. In cases of non-high-dimensional data, in which having enough data that support fitting the model and relying on simulated data, results indicate that only some non-zeros coefficients can be estimated and converged to true values and other coefficients aggressively shrink to zeros (over shrinkage). Thus, models based on both shrinkage priors have similar performance. We also implemneted our MCMC chains for 20 replication samplers and results shows that both t-test and quantiles have the same results. On the other hand, regarding a high-dimensional dataset, in order to select the hyperparameters values for both of our proposed prior and to check the performance of this model, we applied DBN on different case scenarios. Hence, we quantified some measures to evaluate model fitting, such as WAIC, FN, TP and calculating the non-zeros. The findings showed that most of the coefficients shrunk toward zeros aggressively. Thus, if we use some hyperparameter values, we might obtain a few non-zero coefficients, but still the FN values are high. We believe that estimating a few coefficients in the model with FN equal to zero is better than detecting some spurious variables, because indeed, spurious variables may have drawbacks on the disease treatment. We have compared the performance of our modelling with similar structure modelling which is described in R and is called $GRENITS$ package developed by Morrissey (2012). Furthermore, "AthalianaODE-data" was used as standard dataset. Results based on the linear Gaussian model in $GRENITS$ package show out-performance of our results, because our proposed models are detected only by some linkage between the variables.

Finally, in Chapter 6, we applied the sub-sample of the real CVD dataset on our proposed dynamic Bayesian models. We used two different appropriate prior values to show the ability of models based on a specific dataset, the number of coefficients were extremely higher compared to the time points. We conducted the outcome of the MCMC chain, therefore results show that all the coefficients shrink closely to zero. This situation was expected, due to the results of simulated data in the second part of Chapter 5. Above all, the proposed Bayesian shrinkage prior has showed unsatisfactory results based on dynamic Bayesian networks with the measurements error model.

7.2 Future Work

Our research was mainly focused on investigating the problem of high-dimensional data, typically, we were interested in working with various types of linear models. For that purpose, two Bayesian shrinkage prior distributions were proposed. Furthermore, the selection of the tuning parameters and variables were also addressed by sensitivity analysis and coefficients threshold. However, there are some other possible directions to be investigated in the future. The most important task which can possibly be worked on using empirical Bayes (Efron et al., 2001) procedure for selecting the hyperparameters for both shrinkage priors instead of using sensitivity analysis. This is because it will automatically select the hyperparameters based on data, but challenging task is that we have two hyperparameters for each proposed shrinkage prior (c and b_0).

Another possible working direction, is solving the shortage of high-dimensional in dynamic Bayesian networks with the measurements error model. This was which has shown in results in Chapter 5 and the application on real gene expression data in Chapter 6. In this situation, one might explore transformations to linearise the relationship between variables. This is redesign the structure of the model which may lead to better results compared to the current model. Another direction might be to overcome the issues of over-shrinkage of the coefficients, which is proposing the sparse model. Therefore, we would like to handle two types of priors distributions at the same time for the diagonal of coefficients matrix in Equation 5.30, so that, the diagonal has a flatter prior distribution. For example, Gaussian distribution or uniform distribution can be used, in order to keep those diagonal coefficients matrices, $B_{i,j} \neq 0,$ where $i = j,$ because, in reality of gene networks, self-regulation genes usually lead to active genes themselves. For the non-diagonal coefficients matrix, we will use one of our proposed Bayesian shrinkage priors. Furthermore, if the above ideas did not work well, then we will extend the dynamic Bayesian network with measurements error model framework to non-linear functional relationships, even with larger high-dimensional dataset, after modifying the proposed shrinkage priors.

On the other hand, we may also be interested to investigate the setting of the key parameter for inverse Pareto distribution, and use different values of both hyperparameters values a, b for the rescaled Beta distribution, because we assumed that $(a = b = 0.5)$ in Equation 3.12. Furthermore, we will extend and modify the proposed Bayesian shrinkage prior distributions, to

fit the utilised dynamic model that used in this thesis, in order to identifying the best candidate variables (genes) that regulates other genes. Thus, we will use an automatic procedure to set prior parameters using some methods such as, modified Bayesian information creation (mBIC), Bayesian Watanabe-Akaike information creation BWAIC, see Gelman et al. (2014). Also, we would like to implement the proposed shrinkage prior distributions on time-varying dynamic Bayesian networks model. This kind of model is studies the variables selection that may change over time, this will focus on estimating the coefficients for each time slice in the dynamic model. Thus, it is important to identify the coefficients (gene) that differ from zero for every single time of collecting data. Some researchers have worked on the time-varying dynamic Bayesian networks model such as, Kalli and Griffin (2014), Lee et al. (2017) and Song et al. (2009).

Appendix A

A.1 Markov Chain Monte Carlo - MCMC

MCMC algorithm is one of the powerful numerical tools, which plays an important role in many fields such as Statistics, Physics, Computing and Economics. Hastings (1970) has generalised MCMC methods to adopt statistical issues, which is called Metropolis-Hasting algorithm. It is one of the important statistical tools for sampling data from the posterior distribution as in some situations it is difficult to find the full conditional distribution algebraically (Andrieu et al., 2003). A proposal distribution has efficiency in computing MCMC algorithm. It is challenging to use suitable proposal distribution in the case of having the high-dimensional model (Mbalawata et al., 2015). Bayesian Inference is based on this algorithm. In general , we can consider dealing with two main types of MCMC algorithms, Metropolis-Hasting and Gibbs sampler.

A.1.1 Gibbs sampler

Gibbs sampler was first invented by Geman and Geman (1984). It is a simple and special case of MCMC algorithm family, and it relies on the case that can correctly find the full conditional distribution from multivariate joint distribution. Gibbs sampler is appropriate to be utilised the joint probability distribution that has complicated formula and difficult to calculate the full conditional distribution analytically. Gibbs sampler has been applied considerably in literature,

because it gains a sample for target distribution and it is very easy for applying in any statistical software. Suppose that $P(\theta_1, \theta_2, \theta_3, \ldots \theta_p)$ is a joint probability distribution, which we want sampling from for every single parameter, where p represents the number of parameters. Then, we can illustrate the Gibbs sampler for those parameters.

$$\theta_1^{k+1} \sim P(\theta_1|x, \theta_2^k, \theta_3^k, \ldots \theta_p^k)$$

$$\vdots$$

$$\theta_p^{k+1} \sim P(\theta_p|x, \theta_1^{k+1}, \theta_2^{k+1}, \ldots \theta_{p-1}^{k+1})$$

where k is indicated number of the iteration, x is represents of data.

A.1.2 Metropolis-Hastings Algorithm

Metropolis-Hastings algorithm can be used when $P(\theta_i|\theta_j, \text{data}; i \neq j)$ represents the posterior distribution of the unknown parameters, and full conditional distribution is not corresponding to any general probability distribution. In this case, the Gibbs sampler is impossible to use. The choice of selecting the proposal distributions is a very important part in this type of MCMC sampling. We can describe the Metropolis-Hastings through the following steps:

1 Select initial value of $\theta^{(1)}$.

2 Sampling θ^* which is candidate value from a proposal distribution at (k) iteration .

3 Calculate an acceptance ratio (α) :

$$\alpha = min\left\{1, \frac{\Psi(\theta^*)q(\theta^k|\theta^*)}{\Psi(\theta^k)q(\theta^*|\theta^k)}\right\}, \tag{A.1}$$

where $\Psi(\theta^*)$ refers to conditional probability and $q(\theta^*|\theta)$ refer to a proposal distribution.

4 Generate U from Uniform distribution, that is, $U \sim Uni(.|0, 1)$, then we compare U with α. If U is less than α, then $\theta^{k+1} = \theta^*$, otherwise $\theta^{k+1} = \theta^k$.

5 We repeat the steps (2 - 4)for $(k + 1)$ times.

Appendix B

Results of Tuning Parameters Selection and Application on Simulated Data

In the section, we presents some graphs of tuning parameters selection for all candidate hyper-parameters values used with multivariate linear regression.

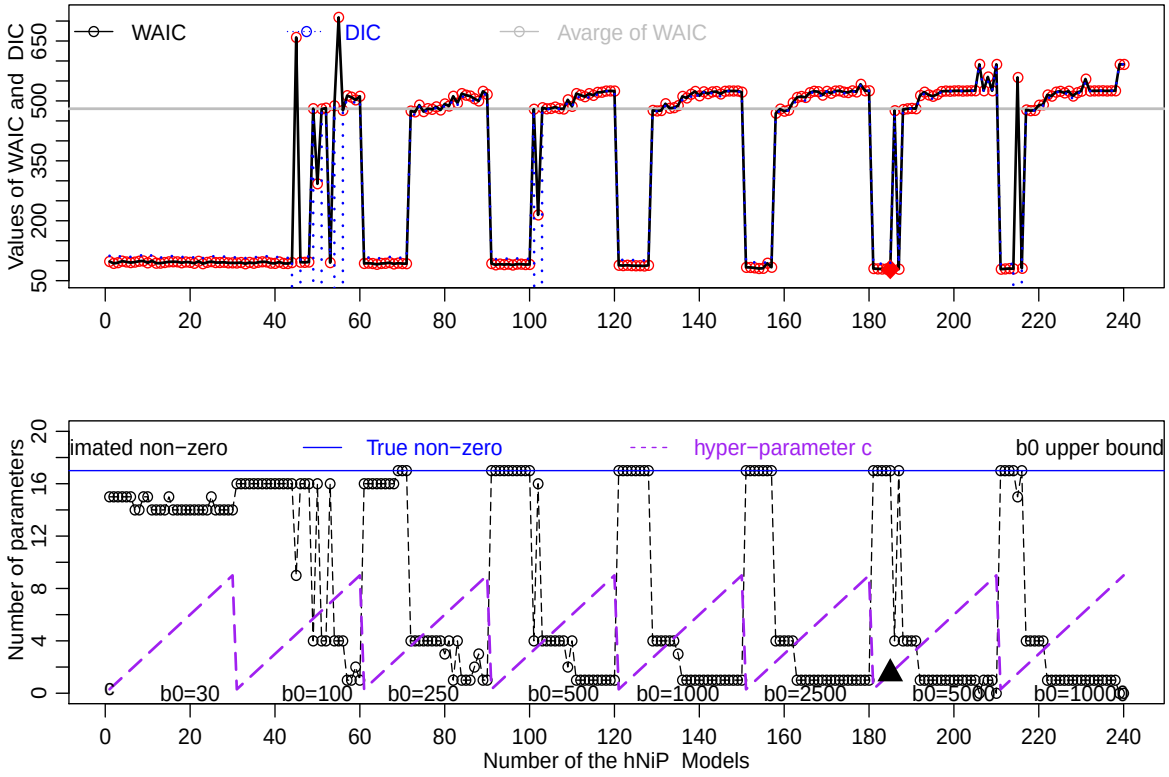


Figure B.1: The relation between upper bounds of hNiP shrinkage prior, hyperparameters and (WAIC and DIC) criteria for all hyperparameters. Top panel shows the values of DIC criteria which illustrated by the dotted blue line and WAIC is represented by a black solid with dotted red line. The grey solid line refers to the average of WAIC for the best number of candidate values of the hyperparameters. The bottom panel displays three lines where the blue solid line indicates the true number of non-zero coefficients in the simulation data of $n = 35, p = 45$. Black dotted dash line represents the number of estimated coefficients. The purple dash line represents used values used for upper-bounds b_0 and for all hyperparameters candidates values.

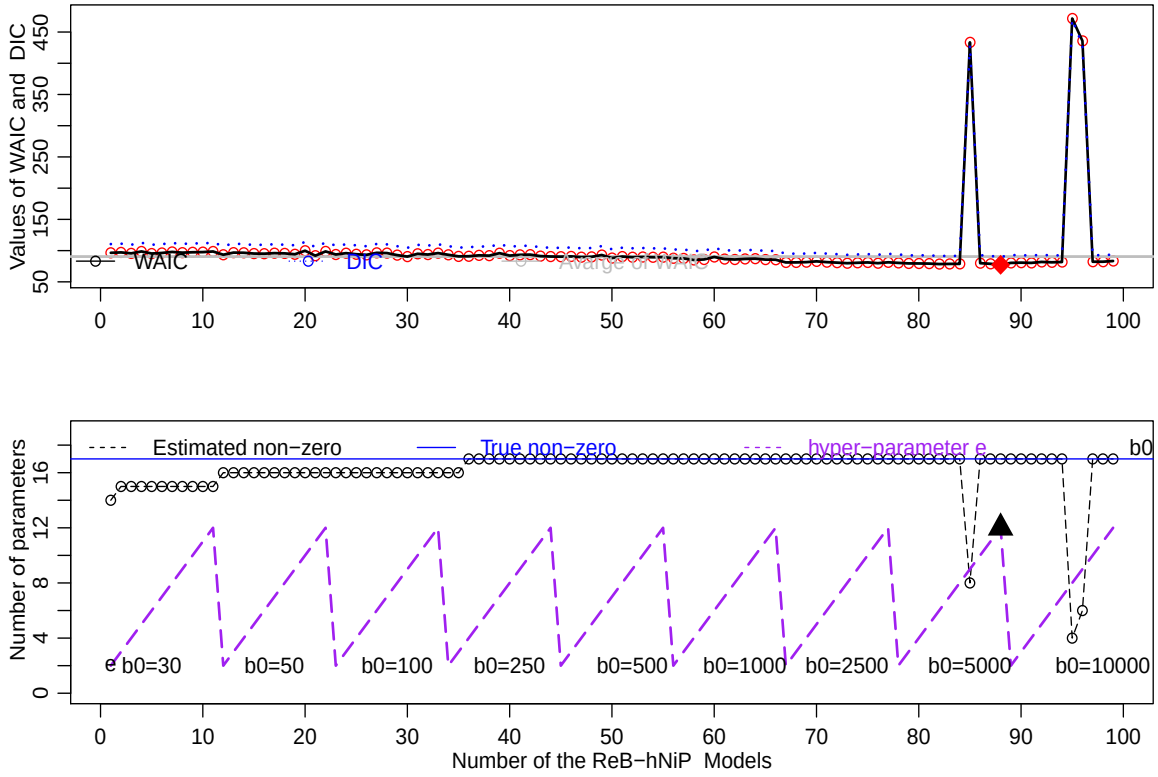


Figure B.2: The relation between upper bounds of ReB-hNiP shrinkage prior, hyperparameters and (WAIC and DIC) criteria for all hyperparameters. Top panel shows the values of DIC criteria which illustrated by the dotted blue line and WAIC is represented by a black solid with dotted red line. The grey solid line refers to the average of WAIC for the best number of candidate values of the hyperparameters. The bottom panel displays three lines where the blue solid line indicates the true number of non-zero coefficients in the simulation data of $n = 35, p = 45$. Black dotted dash line represents the number of estimated coefficients. The purple dash line represents used the values used for upper-bounds b_0 and for all hyperparameters.

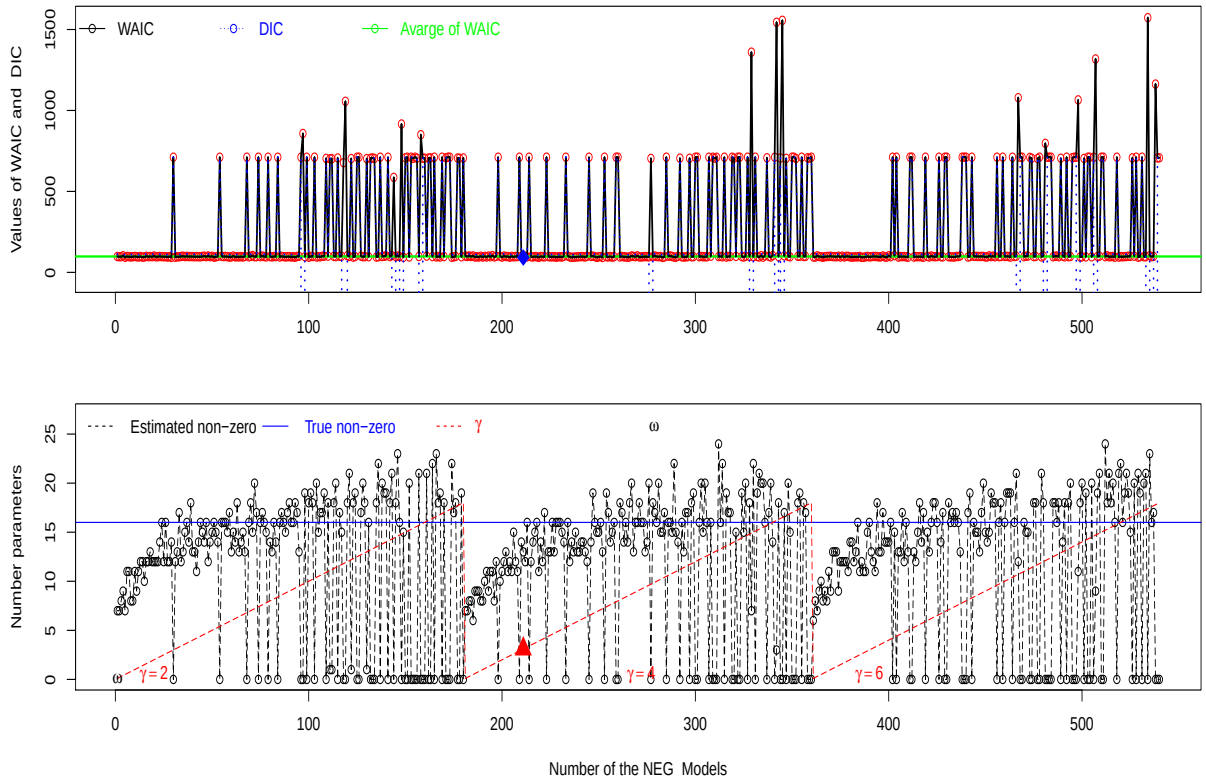


Figure B.3: The relation of the hyperparameter calibration between upper bounds and second-layer hyperparameters using (WAIC and DIC) criteria for model based on NEG shrinkage prior, for all candidate values. The blue dotted line represents the DIC, and the black line with red circle illustrates the WAIC values for the number of the models. The green solid line refers to the average value of the WAIC. In the bottom panel, the blue solid line indicates the number of true non-zero coefficients. The black dashed line with black circle represents the number of estimated regression coefficients. Purple dashed line denotes the several values of candidate upper bounds b_0 and the all range values of hyperparameters for NEG shrinkage prior with ω and γ. The dataset had a size of $n = 35$ and $p = 45$.

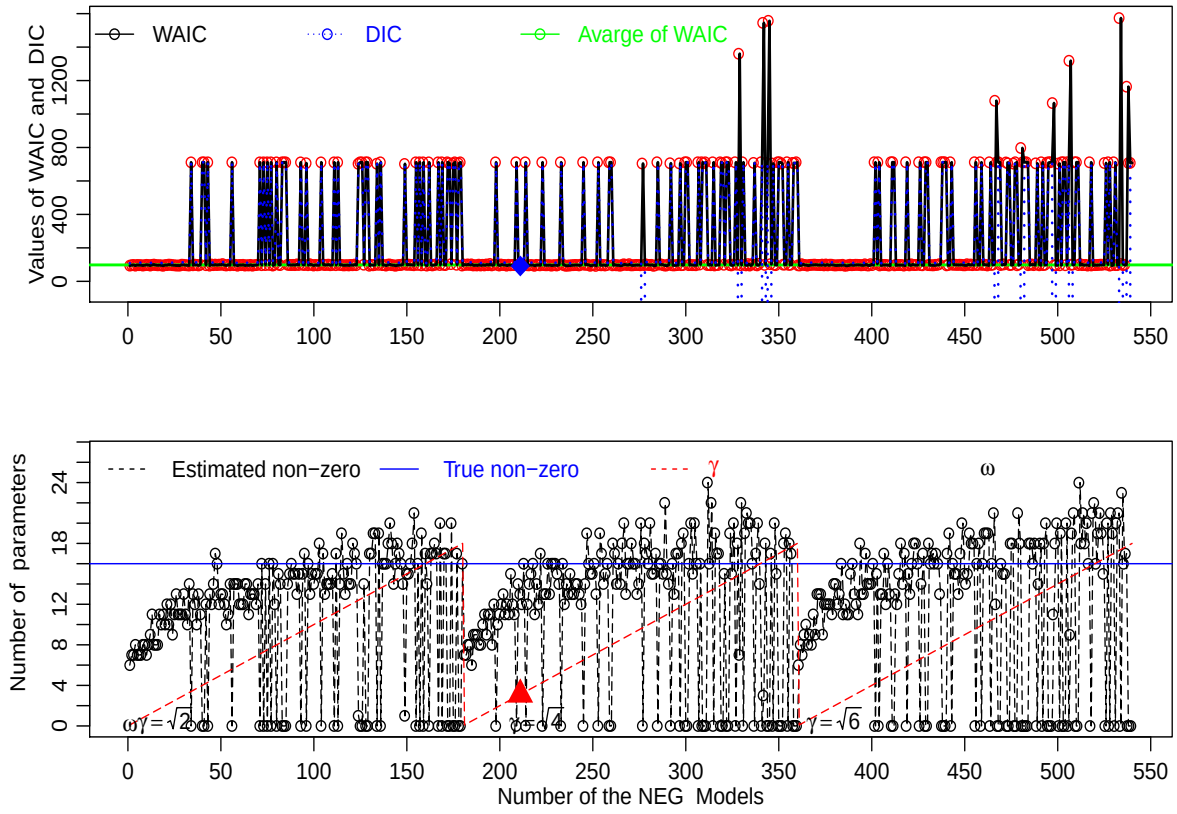


Figure B.4: The relation of the hyperparameter calibration between upper bounds and second-layer hyperparameters using (WAIC and DIC) criteria for model based on NEG shrinkage prior, for all candidate values. The blue dotted line represents the DIC, and the black line with red circle illustrates the WAIC values for the number of the models. The green solid line refers to the average value of the WAIC. In the bottom panel, the blue solid line indicates the number of true non-zero coefficients. The black dashed line with black circle represents the number of estimated regression coefficients. Purple dashed line denotes the several values of candidate upper bounds b_0 and the all range values of hyperparameters for NEG shrinkage prior with ω and γ. The dataset had a size of $n = 35$ and $p = 100$.

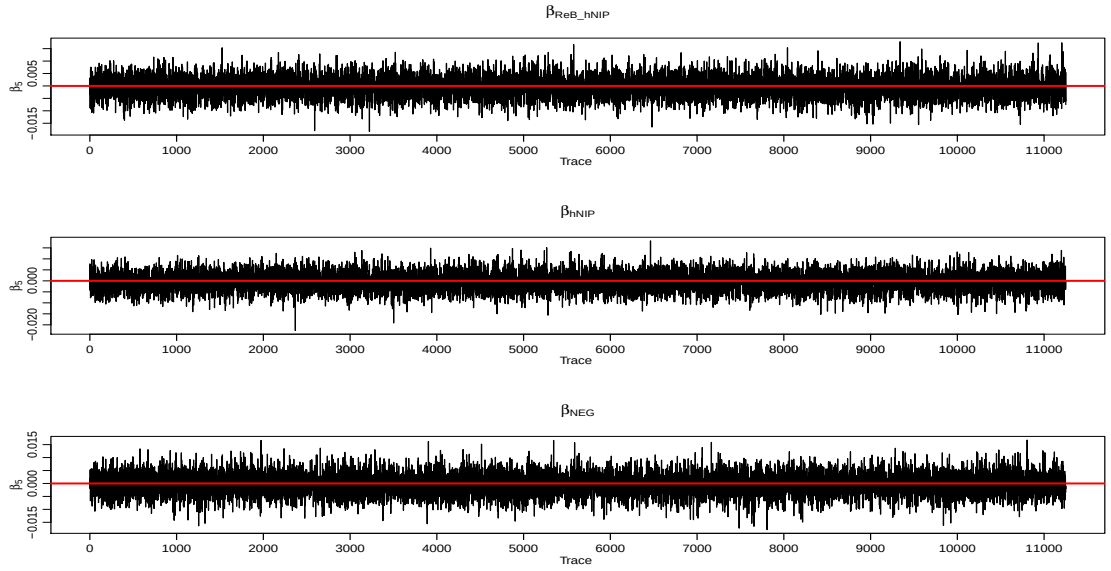


Figure B.5: Trace plots for non-zero coefficient parameter β_5 based on different shrinkage priors: the NEG with $\omega = 2$ and $\gamma = \sqrt{2}$, ReB-hNiP $b_0 = 30, e_0 = 6$, and hNiP with $b_0 = 30, c = 9$ and $d = 6$; where $n = 80$ and $p = 35$. The red line represent the true value of coefficient and the black traces refer to the posterior distribution of the estimated coefficient.

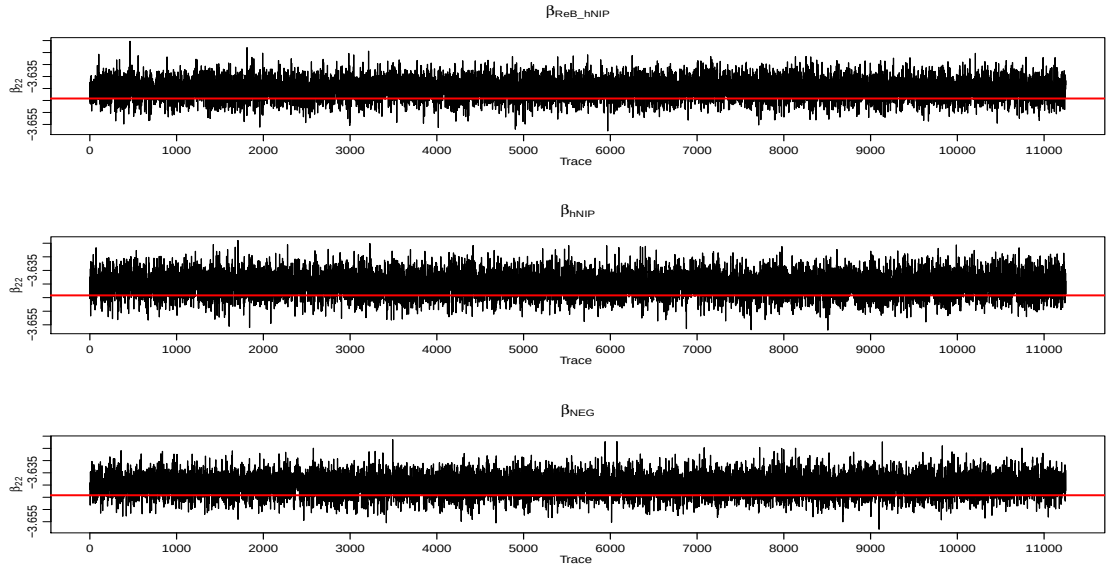


Figure B.6: Trace plots for non-zero coefficient parameter β_{22} based on different shrinkage priors: the NEG with $\omega = 2$ and $\gamma = \sqrt{2}$, the ReB-hNiP with $b_0 = 30, e_0 = 6$, and the hNiP with $b_0 = 30, c = 9$ and $d = 6$; where $n = 80$ and $p = 35$. The red line represent the true value of coefficient and the black traces refer to the posterior distribution of the estimated coefficient.

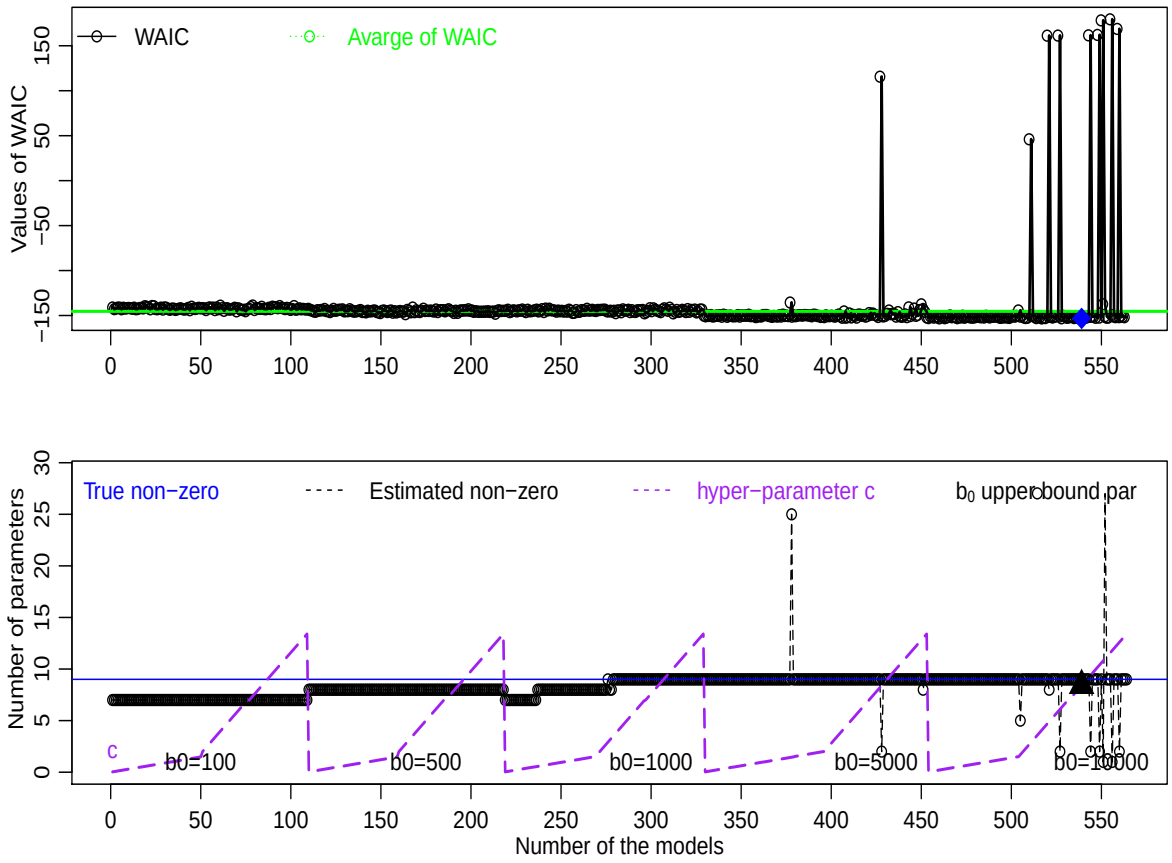


Figure B.7: Tuning parameters selection for hNiP, where the top panel shows two different lies, where WAIC is represented by the black line with red dotted, and the green solid line refer to the average of WAIC. The bottom panel displays three lines, where blue solid line indicates the number of non-zero coefficients in the simulation data when $n = 30$ and $p = 40$. The black circle dotted line represents the number of estimated coefficients. The blue dashed line represents the number of different values of upper-bounds b_0 and parallel with hyperparameters c, d. Both graphs shows the relation between upper bounds of hNiP shrinkage prior, hyperparameters and the best range of WAIC criteria.

Appendix C

Pseudocode

In this sections, we present the pseudocode for all models used in this thesis.

C.1 Pseudocode Bayesian Linear Regression Model

- Input initial values and prior values $\{y, X, a_\lambda, b_\lambda, (\omega, \gamma), \text{or}(b_0, c, d), \text{or}(b_0, e)\}$
- Update precision of error model λ^{k+1}

$$\lambda^{k+1} \sim \text{Ga}\left(.\middle| a_\lambda + \frac{n}{2}, b_\lambda + \frac{1}{2}(\boldsymbol{y} - X\boldsymbol{\beta^k})' I(\boldsymbol{y} - X\boldsymbol{\beta^k})\right)$$

- Update the coefficients $\boldsymbol{\beta^{k+1}}$ using the shrinkage priors as follows;
  - For NEG prior
    - Update hyperparameter ζ

$$\zeta_j^{k+1} \sim \text{Ga}\left(.\middle| \omega + 1, (\phi_j^k + \frac{1}{\gamma^2})\right)$$

* For hyperparameter ϕ, use Metropolis-hasting;

$$\prod_{j=1}^{p+1}\pi(\phi_j^{(k+1)}) \propto \left(\prod_{j=1}^{p+1}\phi_j^{-\frac{1}{2}(k)}\exp\left\{-\left(\frac{\beta_j^{\prime(k)}\beta_j^{(k)}}{2\phi_j^{(k)}}+\zeta_j^{(k+1)}\phi_j^{(k)}\right)\right\}\right),$$

* Update coefficients parameter $\boldsymbol{\beta}$

$$\boldsymbol{\beta^{*(k+1)}} \sim N\big(.\big|\mu_{\boldsymbol{\beta}}, \Omega_{\boldsymbol{\beta}}^{-1}\big).$$

Where $\mu_\beta = \lambda^{(k+1)}\Omega_{\boldsymbol{\beta^k}}^{-1}X'\boldsymbol{y}$, $\Omega_{\boldsymbol{\beta^k}} = (\lambda^{(k+1)}X'X + \Sigma_{\Phi^{(k+1)}})$ and $\Sigma_{\Phi^{(k+1)}}$ is a matrix of $(p+1)\times(p+1)$ dimensional, and the diagonal of this matrix is equal to $\{\frac{1}{\phi_1^{(k+1)}},\ldots,\frac{1}{\phi_{p+1}^{(k+1)}}\}$.

– Update hNiP

– Update For ReB-hNiP

* Update hyperparameter κ^{k+1}
$\pi(\kappa_j) \propto b_0^{\kappa_i}\psi_j^{\kappa_i-1}\frac{e_0^{-1}}{B(a,b)}\left(\frac{\kappa_j}{e_0}\right)^{a-1}\left(1-\frac{\kappa_j}{e_0}\right)^{b-1}$,
where $j = 1, 2\ldots(p+1)$, the values of κ_j should satisfy the condition $(0 < \kappa_j \leq e_0)$ and $a = b = \frac{1}{2}$. The formula in (4.11) has no typical distribution. Therefore, Metropolis-Hasting is used in order to update the posterior distribution.

* Update the parameter of scale mixture Normal distribution ψ_j :

$$\psi_j \sim \text{TrGa}\left(.\big|b_0, \kappa_j+\frac{1}{2}, \frac{\beta_j^2}{2}\right), \tag{C.1}$$

where, $TrGa$ is Truncated Gamma distribution which is upper bounded.

* Updating coefficients parameters $\boldsymbol{\beta}$ as follow :

$$\boldsymbol{\beta} \sim N\big(.\big|\mu_\beta, \Omega_\beta^{-1}\big). \tag{C.2}$$

where $\mu_{\boldsymbol{\beta}} = \lambda\Omega_{\boldsymbol{\beta}}^{-1}X'\boldsymbol{y}$, $\Omega_{\boldsymbol{\beta}} = (\lambda X'X + \Sigma_\psi)$ and Σ_ψ is a matrix of $(p+1)\times(p+1)$ dimensional, and the diagonal of this matrix is equal to $\{\psi_1, \psi_2\ldots,\psi_{p+1}\}$.

C.2 Pseudocode for Measurements Error with Bayesian Regression Model

C.2.1 Pseudocode for Measurements Error With Bayesian Regression Model - NEG Prior

- Input data and prior values $\{y, \mathbf{Z}, a_\lambda, b_\lambda, c_\tau, d_\tau, a_\nu, b_\nu, pb_\nu, a_\phi, b_\phi, Pb_\phi, \gamma, \omega\}$
- Start with initial values $\{\lambda^{(1)}, \beta^{(1)}, \tau^{(1)}, \phi^{(1)}, \nu^{(1)}, \delta^{(1)}, \zeta^1\}$
- Update unobserved variable $\boldsymbol{x_i^{(k+1)}}$

$$N\Bigg(.\Bigg|\big(\lambda^{(k)}\boldsymbol{\beta^{(k)}}\boldsymbol{\beta'^{(k)}} + \tau^{(k)}(\sum_{r=1}^{R}\delta_{ijr}^{(k)})I_p\big)^{-1}\big(\lambda^{(k)}\boldsymbol{\beta^{(k)}}(y_i^{(k)} - \alpha^{(k)}) + \tau^{(k)}(\sum_{r=1}^{R}\delta_{ijr}^{(k)}z_{ijr})\big)$$

$$, (\lambda^{(k)}\boldsymbol{\beta^{(k)}}\boldsymbol{\beta'^{(k)}} + \tau^{(k)}(\sum_{r=1}^{R}\delta_{ijr}^{(k)})I_p)\Bigg).$$

- Update parameters $\zeta_j^{(k+1)}$ by

$$\zeta_j^{(k+1)} \sim \text{Ga}\Bigg(.\Bigg|\omega + 1, (\phi_j^{(k)} + \frac{1}{\gamma^2})\Bigg).$$

- Update parameters $\phi_j^{(k+1)}$ using Metropolis-Hastings algorithm.

  The proposal distribution of $\phi^{(prop)}$ is gamma distribution with shape equal to $(1+\phi_j^{(k)} \times pb_\phi)$ and rate equal to (pb_ϕ), where $\phi = (\phi_1, \phi_2 \dots \phi_{p+1})$

$$\pi(\phi|rest) \propto \Bigg(\prod_{j=1}^{p+1}(\phi_j)^{-\frac{1}{2}} \exp\Bigg\{-(\frac{\beta_j^{(k)}\beta_j^{(k)}}{2\phi_j} + \zeta_j^{(k+1)}\phi_j)\Bigg\}\Bigg).$$

  Metropolis-Hastings algorithm follows the following steps:

  current state is $\Bigg(\phi_j^k, \zeta_j^{k+1}, \beta_i^k\Bigg)$, updating ϕ_j using $q(\phi_j|\phi_j^k)$ which the proposal is $\phi_j^{prop} \sim q(\phi_j|\phi_j^k)$.

Then the acceptance ratio is given as follows:

$$\alpha = min\left\{1, \frac{\pi(\phi_j^{prop}|-)q(\phi_j^k|\phi_j^{prop})}{\pi(\phi_j^k|-)q(\phi_j^{prop}|\phi_j^k)}\right\},$$

if $\alpha > U$ then $\phi_j^{k+1} = \phi_j^{prop}$, otherwise $\phi_j^{k+1} = \phi_j^k$, Where U is random uniform value.

- Update coefficient parameters $\mathcal{B}^{(k+1)} = (\alpha^{(k+1)}, \boldsymbol{\beta^{(k+1)}})$,

$$\mathcal{B}^{(k+1)} \sim N\left(.\middle|\lambda^{(k)}\Omega_\beta^{-1^{(k)}}\mathcal{X}'^{(k+1)}\boldsymbol{y}, (\lambda^{(k)}\mathcal{X}'^{(k+1)}\mathcal{X}^{(k+1)} + \Sigma_\Phi^{(k)})^{-1}\right).$$

- Update parameter precision of error model $\lambda^{(k+1)}$.

$$\lambda^{(k+1)} \sim \text{Ga}\left(.\middle|a_\lambda + \frac{n}{2}, b_\lambda + \frac{1}{2}(\boldsymbol{y} - \mathcal{X}^{*^{(k+1)}}\mathcal{B}^{(k+1)})'I(\boldsymbol{y} - \mathcal{X}^{*^{(k+1)}}\mathcal{B}^{(k+1)})\right).$$

- Update parameters precision of measurement error $\tau^{(k+1)}$.

$$\tau^{k+1} \sim \text{Ga}\left(.\middle|c_\tau + \frac{RNp}{2}, d_\tau + \frac{1}{2}\Big(\sum_{r=1}^{R}\sum_{j=1}^{p}\sum_{i=1}^{n}\delta_{ijr}^{(k)}\nu^{(k)}(z_{ijr} - x_{ij}^{(k+1)})^2\Big)\right).$$

- Update the degree of freedom $\nu^{(k+1)}$

  where Metropolis-Hastings algorithm is used for updating $\nu^{(k+1)}$, the full conditional distribution is as follows:

$$\pi(\nu|rest) \propto \left(\frac{(\frac{\nu}{2})^{\frac{Rnp\nu}{2}}}{(\Gamma(\frac{\nu}{2}))^{Rpn}}\Big(\prod_{i=1}^{n}\prod_{j=1}^{p}\prod_{r=1}^{R}(\delta_{ijr}^{(k)})^{\frac{\nu}{2}-1}\Big)\exp\left\{-\frac{\nu}{2}\sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{r=1}^{R}\delta_{ijr}^{(k)}\right\}\right)$$

$$\left(\frac{b_\nu^{a_\nu}}{\Gamma(a_\nu)}(\nu)^{a_\nu}\exp(-\nu b_\nu)\right).$$

  Metropolis-Hastings algorithm follows the following steps:

  The proposal distribution of $\nu^{(prop)}$ is gamma distribution with shape equal to $(1+\nu^{(k)} \times pb)$ and rate equal to (pb_ϕ) . Current state is $\left(\nu^k, \delta_{ijr}^k\right)$, updating ν using $q(\nu|\nu^k)$ which

the proposal is $\nu^{prop} \sim q(\nu|\nu^k$. Then acceptance ratio is given as follows:

$$\alpha = min\left\{1, \frac{\pi(\nu^{prop}|-)q(\nu^k|\nu^{prop})}{\pi(\nu^k|-)q(\nu^{prop}|\nu^k)}\right\}$$

if $\alpha > U$ then $\nu^{k+1} = \nu^{prop}$ otherwise $\nu^{k+1} = \nu^k$, where U is random value from uniform distribution.

- Update parameters $\delta_{ijr}^{(k+1)}$,

$$\delta_{ijr}^{(k+1)} \sim \text{Ga}\left(.\middle|\frac{\nu^{(k+1)}+1}{2}, \frac{1}{2}\big(\nu^{(k+1)} + \tau^{(k+1)}(z_{ijr} - x_{ij}^{(k+1)})^2\big)\right).$$

C.2.2 Pseudo-Code for Model with hNiP Prior for Measurements Error with Bayesian Regression Model

- Input data and prior values $\{y, \mathbf{Z}, a_\lambda, b_\lambda, c_\tau, d_\tau, a_\nu, b_\nu, pb_\nu, b_0, c, d\}$
- Start with initial values $\{\alpha^{(1)}, \lambda^{(1)}, \beta^{(1)}, \tau^{(1)}, \phi^{(1)}, \nu^{(1)}, \delta^{(1)}, \psi^{(1)}, \kappa^1\}$
- Update Unobserved variable $\boldsymbol{x_i^{(k+1)}}$, where $(i = 1, 2, \ldots, p), (j = 1, 2, \ldots, p)$ and $r = 1, 2, \ldots, R$.
$N\Bigg(.\Bigg|(\lambda^{(k)}\boldsymbol{\beta^{(k)}}\boldsymbol{\beta'^{(k)}} + \tau^{(k)}(\sum_{r=1}^{R}\delta_{ijr}^{(k)})I_p)^{-1}\big(\lambda^{(k)}\boldsymbol{\beta^{(k)}}(y_i^{(k)} - \alpha^{(k)}) + \tau^{(k)}(\sum_{r=1}^{R}\delta_{ijr}^{(k)}z_{ijr})\big), (\lambda^{(k)}\boldsymbol{\beta^{(k)}}\boldsymbol{\beta'^{(k)}} + \tau^{(k)}(\sum_{r=1}^{R}\delta_{ijr}^{(k)})I_p)\Bigg)$
- The full conditional posterior distribution for κ is as follow:

$$\kappa_j^{(k+1)} \sim \text{Ga}\left(.\middle|c+1, \log\left(\frac{b_0}{\psi_j^{(k)}}\right) + d\right)$$

- The TrGa is the truncated Gamma distribution which is bounded from upper. The parameters ψ and κ are hyperparameters prior β_j in the case using hNIP distribution. The full conditional posterior distribution for ψ is given by:

$$\psi_j^{(k+1)} \sim \text{TrGa}\left(.\middle|b_0, \kappa_j^{(k+1)} + \frac{1}{2}, \frac{\boldsymbol{\beta_j^{2(k)}}}{2}\right)$$

- Update coefficient parameters $\mathcal{B}^{(k+1)} = (\alpha^{(k+1)}, \boldsymbol{\beta^{(k+1)}})$.

$$\mathcal{B}^{(k+1)} \sim N\bigg(.\bigg|\lambda^{(k)}\Omega_{\beta}^{-1^{(k)}}\mathcal{X}'^{(k+1)}\boldsymbol{y}, (\lambda^{(k)}\mathcal{X}'^{(k+1)}\mathcal{X}^{(k+1)} + \Sigma_{\psi}^{(k+1)})^{-1}\bigg),$$

where Σ_{ψ} is a matrix of $(p+1) \times (p+1)$ and the diagonal of this matrix is equal to $\{\psi_1^{(k+1)}, \psi_2^{(k+1)} \ldots, \psi_{p+1}^{(k+1)}\}$.

- Update the precision parameter of model $\lambda^{(k+1)}$,

$$\lambda^{(k+1)} \sim \text{Ga}\bigg(.\bigg|a_{\lambda} + \frac{n}{2}, b_{\lambda} + \frac{1}{2}(\boldsymbol{y} - \mathcal{X}^{*^{(k+1)}}\mathcal{B}^{(k+1)})'I(\boldsymbol{y} - \mathcal{X}^{*^{(k+1)}}\mathcal{B}^{(k+1)})\bigg).$$

- Update precision parameter of measurement error $\tau^{(k+1)}$.

$$\tau^{k+1} \sim \text{Ga}\bigg(.\bigg|c_{\tau} + \frac{RNp}{2}, d_{\tau} + \frac{1}{2}\Big(\sum_{r=1}^{R}\sum_{j=1}^{p}\sum_{i=1}^{n}\delta_{ijr}^{(k)}\nu^{(k)}(z_{ijr} - x_{ij}^{(k+1)})^2\Big)\bigg).$$

- Updating degree of freedom parameter $\nu^{(k+1)}$,

where Metropolis-Hastings algorithm is used for updating $\nu^{(k+1)}$, and the full conditional distribution is as follows:

$$\pi(\nu|rest) \propto \bigg(\frac{(\frac{\nu}{2})^{\frac{Rnp\nu}{2}}}{(\Gamma(\frac{\nu}{2}))^{Rpn}}(\prod_{i=1}^{n}\prod_{j=1}^{p}\prod_{r=1}^{R}(\delta_{ijr}^{(k)})^{\frac{\nu}{2}-1})\exp\left\{-\frac{\nu}{2}\sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{r=1}^{R}\delta_{ijr}^{(k)}\right\}\bigg)$$
$$\bigg(\frac{b_{\nu}^{a_{\nu}}}{\Gamma(a_{\nu})}(\nu)^{a_{\nu}}\exp(-\nu b_{\nu})\bigg).$$

Metropolis-Hastings algorithm has the following steps:

The proposal distribution of $\nu^{(prop)}$ is gamma distribution with shape equal to $(1+\nu^{(k)} \times pb)$ and rate equal to (pb_{ϕ}) .

Current state is $\big(\nu^k, \delta_{ijr}^k\big)$, updating ν using $q(\nu|\nu^k)$ which the proposal is $\nu^{prop} \sim q(\nu|\nu^k)$

Then the acceptance ratio is given as follows:

$$\alpha = min\left\{1, \frac{\pi(\nu^{prop}|-)q(\nu^k|\nu^{prop})}{\pi(\nu^k|-)q(\nu^{prop}|\nu^k)}\right\}$$

$\Rightarrow \nu^{k+1} = \nu^{prop}$, otherwise $\nu^{k+1} = \nu^k$, where U is a random value from the uniform distribution.

- Update parameters $\delta_{ijr}^{(k+1)}$,

$$\delta_{ijr}^{(k+1)} \sim \text{Ga}\left(.\middle| \frac{\nu^{(k+1)}+1}{2}, \frac{1}{2}\left(\nu^{(k+1)} + \tau^{(k+1)}(z_{ijr} - x_{ij}^{(k+1)})^2\right)\right)$$

C.2.3 Pseudocode for Model with ReB-hNiP Prior for Measurements Error with Bayesian Regression Model

- Input data and prior values $\{y, \text{Z}, a_\lambda, b_\lambda, c_\tau, d_\tau, a_\nu, b_\nu, pb_\nu, b_0, e\}$
- Start with initial values $\{\alpha^{(1)}, \lambda^{(1)}, \beta^{(1)}, \tau^{(1)}, \phi^{(1)}, \nu^{(1)}, \delta^{(1)}, \psi^{(1)}, \kappa^1\}$
- Update unobserved variable $\boldsymbol{x_i^{(k+1)}}$, where $(i = 1, 2, \ldots, p), (j = 1, 2, \ldots, p)$ and $r = 1, 2, \ldots, R$.
$N\Bigg(.\Bigg|(\lambda^{(k)}\boldsymbol{\beta^{(k)}}\boldsymbol{\beta'^{(k)}} + \tau^{(k)}(\sum_{r=1}^{R}\delta_{ijr}^{(k)})I_p)^{-1}\left(\lambda^{(k)}\boldsymbol{\beta^{(k)}}(y_i^{(k)} - \alpha^{(k)}) + \tau^{(k)}(\sum_{r=1}^{R}\delta_{ijr}^{(k)}z_{ijr})\right),\ (\lambda^{(k)}\boldsymbol{\beta^{(k)}}\boldsymbol{\beta'^{(k)}} + \tau^{(k)}(\sum_{r=1}^{R}\delta_{ijr}^{(k)})I_p)\Bigg)$.
- The full conditional posterior distribution of hyperparameter of the coefficients κ_j is as follows:

$$\pi(\kappa_j) \propto b_0^{\kappa_i}\psi_j^{\kappa_i - 1}\frac{e_0^{-1}}{B(a,b)}\left(\frac{\kappa_j}{e_0}\right)^{a-1}\left(1 - \frac{\kappa_j}{e_0}\right)^{b-1},$$

where $j = 1, 2 \ldots (p+1)$, the values of κ_j should satisfy the condition $(0 < \kappa_j \leq e_0)$ and $a = b = \frac{1}{2}$. Metropolis-Hastings algorithm is applied to compute this parameter.

- TrGa is the Truncated Gamma distribution which is bounded from upper. The parameters ψ and κ are hyperparameters prior β_j in the case using ReB-hNIP distribution. The full conditional posterior distribution of ψ is given by:

$$\psi_j \sim \text{TrGa}\left(.\middle| b_0, \kappa_j + \frac{1}{2}, \frac{\beta_j^2}{2}\right).$$

- Updating coefficient parameters $\mathcal{B}^{(k+1)} = (\alpha^{(k+1)}, \boldsymbol{\beta^{(k+1)}})$,

$$\mathcal{B}^{(k+1)} \sim N\bigg(.\bigg|\lambda^{(k)}\Omega_{\beta}^{-1^{(k)}}\mathcal{X}'^{(k+1)}\boldsymbol{y}, (\lambda^{(k)}\mathcal{X}'^{(k+1)}\mathcal{X}^{(k+1)} + \Sigma_{\psi}^{(k+1)})^{-1}\bigg),$$

where Σ_{ψ} is a matrix of $(p+1)\times(p+1)$ dimensional, and the diagonal of this matrix is equal to $\{\psi_1^{(k+1)}, \psi_2^{(k+1)} \ldots, \psi_{p+1}^{(k+1)}\}$.

- Update precision parameter of model $\lambda^{(k+1)}$.

$$\lambda^{(k+1)} \sim \text{Ga}\bigg(.\bigg|a_{\lambda} + \frac{n}{2}, b_{\lambda} + \frac{1}{2}(\boldsymbol{y} - \mathcal{X}^{*^{(k+1)}}\mathcal{B}^{(k+1)})'I(\boldsymbol{y} - \mathcal{X}^{*^{(k+1)}}\mathcal{B}^{(k+1)})\bigg).$$

- Update precision parameter of measurement error $\tau^{(k+1)}$.

$$\tau^{k+1} \sim \text{Ga}\bigg(.\bigg|c_{\tau} + \frac{RNp}{2}, d_{\tau} + \frac{1}{2}\Big(\sum_{r=1}^{R}\sum_{j=1}^{p}\sum_{i=1}^{n}\delta_{ijr}^{(k)}\nu^{(k)}(z_{ijr} - x_{ij}^{(k+1)})^2\Big)\bigg).$$

- Update the degree of freedom parameter $\nu^{(k+1)}$

  Where Metropolis-Hastings algorithm is used for updating $\nu^{(k+1)}$ the full conditional distribution as follows:

$$\pi(\nu|rest) \propto \left(\frac{(\frac{\nu}{2})^{\frac{Rnp\nu}{2}}}{(\Gamma(\frac{\nu}{2}))^{Rpn}}\Big(\prod_{i=1}^{n}\prod_{j=1}^{p}\prod_{r=1}^{R}(\delta_{ijr}^{(k)})^{\frac{\nu}{2}-1}\Big)\exp\left\{-\frac{\nu}{2}\sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{r=1}^{R}\delta_{ijr}^{(k)}\right\}\right)$$
$$\left(\frac{b_{\nu}^{a_{\nu}}}{\Gamma(a_{\nu})}(\nu)^{a_{\nu}}\exp(-\nu b_{\nu})\right).$$

  Metropolis-Hastings algorithm follow the bellow steps:

  The proposal distribution of $\nu^{(prop)}$ is gamma distribution with shape equal to $(1+\nu^{(k)}\times pb)$ and rate equal to (pb_{ϕ}) . Current state is $\left(\nu^k, \delta_{ijr}^k\right)$, updating ν using $q(\nu|\nu^k)$ which the proposal is $\nu^{prop} \sim q(\nu|\nu^k)$. Then the acceptance ratio is given as follows:

$$\alpha = min\left\{1, \frac{\pi(\nu^{prop}|-)q(\nu^k|\nu^{prop})}{\pi(\nu^k|-)q(\nu^{prop}|\nu^k)}\right\},$$

  if $\alpha > U$, then $\nu^{k+1} = \nu^{prop}$, otherwise $\nu^{k+1} = \nu^k$, where U is a random uniform value.

- Update parameters $\delta_{ijr}^{(k+1)}$,

$$\delta_{ijr}^{(k+1)} \sim \text{Ga}\left(\delta_{ijr}^{(k+1)} \middle| \frac{\nu^{(k+1)}+1}{2}, \frac{1}{2}\big(\nu^{(k+1)} + \tau^{(k+1)}(z_{ijr} - x_{ij}^{(k+1)})^2\big)\right),$$

C.3 Pseudocode of Dynamic Bayesian Networks with Measurements Error Model

- Setup the initial values for $\{y, \mathbf{Z}, a_\lambda, b_\lambda, c_\tau, d_\tau, a_\nu, b_\nu, pb_\nu, b_0, e\}$
- Start with initial values $\{\alpha^{(1)}, \lambda^{(1)}, B^{(1)}, \tau^{(1)}, \phi^{(1)}, \nu^{(1)}, \delta^{(1)}, \psi^{(1)}, \kappa^1\}$
- Update unobserved variables Y^{k+1} .

  Y for $t = 2 \dots (T-1)$ as follows:

  $$Y_i^{t(k+1)} \sim N\left(.\middle| \frac{\lambda\Big(\alpha_i^{(k)} + \sum_{j=1}^{p} \beta_{ij}^{(k)} Y_j^{t-1} + \beta_{ii}^{(k)}(Y_i^{t+1} - \alpha_i^{(k)} - \sum_{i\neq j}^{p} \beta_{ij}^{(k)} Y_j^t)\Big) + \tau^{(k)} R \bar{X}_i^t}{\lambda^{(k)}(1+\beta_{ii}^{2(k)}) + \tau^{(k)} R}, \lambda(1+\beta_{ii}^{2(k)}) + \tau^{(k)} R\right),$$

  where $i = 1, 2, 3 \dots p$ and $t = 2, \dots (T-1)$.

  Regarding to compute for unobserved variables Y^{T+1} the $t = T+1$, we used the following formula:

  $$Y_i^{t=T^{(k+1)}} \sim N\left(.\middle| \frac{\lambda(\alpha_i^{(k)} + \sum_{i\neq j}^{p} \beta_{ij}^{(k)} Y_i^{t=T-1} + \beta_{ii}^{(k)} Y_i^{t=T-1}) + \tau^{(k)} R \bar{X}^{t=T}}{(\lambda^{(k)} + R\tau^{(k)})}, (\lambda^{(k)} + R\tau^{(k)})\right).$$

  We supposed that $Y^1 \sim N(.|0, 1)$ overall number of iteration.

- The intercept parameter of the model is represented by $\alpha_i^{(k+1)}$ and it is computed as follow:

$$\alpha_i^{(k+1)} \sim N\left(.\middle| \frac{\lambda(\sum_{t=1}^{T} Y_i^{t(k+1)} - \sum_{t=1}^{T}\sum_{j=1}^{p} \beta_{ij}^{(k)} Y_j^{t-1^{(k+1)}}) + A^{*(k)}}{(T\lambda^{(k+1)} + p_\alpha)}, (T\lambda^{(k+1)} + \sigma_\alpha)\right),$$

$$A^{*(k+1)} \sim \text{N}\left(\frac{pP_\alpha \bar{\alpha}}{(q_a + pP_\alpha)}, (q_a + pP_\alpha)\right).$$

- The precision λ^{k+1} of the model can be computed by;

$$\lambda^{k+1} \sim \text{Ga}\bigg(.\bigg|c_\lambda+\frac{pT}{2}, d_\lambda+\frac{1}{2}\sum_{t=2}^{T}(Y^{t(k+1)}-\alpha-BY^{t-1(k+1)})' I(Y^{t(k+1)}-\alpha-BY^{t-1(k+1)})\bigg).$$

- Precision of measurement error model τ^{k+1} is as follow:

$$\tau^{k+1} \sim \text{Ga}\bigg(.\bigg|a_\tau + \frac{RT(G-1)}{2}, b_\tau + \frac{1}{2}\sum_{i=1}^{p}\sum_{t=1}^{T}(X_i^t - Y_i^{t(k+1)})^2\bigg),$$

  where: X_i^t is a vector of size $(R \times 1)$.

- For updating the coefficients matrix B, the Dynamic Bayesian Networks with measurement error model is divided into two part in tersm of the shrinkage prior distribution of coefficient parameters.

  - **For ReB-hNiP prior**

    * The parameters ψ and κ are hyperparameters prior β_{ij} in the case utilising ReB-hNIP distribution.
      *Update the full conditional posterior distribution of hyperparameter of the coefficients κ_{ij} is as follows: $TrGa$ is Truncated Gamma distribution which is bounded from above. We let $TrGa(.|b_0, a, b) = Ga(.|a, b)I_{0<b_0}$.
      **Metropolis-Hasting algorithm is used to update the hyperparameter of the coefficients $\kappa_{ij}^{(k+1)}$ as it is impossible update it directly:

$$\pi(\kappa_{ij}^{(k+1)}) \propto b_0^{\kappa_{ij}^{(k)}} \psi_{ij}^{(k+1)\kappa_{ij}^{(k)}-1} \frac{e_0^{-1}}{B(a,b)}\left(\frac{\kappa_{ij}^{(k)}}{e_0}\right)^{a-1}\left(1-\frac{\kappa_{ij}^{(k)}}{e_0}\right)^{b-1},$$

      where $j = 1, 2 \ldots p$ and value of $\kappa_{ij}^{(k+1)}$ should satisfy the condition $(0 < \kappa_{ij}^{(k+1)} \leq e_0)$ and $a = b = \frac{1}{2}$.

    * The full conditional posterior distribution of $\psi_{ij}^{(k+1)}$ which is scale mixture of Normal distribution , is given by:

$$\psi_{ij}^{(k+1)} \sim \text{TrGa}\bigg(.\bigg|b_0, \kappa_{ij}^{(k)} + \frac{1}{2}, \frac{\beta_{ij}^{2(k)}}{2}\bigg).$$

– **For hNiP prior**

∗ The parameters ψ and κ are hyperparameters prior β_{ij} are updating using Gibbs sampler.

* Update of hyperparameter of the regression coefficients κ_{ij} as follows:

$$\kappa_{ij}^{(k+1)} \sim \text{Ga}\left(.\middle| c+1, \log\left(\frac{b_0}{\psi_{ij}^{(k)}}\right)+d\right)$$

where $(j, i = 1, 2 \ldots, p.$

We assumed Truncated Gamma distribution is symbolized as $TrGa(.|b_0, a, b) = Ga(.|a, b)I_{0<b_0}$ and it is bounded from above.

∗ Update hyperparameters prior ψ_{ij} is given by:

$$\psi_{ij}^{(k+1)} \sim \text{TrGa}\left(.\middle| b_0, \kappa_j^{(k+1)}+\frac{1}{2}, \frac{\beta_{ij}^{2(k+1)}}{2}\right)$$

- Updating the coefficients parameters B, where every single vector $(\boldsymbol{\beta_i^{(k+1)}})$ update individually either using the prior distribution of ReB-hNip or hNiP as follows:

$$\boldsymbol{\beta_i^{(k+1)}} \sim N\big(.\big|\mu_{\boldsymbol{\beta_i^{(k)}}}, \Omega^{-1}_{\boldsymbol{\beta_i^{(k)}}}\big)$$

The $\boldsymbol{\beta_i^{(k+1)}}$ represents a single vector in B, where $\mu_{\boldsymbol{\beta_i}} = \lambda^{(k+1)}\,\Omega^{-1}_{\boldsymbol{\beta_i}}\big(\sum_{t=2}^{T} Y^{t-1^{(k+1)}}(Y_i^{t^{(k+1)}} - \alpha_i^{(k+1)})\big)$, $\Omega_{\boldsymbol{\beta_i}} = \big(\lambda^{(k+1)}\big[\sum_{t=2}^{T}(Y^{t-1^{(k+1)}}Y^{t-1^{(k+1)}}\big] + \Sigma_{\psi_i^{(k+1)}}\big)$ and $\Sigma_{\psi_i^{(k+1)}}$ is a matrix of $(p\times p)$ diagonal, and the diagonal of this matrix is equal to $\psi_i^{(k+1)} = \{\psi_1^{(k+1)}, \psi_2^{(k+1)} \ldots, \psi_p^{(k+1)}\}$.

Appendix D

Some Statistical Distribution

In this section, we present some common statistical distributions that have been used in this thesis, and we also present some properties of them.

1 - **(Gaussian) Normal Distribution.** Normal distribution is one of the most important statistical distribution, which x is having a random variable with two parameters the mean μ and the precision λ. It is denoted by $x \sim N(x|\mu, \lambda^{-1})$, and the densityof normal distribution is as:

$$N(x|\mu, \lambda^{-1}) = \frac{\lambda}{\sqrt{2\pi}} e^{\frac{-\lambda}{2}(x-\mu)^2} \quad -\infty < x < \infty$$

The main properites of this distribution is the expacted value is $E(x) = \mu$ and the Variance value is $Var(x) = \lambda^{-1}$

2 - **Gamma Distribution.**

In this distribution the random variable x has two parameters, represent the shape ω parameter and the rate γ parameter. This distribution denoted by x $\sim$ Ga$(x|\omega, \gamma)$. The denisty of Gamma distrubution is:

$$\text{Ga}(x|\omega, \gamma) = \frac{\gamma^{\omega}}{\Gamma(\omega)} \text{x}^{\omega-1} e^{-\gamma \text{x}} \quad \omega, \gamma > 0$$

The mean and the variance of Gamma distributios is given by expacted value is $E(x) =$

$\frac{\omega}{\gamma}$ and Variance value is $Var(x) = \frac{\omega}{\gamma^2}$.

* Exponetional distrbution is the special case of Gamma distribution if $(\omega = 1)$.

3 - **Inverse Pareto Distrbution.**

The density of Inverse Pareto (iP) distribution is:

$$\text{iP}(X|\kappa, b) = \kappa b^{-\kappa} X^{\kappa-1} \ , \ X \leq b, (\kappa, b > 0)$$

The distribution function denoted by $X \sim \text{iP}(X|\kappa, b)$, where X is a random variable, and the upper bound parameter represtnted by b and κ is rate parameter. In addition, expacted value is $E(X) = \frac{\kappa b}{\kappa+1}$, and variance value is $Var(X) = \frac{\kappa b^2}{(\kappa+1)^2(\kappa+2)}$.

4 - **Rescaled Beta distrbution.** The density of Rescaled Beta is:

$$\text{ReB}(x\big|e_0, a, b) = \frac{e_0^{-1}}{B(a,b)} \left(\frac{x}{e_0}\right)^{a-1} \left(1 - \frac{x}{e_0}\right)^{b-1} \quad a, b, e_0 > 0$$

where $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a,b)}$, and it is equal to $B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx$. The random variable x has three parameters (a, b, e_0) where a, b are shape parameters and e_0 is rescaled parameter, This distrubution is denoted by $\text{ReB}(x\big|e_0, a, b)$.

In addition, mean $E(x) = \frac{a\ e_0}{a+b}$ and the variance $Var(x) = \frac{a\ b\ e_0^2}{(a+b)^2\ (a+b+1)}$

5 - **Truncated Gamma Distribution.**

Truncated Gamma distribution which is bounded from above is denoted by $\text{TrGa}(x|b, \omega, \gamma)$ where $I_{x<b}$, the density function of this distribution is given:

$$\text{TrGa}(x|b, \omega, \gamma) = C\text{x}^{\omega-1}e^{-\gamma x}, \qquad 0 < x < b \quad \omega, \gamma > 0$$

where $C = \int_0^b \text{X}^{\omega-1}e^{-\gamma x}dx$, and also $\omega > 0$ is the scale and $\gamma > 0$ represents the shape,

6 - Student's t distribution $P(Y|\mu, \lambda, \psi) = \int_0^\infty N\Big(Y|\mu, \psi^{-1}\lambda^{-1}\Big) \quad \text{Ga}\Big(\psi|\frac{\nu}{2}, \frac{\nu}{2}\Big)d\psi$

$\text{St}(x|\mu, \lambda, \nu) = N\big(x|\mu, \lambda^{-1}\big) \quad \text{Ga}\big(\lambda|\frac{\nu}{2}, \frac{\nu}{2}\big)$ where μ indicates the mean, degree of reedom denoted by ν and λ is precision.

Appendix E

Prior and Posterior Distrbution for NEG Model

In this section we present the prior and full conditional distribution for Bayesian regression model when using NEG prior distribution. The prior and posterior distributions, is presents in only the main part that related to NEG prior distributions without repeating the full posterior distribution of the coefficients, precision parameters. Due to have similar formula and formatting to the proposed priors distribution under same linear model. The prior distribution is as follows:

$$\pi(\phi,\beta,\zeta)=\prod_{j=1}^{p+1}\frac{1}{\phi_j^{\frac{1}{2}}}\exp\Big\{-\frac{1}{2}\boldsymbol{\beta_j'}\frac{1}{\phi_j}\boldsymbol{\beta_j}\Big\}\prod_{j=1}^{p+1}\Big(\zeta_j\exp(-\phi_j\zeta_j)\Big)\prod_{j=1}^{p+1}\frac{(\frac{1}{\gamma^2})^{\omega}}{\Gamma(\omega)}\zeta_j^{\omega-1}\exp\Big\{-\frac{\zeta_j}{\gamma^2}\Big\}. \tag{E.1}$$

The full Bayesian regression model relies on the NEG prior distribution for reduction the coefficient parameters β_i and it is given as follows:

$$\begin{aligned}\pi(\Theta|y,X)=(\lambda)^{n/2}\exp\Big\{-\frac{\lambda}{2}(\boldsymbol{y}-X\boldsymbol{\beta})'I(\boldsymbol{y}-X\boldsymbol{\beta})\Big\}\frac{b_\lambda^{a_\lambda}}{\Gamma(a_\lambda)}\lambda^{a_\lambda-1}\exp\Big\{-b_\lambda\lambda\Big\}\\ \prod_{j=1}^{p+1}\frac{1}{\phi_j^{\frac{1}{2}}}\exp\Big\{-\frac{1}{2}\boldsymbol{\beta_j'}\frac{1}{\phi_j}\boldsymbol{\beta_j}\Big\}\prod_{j=1}^{p+1}\Big(\zeta_j\exp(-\phi_j\zeta_j)\Big).\prod_{j=1}^{p+1}\frac{(\frac{1}{\gamma^2})^{\omega}}{\Gamma(\omega)}\zeta_j^{\omega-1}\exp\Big\{-\frac{\zeta_j}{\gamma^2}\Big\}.\end{aligned} \tag{E.2}$$

E.1 The Full Conditional Distributions.

We have showed the full conditional distribution for precision of the model λ in 4.5. In the following formulation, the full conditional distributions are displayed. The first parameter ζ describe the hyper-parameter of scale mixture normal prior of regression model and sit is refers to the rate of gamma distribution of the second order of the hierarchal NEG prior distribution,

$$\zeta_j \sim \text{Ga}\bigg(\zeta_j|\omega+1,(\phi_j+\frac{1}{\gamma^2})\bigg). \tag{E.3}$$

Metropolis-Hastings algorithm is used for updating ϕ_j. The proposal distribution is gamma with shape equal to $(1+\phi_j \times pb_\phi)$ and rate equal to (pb_ϕ). We are fixing the mode at the current value and use pb_ϕ to control the acceptance rate.

$$\pi(\phi_j) \propto \bigg(\phi_j^{-\frac{1}{2}} \exp\bigg\{-(\frac{\beta_j^{'}\beta_j}{2\phi_j}+\zeta_j\phi_j)\bigg\}\bigg). \tag{E.4}$$

In this cases, we update a vector of ϕ_j where $(j=1,2,\ldots,p+1)$ via using Metropolis-Hastings algorithm.

$$\prod_{j=1}^{p+1}\pi(\phi_j) \propto \bigg(\prod_{j=1}^{p+1}\phi_j^{-\frac{1}{2}} \exp\bigg\{-(\frac{\beta_j^{'}\beta_j}{2\phi_j}+\zeta_j\phi_j)\bigg\}\bigg). \tag{E.5}$$

Then, we compute the full conditional distribution of the coefficient parameters $\boldsymbol{\beta}$.

$$\boldsymbol{\beta}^{*} \sim N\big(\boldsymbol{\beta}|\mu_{\boldsymbol{\beta}},\Omega_{\boldsymbol{\beta}}^{-1}\big) \tag{E.6}$$

Where $\mu_\beta = \lambda\Omega_{\boldsymbol{\beta}}^{-1}X^{'}\boldsymbol{y}$, $\Omega_{\boldsymbol{\beta}} = (\lambda X^{'}X+\Sigma_\Phi)$ and Σ_Φ is a matrix of $(p+1)\times(p+1)$ diagonal and the diagonal of this matrix is equal to $\{\frac{1}{\phi_1},\frac{1}{\phi_2}\ldots,\frac{1}{\phi_{p+1}}\}$.

Bibliography

Abegaz, F., Wit, E., 2013. Sparse time series chain graphical models for reconstructing genetic networks. Biostatistics 14, 586–599.

Abramowitz, M., Stegun, I.A., et al., 1964. Handbook of mathematical functions. Applied Mathematics Series 55, 39.

Aijun, Y., Xuejun, J., Liming, X., Jinguan, L., 2017. Sparse bayesian variable selection in multinomial probit regression model with application to high-dimensional data classification. Communications in Statistics-Theory and Methods 46, 6137–6150.

AKAIKE, H., 1973. Information theory and an extension of the maximum likelihood principle, in: 2nd Inter. Symp. on Information Theory, Akademiai Kiado. pp. 267–281.

Andrews, D.F., Mallows, C.L., 1974. Scale mixtures of normal distributions. Journal of the Royal Statistical Society. Series B (Methodological) , 99–102.

Andrieu, C., De Freitas, N., Doucet, A., Jordan, M.I., 2003. An introduction to mcmc for machine learning. Machine Learning 50, 5–43.

Armagan, A., Clyde, M., Dunson, D.B., 2011. Generalized beta mixtures of gaussians, in: Shawe-Taylor, J., Zemel, R.S., Bartlett, P.L., Pereira, F., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 24. Curran Associates, Inc., pp. 523–531.

Armagan, A., Dunson, D.B., Lee, J., 2013a. Generalized double pareto shrinkage. Statistica Sinica 23, 119.

Armagan, A., Dunson, D.B., Lee, J., Bajwa, W.U., Strawn, N., 2013b. Posterior consistency in linear models under shrinkage priors. Biometrika 100, 1011–1018.

Bani, K.M., 2009. Bayesian Analysis of Gene Expression Data. John Wiley & Sons.

Bar-Joseph, Z., 2004. Analyzing time series gene expression data. Bioinformatics 20, 2493–2503.

Beal, M.J., Falciani, F., Ghahramani, Z., Rangel, C., Wild, D.L., 2005. A bayesian approach to reconstructing genetic regulatory networks with hidden factors. Bioinformatics 21, 349–356.

Ben-Gal, I., 2008. Bayesian networks. Encyclopedia of Statistics in Quality and Reliability 1.

Bernardo, J., Smith, A., 2000. Bayesian Theory. Wiley Series in Probability and Statistics, Wiley.

Breheny, P., Breheny, M.P., LazyData, T., 2018. Package ncvreg .

Breiman, L., et al., 1996. Heuristics of instability and stabilization in model selection. The Annals of Statistics 24, 2350–2383.

Brimacombe, M., 2014. High-dimensional data and linear models: a review. Open Access Med Stat 4, 17–27.

Buonaccorsi, J.P., 2010. Measurement error: models, methods, and applications. Chapman and Hall/CRC.

Cao, C., Lin, J., Shi, J.Q., Wang, W., Zhang, X., 2015. Multivariate measurement error models for replicated data under heavy-tailed distributions. Journal of Chemometrics 29, 457–466.

Carroll, R.J., Ruppert, D., Stefanski, L.A., Crainiceanu, C.M., 2006. Measurement error in nonlinear models: a modern perspective. CRC press.

Carvalho, C.M., Polson, N.G., Scott, J.G., 2009. Handling sparsity via the horseshoe., in: AISTATS, pp. 73–80.

Carvalho, C.M., Polson, N.G., Scott, J.G., 2010. The horseshoe estimator for sparse signals. Biometrika 97, 465–480.

Causton, H., Quackenbush, J., Brazma, A., 2009. Microarray gene expression data analysis: a beginner's guide. Wiley. com.

Chai, L.E., Mohamad, M.S., Deris, S., Chong, C.K., Choon, Y.W., Ibrahim, Z., Omatu, S., 2012. Inferring gene regulatory networks from gene expression data by a dynamic bayesian network-based model, in: Distributed Computing and Artificial Intelligence, Springer. pp. 379–386.

Chen, J., Chen, Z., 2012. Extended bic for small-n-large-p sparse glm. Statistica Sinica , 555–574.

Chen, R., Resnick, S.M., Davatzikos, C., Herskovits, E.H., 2012. Dynamic bayesian network modeling for longitudinal brain morphometry. NeuroImage 59, 2330 – 2338. doi:doi: http://dx.doi.org/10.1016/j.neuroimage.2011.09.023.

Dallas, A., 1976. Characterizing the pareto and power distributions. Annals of the Institute of Statistical Mathematics 28, 491–497.

Datta, J., Ghosh, J.K., 2013. Asymptotic properties of bayes risk for the horseshoe prior. Bayesian Anal. 8, 111–132. URL: https://doi.org/10.1214/13-BA805, doi:doi: 10.1214/13-BA805.

David, L.A., Wiggins, C.H., 2007. Benchmarking of dynamic bayesian networks inferred from stochastic time-series data. Annals of the New York Academy of Sciences 1115, 90–101.

Dondelinger, F., Lèbre, S., Husmeier, D., 2013. Non-homogeneous dynamic bayesian networks with bayesian regularization for inferring gene regulatory networks with gradually time-varying structure. Machine Learning 90, 191–230.

Doshi, F., Wingate, D., Tenenbaum, J., Roy, N., 2011. Infinite dynamic bayesian networks, in: Proceedings of the 28th International Conference on Machine Learning (ICML-11), pp. 913–920.

Efron, B., Tibshirani, R., Storey, J.D., Tusher, V., 2001. Empirical bayes analysis of a microarray experiment. Journal of the American Statistical Association 96, 1151–1160.

En Chai, L., Saberi Mohamad, M., Deris, S., Khim Chong, C., Wen Choon, Y., Omatu, S., 2014. Current development and review of dynamic bayesian network-based methods for inferring gene regulatory networks from gene expression data. Current Bioinformatics 9, 531–539.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96, 1348–1360.

Fan, Y., Tang, C.Y., 2013. Tuning parameter selection in high dimensional penalized likelihood. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75, 531–552.

Fan, Y., Wang, X., Peng, Q., 2017. Inference of gene regulatory networks using bayesian nonparametric regression and topology information. Computational and Mathematical Methods in Medicine 2017.

Fawcett, T., 2006. An introduction to roc analysis. Pattern Recognition Letters 27, 861–874.

Friedman, N., Linial, M., Nachman, I., Pe'er, D., 2000. Using bayesian networks to analyze expression data. Journal of Computational Biology 7, 601–620.

Friedman, N., Murphy, K., Russell, S., 1998. Learning the structure of dynamic probabilistic networks, in: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers Inc.. pp. 139–147.

Fu, F., Zhou, Q., 2013. Learning sparse causal gaussian networks with experimental intervention: Regularization and coordinate descent. Journal of the American Statistical Association 108, 288–300. doi:doi: 10.1080/01621459.2012.754359.

Fuller, W., 1987. Measurement Error Models. Wiley Series in Probability and Statistics, Wiley.

Gabry, J., 2017. bayesplot: Plotting for bayesian models. URL: `http://mc-stan.org/`. r package version 1.2.0.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., Gelman, A., 2017. Visualization in bayesian workflow. arXiv preprint arXiv:1709.01449 .

Gelman, A., Hwang, J., Vehtari, A., 2014. Understanding predictive information criteria for bayesian models. Statistics and Computing 24, 997–1016.

Gelman, A., Shalizi, C.R., 2013. Philosophy and the practice of bayesian statistics. British Journal of Mathematical and Statistical Psychology 66, 8–38.

Gelman, A., Stern, H.S., Carlin, J.B., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013. Bayesian data analysis. Chapman and Hall/ CRC.

Gelman, A., Van Mechelen, I., Verbeke, G., Heitjan, D.F., Meulders, M., 2005. Multiple imputation for model checking: completed-data plots with missing and latent data. Biometrics 61, 74–85.

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

Ghosh, J.K., Marco, V., 1999. Probabilistic Bayesian Network Model Building of Heart Disease. Technical Report. University of South Carolina.

Gilks, W.R., 1999. Markov Chain Monte Carlo In Practice. Chapman and Hall/CRC.

Godsey, B., 2013. Improved inference of gene regulatory networks through integrated bayesian clustering and dynamic modeling of time-course expression data. PloS one 8, e68358.

Gonçalves, F.B., Prates, M.O., Lachos, V.H., 2015. Robust bayesian model selection for heavy-tailed linear regression using finite mixtures. arXiv preprint arXiv:1509.00331 .

Gradshteyn, I.S., Ryzhik, I.M., 2014. Table of integrals, series, and products. Academic Press.

Gramacy, R.B., Gramacy, M.R.B., data augmentation extends this Bayesian, M., 2019. Package monomvn .

Griffin, J.E., Brown, P., 2011. Bayesian hyper-lassos with non-convex penalization. Australian & New Zealand Journal of Statistics 53, 423–442.

Griffin, J.E., Brown, P.J., 2007. Bayesian adaptive lassos with non-convex penalization. Centre for Research in Statistical Methodology, University of Warwick, Coventry, UK, Technical Report , 07–2.

Griffin, J.E., Brown, P.J., 2010. Inference with normal-gamma prior distributions in regression problems. Bayesian Anal. 5, 171–188. doi:doi: 10.1214/10-BA507.

Griffiths, T.L., Chater, N., Kemp, C., Perfors, A., Tenenbaum, J.B., 2010. Probabilistic models of cognition: Exploring representations and inductive biases. Trends in Cognitive Sciences 14, 357–364.

Hans, C., 2009. Bayesian lasso regression. Biometrika 96, 835–845.

Hastie, T., Efron, B., Hastie, M.T., 2007. The lars package .

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

Hill, S., 2012. Sparse graphical models for cancer signalling. Ph.D. thesis. University of Warwick.

Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12, 55–67.

Hoggart, C.J., Whittaker, J.C., De Iorio, M., Balding, D.J., 2008. Simultaneous analysis of all snps in genome-wide and re-sequencing association studies. PLoS genetics 4, e1000130.

Huang, X., Pan, W., Grindle, S., Han, X., Chen, Y., Park, S.J., Miller, L.W., Hall, J., 2005. A comparative study of discriminating human heart failure etiology using gene expression profiles. BMC Bioinformatics 6, 205.

Husmeier, D., Dybowski, R., Roberts, S., 2005. Probabilistic modeling in bioinformatics and medical informatics. Springer.

Imoto, S., Goto, T., Miyano, S., et al., 2002. Estimation of genetic networks and functional structures between genes by using bayesian networks and nonparametric regression., in: Pacific Symposium on Biocomputing, World Scientific. pp. 175–186.

Imoto, S., Kim, S., Goto, T., Aburatani, S., Tashiro, K., Kuhara, S., Miyano, S., 2003. Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. Journal of Bioinformatics and Computational Biology 1, 231–252.

Jung, Y., 2016. Efficient tuning parameter selection by cross-validated score in high dimensional models. International Journal of Mathematical and Computational Sciences 10, 19 – 25.

Kalli, M., Griffin, J.E., 2014. Time-varying sparsity in dynamic regression models. Journal of Econometrics 178, 779–793.

Karlebach, G., Shamir, R., 2008. Modelling and analysis of gene regulatory networks. Nature Reviews Molecular Cell Biology 9, 770–780.

Kifayat, T., Aslam, M., Ali, S., 2012. Bayesian inference for the parameter of the power distribution. Journal of Reliability and Statistical Studies 5, 45–58.

Kim, S., Imoto, S., Miyano, S., 2004. Dynamic bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data, in: Biosystems, pp. 57–65.

Kuznetsov, A., Kærn, M., Kopell, N., 2004. Synchrony in a population of hysteresis-based genetic oscillators. SIAM Journal on Applied Mathematics 65, 392–425.

Lachos, V., Angolini, T., Abanto-Valle, C., 2011. On estimation and local influence analysis for measurement errors models under heavy-tailed distributions. Statistical Papers 52, 567–590.

Laird, N.M., Lange, C., 2010. The fundamentals of modern statistical genetics. Springer Science & Business Media.

Lèbre, S., 2009. Inferring dynamic genetic networks with low order independencies. Statistical Applications in Genetics and Molecular Biology 8, 1–38.

Lee, G., Lee, H., Sohn, K.A., 2017. Generating Time Series Simulation Dataset Derived from Dynamic Time-Varying Bayesian Network. Springer Singapore, Singapore. doi:doi: 10.1007/978-981-10-4154-9_7.

Lee, M.L.T., 2007. Analysis of microarray gene expression data. Springer Science & Business Media.

Leng, C., Tran, M.N., Nott, D., 2014. Bayesian adaptive lasso. Annals of the Institute of Statistical Mathematics 66, 221–244.

Li, L., Yao, W., 2014. Fully bayesian logistic regression with hyper-lasso priors for high-dimensional feature selection. arXiv preprint arXiv:1405.3319 .

Liang, H., Li, R., 2009. Variable selection for partially linear models with measurement errors. Journal of the American Statistical Association 104, 234–248.

Lin, J.G., Cao, C.Z., 2013. On estimation of measurement error models with replication under heavy-tailed distributions. Computational Statistics 28, 809–829.

Locke, J.C., Kozma-Bognár, L., Gould, P.D., Fehér, B., Kevei, E., Nagy, F., Turner, M.S., Hall, A., Millar, A.J., 2006. Experimental validation of a predicted feedback loop in the multi-oscillator clock of arabidopsis thaliana. Molecular Systems Biology 2.

Lusis, A.J., Weiss, J.N., 2010. Cardiovascular networks systems-based approaches to cardiovascular disease. Circulation 121, 157–170.

Lutful Kabir, A., Ahsanullah, M., 1974. Estimation of the location and scale parameters of a power-function distribution by linear functions of order statistics. Communications in Statistics-Theory and Methods 3, 463–467.

Ma, Y., Li, R., 2010. Variable selection in measurement error models. Bernoulli: Official Journal of the Bernoulli Society for Mathematical Statistics and Probability 16, 274.

Madan, D.B., Seneta, E., 1990. The variance gamma (vg) model for share market returns. Journal of Business , 511–524.

Mallick, H., Yi, N., 2013. Bayesian methods for high dimensional linear models. Journal of Biometrics & Biostatistics 1, 005.

Marshall, A.H., Hill, L.A., Kee, F., 2010. Continuous dynamic bayesian networks for predicting survival of ischaemic heart disease patients. Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems 0, 178–183.

Mbalawata, I.S., Särkkä, S., Vihola, M., Haario, H., 2015. Adaptive metropolis algorithm using variational bayesian adaptive kalman filter. Computational Statistics & Data Analysis 83, 101–115.

Meinshausen, N., Bühlmann, P., 2006. High-dimensional graphs and variable selection with the lasso. The Annals of Statistics 34, 1436–1462.

Mitchell, T.J., Beauchamp, J.J., 1988. Bayesian variable selection in linear regression. Journal of the American Statistical Association 83, 1023–1032.

Moothathu, T., 1993. A characterization of the α -mixture pareto distribution through a property of lorenz curve. Sankhyā: The Indian Journal of Statistics, Series B , 130–134.

Morris, R.W., Cooper, J.A., Shah, T., Wong, A., Drenos, F., Engmann, J., McLachlan, S., Jefferis, B., Dale, C., Hardy, R., Kuh, D., Ben-Shlomo, Y., Wannamethee, S.G., Whincup, P.H., Casas, J.P., Kivimaki, M., Kumari, M., Talmud, P.J., Price, J.F., Dudbridge, F., Hingorani, A.D., Humphries, S.E., 2016. Marginal role for 53 common genetic variants in cardiovascular disease prediction. Heart 102, 1640–1647. doi:doi: 10.1136/heartjnl-2016-309298.

Morrissey, E., 2012. Grenits: Gene regulatory network inference using time series R package version 1.26.0.

Morrissey, E.R., Juárez, M.A., Denby, K.J., Burroughs, N.J., 2011. Inferring the time-invariant topology of a nonlinear sparse gene regulatory network using fully bayesian spline autoregression. Biostatistics 12, 682–694.

Morrissey, E.R., Jurez, M.A., Denby, K.J., Burroughs, N.J., 2010. On reverse engineering of gene interaction networks using time course data with repeated measurements. Bioinformatics 26, 2305–2312. URL: `http://dblp.uni-trier.de/db/journals/bioinformatics/bioinformatics26.html#MorrisseyJDB10`.

Murphy, K., Mian, S., et al., 1999. Modelling gene expression data using dynamic Bayesian networks. Technical Report. Technical report, Computer Science Division, University of California, Berkeley, CA.

Murphy, K.P., 2002. Dynamic Bayesian Networks: Representation, Inference and Learning. Ph.D. thesis. University of California, Berkeley.

Naldi, A., Berenguier, D., Fauré, A., Lopez, F., Thieffry, D., Chaouiya, C., 2009. Logical modelling of regulatory networks with ginsim 2.3. Biosystems 97, 134–139.

Napoli, C., Lerman, L.O., Sica, V., Lerman, A., Tajana, G., de NIGRIS, F., 2003. Microarray analysis: a novel research tool for cardiovascular scientists and physicians. Heart 89, 597–604.

Ndiaye, N.C., Azimi Nehzad, M., El Shamieh, S., Stathopoulou, M.G., Visvikis-Siest, S., 2011. Cardiovascular diseases and genome-wide association studies. Clinica Chimica Acta 412, 1697–1701.

Ni, Y., Stingo, F.C., Baladandayuthapani, V., 2015. Bayesian nonlinear model selection for gene regulatory networks. Biometrics 71, 585–595.

Nick, Townsend, P., Bhatnagar, W., E, K., Wickramasinghe, M., Rayner, 2015. CARDIOVASCULAR DISEASE STATISTICS 2015 United kingdom. British Heart Foundation, London.

O'Hara, R.B., Sillanpää, M.J., et al., 2009. A review of bayesian variable selection methods: what, how and which. Bayesian Analysis 4, 85–117.

Ong, I.M., Glasner, J.D., Page, D., 2002. Modelling regulatory pathways in e. coli from time series expression profiles. Bioinformatics 18, S241–S248.

Opgen-Rhein, R., Strimmer, K., 2007. Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process. BMC Bioinformatics 8, S3.

Park, T., Casella, G., 2008. The bayesian lasso. Journal of the American Statistical Association 103, 681–686.

Pawlas, P., Szynal, D., 2000. Recurrence relations for single and product moments of k-th lower record values from the inverse distributions of pareto's type and characterizations. Discussiones Mathematicae Probability and Statistics 20, 223–231.

Pearson, R., Liu, X., Sanguinetti, G., Milo, M., Lawrence, N., Rattray, M., 2009. puma: a bioconductor package for propagating uncertainty in microarray analysis. BMC Bioinformatics 10, 211.

Penfold, C.A., Wild, D.L., 2011. How to infer gene networks from expression profiles, revisited. Interface Focus 1, 857–870.

Perrin, B.E., Ralaivola, L., Mazurie, A., Bottani, S., Mallet, J., dAlche Buc, F., 2003. Gene networks inference using dynamic bayesian networks. Bioinformatics 19, ii138–ii148.

Radhakrishnan, N., Marco, L., Marco, Scutari, S.L., 2013. Bayesian Networks in R with Applications in Systems Biology. Springer, New York. ISBN 978-1461464457.

Rajaratnam, B., Sparks, D., 2015. Fast bayesian lasso for high-dimensional regression. arXiv preprint arXiv:1509.03697 .

Robinson, T.R., 2010. Genetics for dummies. John Wiley & Sons.

Rockova, V., Lesaffre, E., et al., 2014. Incorporating grouping information in bayesian variable selection with applications in genomics. Bayesian Analysis 9, 221–258.

Rogers, W.H., Tukey, J.W., 1972. Understanding some long-tailed symmetrical distributions. Statistica Neerlandica 26, 211–226.

Rosenbaum, M., Tsybakov, A.B., et al., 2010. Sparse recovery under matrix uncertainty. The Annals of Statistics 38, 2620–2651.

Rudelson, M., Zhou, S., 2015. High dimensional errors-in-variables models with dependent measurements. arXiv preprint arXiv:1502.02355 .

Rudelson, M., Zhou, S., et al., 2017. Errors-in-variables models with dependent measurements. Electronic Journal of Statistics 11, 1699–1797.

Russe, A., 2009. Computational Biology: New Research. Nova Science Publishers, New York.

Schwarz, G., et al., 1978. Estimating the dimension of a model. The Annals of Statistics 6, 461–464.

Seo, D., Ginsburg, G.S., Goldschmidt-Clermont, P.J., 2006. Gene expression analysis of cardiovascular diseases: Novel insights into biology and clinical applications. Journal of the American College of Cardiology 48, 227 – 235.

Shakeel, M., ul Haq, M.A., Hussain, I., Abdulhamid, A.M., Faisal, M., 2016. Comparison of two new robust parameter estimation methods for the power function distribution. PloS one 11, e0160692.

Shiffman, D., Porter, J.G., 2000. Gene expression profiling of cardiovascular disease models. Current Opinion in Biotechnology 11, 598–601.

Shojaie, A., Michailidis, G., 2010. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. Biometrika 97, 519–538.

Song, L., Kolar, M., Xing, E.P., 2009. Time-varying dynamic bayesian networks, in: Bengio, Y., Schuurmans, D., Lafferty, J.D., Williams, C.K.I., Culotta, A. (Eds.), Advances in Neural Information Processing Systems 22. Curran Associates, Inc., pp. 1732–1740.

Song, Q., Liang, F., 2017. Nearly optimal bayesian shrinkage for high dimensional regression. arXiv preprint arXiv:1712.08964 .

Sørensen, Ø., Frigessi, A., Thoresen, M., 2014. Covariate selection in high-dimensional generalized linear models with measurement error. arXiv preprint arXiv:1407.1070 .

Sørensen, Ø., Frigessi, A., Thoresen, M., 2015. Measurement error in lasso: Impact and likelihood bias correction. Statistica Sinica , 809–829.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64, 583–639.

Strimbu, K., Tavel, J.A., 2010. What are biomarkers? Current Opinion in HIV and AIDS, NIH Public Access 5, 463.

Sugimoto, N., Iba, H., 2004. Inference of gene regulatory networks by means of dynamic differential bayesian networks and nonparametric regression. Genome Informatics Series 15, 121.

Suzuki, D.T., Griffiths, A.J., et al., 1976. An introduction to genetic analysis. WH Freeman and Company.

Thornton-Wells, T.A., Moore, J.H., Haines, J.L., 2004. Genetics, statistics and human disease: analytical retooling for complexity. TRENDS in Genetics 20, 640–647.

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

Tobon-Mejia, D., Medjaher, K., Zerhouni, N., 2012. Cnc machine tool's wear diagnostic and prognostic by using dynamic bayesian networks. Mechanical Systems and Signal Processing 28, 167–182.

Vasan, R.S., 2006. Biomarkers of cardiovascular disease molecular basis and practical considerations. Circulation 113, 2335–2362.

Vehtari, A., Gelman, A., 2014. Waic and cross-validation in stan. Submitted. http://www. stat. columbia. edu/˜ gelman/research/unpublished/waic_ stan. pdf Accessed 27, 5.

Vijesh, N., Chakrabarti, S.K., Sreekumar, J., 2013. Modeling of gene regulatory networks: a review. Journal of Biomedical Science and Engineering 6, 223.

West, M., 1987. On scale mixtures of normal distributions. Biometrika 74, 646–648.

WHO, 2017. World health organization statistical information system. http://www.who.int/cardiovascular_diseases/world-heart-day-2017/en/. Accessed: 2018-04-01.

Wilson, P.W., DAgostino, R.B., Levy, D., Belanger, A.M., Silbershatz, H., Kannel, W.B., 1998. Prediction of coronary heart disease using risk factor categories. Circulation 97, 1837–1847.

Wit, E.C., Abbruzzo, A., 2015. Inferring slowly-changing dynamic gene-regulatory networks. BMC Bioinformatics 16, 1.

Xu, Q., You, J., 2007. Covariate selection for linear errors-in-variables regression models. Communications in Statistics Theory and Methods 36, 375–386.

Yang, S., Khot, T., Kersting, K., Natarajan, S., 2016. Learning continuous-time bayesian networks in relational domains: A non-parametric approach, in: Thirtieth AAAI Conference on Artificial Intelligence.

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68, 49–67.

Zhu, S., Wang, Y., 2012. Modelling non-stationary gene regulatory process with hidden markov dynamic bayesian network, in: Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on, IEEE. pp. 1–4.

Zou, C., Feng, J., 2009. Granger causality vs. dynamic bayesian network inference: a comparative study. BMC Bioinformatics 10, 122.

Zou, H., 2006. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101, 1418–1429.

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

Zou, M., Conzen, S.D., 2005. A new dynamic bayesian network (dbn) approach for identifying gene regulatory networks from time course microarray data. Bioinformatics 21, 71–79.