ABSTRACT

TIAN, YIQING. Variable Selection in Logistic Regression with Applications. (Under the direction of Howard Bondell and Alyson Wilson.)

Logistic regression plays important roles in statistical methods development and application. Logistic model has been widely used in disease prediction, reliability prediction, and credit risk analysis. People are also interest in the situation that we have more predictors than the sample size. With high dimension situation, we would expect to perform variable selection before continuing with the prediction. There are extensive variable selection method focusing on linear model, some of them have been extended to logistic model variable selection. In this thesis, we mainly focus on three parts and focus on two areas. The first part is methodology development. We extend the Joint Credible Region approach which is first applied in linear model to logistic model variable selection. The next two parts focus on the application of the logistic model: reliability variable selection and reliability assessment.

The first chapter detailed the logistic model variable selection. We extend the Joint Credible Region approach to the logistic model. The idea behind joint credible region approach is, based on a chosen prior, we build a sequence of posterior credible region. Given a credible region at a particular level, any points within the region can be considered as the potential values for the parameters. The method chooses the model represented by the points within the region that is the most sparse. As the coverage level increases, the model becomes more sparse as the credible region covers more and more space. The expansion of the credible region yields a sequence of credible sets that can be seen as a sequence of models. We used Normal-Gamma prior to construct joint credible region. The Normal-Gamma prior has nice shrinkage property that by some tuning process, it

would be able to shrinkage the posterior mean of non-important predictor to a value close to zero,while the important ones significant different from zero. We developed ways of tuning hyper-parameter in Normal-Gamma prior which is based on the prior belief of important predictors. It shows the proposed method outperform existing logistic model variable selection methods including LASSO, forward selection and screening. We also examined the association between the correlation of important predictors and prior that may have on the selection performance. Simulation results show, with highly correlated predictors, a tight prior with fewer important predictor belief would do better; vice versa.

The second problem is an application of the method developed in the first chapter. We extend the method to the reliability variable selection. Reliability is the probability of a system pass a test. There are a bunch of factors affect the system's reliability, and the goal is to identify the important ones. Thus, the problem can be formalized by logistic model variable selection. What we observe are both component, sub-component and system level data, and we want to simultaneously modeling the both source of data. For reliability variable selection, the challenge is, with a complex system, the likelihood function can be complicated, which add more difficulties to the computation of posterior distribution. In this chapter, we developed a new approach to compute the posterior, without the facilitation of the full likelihood function. The new approach is called Joint Credible Region within Imputation. In this approach, if we observe only system, then we treat the component information is missing, and update the missing component conditional on the status of system and the cut sets. After recovery the missing component data, we can sample from the posterior distribution of each component.

The last problem discussed is reliability assessment. For reliability assessment, the response is still binary, but we only consider one covariate, the time. We focus on the situation that the binary outcome is unbalanced. That, for a test, we observe fewer

or even no failures. The default sets of priors explored in previous research leads to unreasonable result. One set of default prior lead to the lower posterior confidence interval drop immediately past the last observation we have; the other set of default prior leads to the posterior confidence interval stays high as 1. In this chapter, we focus on prior development, in other words, prior elicitation to remedy the outcome. We introduced two sets of prior: Fixed Reliability Prior and Fixed Time Prior. The fixed reliability prior choose two reliability values and specify a time interval in which we think those reliabilities will occur. The fixed time prior would choose two times and then specify a range of possible reliabilities. The new prior demonstrate its superiority in reliability assessment. We also examined the sensitivity of the proposed prior and form our recommendation.

Variable Selection in Logistic Regression with Applications

by
Yiqing Tian

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2017

APPROVED BY:

Rui Song	Ralph Smith
Howard Bondell Co-chair of Advisory Committee	Alyson Wilson Co-chair of Advisory Committee

DEDICATION

To my parents.

BIOGRAPHY

Yiqing Tian was born in Tianjin, China. She obtained a Bachelors degree in Economics from the School of Economics, Nankai University, in 2005, and then obtained a master degree in Economics from Nankai University. Then, she decided to pursue her Ph.D. in Statistics at North Carolina State University in United States. Her doctoral dissertation research is under the valuable guidance of her two advisors, Dr. Howard Bondell and Dr. Alyson Wilson. During her graduate studies, she worked as a Teaching Assistant, Research Assistant, Graduate Industrial Trainee at UCB Bioscience Inc from year 2015-2017.

ACKNOWLEDGEMENTS

First, I would like to express my sincerest attitude to my advisors, Dr.Bondell and Dr.Wilson, for all their great help, to bring me into the academic world of Statistics. Dr.Bondell and Dr.Wilson not only taught me how to do the research, but also guide me how to conduct the research. I still remember the first day, Dr.Bondell said, I am happy to take you into the research. And Dr.Wilson said, it is not important what you did, but it is important to help you grow into independent researcher. I learned a lot from them, from capability of critical thinking to technical writing skills. These are very important to me, and helped me grow up as an independent researcher and have lasting influence on my future career development. I would also like to thank Dr.Rui Song and Dr.Ralph Smith taking precious time to be on my committee and providing valuable suggestions.

I would like to thank the Department and UCB Bioscience to provide me a great opportunity to work as Graduate Assistant Trainee. I appreciate the support from my three supervisors: Tim Williams, Debra Rubin and Matt Camplin. I also want to thank Dr.Guochen Song, supervisor I interned at Biogen Company, for his great guidance. I always learned new things from him every time we have regular meeting.

I want to express my great thank to my parents, especially my father, Jianyong Tian. Without him, I could not be what I am today. He tried very effort to bring me up, build me the best education environment. I am now growing up, but I can not be at your side and take care of you. Please let wind bring me best wishes to you.

I would also want to say thank you for all my friends, Ms. Yi Chen, Dr. Haoshi Yang, Dr. Junjie Zhao, Ms. Zhidong Chen, Dr. Na Zhang, and the family of Millet, for all your support during my PhD study.

Last, I want to say thank you to my beloved husband, Dr. Gangyao Wang, and my

lovely son, Mr. Andy Wang. You bring me endless support and happiness, and make Raleigh my home.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Logistic regression has been widely used in statistical modeling. Consider the following logistic regression model,

$$y_i|\boldsymbol{\beta} \sim \text{binomial}(n_i, \pi_i(\boldsymbol{\beta})),\ \pi_i = \frac{e^{\boldsymbol{x_i}'\boldsymbol{\beta}}}{1 + e^{\boldsymbol{x_i}'\boldsymbol{\beta}}},\ i = 1, ..., n, \tag{1.1}$$

with a data set of n observations and p predictors. The response y_i is the number of "successes" observed in n_i independent Bernoulli trails with success probability π_i, and x_i is a predictor vector with length p. If $n_i = 1$, all of our data are Bernoulli trials, with either $y_i = 1$ or $y_i = 0$. In this dissertation, we primarily focus on variable selection in logistic regression and an application to reliability assessment.

For variable selection in logistic regression, we are particularly interested in the high-dimensional case where the number of predictors exceeds the sample size, $p > n$. With high-dimensional data, maximum likelihood estimates do not exist, and over-fitting can

be a problem. For reliability prediction, we are interested in estimating the failure probability, π_i, especially in the case that failures are rare. We first address the variable selection problem in logistic regression, focusing on Bayesian approaches, and we then apply this method to reliability assessment.

1.1 Overview of Methods for Model Variable Selection

There has been extensive work on variable selection for linear regression models. One set of methods uses stepwise selection with information criteria such as AIC and BIC [1, 25]. At each step, these approaches assess the model fit penalized by the number of estimated predictors, and thereby trade off goodness of fit and model complexity.

In parallel, other methods, based on penalized likelihood functions, have been proposed. The key idea of penalization is to impose regularization on the likelihood function in order to avoid over-fitting. The parameters $\boldsymbol{\beta}$ are chosen to minimize

$$l(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \nu P_\alpha(\boldsymbol{\beta}) \,, \tag{1.2}$$

where $P_\alpha(\boldsymbol{\beta})$ is a regularization penalty and ν is a tuning parameter that is often chosen by cross-validation, but that may be chosen in many other ways, such as AIC or BIC. With certain choices of regularization, some coefficients may collapse to zero, which makes these methods useful for variable selection. Examples include LASSO [26], group LASSO [27], adaptive LASSO [30], SCAD [8], and the Dantzig selector [6].

Many of these variable selection methods have been extended to logistic regression models. For example, in forward selection [28], variables are added into the model one-

at-a-time, and the score statistic is used to select the next variable to add. Regularization approaches have also been extended for use with generalized linear models. [9] developed a cyclical coordinate descent method to fit the generalized linear model and computed a regularization path. Penalties include the l_1 (LASSO), l_2 (ridge), and a combination of the two, which is known as the elastic net penalty [31]. Formally, the elastic net penalty is, $P_\alpha(\beta) = (1-\alpha)\frac{1}{2}\|\boldsymbol{\beta}\|_2^2 + \alpha\|\boldsymbol{\beta}\|_1$, where $\|\boldsymbol{\beta}\|_2^2 = \sum_{j=1}^{p}\beta_j^2$ is the ridge penalty and $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p}|\beta_j|$ is the LASSO penalty. The LASSO penalty expects more coefficients close to zero, while the ridge penalty shrinks the coefficients toward each other and towards zero. The elastic net can be seen as a compromise between the LASSO ($\alpha = 1$) and ridge penalties ($\alpha = 0$).

Alternatively, Bayesian variable selection methods take a different approach for both linear and logistic regression models. Rather than specifying a single optimal model, they estimate the posterior probability of each possible model. Many Bayesian variable selection methods have been proposed, including the Bayes factor [18, 11] and stochastic search variable selection [10].

Bayes factors are a common metric in Bayesian variable selection [18, 11]. They measure the evidence from the data in support of the null model. However, as in exhaustive search methods, a limitation of Bayes factor methods is that they require searching over a 2^p model space, which is usually not feasible when p is large.

To avoid searching over 2^p sub-models, [10] proposed stochastic search variable selection (SSVS). The basic idea of SSVS is to identify the subsets of predictors having the highest posterior probabilities. Latent variables, $\boldsymbol{I} = (I_1, ..., I_p)$, are introduced into the model to identify subsets: $I_j = 1$ indicates x_j is included in the model. Conditional on the latent variables, $\boldsymbol{I}$, a mixture normal prior is placed on the regression coefficients β_j: $P(\beta_j|I_j) = (1-I_j)N(0, c^2) + I_j N(0, \tau^2)$, with $P(I_j = 1) = p_j$. Here τ^2 is set to a large

value so that the non-zero β_j are included in the final model, and σ_j is set to a small value so that β_j is estimated to be zero if it is not important. SSVS fits the model using Gibbs sampling to find the posterior distribution over the possible subsets.

Another strategy is to specify a shrinkage prior for the model coefficients. Although this does not perform variable selection, since all variables are still in the model, it allows for better estimation and thresholding of small coefficients. As an example, consider the Laplace prior [21], which is a member of the "scale mixture of normals" family. It assumes, conditional on τ_j, that $\beta_j \sim N(0, \tau_j)$, and τ_j follows an exponential distribution with $\tau_j \sim \exp(d)$. The degree of shrinkage is tuned by the hyper-parameter d.

Since the Laplace prior only involves one parameter, it limits the adaptability of the shrinkage. [12] argue that the Laplace prior with a single hyper-parameter fixes the rate of decay of the ordered regression coefficients and leads to unreasonable shrinkage of the parameters. In order to shrink the non-important predictors to zero, one needs to set the mean of the exponential prior to a small value around zero. However, with a fixed decay rate, this may also shrink the true non-zero elements to a value close to zero. More specifically, they show that if $\zeta_j = \frac{\tau_j}{\sum_{j=1}^{p} \tau_j}$ denotes the proportion of total variability explained by the j^{th} predictor, and $\zeta_{(j)}$ is the ordered version of ζ_j, then the ordered proportion decay rate, $\gamma_j = \log \zeta_{(j)} - \log \zeta_{(j+1)}$ does not depend on the hyper-parameter d. This means that the decay rate of the ordered regression coefficients is fixed.

Given this critique, they propose a more generalized form of Laplace prior – the Normal-Gamma prior – to overcome the fixed decay rate problem. The Normal-Gamma prior has the form of $\beta_j \,|\, \tau_j \sim N(0, \tau_j)$ and $\tau_j \sim \text{Gamma}(\lambda, d)$. In this case, the decay of the ordered regression coefficients is determined by the shape parameter λ, which provides more flexibility on the shrinkage of the parameters. Recently a number of global-local shrinkage priors have been proposed. Some examples include the Student-t,

Strawderman-Berger, and Horseshoe prior [23]. Similar to the Normal-Gamma prior, these global-local shrinkage priors allow for varying shrinkage on different coefficients.

1.2 Brief Introduction to Bayesian Reliability Assessment

System reliability assessment is often conducted by considering the reliability of the components and subsystems [14, 16]. We consider the case where binary data are collected for a series system, where the system works only if all components are functioning. We consider a Bayesian approach to assessing system reliability that simultaneously models system- and component-level reliability data. [17] and [13] were the first papers to suggest a fully Bayesian solution to the problem of simultaneously modeling both component- and system-level binary data.

Start by considering system reliability as a single point in time. Consider the example introduced in Chapter 5 from [14] (Table 1.1). There are three components in series. Data are observed for each component and for the system. Let (n_i, x_i) be the binomial observation for component i, and denote the component reliability as R_i; let (n_s, x_s) be the binomial observation for the system, and R_s is the system reliability. Then the likelihood function is written as

$$p(R_1, R_2, R_3) = R_1^{x_1}(1-R_1)^{n_1-x_1} R_2^{x_2}(1-R_2)^{n_2-x_2} R_3^{x_3}(1-R_3)^{n_3-x_3} R_s^{x_s}(1-R_s)^{n_s-x_s} \tag{1.3}$$

where $R_s = R_1 R_2 R_3$. Using the data in Table 1.1, the likelihood function is

$$p(R_1, R_2, R_3) = R_1^8(1-R_1)^2 R_2^7(1-R_2)^2 R_3^3(1-R_3)^1 R_s^{10}(1-R_s)^2 \tag{1.4}$$

Table 1.1: A series system example from [14]

component	Success(x_i)	Failures	Unites Tested (n_i)
1	8	2	10
2	7	2	9
3	3	1	4
System	10 (x_s)	2	12 (n_s)

A fully Bayesian approach requires the specification of a prior $\pi(R_1, R_2, R_3)$. One common approach is to use a product of independent beta distributions [29]. However, the induced system distribution is not in a closed form, and working the inverse problem, where a prior is specified on the system and the component priors are specified as independent beta distributions requires simulation. Another possible prior distribution is the negative log-gamma (NLG) prior [22, 19, 20, 2, 3]. If each component's reliability is specified by negative log-gamma prior, the resulting system prior is available in closed form and is also negative log-gamma distribution.

More specifically, if component reliability $R_i \sim \text{NLG}(\alpha_i, \beta)$, the induced system reliability also follows a NLG prior, $\text{NLG}(\alpha, \beta)$, with

$$\pi(R_s) = \frac{1}{\Gamma(\alpha)\beta^\alpha}[-log(R_s)]^{\alpha-1} R_s^{\frac{1}{\alpha}-1} \tag{1.5}$$

where $\sum_{i=1}^{c} \alpha_i = \alpha$. The closed form of the NLG prior allows it to be used in a straightforward way to specific informative priors for a series system with larger number of

components.

Now suppose that we want to model reliability as changing with time. One approach is to model the reliability of each component using a logistic regression model. For a single component i, let $X_{i,k}$ be the binary response at time point k, where $k = 1, 2, \ldots, T_i$. Let $R_{i,k}(t_{i,k})$ denote the corresponding reliability. Then

$$X_{i,k} \sim \text{Bernoulli}(R_{i,k}(t_{i,k})) \text{ with } \text{logit}(R_{i,k}(t_{i,k})) = a_i + b_i t_{i,k} \tag{1.6}$$

For a Bayesian approach, we specify a prior for a_i and b_i. Priors can be directly specified on (a_i, b_i), or they can be elicited by considering other prior information. [29] considers two approaches for using the NLG distribution for specifying prior information in modeling reliability changing with time.

In Chapter 3, we consider the situation where few (or no) failures are observed for a component and describe a number of prior distributions useful for the analysis.

1.3 Outline

We focus on logistic regression variable selection and reliability assessment from the Bayesian perspective. Chapter 2 extends the Posterior Joint Credible Region approach developed by [5] to the logistic regression model. We develop a Normal-Gamma prior for the regression coefficients. We discuss the shrinkage properties of the prior and propose methods for tuning the hyper-parameters of the prior based on prior beliefs about of the number of important predictors. AIC/BIC is used as the inclusion criterion to select the optimal subset model.

In Chapter 3, we focus on an application of the proposed method to selecting variables

for system reliability assessment. Several factors may affect the system's reliability, and the goal is to identify the important ones. System reliability assessment is viewed as logistic regression variable selection. In our application, we consider the use of both component- and system-level data. With system data, we consider the component data as "missing" and develop methods called *Joint Credible Region within Bayesian Imputation*.

Chapter 4 considers an application of reliability assessment with unbalanced responses. We focus on developing Bayesian prior distributions to make reasonable prediction. We propose two methods of prior elicitation and evaluate the sensitivity of subsequent inference and prediction. Chapter 5 provides conclusions and discussion.

Chapter 2

Bayesian Variable Selection via Joint Credible Region

2.1 Joint Credible Region

The joint credible region approach is proposed for variable selection in the linear model by [5]. In this chapter, we extend the method to the logistic regression model. The idea behind the approach is to build a sequence of posterior credible regions based on a chosen prior. Given a credible region at a particular level, any points within the region can be considered as the potential values for the parameters. The method chooses the model represented by the points within the region that is the most sparse. As the coverage level increases, the model becomes more sparse as the credible region covers more and more space. The expansion of the credible region yields a sequence of credible sets that can be seen as a sequence of models.

Consider the linear regression model, $\boldsymbol{Y} = \boldsymbol{x}^T\boldsymbol{\beta} + \varepsilon$ with a data set of n observations and p predictors. The errors are assumed to be independent and identically

distributed with variance σ^2. The first step for the approach is to build a sequence of posterior credible region by specifying a prior on the full model space. For simplicity, suppose a normal prior is specified for the full model: $\boldsymbol{\beta} \,|\, \sigma^2, \tau \sim N(\mathbf{0}, (\sigma^2/\tau) I_p)$ with τ fixed. The induced posterior distribution on β is elliptical, with the form $P(\boldsymbol{\beta}|\text{Data}) = H[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \Sigma^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})]$, where H is a monotone decreasing function, $\hat{\boldsymbol{\beta}}$ and Σ are the corresponding posterior mean and covariance matrix, obtained by MCMC. If C_α denotes a credible region containing (1-α)×100% of the posterior probability, [5] suggested using a elliptical credible region of the form

$$C_\alpha = \{\boldsymbol{\beta} : (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \Sigma^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \leq K_\alpha\} \text{ for some } K_\alpha, \tag{2.1}$$

to represent the proposed posterior credible region at the α level. They also argued that, although placing a prior distribution on τ no longer maintains the elliptical shape for the posterior distribution, the joint credible region still can be constructed from an elliptical contour having the same form of Equation (2.1).

The next step is to find $\tilde{\boldsymbol{\beta}}$ such that

$$\tilde{\boldsymbol{\beta}} = \text{argmin} \|\boldsymbol{\beta}\|_0, \text{ subject to } \boldsymbol{\beta} \in C_\alpha, \tag{2.2}$$

which gives a candidate model $A_n = \{j | \tilde{\beta}_j \neq 0\}$. However, as pointed out in [5], using the L_0 norm, the optimization solution is not unique. Thus, they replace the L_0 criterion by a combination of the L_0 and L_1 criteria. Then by a local linear approximation [32], the optimization problem becomes

$$\tilde{\boldsymbol{\beta}} = \text{ arg min } \{(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \Sigma^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \lambda_\alpha \sum_{j=1}^{p} \frac{|\beta_j|}{(|\hat{\beta}_j|)^2}\} \,. \tag{2.3}$$

The proposed solution can be considered as a function of λ_α, where there is a one-to-one transformation from λ_α to α. The sequence of solution can be obtained via the LARS algorithm [7] or coordinate optimization [9].

This method is consistent in selection for the fixed-dimensional case, which suggests that the sequence of credible sets will contain the true model with high probability. In addition, it avoids an exhaustive search over the 2^p model space.

2.2 Generalization to Logistic Regression

We extend the method proposed in Section 2.1 to logistic regression variable selection. We propose an elliptical contour for the joint credible region, compute a posterior mean and variance, then find the sparse solution by applying Equation (2.3). The issue we consider is the choice of prior distributions for the logistic regression.

2.2.1 Choice of Prior

The choice of the prior distribution for the parameters of the logistic regression is key for Bayesian inference. As the dimension increases, especially in the case where $p > n$, a non-informative prior would result in unstable performance. Thus, we would like to explore an adaptive prior that has the property of being able to shrink some of the regression coefficients towards zero. A widely used class of priors is the family of *scale mixture of normals* priors. We consider a common member of the family, the Normal-Gamma prior, and a special case of Normal-Gamma prior, the Laplace prior. By varying the choice of hyper-parameters in the Normal-Gamma prior, we can adjust the sparseness of the regression coefficients, which helps achieve the goal of varying shrinkage. Compared with the Laplace prior, the Normal-Gamma prior is more flexible and can achieve better

performance.

The most general form of a scale mixture of normals family can be written as

$$f(\beta \,|\, \xi) = \int \phi(\beta \,|\, \mu, \Sigma) g(\Sigma \,|\, \xi) d\Sigma \,, \tag{2.4}$$

where β are the unknown coefficients, ϕ denotes the normal density, and g is any proper prior density for the variance matrix Σ with hyperparameter ξ. Our choice of Normal-Gamma prior arises if g is a gamma distribution.

We use the following hierarchical prior specification for the logistic regression coefficients $\boldsymbol{\beta}$ in Equation (1):

$$\begin{aligned} \boldsymbol{\beta} \,|\, \boldsymbol{\tau} &\sim N(0, \boldsymbol{B}), \\ \boldsymbol{B} &= diag(\tau_0, \tau_1, ..., \tau_p), \\ \tau_j \,|\, \lambda, d &\sim \text{Gamma}(\lambda, d), j = 1, ..., p, \text{ and} \\ \tau_0, \tau_1, ..., \tau_p, \lambda, d &> 0 \end{aligned} \tag{2.5}$$

where $\text{Gamma}(x \,|\, \lambda, d)$ represents the density of gamma distribution with shape parameter λ and rate parameter d:

$$\text{Gamma}(x \,|\, \lambda, d) = \frac{d^\lambda}{\Gamma(\lambda)} x^{\lambda - 1} \exp\{-dx\} \tag{2.6}$$

Equation (2.5) shows that each regression coefficient β_j has a normal prior distribution conditional on the scale parameter τ_j, which has a gamma distribution specified for the prior. We include an intercept in the model and fix its variance parameter at a known constant ($\tau_0 = 5$), which is flat for a logistic regression model.

The choice of hyper-parameters, λ and d, in the Normal-Gamma prior is important since they play key roles in determining the amount of shrinkage, and thus affect the variable selection results. We explicitly tune the hyper-parameters λ and d by considering their effects on the marginal prior distribution of $\boldsymbol{\beta}$.

Note that the ratio λ/d is the prior mean of τ_j, and λ/d^2 is the prior variance of τ_j. These choices explicitly affect the marginal distribution of β_j. We consider these two effects and want to choose λ and d such that non-zero β_j are significantly different from 0 with high probability, while the "noise" parameters shrink to values close to 0. Recall that $\tau_j \sim \text{Gamma}(\lambda, d)$ and that the prior distribution for β_j is $N(0, \tau_j)$ conditional on τ_j. We want to have large τ_j for non-zero/important coefficients and small values for non-important variables.

2.2.2 Calibration of the Prior Mean

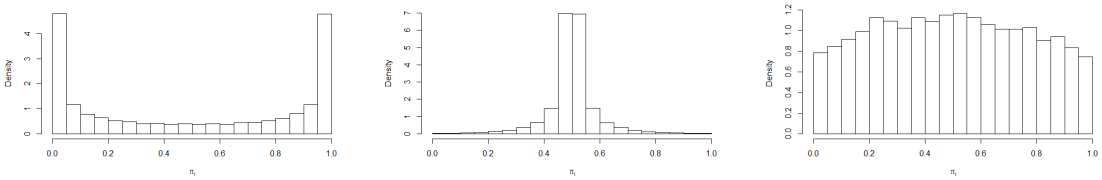
It follows from Equation (1) that the predictive probability for the i^{th} subject is

$$\pi_i = \frac{\exp(x_i'\boldsymbol{\beta})}{1 + \exp(x_i'\boldsymbol{\beta})}$$

The distribution of $\boldsymbol{\pi}$ is an artifact of the prior for $\boldsymbol{\beta}$. There is often some prior knowledge about the success probability of the population. For example, [15] presents a version of the g-prior for $\boldsymbol{\beta}$ such that the prior distribution on the overall population logistic predictive probabilities of success can be set to match a beta distribution Beta(a, b). We adopt the same strategy, and our goal is to model the marginal prior on β to induce a distribution on π_i that matches some Beta(a, b). If no prior information is given on the predictive probability π_i, we use a Beta$(1, 1)$ = Uniform$(0, 1)$ distribution for π_i. In [15], they demonstrated that setting the variance component, τ_j, at $3/p$ induced an

approximately Uniform$(0, 1)$ distribution on the predictive probability. In our setting, with a hyper-prior assigned to τ_j, the variance component itself is a random quantity. The value λ/d determines the mean of this variance component in the prior of $\boldsymbol{\beta}$. We thus set $\lambda/d = 3/p$ to approximately induce a Uniform$(0, 1)$ distribution on π_i. Note that $\tau_j \neq 3/p$, but that it varies around that value.

This can be illustrated by a simple simulation. Figure 2.1 shows the effect of different choices of λ/d on the induced predictive probability distribution. With $n = 200$ and $p = 200$, three choices of λ/d are considered: $\lambda/d = (0.1, 0.015, 0.001)$. A vector of coefficients $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)$ is generated from a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\tau})$, where the variance component $\boldsymbol{\tau} = (\tau_0, ..., \tau_p)$ is drawn independently from a Gamma(λ, d). We repeated this 10,000 times and computed the predictive probabilities. Figure 2.1a corresponds to $\lambda/d = 0.1 > 3/p$, where the induced predictive probability has a U shape, while Figure 2.1b shows $\lambda/d = 0.001 < 3/p$ where the simulated predictive probability has a inverted U shape. Figure 2.1c is $\lambda/d = 3/p$, the case that the predictive probability is approximately uniform.



(a) $\lambda/d = 0.1$ (b) $\lambda/d = 0.001$ (c) $\lambda/d = 0.015$

Figure 2.1: The induced prior predictive probability distribution π_i based on x_i with d=2, $p = 200$

2.2.3 Calibration of the Prior Variance

Varying shrinkage of the coefficients is another consideration for tuning. [12] discussed the effect of changing the prior variance, $\text{Var}(\tau_j) = \lambda/d^2$, on prior beliefs about the proportion of important variables. Let φ be the total variation in $\boldsymbol{\beta}$, where $\varphi = \tau_1 + ... + \tau_p$. Since each τ_j represents the variation in the j^{th} predictor, for the proposed model (Equation 2.5), a small τ_j is likely to shrink the corresponding regressor β_j to a value close to 0, and a large τ_j would suggest a regressor significantly different from 0. If the total variance φ and the prior mean of the variance component τ_j is fixed, then we can adjust each τ_j to control the shrinkage. Note that since the mean is λ/d and variance is λ/d^2, the pair $(\lambda/d, d)$ fully specifies the gamma distribution. We set $\lambda/d = 3/p$ to control the expectation, and now have d to control the variance.

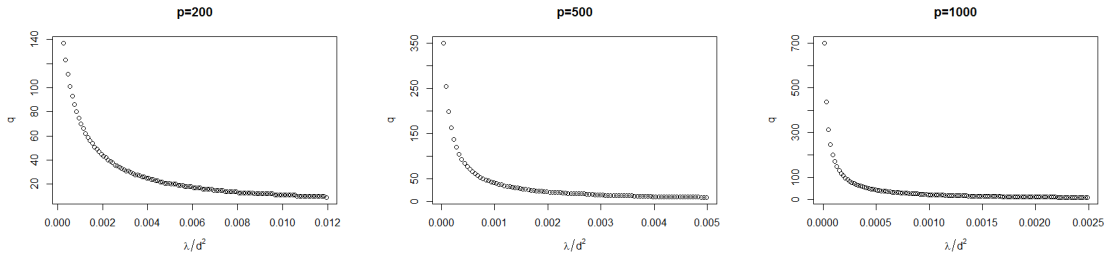
Suppose that we believe the number of important variables is q. If $\tau_{(1)}, ..., \tau_{(p)}$ are the ordered variance components, then $\delta = \sum_{j=1}^{q} \tau_{(j)} / \sum_{j=1}^{p} \tau_j$ can be interpreted as the proportion of the total prior variance in $\boldsymbol{\beta}$ explained by the q largest τ_j. Therefore, to control sparseness, we want most of the variation in $\boldsymbol{\beta}$ to be explained by a small proportion of τ. For a fixed q, we want to choose d satisfy the following equation

$$\text{mean}\left(\frac{\sum_{j=1}^{q} \tau_{(j)}}{\sum_{j=1}^{p} \tau_j}\right) \approx 1 - \varepsilon \ , \tag{2.7}$$

where ε is a small prespecified value and the prior mean of τ_j is fixed at $3/p$.

This technique can be used to specify both "tight" and "smooth" priors. If the choice of hyper-parameters leads to most variance in $\boldsymbol{\beta}$ concentrating on only a few regressors, then we call it a "tight" prior. On the other hand, if the variance in $\boldsymbol{\beta}$ is spread out among more regressors, we consider it a "smooth" prior. Note that, for any p and ε, there is one-to-one correspondence between the degree of sparseness and the value of λ/d^2, and

this relationship is data-independent and monotonic (Figure 2.2).



(a) (b) (c)

Figure 2.2: Relationship of important proportion with prior variance

2.2.4 Laplace Prior

The Laplace prior is a special case of the Normal-Gamma prior with $\lambda = 1$. However, it has less flexibility in determining the degree of sparseness. Since λ is fixed, the only flexibility is determined by d. We choose d to induce an approximately Uniform$(0, 1)$ distribution on the predictive probability by setting $1/d = 3/p$. Then both prior mean $1/d$ and prior variance $1/d^2$ are fixed, which induces a prior on the corresponding number of important variables. Our tuning results show that the Laplace prior is automatically a smooth prior and applies almost equal shrinkage to all of the coefficients, which means that the number of induced important variables q is close to dimension of the predictors.

2.2.5 Inference

Even with the choice of the Normal-Gamma prior, computation is not straightforward. Introducing a latent Pólya-Gamma random variable into the model [24] allows the in-

ference to be directly implemented. With a normal prior specified on the regression coefficients and the latent variable, all updates can be made using Gibbs sampling with closed-form full conditional distributions.

The induced Pólya-Gamma variable is defined as follows. Let $X \sim PG(b, c)$ represent a Pólya-Gamma random variable. Then for $b > 0$ and $c \in \Re$ we have

$$X \stackrel{D}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k - 1/2)^2 + c^2/(4\pi^2)} , \tag{2.8}$$

where $g_k \sim Gamma(b, 1)$ are independent gamma random variables. To write the Pólya-Gamma random variable density function, we have

$$p(x \,|\, b, c) = \frac{\exp(-\frac{c^2}{2}x)p(x|b, 0)}{\mathrm{E}\{\exp(-\frac{c^2}{2}x)\}} , \tag{2.9}$$

where $p(x \,|\, b, 0)$ is the density of $PG(b, 0)$ random variable.

Consider a Bernoulli trial, $n_i = n = 1$, and let y_i denote the trial result, with $y_i \in \{0, 1\}$. Let $x_i = (x_{i1}, x_{i2}, ..., x_{ip})$ be a vector of regression coefficients for $i = 1, ..., n$. Then we have $y_i \sim binomial(1, \pi_i(\boldsymbol{\beta}))$, where $\pi_i = \exp(x_i'\boldsymbol{\beta})/\{1 + \exp(x_i'\boldsymbol{\beta})\}$. To use the Pólya-Gamma strategy, we introduce a sequence of latent Pólya-Gamma random variables $\{\omega_i\}_{i=1}^n$ that are generated independently from the observed response $\{y_i\}_{i=1}^n$ with distribution

$$\omega_i \sim PG(n_i, 0) . \tag{2.10}$$

To sample from the posterior distribution, iterate

$$\begin{aligned} \omega_i \,|\, \boldsymbol{\beta} &\sim PG(n_i, \boldsymbol{x_i}^T \boldsymbol{\beta}), \\ p(\boldsymbol{\beta} \,|\, \boldsymbol{y}, \omega_1, ..., \omega_n, \boldsymbol{\tau}) &\sim N(m, V), \\ p(\tau_j \,|\, \beta_j) &\sim \text{GIG}(\lambda - \frac{1}{2}, 2d, \beta_j^2) \end{aligned} \tag{2.11}$$

where

$$\begin{aligned} V &= (\boldsymbol{B}^{-1} + X^{'} \Omega X)^{-1} \\ m &= V(\boldsymbol{B}^{-1} b + X^{'} \Omega z) \end{aligned} \tag{2.12}$$

with $z = (k_1/\omega_1, ..., k_N/\omega_N)$, $k_i = y_i - n_i/2$, and $\Omega = diag(\omega_1, ..., \omega_N)$. The Generalized Inverse Gamma distribution denotes as $\text{GIG}(p, a, b)$, has the density

$$f(x \,|\, p, a, b) = \frac{(a/b)^{p/2}}{2K_p(\sqrt{ab})} x^{p-1} \exp\{-\frac{1}{2}(ax + b/x)\}, x > 0 \,. \tag{2.13}$$

Because it has two free parameters, the Normal-Gamma prior is more flexible in determining the sparseness of the regression coefficients than Laplace prior. For comparison, we will also consider using a Laplace prior for the regression coefficient. For a Laplace prior, the Gibbs updates for $\boldsymbol{\omega}$ and $\boldsymbol{\beta}$ are the same as those for the Normal-Gamma prior. The updates for τ_j are simplified by sampling $1/\tau_j$ conditional on β_j from an inverse-Gaussian distribution,

$$\frac{1}{\tau_j} \,|\, \beta_j \sim \text{InvGaussian}(\sqrt{\frac{2d}{(\beta_j)^2}}, 2d) \,. \tag{2.14}$$

If $\mu = \sqrt{\frac{2d}{(\beta_j)^2}}$ and $\nu = 2d$, then the inverse-Gaussian distribution can be written as

$$f(x \mid \mu, \nu) = \{\frac{\nu}{2\pi x^3}\}^{1/2} \exp\{-\frac{\nu(x-\mu)^2}{2\mu^2 x}\} \ . \tag{2.15}$$

2.3 Simulation Results

2.3.1 Methods and Metrics

We use several simulation studies to examine the performance of the joint credible region approach and compare it with other variable selection methods: LASSO, forward selection, and screening. We also examine the impact of hyper-parameter selection for the joint credible region approach.

With LASSO, we used a L_1 norm penalized function in the form of $\|\boldsymbol{\beta}\|_1$ to obtain the solution path. Forward selection uses the score test to perform variable selection. More specifically, in the first step, it chooses the variable with largest univariate score statistic. Then the next step is to find the variable that gives the largest score assuming the variable chosen in the previous step stays in the model. The process is repeated until all variables are included in the model. Screening is implemented by marginally performing a two-sample t-test for each predictor with groups defined by the binary response and then ordering the variables based on the absolute magnitude of the t statistics. The joint credible region approach uses both Normal-Gamma and Laplace priors. The posterior is approximated by 10,000 MCMC iterations following 5000 burn-in samples for 15,000 iterations in total. The ordered solution path is computed via obtaining the solution path to the optimization in (2.3).

The hyper-parameters λ and d in the Normal-Gamma prior are tuned as discussed in

Table 2.1: Hyper-Parameter tuning results. The first two columns are tuning result for the Normal-Gamma prior, while the last column is the result for the Laplace prior. The numbers in the table are the prior variance λ/d^2 corresponding to each p, q, while in parenthesis is the prior belief on the important fraction.

λ/d^2(q/p)	Normal-Gamma	Normal-Gamma	Laplace
p=200	0.012 (9/200)	0.002 (40/200)	0.000225 (140/200)
p=500	0.005 (9/500)	0.00102 (40/500)	0.000036 (350/500)
p=1000	0.0025 (9/1000)	0.00054 (40/1000)	0.000009 (700/1000)

Section 2.2. We specify $\varepsilon = 0.05$ so that 95% of the variance in β is to be explained by the q largest regressors. Two beliefs on the number of important variables are considered, $q = 9, 40$ where $q = 9$ is true number of important variables. The tuning results are shown in Table 2.1 for $p = 200, 500, 1000$. From the table, we can see that a small prior variance λ/d^2 is associated with a large number of important variables q.

In order to assess the performance of the approaches, we consider the ordered solution path of predictors. One issue that arises is due to separation, particularly when $p \geqq n$, although it commonly occurs even before that point. Separation is usually associated with categorical data in logistic or probit regression and occurs when only one category's outcome is observed for a region in predictor space. For example, for a continuous predictor x, we might always observe $y = 1$ when $x \geq 0$. This is more likely to happen in high dimensions. When separation occurs, the maximum likelihood estimate does not exist, and both LASSO and forward selection rely on maximum likelihood estimates. Due to the infinite maximum likelihood estimates, the solution path breaks and results in not all predictors being included in the model in the last step. We then use screening to order the remaining variables once this happens.

To compare performance, we use Receiver Operating Characteristic (ROC) and Precision-

Recall (PRC) curves for the ordered solution paths. Let true positives (TP) denote the predictors that are included in the model correctly, and false positives (FP) denote the predictors that are incorrectly included in the model. True negatives (TN) are the variables that are correctly excluded from model, and false negatives (FN) are the variables that are wrongly omitted from the model. The ROC curve describes the trade off between type I error and power. It plots the false positive rate (1-specificity) on the x-axis and true positive rate (sensitivity) on the y-axis, where sensitivity measures the proportion of positives that are correctly identified and specificity is the proportion of negatives that are correctly identified. The PRC shows the trade off between power and false discovery rate and plots the true positive rate (sensitivity) on the x-axis and precision on y-axis, where precision is defined as the ratio of true positives to the total number declared as positives.

In the simulation studies, two different data generating mechanisms are considered: prospective and retrospective data generation. We describe these mechanisms in the following sub-sections.

2.3.2 Prospective Data Generation

For prospective data generation, given the true $\boldsymbol{\beta}$, we use $n = 200$ and vary the dimension of the predictors $p \in \{200, 500, 1000\}$. The correlation between predictors, $x_{i,j}$ and $x_{i,k}$ follows an AR(1) structure, with $Cor(x_{ij}, x_{ik}) = \rho^{|j-k|}$. We choose $\rho = \{0.5, 0.8\}$ to represent relatively low and high correlation. We denote the corresponding variance-covariance matrix by $\Sigma_{p\times p}(\rho)$. For each setting, we generate 250 datasets from a logistic regression model. For a given true $\boldsymbol{\beta}$ and $\Sigma_{p\times p}(\rho)$, we generate $\boldsymbol{x_i} \sim N(\mathbf{0}_{p\times 1}, \Sigma_{p\times p}(\rho))$. Then we have $\pi_i = \exp(x_i^{'}\boldsymbol{\beta})/\{1 + \exp(x_i^{'}\boldsymbol{\beta})\}$ and generate y_i conditional on π_i from a

Bernoulli distribution with success probability $\pi_i(\boldsymbol{\beta})$.

The first case with prospective data generation has the important variables spread out: the important and non-important variables alternate. The important variables occur every eight places starting from the second position with $\beta_j = 1.8$ for $j \in \{2, 10, 18, 26, 34, 42, 50, 58, 66\}$. We have 9 important variables in total.

Figure 2.3 displays the ROC curve and PRC curve for $p = 200$ predictors with a spread signal. The upper two plots show $\rho = 0.8$ and lower two corresponding to $\rho = 0.5$. Under high correlation ($\rho = 0.8$), the joint credible region method using the Normal-Gamma prior shows substantial improvement over the other methods. For low correlation, the performance of all methods improves and gives similar performance except for screening, which does not perform well in either case. This is to be expected, since screening does selection based on the marginal correlation between predictors. High correlation between important predictors helps detect other variables while low correlation does the reverse. When signals are spread out, the correlation between important variables is lower under the AR(1) structure, which affects the performance of screening.

When comparing the choice of priors, both for low and high correlation, the Normal-Gamma priors specified with $q = 9$ perform best, followed by $q = 40$. Although the use of the Normal-Gamma prior tuned with $q = 40$ does not perform as well as the case of $q = 9$, it still outperforms other non-Bayesian methods. The performance of Laplace prior, which suggests more important variables, is not as good as the Normal-Gamma priors with $q = 9$ or $q = 40$.

Figure 2.4 gives the results for $p = \{500, 1000\}$ predictors under high correlation $\rho = 0.8$. The plot shows results consistent with the $p = 200$ case for both $p = \{500, 1000\}$. The joint credible region method demonstrates substantial improvement over the non-Bayesian methods. The prior tuned with $q = 9$ performs best, while Laplace prior does

poorly.

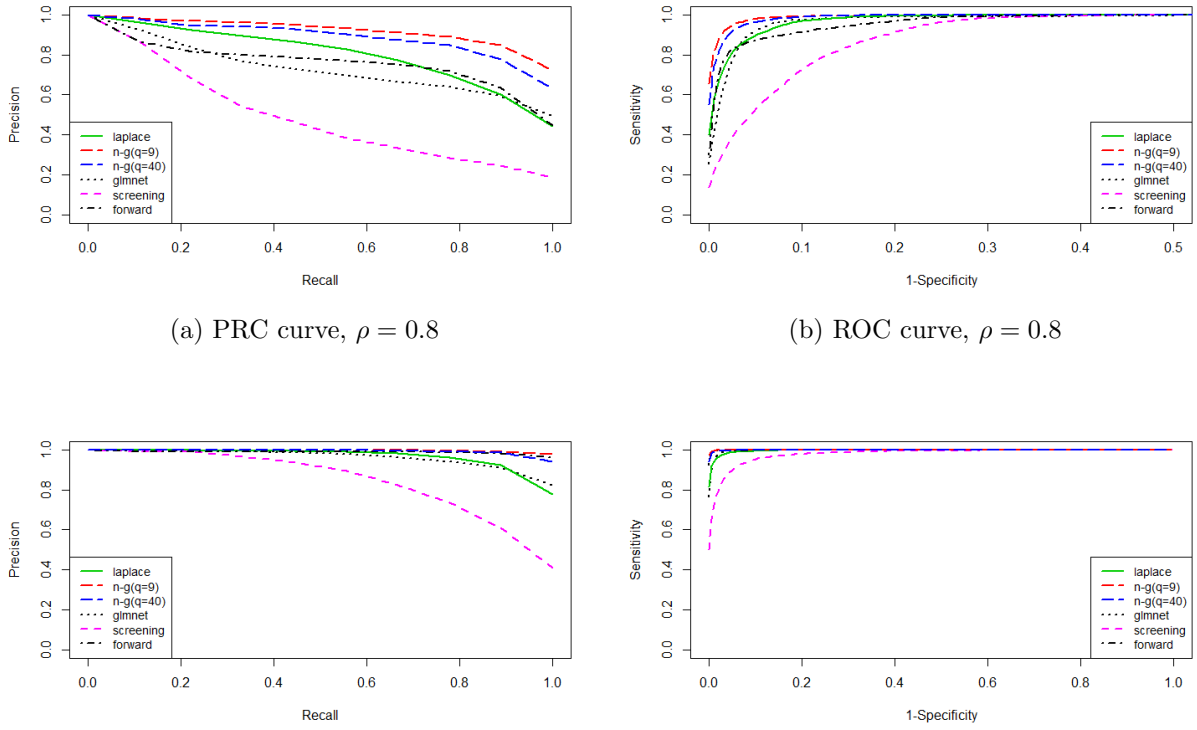
Since the important predictors/signals are spread out, the use of Normal-Gamma prior with $q = 9$, which suggest fewer important predictors helps to push all effects to the true important predictor among variables that are highly correlated. As the shrinkage is not identical across all predictors, it help to identify the important variables while shrinking the non-important predictors around it to a value close to zero. However, the Laplace prior put equal shrinkage to both important and non-important predictors, which results in a worse performance.

We also performed the simulation study for $p = \{500, 1000\}$ in the relatively easier case of $\rho = 0.5$ (results not shown). With low correlation, although all methods perform better, the results are similar to those seen with high correlation.

The other case we consider with prospective data generation uses blocks of important predictors. In this setting, we used three blocks with three non-zero coefficients in each. The true coefficients are given by $(0, 0, 1.8\mathbf{3}^T, \mathbf{0}_{18}^T, 1.8\mathbf{3}^T, \mathbf{0}_{18}^T, 1.8\mathbf{3}^T, \mathbf{0}_{p-46}^T)$, where the subscript denotes the length of the vector. The magnitude of signal we adopt here is same as in the spread signal, with all coefficients positive. We again have 9 important variables.

Figure 2.5 displays the results in this setting with $p = 200$. The upper two plots are for high correlation, while the lower two show results for low correlation. In this case, the Laplace prior gives the best results for both high and low correlation, followed by the Normal-Gamma prior tuned with $q = 40$. Forward selection does poorly for each case.

These results are to be expected. Since important variables are blocked together and highly correlated, the use of the Laplace prior, which is more smooth, gives approximately equal shrinkage to adjacent predictors, avoiding pushing all effect to a single variable. Thus, it can simultaneously pick all the important variables that are correlated with each



(a) PRC curve, $\rho = 0.8$

(b) ROC curve, $\rho = 0.8$

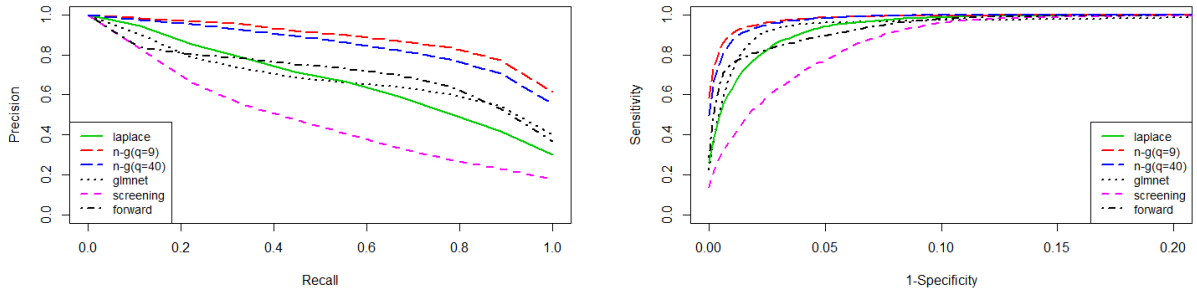
(c) PRC curve, $\rho = 0.5$

(d) ROC curve, $\rho = 0.5$

Figure 2.3: Plot of mean PRC and ROC curves from prospectively generating 250 datasets for $n = 200$, $p = 200$, $\rho = 0.8, 0.5$ with spread signal.

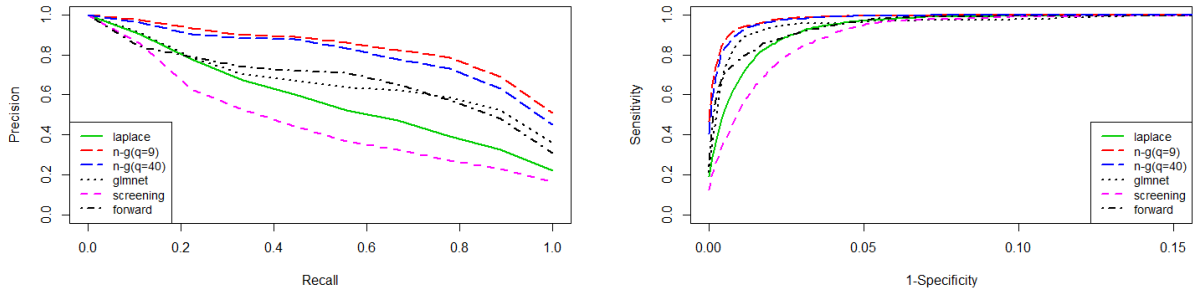
other. The poor performance of forward selection is perhaps due to the correlation of important predictors, since in each step, the method selects variables conditional on the predictors already in the model. It may be hard to select other important variables that highly correlated with the one already chosen.

Figure 2.6 displays the results for $p = \{500, 1000\}$ with high correlation ($\rho = 0.8$). Compared to the $p = 200$ case, we can observe improvement in the performance of the joint credible region method. Again, the Laplace prior outperform other methods. Due to the correlation between important predictors, we expect the poor performance of forward



(a) PRC curve, $\rho = 0.8$, $p = 500$

(b) ROC curve, $\rho = 0.8$, $p = 500$



(c) PRC curve, $\rho = 0.5$, $p = 1000$

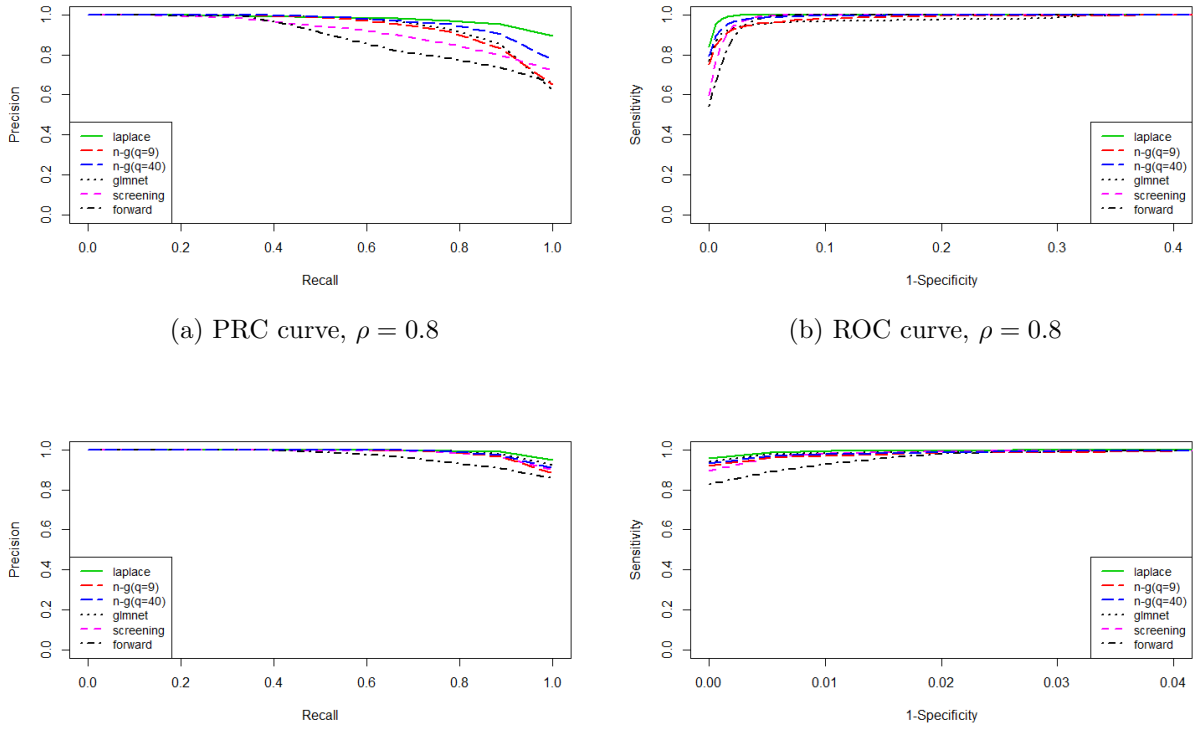
(d) ROC curve, $\rho = 0.5$, $p = 1000$

Figure 2.4: Plot of mean PRC and ROC curve from prospectively generating 250 datasets for $n = 200$, $p = \{500, 1000\}$, $\rho = 0.8$ with spread signal.

selection. We observed consistent result with $\rho = 0.5$ for both $p = \{500, 1000\}$.

2.3.3 Retrospective Data Generation

An alternative way of generating data is to generate predictors conditional on the response, known as retrospective, or case control, data generation. Conditional on the response of $y = 1$ or $y = 0$, we generate two groups of predictors with the same correlation but different mean vectors from multivariate normal distributions. Although the



(a) PRC curve, $\rho = 0.8$

(b) ROC curve, $\rho = 0.8$

(c) PRC curve, $\rho = 0.5$

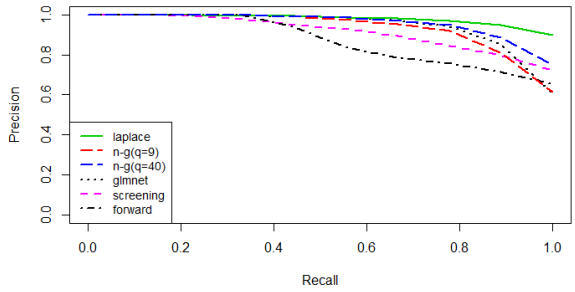
(d) ROC curve, $\rho = 0.5$

Figure 2.5: Plot of mean PRC and ROC curve from prospectively generating 250 datasets for $n = 200$, $p = 200$, $\rho = 0.8, 0.5$ with block signal.

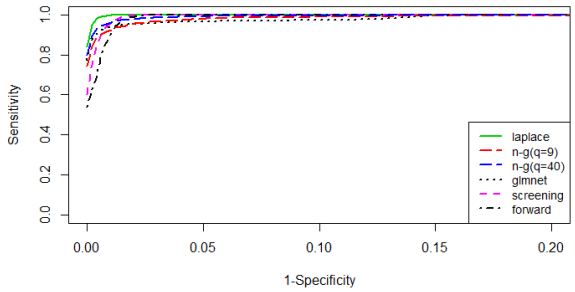
correlation in each group is the same, the marginal distribution of the predictors can be changed according to the magnitude of the shift of the mean vector.

For a given correlation ρ, we again use an AR(1) structure and denote the variance-covariance matrix by $\Sigma_{p\times p}(\rho)$. We generate $\boldsymbol{x_i} \,|\, y = 1 \sim N(\boldsymbol{\mu_1}, \Sigma_{p\times p}(\rho))$, and $\boldsymbol{x_i} \,|\, y = 0 \sim N(\boldsymbol{\mu_0}, \Sigma_{p\times p}(\rho))$, with $\mu_0 = \mathbf{0}_{p\times 1}$. We use an equal proportion of $y = 0$ and $y = 1$, with $\boldsymbol{y} = (\mathbf{1}_{n/2}, \mathbf{0}_{n/2})$. The true $\beta_j, j = 2, ..., p+1$ are given by $\Sigma^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_0})$.

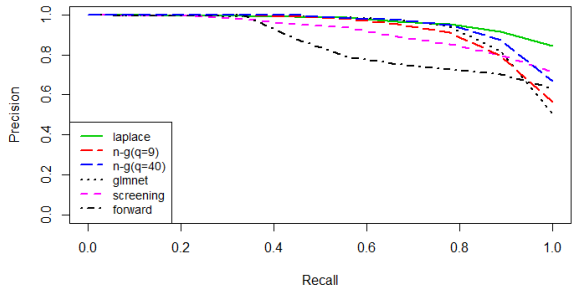
In order to generate sparse coefficients, we randomly choose three non-zero elements in the mean vector $\boldsymbol{\mu_1}$. Due to the AR(1) correlation structure, Σ^{-1} is a banded ma-



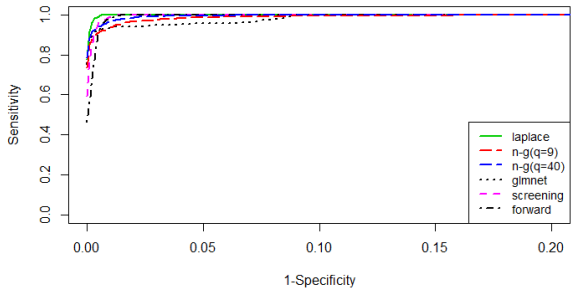
(a) PRC curve, $\rho = 0.8$, $p = 500$



(b) ROC curve, $\rho = 0.8$, $p = 500$



(c) PRC curve, $\rho = 0.8$, $p = 1000$



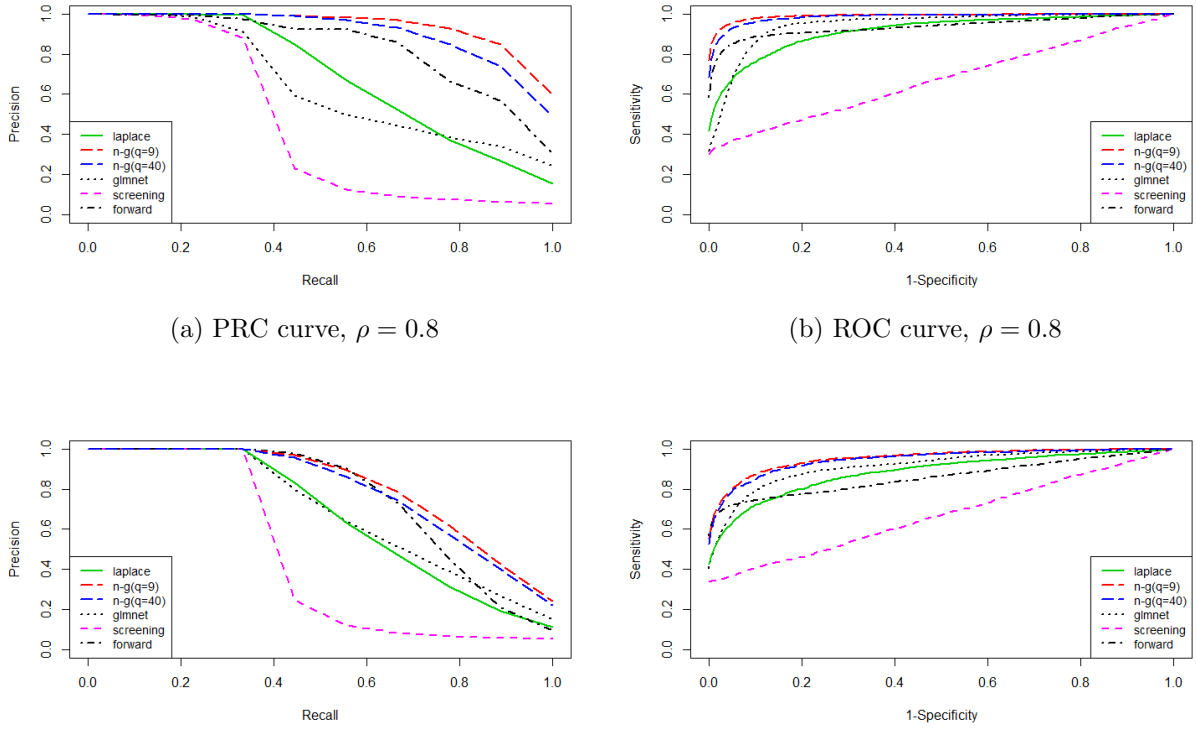
(d) ROC curve, $\rho = 0.8$, $p = 1000$

Figure 2.6: Plot of mean PRC and ROC curve from prospectively generating 250 datasets for $n = 200$, $p = 500, 1000$, $\rho = 0.8$ with block signal.

trix, having non-zero elements on the diagonal and one step off of the diagonal. When multiplied by the mean vector, this results in a coefficient vector with 9 important variables. With $\rho = 0.5$, we use $\boldsymbol{\mu_1} = (0, 0, 1, \mathbf{0}_{20}^T, 1, \mathbf{0}_{20}^T, 1, \mathbf{0}_{p-45}^T)$, which results in $\boldsymbol{\beta} = (0, -1.33, 2.73, -1.33, \mathbf{0}_{18}, -1.33, 2.73, -1.33, \mathbf{0}_{18}, -1.33, 2.73, -1.33, \mathbf{0}_{p-46})$. For $\rho = 0.8$, we use $\boldsymbol{\mu_1} = (0, 0, 0.6, \mathbf{0}_{20}^T, 0.6, \mathbf{0}_{20}^T, 0.6, \mathbf{0}_{p-45}^T)$, which results in $\boldsymbol{\beta} = (0, -0.67, 1.67, -0.67, \mathbf{0}_{18}, -0.67, 1.67, -0.67, \mathbf{0}_{18}, -0.67, 1.67, -0.67, \mathbf{0}_{p-46})$.

When data are generated retrospectively, the coefficients are created in blocks automatically according to the mean vector. It can be observed that although the important

variables are formed in blocks, the induced adjacent variables have opposite signs. Therefore, the effect of adjacent variables may cancel each other out. We would expect the similar selection result as in prospective data generation with spread signal. Figure 2.7 plots the ROC and PRC curves for $p = 200$. The upper two plots are for high correlation ($\rho = 0.8$) and lower two are for low correlation ($\rho = 0.5$). We observe that the Normal-Gamma prior tuned using $q = 9$ performs best. The benefit of the joint credible region approach increases substantially for $\rho = 0.8$. When the dimension of p is increased to $p = \{500, 1000\}$, we see consistent results (Figure 2.8).



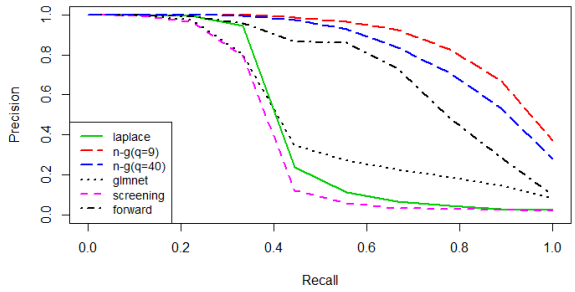
(a) PRC curve, $\rho = 0.8$

(b) ROC curve, $\rho = 0.8$

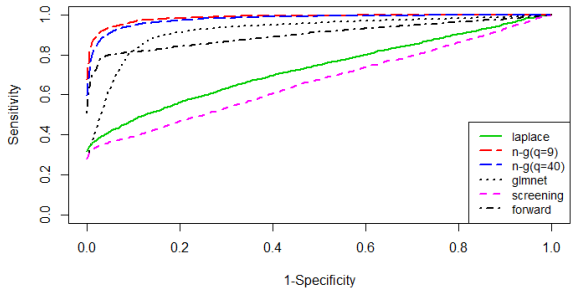
(c) PRC curve, $\rho = 0.5$

(d) ROC curve, $\rho = 0.5$

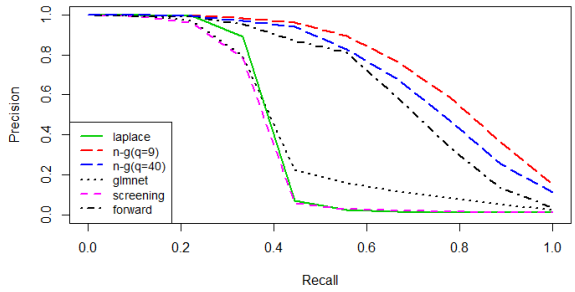
Figure 2.7: Plot of mean PRC and ROC curve from retrospectively generating 250 datasets for $n = 200$, $p = 200$, $\rho = 0.8, 0.5$.



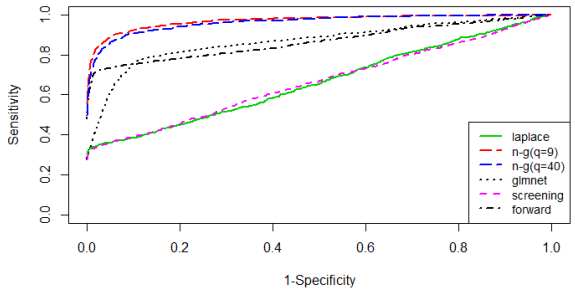
(a) PRC curve, $\rho = 0.8$, p=500



(b) ROC curve, $\rho = 0.8$, p=500



(c) PRC curve, $\rho = 0.8$, p=1000



(d) ROC curve, $\rho = 0.8$, p=1000

Figure 2.8: Plot of mean PRC and ROC curve from retrospectively generating 250 datasets for $n = 200$, $p = 500, 1000$, $\rho = 0.8$.

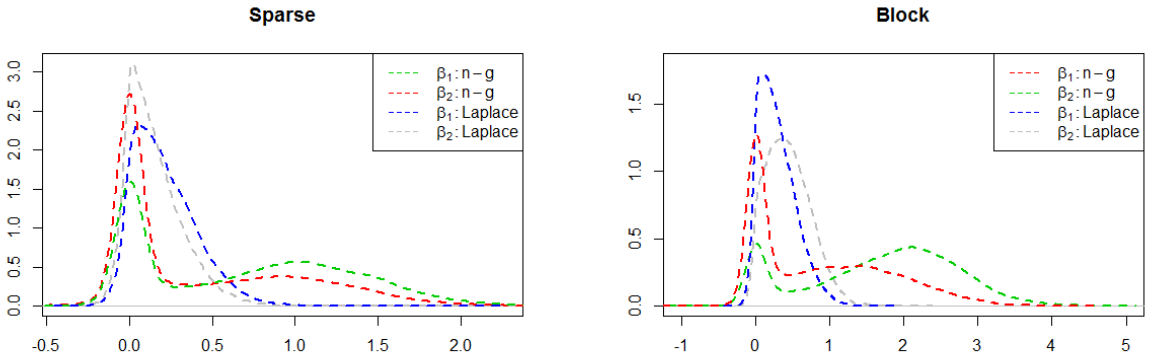
2.3.4 Further Discussion of Adaptive Shrinkage

From the simulation study, we observe that the adaptive shrinkage within the model is determined by both the data and the choice of prior. For the same degree of sparseness, the shrinkage can differ according to the data structure. In this section, we will discuss the effect that the data structure and the prior choice have on the identification of important variables.

With the same number of important variables, we investigated two data structures. The first was "spread," where there is low to no correlation between the important predictors; the second is "block," where the important predictors are highly correlated. A tight prior, developed by tuning a Normal-Gamma distribution, will try to select a single variable from among several important variables that are highly correlated with each other (in a block). This will cause an error, as only one predictor will be selected. By contrast, if there is low correlation among the important variables, a tight prior would be able to make reasonable inference. On the other hand, since the Laplace prior is fairly smooth, it can be expected to work well when the important predictors are in a block.

These effects can also be observed from the posterior distributions in Figure 2.9. This figure shows the posterior densities of two highly correlated predictors, β_1 and β_2, using both the Normal-Gamma and Laplace prior with $n = 200, p = 200$ under different data structures. The Normal-Gamma prior has $q = 9$. Figure 2.9(a) shows the spread situation, where β_1 is important and β_2 is noise. From the figure, it can be observed that the Normal-Gamma prior gives support for β_1 being significantly different from 0, while β_2 has been substantially shrunk toward 0. With the Laplace prior, since it is more smooth, the posterior modes for both β_1 and β_2 are shrunk towards zero. Figure 2.9(b) shows the situation where the important variables are blocked together and both β_1 and β_2 are important. The Normal-Gamma prior selects only one of the important predictors β_1; the posterior mode of β_2 has a peak around zero. However, the Laplace prior gives more even shrinkage to both of β_1 and β_2. In this case, the Laplace prior is preferred.

A tight prior implemented using a tuned Normal-Gamma prior is preferred if the important variables are likely to be spread out; otherwise, the Laplace prior is a better alternative.



(a) (b)

Figure 2.9: Prior effect on posterior distribution. (a): Spread signals. (b): Block signals

2.4 Model Selection via AIC/BIC

In the previous simulation studies, we considered the entire solution path for each method. However, in practice, we want to use the variable selection method to choose a reduced set of variables. For each method, we considered AIC and BIC as criteria for selecting a reduced set and ran an additional set of simulations. We consider both $q = 9$ and $q = 3$ important predictors. The results for $p = 200$ and $\rho = 0.8$, based on 250 datasets, are shown in Tables 2.2-2.5, which report the coverage proportion (the proportion of times that the selected model covers the true model), average model size, and average number of important predictors.

From Tables 2.2-2.5, we can see that for either $q = 9$ or $q = 3$, AIC includes more predictors in the subset than BIC. With $p = 200$, the results show that with spread signals, the joint credible region with a tight prior has better selection properties among all the other methods as measured by smaller average model size, larger coverage propor-

Table 2.2: BIC- and AIC-based selection performance with sparse signal for $n = 200$, p $\in \{200, 1000\}$ based on 250 datasets. The entries in the table are coverage proportion (COV), average model size (MS), and average number of important predictors out of the 9 included (IP).

	BIC						AIC					
	p=200			p=1000			p=200			p=1000		
	COV	MS	IP	COV	MS	IP	COV	MS	IP	COV	MS	IP
n-g(q=9)	49.2	10.18	8.22	39.2	15.0	7.97	70.8	14.74	8.64	44.0	16.56	8.18
n-g(q=40)	47.6	10.88	8.21	29.6	13.67	7.68	73.6	15.11	8.66	38.0	16.32	8.09
Laplace	14.4	11.16	7.29	2.4	12.18	5.6	29.6	14.41	7.9	1.6	15.97	6.33
Foward Selection	29.6	12.42	7.59	9.6	12.1	6.47	34.8	13.53	7.74	14.8	13.82	6.67
Screening	2.0	12.28	4.8	0.0	11.16	4.39	4.0	24	6.7	0.0	15.46	5.13
Glmnet	29.6	12.95	7.59	30.8	15.98	7.6	46.8	15.77	8.19	31.2	17.28	7.82

tion, and average number of important predictors. Screening performs the worst, with a larger model size and fewer important predictors. With block signals, as expected, the Laplace prior displays best selection properties, while forward selection does worse, with low coverage proportion and average important predictors.

When we increase the dimension to $p = 1000$, the selection results are consistent with the $p = 200$ case for both the sparse and block case. Overall, the selection results are consistent with the variable selection results.

2.5 Discussion

Overall, variable selection using the joint credible region approach results in a better performance than other methods, as measured by ROC and PRC curves, for both low and high dimensional cases. The improvement of the joint credible region method over existing methods increases significantly with higher correlation. Specifically, in the situation when data are generated prospectively with a spread signal, we would expect a joint

Table 2.3: BIC- and AIC-based selection performance with sparse signal for $n = 200$, p $\in \{200, 1000\}$ based on 250 datasets. The entries in the table denote coverage proportion (COV), average model size (MS), and average number of important predictors out of the 3 included (IP).

	BIC						AIC					
	p=200			p=1000			p=200			p=1000		
	COV	MS	IP	COV	MS	IP	COV	MS	IP	COV	MS	IP
n-g(q=3)	99.6	5.66	2.99	99.2	7.59	2.99	100.0	12.06	3	100.0	8.72	3
n-g(q=40)	97.6	4.7	2.97	98.8	6.87	2.99	100.0	11.16	3	100.0	8.54	3
Laplace	94.8	3.27	2.9	95.2	4.1	2.95	99.2	8.0	2.99	99.6	4.9	2.99
Foward Selection	96.0	9.9	2.96	93.6	8.72	2.93	100.0	13.31	2.96	96.0	11.49	2.96
Screening	86.8	6.19	2.86	72.0	5.16	2.69	97.2	7.6	2.97	99.6	8.57	2.99
Glmnet	96.8	4.1	2.97	98.8	6.1	2.98	100.00	12.65	3	100.0	16.21	3

Table 2.4: BIC- and AIC-based selection performance with block signal for $n = 200$, p $\in \{200, 1000\}$ based on 250 datasets. The entries in the table denote coverage proportion (COV), average model size (MS), and average number of important predictors out of the 9 included (IP).

	BIC						AIC					
	p=200			p=1000			p=200			p=1000		
	COV	MS	IP	COV	MS	IP	COV	MS	IP	COV	MS	IP
n-g(q=9)	38.4	11.94	7.8	25.2	10.05	7.48	56.0	14.79	8.41	36.4	12.12	8.1
n-g(q=40)	42.0	11.93	7.84	32.4	9.44	7.54	57.6	14.71	8.44	52.8	12.04	8.4
Laplace	46.0	9.48	7.7	29.6	8.29	7.6	88.0	14.47	8.86	72.8	10.59	8.64
Foward Selection	17.6	9.43	6.59	7.6	8.15	5.58	61.6	14.25	8.24	24.0	10.87	7.39
Screening	42.4	11.25	8	19.6	9.05	7.43	77.2	14.14	8.74	54.0	11.17	8.3
Glmnet	44.8	12.39	7.95	23.2	9.56	7.49	60.0	15.43	8.51	40.0	12.28	8.2

Table 2.5: BIC- and AIC-based selection performance with block signal for $n = 200$, p $\in \{200, 1000\}$ based on 250 datasets. The entries in the table denote coverage proportion (COV), average model size (MS), and average number of important predictors out of the 3 included (IP).

	BIC						AIC					
	p=200			p=1000			p=200			p=1000		
	COV	MS	IP	COV	MS	IP	COV	MS	IP	COV	MS	IP
n-g(q=3)	79.6	4.4	2.79	88.4	7.07	2.88	92.4	5.68	2.94	93.6	8.7	2.94
n-g(q=40)	84.8	4.19	2.85	93.6	6.8	2.94	96.8	5.6	2.97	97.6	8.58	2.98
Laplace	84.0	3.2	2.84	88.8	3.22	2.88	99.2	4.76	2.99	99.2	7.14	2.99
Foward Selection	74.8	5.4	2.73	60.8	7.0	2.57	77.2	5.93	2.76	76.0	7.45	2.74
Screening	86.0	3.24	2.86	81.6	3.14	2.82	98.0	3.74	2.98	97.6	4.82	2.98
Glmnet	92.4	4.07	2.92	98.0	6.82	2.98	98.8	5.68	2.98	98.8	8.68	2.98

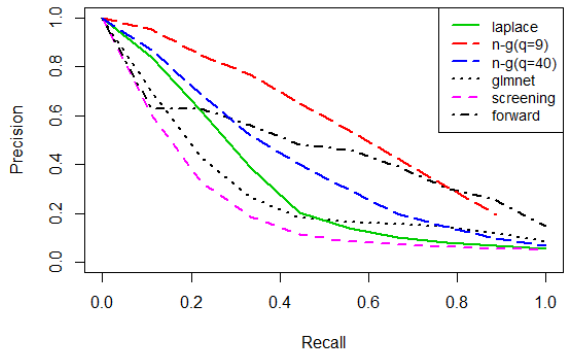
credible region approach with a tight prior to work well, followed by forward selection. The poor performance of screening is due to the low marginal correlation. When the signals are in blocks, the selection results reverse. In this case, we would expect the smooth priors to perform best, while forward selection does poorly. Retrospectively generated data shows similar selection properties as the spread signal case; again, we expect the tight prior and forward selection do the best. This is because the opposite signs of the important variables enable cancellation of much of effect of the adjacent signals.

These results are sensitive to the choice of prior and hyper-parameters. The block signals are better detected using the Laplace prior, while the spread signals prefer a tight prior. In reality, since we don't know which variables are important, we want to try both priors to see if they give similar results. If not, we would rely on expert knowledge to make the selection.

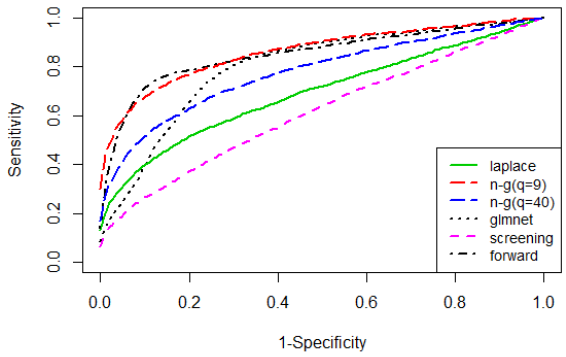
2.6 Additional Simulations

In the previous section, we discussed two methods of data generation – prospective and retrospective. We notice that the retrospective data generation induced opposite signs for adjoining important predictors. We are also interested in an additional situation. Suppose that we prospectively simulate data with positive correlated block signals, but that the signals have opposite signs. Intuitively, we would expect similar performance as retrospective data generation, where we have low correlation between important predictors due to the opposite signs. In this section, we consider additional simulation to examine this hypothesis.

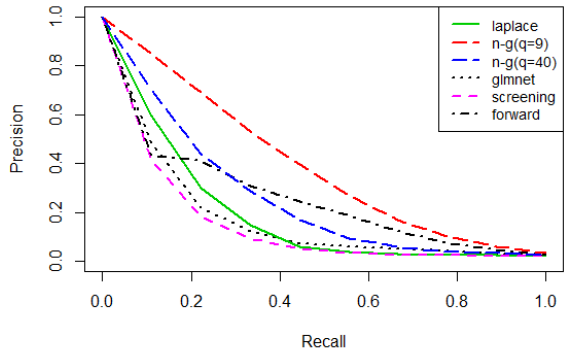
For the simulation design, we again prospectively generate data, but for the important signals, we use exactly the same layout as we used previously for retrospective data generation, with the true coefficients given by $(0, 0, -0.9, 1.8, -0.9, \mathbf{0}_{18}^T, -0.9, 1.8, -0.9, \mathbf{0}_{18}^T, -0.9, 1.8, -0.9, \mathbf{0}_{p-46}^T)$. Here we only consider $p \in (200, 500)$. Figure 2.10 shows the ROC and PRC curves. The simulation results confirm our initial hypothesis, that for both high and low correlation, a tight prior using Normal-Gamma ($q = 9$) performs best, and screening does poorly.



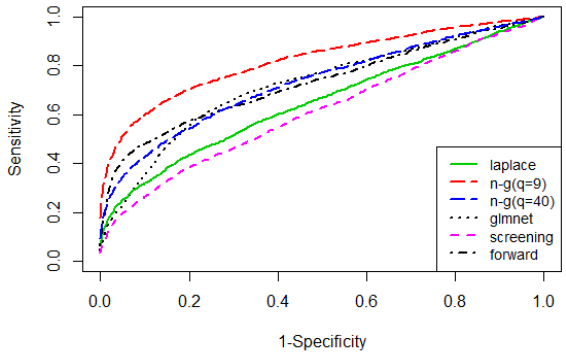
(a) PRC curve, $\rho = 0.8$



(b) ROC curve, $\rho = 0.8$



(c) PRC curve, $\rho = 0.5$



(d) ROC curve, $\rho = 0.5$

Figure 2.10: Plot of mean PRC and ROC curves from prospectively generating 250 datasets with alternating signs for the important variables for $n = 200$, $p = 200, 500$, $\rho = 0.8$. The upper two plots are for $p = 200$ and the lower two plots are for $p = 500$

Chapter 3

Bayesian Reliability Variable Selection with Missing information

3.1 Introduction and Motivating Problem

In this chapter, we consider an application of the method developed in Chapter 2: specifically, variable selection within the assessment of system reliability.

Suppose that we are interested in assessing the reliability of a system. A common approach is to understand the system's structure – frequently series, parallel, or a combination of the two – and to estimate the system's reliability using the components' reliability [14, 16]. A common hypothesis is that only a few of the components contribute to how the system's reliability might change over time.

Suppose that we observe binary (pass/fail) data about the system and the components along with explanatory variables. We are interested in identifying the important explanatory variables for each component – or possibly noticing that none of the explanatory variables are non-zero, which means that a component's reliability is roughly

constant.

If we only have data about the components, several logistic regression models could be fit to the component data. However, if we also have pass/fail data on the full system, and want to use both of component level and system level data, computation becomes more complicated due to the form of the likelihood. In this chapter, we propose a Bayesian method to perform variable selection using both component-level and system-level data that helps address the computational challenges.

3.1.1 Reliability Framework

In order to introduce the problem more specifically, consider the example in Figure 3.1 [13]. This system is comprised of five components. The ways in which the system might fail as a function of these components are represented by a fault tree. The fault tree describes how the system might fail using AND and OR logic gates.

In "OR" logic, if either of the *input* events occurs, the *output event* occurs; in "AND" logic, both of the input events must occur for the output event to occur. The failure of a component is a *basic event*; the event at the top of the tree, in our case system failure, is called the *top event*. In Figure 3.1, there are five basic events that are connected using AND and OR gates to describe how the top event, a system failure, might occur.

It can be difficult to tell from inspection of the fault tree what combinations of event lead to system failure. Mathematically, we can use *minimum cut sets* to represent the fault tree. A *minimum cut set* is a set of components such that if all of the components fail, the system fails; however, if any one of the components functions, the system functions. In the following section, all of the cut sets we consider are minimum cut sets.

To build a representation of the system equivalent to Figure 3.1, we use a reliabil-

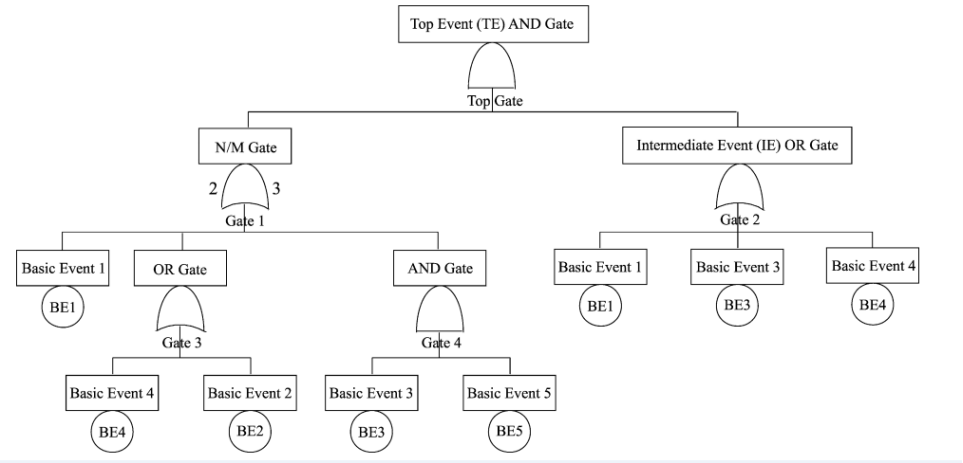


Figure 3.1: An Example of a Fault Tree from [13]

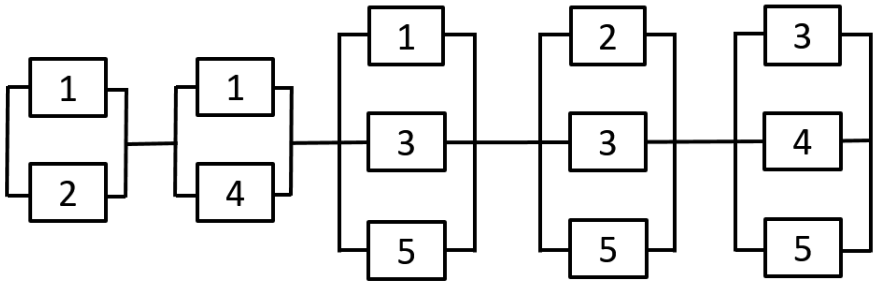


Figure 3.2: Reliability Block Diagram using Minimum Cut Sets

ity block diagram. A *reliability block diagram* is a series of blocks, connected in series or parallel configurations. Parallel path indicate that all components must fail for the system to fail; series paths indicate that any component failure causes a system failure. Our blocks are minimum cut sets. In the example, the system can be decomposed into the following minimum cut sets: $\{BE1, BE2\}, \{BE1, BE4\}, \{BE1, BE3, BE5\}$, $\{BE2, BE3, BE5\}, \{BE3, BE4, BE5\}$. The components in each cut set are put in parallel with each other and then connected as blocks in series (Figure 3.2). Figures 3.1 and 3.2 give the same representation of system failure.

3.1.2 The Model

Continuing with our example, suppose each component has p explanatory variables or predictors. We observe some component-level and some system-level data. Among the $5p$ predictors, only a few are important. Our goal is to develop a method to perform variable selection simultaneously using both the system and component data.

Let $\boldsymbol{x_{i,c}}$ denote the covariates matrix for the i^{th} observation of the c^{th} component. Since for each component, there are p predictors, the component reliability ($\pi_{i,c}$) can be modeled as

$$\pi_{i,c} = \frac{e^{x_{i,c}^{'}\boldsymbol{\beta}_c}}{1 + e^{x_{i,c}^{'}\boldsymbol{\beta}_c}},\ c = 1, ..., 5,\ \boldsymbol{\beta}_c = (\beta_{c,0}, \beta_{c,1}, ..., \beta_{c,p})$$

$\boldsymbol{\beta}_c$ is a p-dimensional vector of regression coefficients for component c. Using the minimum cut set representation from Figure 3.2, system reliability ($\pi_{i,R}$) for observation i is

$$\begin{aligned}\pi_{i,R} = &\{1 - (1 - \pi_{i,1})(1 - \pi_{i,2})\}\{1 - (1 - \pi_{i,1})(1 - \pi_{i,4})\} \\ &\{1 - (1 - \pi_{i,1})(1 - \pi_{i,3})(1 - \pi_{i,5})\} \\ &\{1 - (1 - \pi_{i,2})(1 - \pi_{i,3})(1 - \pi_{i,5})\} \\ &\{1 - (1 - \pi_{i,3})(1 - \pi_{i,4})(1 - \pi_{i,5})\}\end{aligned}$$

Let C denote component-level observations and S denote system-level observations, with $y_{i,c}$ and $y_{i,s}$ denoting the observed outcome (pass/fail) for component and system data. $y_{i,\cdot} = 1$ denoting a "pass" and $y_{i,\cdot} = 0$ denoting a "fail". Then the full likelihood

function can be written as

$$L = \prod_{i \in C} \prod_{c=1}^{5} (\pi_{i,c})^{y_{i,c}} (1 - \pi_{i,c})^{1-y_{i,c}} \prod_{i \in S} (\pi_{i,s})^{y_{i,s}} (1 - \pi_{i,s})^{1-y_{i,s}} \tag{3.1}$$

3.2 Reliability Variable Selection via Penalized Credible Regions

We want to develop a variable selection method for the $5p$ predictors. The reliability variable selection can be framed as a logistic regression variable selection problem where the response is binary to indicate a "pass" or "fail." We will adapt the joint credible region approach proposed in Chapter 2.

For model selection, following the logic in Chapter 2, we again use a shrinkage prior to construct the posterior credible region. In particular, we use the Normal-Gamma prior given as in equation (2.5), but place a prior on each component:

$$\begin{aligned}
\boldsymbol{\beta_c} \,|\, \boldsymbol{\tau} &\sim N(0, \boldsymbol{B}), c = 1, \ldots, 5 \\
\boldsymbol{B} &= diag(\tau_0, \tau_1, \ldots, \tau_p), \\
\tau_j \,|\, \lambda, d &\sim \text{Gamma}(\lambda, d), j = 1, ..., p, \text{ and} \\
\tau_0, \tau_1, \ldots, \tau_p, \lambda, d &> 0
\end{aligned} \tag{3.2}$$

We follow the same principles as in Section 2.2.1 to tune the hyper-parameters λ and d, which control the sparsity of the model. We set $\lambda/d = 3/p$ to match a Uniform $(0, 1)$ distribution on overall predictive probability (Section 2.2.2). Setting $\lambda/d^2 = 1$ captures a prior belief that the number of important predictor is $q = 2$ (Section 2.2.3). The prior

belief of number of important predictor is critical here since it adjusts the sparsity of the model. If we assume only a few important predictors, it leads a sparse model, and vice versa.

With this prior specification and the likelihood function in Equation (3.1), we can sample from the posterior distribution. With the posterior mean $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1, ..., \hat{\boldsymbol{\beta}}_5)$ and variance-covariance $\Sigma = (\Sigma_1, ..., \Sigma_5)$ estimates, we can build a sequence of joint credible region as in Section 2.2 to perform variable selection [5] by solving the following optimization problem using LARS algorithm:

$$\tilde{\boldsymbol{\beta}} = \text{ arg min } \{(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \Sigma^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \lambda_\alpha \sum_{j=1}^{5p} \frac{|\beta_j|}{(|\hat{\beta}_j|)^2}\} \, . \tag{3.3}$$

The proposed solution can be considered as a function of λ_α, where there is a one-to-one transformation from λ_α to α. However, since convergence for a standard MCMC algorithm is slow, we develop another computational approach.

3.2.1 Joint Credible Regions via Imputation

Instead of working with the likelihood function as shown in Equation (3.1), we develop an approach where we consider the system data as observations that are missing component information. If we can impute the component-level data, we can fit a Bayesian logistic regression model in a straightforward way. In particular, by introducing a Polya-Gamma random variable [24] for each component, the computation is simplified to an efficient Gibbs sampler, as in Equation (2.11), for each component.

We impute the missing component data via a Bayesian method. Within the Gibbs sampler, this approach essentially treats the missing data as a model parameter, and iteratively updates the missing data based on its full conditional distribution.

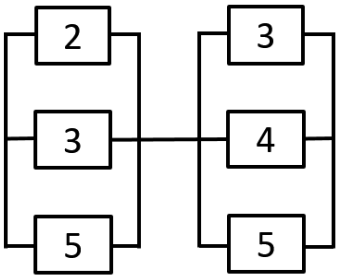


Figure 3.3: The "No-C system" for imputation of component 1

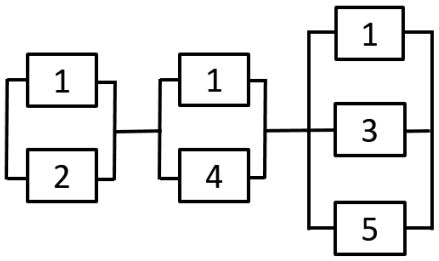


Figure 3.4: The "C system" for imputing component 1

To illustrate the imputation method, we will need three subsystems, which we call the "No-C", "C," and "Hidden-C" systems. Consider the example in Figure 3.2, and suppose we want to impute a value for component c. Let the "No-C system" denote the system that does not include any cuts sets containing component c. Any failure of the "No-C system" implies a system failure.

For example, using Figure 3.2 as an example, if we want to impute a value for component 1, the corresponding No-C system is comprised of $\{BE2, BE3, BE5\}$ and $\{BE3, BE4, BE5\}$ in parallel, as given in Figure 3.3.

Let the "C system" denote the minimum cut sets containing component c in parallel. Figure 3.4 shows the "C system" of component 1.

Let the "Hidden-C system" denote the cut sets containing component c linked in parallel with component c omitted. Consider Figure 3.2 as an example, and suppose

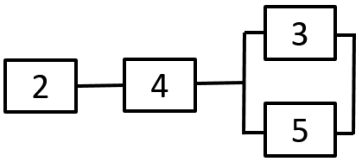


Figure 3.5: The "Hidden-C system" for imputing component 1

that we want to impute component 1. The Hidden-C system is shown in Figure 3.5. If the Hidden-C system fails, but component c passes, then the C-system passes; if the Hidden-C system fails and component c fails, then the C-system fails.

Let $y_{i,c}^t$ be the i^{th} observation of component c in t^{th} iteration of our MCMC algorithm, and β_c^{t-1} the $t-1$st draw of the parameter β_c. We want to update $y_{i,c}^t$. We have the following four scenarios for imputation.

Situation 1. We have observed a system failure. Consider the No-C system, comprised of the cut sets that do not include component c. If all of these cut sets "pass," since they are connected in parallel, the No-C system passes. This means that at least one of the cut sets containing component c failed, which means that component c has failed. Consequently,

$$Pr(y_{i,c}^t = 1) = 0 \tag{3.4}$$

Situation 2. We have observed a system failure. Consider the No-C system. If at least one of these cut sets failed, which means the smaller system failed, we have no information about whether component c failed or not. Consequently, we update using the marginal probability, so

$$Pr(y_{i,c}^t = 1) = \pi_{i,c} = \frac{e^{x_{i,c}^{'}\boldsymbol{\beta}_c^{t-1}}}{1 + e^{x_{i,c}^{'}\boldsymbol{\beta}_c^{t-1}}}, \quad \boldsymbol{\beta}_c^{t-1} = (\beta_{c,0}, \beta_{c,1}, ..., \beta_{j,p}) \tag{3.5}$$

Situation 3. We have observed a success system. This implies that the C-system must pass. Consider the Hidden-C system. If the Hidden-C system fails, then component c must pass for the C-system to pass. This implies that

$$Pr(y_{i,c}^t = 1) = 1 \tag{3.6}$$

Situation 4. We have observed a success system. Consider the Hidden-C system. If the Hidden-C system passes, then we have no information about component c. We update using the marginal probability,

$$Pr(y_{i,c}^t = 1) = \pi_{i,c} = \frac{e^{x_{i,c}^{'}\boldsymbol{\beta}_c^{t-1}}}{1 + e^{x_{i,c}^{'}\boldsymbol{\beta}_c^{t-1}}}, \;\; \boldsymbol{\beta}_c^{t-1} = (\beta_{c,0}, \beta_{c,1}, ..., \beta_{j,p}) \tag{3.7}$$

As an example, consider Situations 3 and 4. We observe that the system passes the test, and we impute the component information for component 1. The cut sets for the entire system are $\{BE1, BE2\}, \{BE1, BE4\}, \{BE1, BE3, BE5\}$, $\{BE2, BE3, BE5\}$, $\{BE3, BE4, BE5\}$. We consider the Hidden-C system, which is shown in Figure 3.5. If BE2, BE4, or BE3 and BE5 fail, then component 1 must pass for the system to pass. If BE2, BE4, and BE3 or BE5 pass, then BE1 can either pass or fail for a system success, and we update its value using its marginal probability.

Then Bayesian updates for $\boldsymbol{\beta}_c, \;\; c = 1, ..., 5$ are followed by the updates for each missing component. We fit the logistic regression model for each component using the Normal-Gamma prior from Equation (3.2). We assign an initial value of either 0 or 1 for each component based on the system's status. If we observe a system success, then we would assign $y_{i,c}^0 = 1$ for $c = 1, \ldots, 5$; if we observe a system failure, we use $y_{i,c}^0 = 0$ for $c = 1, \ldots, 5$.

3.2.2 Selecting Variables

After fitting the model using MCMC, we use the posterior mean and variance-covariance matrix estimates to compute the ordered solution path using Equation (3.3). Since the joint credible regions approach gives us a sequence of models, we need to determine which model to choose, i.e., the stopping criterion. As in Section 2.4, we use AIC/BIC [1, 25] as the criterion, which gives us a sequence of nested models. However, when treating the component data as missing, MCMC is required to re-fit each model to get posterior estimates for the selected predictors. This is time consuming and inefficient.

A more efficient way to compute the AIC/BIC is to us an *approximate AIC/BIC*. Instead of refitting the model every time, we use the full posterior estimates of the model to approximate the AIC/BIC. Given the posterior mean and variance-covariance matrix of the full model, we can approximate the posterior mean of the selected predictors by assuming the other parameters are equal to zero.

The derivation uses the conditional distributions of the multivariate normal distribution. We have $\boldsymbol{\beta}_{1*5p} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let $\boldsymbol{\beta_1}$ denote the set of predictors included in the model and $\boldsymbol{\beta_2}$ corresponds to the predictors temporarily excluded from the model.Write $\boldsymbol{\beta} = (\boldsymbol{\beta_1}, \boldsymbol{\beta_2})$, $\boldsymbol{\mu} = (\boldsymbol{\mu_1}, \boldsymbol{\mu_2})$, and

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

We are interested in the AIC/BIC estimates of $\boldsymbol{\beta}_1$. We have $(\boldsymbol{\beta_1}|\boldsymbol{\beta_2} = \mathbf{0}) \sim N(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}})$,

where

$$\bar{\boldsymbol{\mu}} = \boldsymbol{\mu_1} + \Sigma_{12}\Sigma_{22}^{-1}(0 - \boldsymbol{\mu_2}) \tag{3.8}$$

$$\bar{\boldsymbol{\Sigma}} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \tag{3.9}$$

$\boldsymbol{\mu_1}$, $\boldsymbol{\mu_2}$, and Σ can be estimated using the posterior distribution obtained by MCMC. We can compute the approximate AIC/BIC as

$$AIC = 2 * 5p - 2ln(L(\bar{\boldsymbol{\mu}})) \tag{3.10}$$

$$BIC = ln(n) * 5p - 2ln(L(\bar{\boldsymbol{\mu}}))$$

where $L(\cdot)$ is of the same form as Equation (3.1), by excluding the predictors corresponding to $\boldsymbol{\beta_2}$ from the model.

3.3 Simulation Study and Result

3.3.1 Simulation Settings

We conduct a simulation study to examine the proposed method for variable selection. We generate data from the reliability block diagram in Figure 3.2. We consider both a relatively small and large sample size, with $n = 50, 200$, and let $p = 10$ for each component. Given the truth $\boldsymbol{\beta}_c$, $c = 1, 2, 3, 4, 5$, for each component, we generate predictors $x_{i,j,c}$ following a AR(1) structure, with the correlation between $x_{i,j,c}$ and $x_{i,k,c}$ given by $\rho^{|j-k|}$, with $\rho = 0.8$. Let $\Sigma_{p\times p,c}(\rho)$ denote the variance-covariance matrix for component c. Then we can generate $\boldsymbol{x_{i,c}} \sim N(\mathbf{0}_{p\times 1}, \Sigma_{p\times p}(\rho))$.

We consider two different scenarios of important predictors, with $s = 2, 8$. With

$s = 2$, we use $\beta_{j=2,c} = 1.8$ for $c = 1$, and $\beta_{j=6,c} = 1.8$ for $c = 4$. For $s = 8$, we use $\beta_{j=2,c} = 1.8$ for $c = 1$, $\beta_{j=1,8,c} = 1.8$ for $c = 2$, $\beta_{j=3,c} = 1.8$ for $c = 3$, $\beta_{j=6,c} = 1.8$ for $c = 4$, and $\beta_{j=5,c} = 1.8$ for $c = 5$. For each situation, 250 data sets are generated.

3.3.2 Metrics

In our simulation study, we compare our approach, JCR via imputation, with variable selection by screening. For JCR via imputation, we randomly remove 30% of the component data, and use the corresponding system data together with the "non-missing" component data. For screening, we compute maximum likelihood estimates using Equation (3.1), using the same data as JCR via imputation. We consider the maximum likelihood estimate for the response with one predictor each time, then compute the z statistics. The resulting solution path is ordered by the absolute magnitude of the z statistics.

In order to compare the performance of the two methods, we use Receiver Operating Characteristic (ROC) curves to plot the false positive rate (1-specificity) on the x-axis and true positive rate (sensitivity) on the y-axis. The Receiver Operating Characteristic (ROC) curves are described in Section 2.3.1. We also investigate the selection properties for both JCR via imputation and screening by computing the approximate AIC and BIC along the solution path (Section 3.2.2). We choose the minimum approximate AIC/BIC as the stopping criterion for each data set and compute the correct selection proportion (CS), coverage proportion (COV, the proportion of times that the selected model covers the true model), average model size (MS), and average number of important predictors (IP).

In addition to screening, we also fit the model using JCR with two partial datasets:

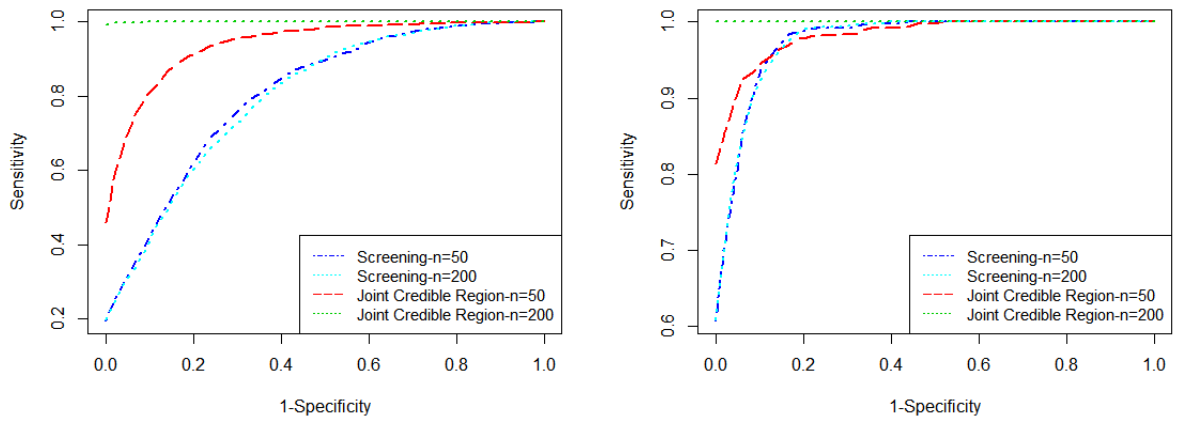
one using only the 70% "non-missing" component data, and one using the full dataset, with all of the component data.

We use a Normal-Gamma prior, as specified in Equation 3.2 for the JCR approach.

3.3.3 Results

Figure 3.6 plots the ROC curve for JCR via imputation and the screening approach with either 2 or 8 predictors based on 250 data sets. From Figure 3.6, we see the superiority of JCR via imputation over screening in either case, and the improvement is even more significant with 8 important predictors (left panel) over 2 important predictors. Both of the methods' performance improved as we increased the sample size from $n = 50$ to $n = 200$. Table 3.1 gives the mean area under the ROC curve and the standard deviation. The result also confirms the better performance of JCR via imputation over screening, in terms of larger mean area under the ROC curve and smaller standard deviation. For example, with $q = 8, n = 50$ important predictors, the area under ROC curve is 0.93 for JCR via imputation versus 0.76 via screening.

Tables 3.2-3.5 show the metrics CS, COV, MS, IP for JCR via Imputation, JCR using partial data, JCR using full data, and screening under different scenarios. We use approximate AIC and BIC as the stopping criterion. From the tables, JCR using full data set performs best, in terms of higher CS and COV and lower MS and IP; followed by JCR via Imputation, JCR using partial data, and screening. This is to be expected, since JCR using full data uses more data than the other methods, followed by JCR via imputation and JCR using partial data. Since the correlation between important predictors is low, and screening performs variable selection based on the marginal correlation, we would expect its poor performance.



(a) 8 important predictors

(b) 2 important predictors

Figure 3.6: ROC curve of JCR via imputation versus Screening. (a) 8 important predictors. (b) 2 important predictors.

3.4 Discussion

In this chapter, we proposed a new method to address variable selection for reliability simultaneously modeling component and system data from a Bayesian perspective. If we only observe system data, we treat the "component data" as missing. We update the "missing" component data iteratively, conditional on the status of the system and other (imputed) components. We derived the full conditional distribution for updating the missing component under all possible situations. We then use the Joint Credible Region approach to perform variable selection. The simulation illustrates the superiority of the proposed method. In the simulation design, we consider a five component system, but the method can be extended to more complicated situations.

Table 3.1: Selection performance for p = 10 (each component) for various choices based on 250 data sets. The entries in the table denote Correct Selection Proportion (CS), Coverage Proportion (COV), Average Model Size (MS), and Average Number of Important Predictors out of the 2 Included (IP).

	8 important predictors		2 important predictors	
	n=50	n=100	n=50	n=100
JCR	0.93 (0.05)	0.99 (0.001)	0.976 (0.045)	1.00 (0.00)
Screening	0.79 (0.06)	0.78 (0.06)	0.96 (0.036)	0.96 (0.035)

Table 3.2: Selection performance for p = 10 (each component) for various choices based on 250 data sets. The entries in the table denote Correct Selection Proportion (CS), Coverage Proportion (COV), Average Model Size (MS), and Average Number of Important Predictors out of the 2 Included (IP).

q=2,n=50	Imputation				Partial				Full				Screening			
	CS	COV	MS	IP	CS	COV	MS	IP	CS	COV	MS	IP	CS	COV	MS	IP
AAIC	9.6	72.0	4.14	1.7	10.8	70.4	4.07	1.65	9.2	84.4	4.14	1.84	2.4	62.0	5.96	1.58
ABIC	39.6	59.2	2.35	1.53	38.0	54.5	2.25	1.46	53.6	72.8	2.29	1.71	7.2	24.8	2.54	1.13

Table 3.3: Selection performance for p = 10 (each component) based on 250 data sets. The entries in the table denote Correct Selection Proportion (CS), Coverage Proportion (COV), Average Model Size (MS), and Average Number of Important Predictors out of the 2 Included (IP).

q=2,n=200	Imputation				Partial				Full				Screening			
	CS	COV	MS	IP	CS	COV	MS	IP	CS	COV	MS	IP	CS	COV	MS	IP
AAIC	8.0	100.0	4.49	2.0	10.0	100.0	4.46	2.0	9.2	100.0	4.6	2.0	6.8	38.8	4.34	1.3
ABIC	84.4	98.8	2.17	1.99	81.2	98.0	2.22	1.98	79.2	100.0	2.26	2.0	12.4	17.6	2.03	1.01

Table 3.4: Selection performance for p = 10 (each component) for various choices based on 250 data sets. The entries in the table denote Correct Selection Proportion (CS), Coverage Proportion (COV), Average Model Size (MS), and Average Number of Important Predictors out of the 8 Included (IP).

q=8,n=50	Imputation				Partial				Full				Screening			
	CS	COV	MS	IP	CS	COV	MS	IP	CS	COV	MS	IP	CS	COV	MS	IP
AAIC	0.8	11.2	9.38	6.14	0.4	8.4	9.13	5.99	9.2	36.0	9.71	6.99	0	1.6	10.38	2.65
ABIC	2.4	4.4	7.09	5.3	1.6	2.0	6.94	5.1	15.2	17.6	7.71	6.32	0	0	6.94	2.96

Table 3.5: Selection performance for p = 10 (each component) for various choices based on 250 data sets. The entries in the table denote Correct Selection Proportion (CS), Coverage Proportion (COV), Average Model Size (MS), and Average Number of Important Predictors out of the 8 Included (IP).

q=8,n=200	Imputation				Partial				Full				Screening			
	CS	COV	MS	IP	CS	COV	MS	IP	CS	COV	MS	IP	CS	COV	MS	IP
AAIC	11.6	98.4	10.35	7.98	14.4	96.8	10.25	7.97	9.2	100.0	10.44	8	0	0	8.43	3.45
ABIC	80.8	91.2	8.15	7.9	76.4	88.8	8.19	7.88	82.2	98.8	8.18	7.98	0	0.4	11.47	4.16

Chapter 4

Prior Distributions for Bayesian System Reliability Assessment

4.1 Example and Motivation

Consider a system where we collect data about the full system, subsystems (collections of components), or individual components. Our data are binary pass/fail measurements made at a known time. Our goals are to assess system reliability and predict it for a short time past the last observation. We model the data using Bayesian logistic regression, which requires that we specify a prior distribution for the regression coefficients. Naive specification of the prior can lead to poor prediction properties both in terms of the point estimates and posterior credible intervals.

In addition, we are interested in the case where our data are unbalanced; specifically, there may be very few failures observed for a particular component. In this case, both non-Bayesian methods and Bayesian inference with naive non-informative priors can lead to poor results. We focus on developing a class of priors that avoids this problem.

To motivate the problem, consider the following example that is derived from the assessment of a weapons system. We have a six-component series system, where the system fails if any component fails. Binary data are collected on each of the six components $C_i, i = 1, \ldots, 6$ at several different times. The data ([29]) are summarized in Table 4.1, where the number of tests for the component/system is followed by the number of failures in parentheses. Notice that much of the data are collected at $t = 0$. In addition to the component data, 72 systems are tested between $t = 0$ and $t = 104$. For the system data, there are 12 failures out of 72 tests. Since we are working with a series system, note that for each of the 12 system failures, we know that at least one of the C_i failed.

Table 4.1: Summary of Test Data. The notation 190(0) means that we observed 0 failures in 190 observations.

component	Age=0	Age 130-154
1	190(0)	92(2)
2	196(0)	77(0)
3	123(2)	82(2)
4	197(1)	79(54)
5	136(3)	82(21)
6	136(1)	82(18)
System	Age 0-104	72(12)

For component data, let y_{ijk} denote the test result for the jth test unit of component i at time point k, where $i = 1, \ldots, 6$; $j = 1, 2, ..., N_i$; $k = 1, 2, ..., T_{ij}$. Let t_{ijk} denote the time at which the k^{th} test on unit j of component i is performed, so there are N_i units of component i, and the j^{th} unit is tested at T_{ij} different time points. Let $p_i(t_{ijk})$ denote the reliability of i^{th} component at time t_{ijk}. Then we have

$$y_{ijk} \sim \text{Bernoulli}(p_i(t_{ijk})) \; . \tag{4.1}$$

We model the reliability $p_i(t_{ijk})$ as

$$\log\left(\frac{p_i(t_{ijk})}{1-p_i(t_{ijk})}\right) = \text{logit}(p_i(t_{ijk})) = \beta_{0i} + \beta_{1i}t_{ijk} \text{ for } i = 1, 2, \ldots, 6 \tag{4.2}$$

For system data, let $y_{s,l}$ denote the test result for the s^{th} unit of system at time point l, $s = 1, ..., 72; l = 1, 2, ..., T_s$. Let $p_s(t_s)$ denote the reliability of s^{th} system at time t_s. Then we have system reliability as

$$p_s(t_s) = \prod_{i=1}^{6} p_i(t_s) \tag{4.3}$$

that $p_i(t_s)$ is modeled through $\log\left(\frac{p_i(t_s)}{1-p_i(t_s)}\right) = \text{logit}(p_i(t_s)) = \beta_{0i} + \beta_{1i}t_s$ The likelihood function is given by

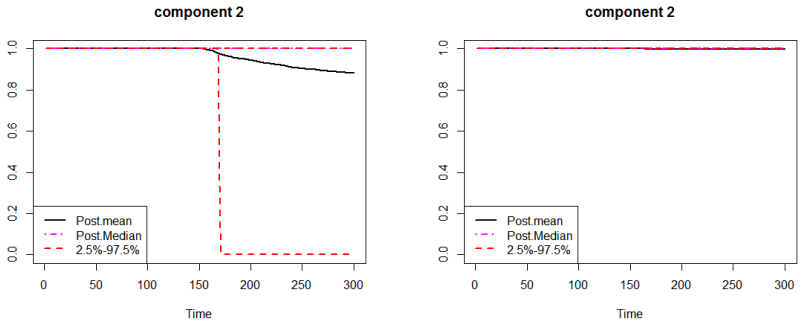
$$L = \prod_i \prod_j (p_i(t_{ijk})^{y_{ijk}}(1 - p_i(t_{ijk})^{1-y_{ijk}} \prod_s (p_s(t_s))^{y_{s,l}}(1 - p_s(t_s))^{1-y_{s,l}} \tag{4.4}$$

4.1.1 The Default Prior

To perform a Bayesian analysis, we must specify a prior distribution for β_{0i} and β_{1i}. Suppose that we choose common non-informative priors, with $\beta_{0i} \sim \text{Normal}(0, 1000^2)$ and either $\beta_{1i} \sim \text{Normal}(0, 10^2)$ or $\beta_{1i} \sim \text{Normal}(0, 1000^2)$.

We fit the model using both the component and the system data. We sample using the Metropolis-Hastings algorithm with a normal proposal distribution using the likelihood function in Equation (4.4). The normal proposal distribution is centered at previous draw. As an example, consider the posterior distribution for the reliability of component 2, $p_2(t)$. Note that for component 2, no failures were observed. However, the 12 system failures may have contained one or more from component 2. Figure 4.1 shows

the mean, median, and 95% central credible intervals, respectively, for the posterior for the reliability of the two priors. From the plot, we can see that the two different priors lead to a considerably different posterior credible intervals for the prediction. With $\beta_{12} \sim \text{Normal}(0, 10^2)$, the credible interval quickly becomes $(0, 1)$ after the last observation at time $t = 154$. With the non-informative prior $\beta_{12} \sim \text{Normal}(0, 1000^2)$, the credible interval stays very close to 1.



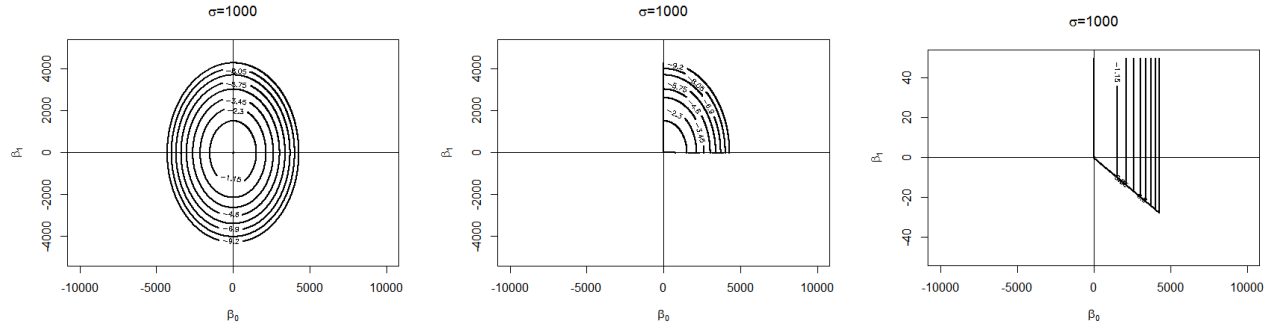
(a) $\beta_{12} \sim \text{Normal}(0, 10^2)$ (b) $\beta_{12} \sim \text{Normal}(0, 1000^2)$

Figure 4.1: Posterior Confidence Interval of Reliability of component 2 using the two priors

4.1.2 Discussion on the Default Prior

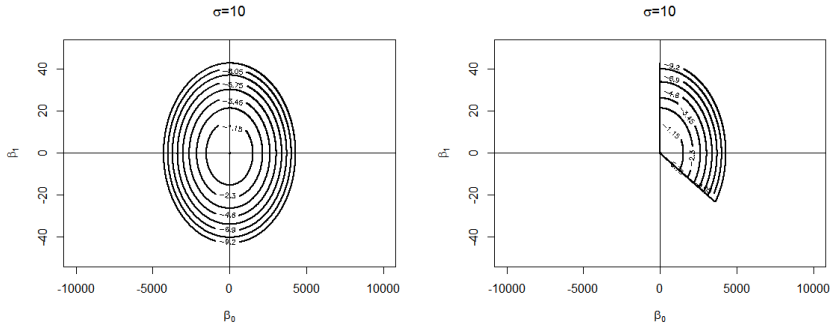
The two "default" non-informative prior distributions lead to very different solutions – neither of which matches our intuition. It does not make sense, given the data, that we cannot predict beyond the last observation, nor does it make sense that the reliability is close to 1 with very high confidence.

In order to understand the results, we plot the joint prior and joint posterior contours



(a) $\beta_{12} \sim \text{Normal}(0, 1000^2)$ (b) $\beta_{12} \sim \text{Normal}(0, 1000^2)$ (c) $\beta_{12} \sim \text{Normal}(0, 1000^2)$

Figure 4.2: Prior and Posterior Distribution with $\beta_{12} \sim \text{Normal}(0, 1000^2)$. (a) Prior Contour. (b) Posterior Contour.(c) Zoomed in Posterior Contour.



(a) $\beta_{12} \sim \text{Normal}(0, 10^2)$ (b) $\beta_{12} \sim \text{Normal}(0, 10^2)$

Figure 4.3: Prior and Posterior Distribution with $\beta_{12} \sim \text{Normal}(0, 10^2)$. (a) Prior Contour. (b) Posterior Contour.

for the regression coefficients β_{02} and β_{12}. Figures 4.2 and 4.3 show the joint prior and joint posterior distribution for the two priors.

From the prior contours, Figures 4.2(a) and 4.3(a), there is mass in all four quadrants. The difference comes in how diffuse the prior is for β_{12}. When $\beta_{12} \sim \text{Normal}(0, 1000^2)$, it has mass over a larger region.

For the posterior contours, Figures 4.2(b) and 4.3(b), we do not expect $\beta_{02} < 0$. Although there is prior mass with $\beta_{02} < 0$, $\beta_{02} < 0$ is inconsistent with the observed data. When $\beta_{02} < 0$, it indicates the reliability at $t = 0$, given by $\exp(\beta_{02})/\{1 + \exp(\beta_{02})\}$, is less than 0.5. For the component-level data, we observe no failures, which suggests high reliability; we observe only 12 system failures in 262 trials, and even assuming they all come from component 2, its reliability is unlikely to be as low as 0.5.

Now consider β_{12}, the slope. The data supports either $\beta_{12} < 0$ or $\beta_{12} > 0$. Examining the posterior contours, we see that points with $\beta_{02} > 0, \beta_{12} > 0$ are consistent with data since the linear combination of $\beta_{02} + \beta_{12}t$ support the reliability having a value around 1.

For $\beta_{12} < 0$, we would expect a large positive β_{02} so that the linear combination of $\beta_{02} + \beta_{12}t$ is sufficiently large at $t = 154$ so that reliability, given as which is $\exp(\beta_{02} + \beta_{12}t)/\{1 + \exp(\beta_{02} + \beta_{12}t)\}$, is consistent with the observed data. If β_{02} is too small, the reliability drops "early" which is not consistent with the data. For both posteriors, we observe the same "fan" shape where $\beta_{12} < 0$ (See Figure 4.2(c).) A difference exists between the two posteriors in the mass in the area where $\beta_{12} < 0$.

Since the $\beta_{12} \sim \text{Normal}(0, 1000^2)$ prior covers more space, the amount of posterior mass in the area with $\beta_{12} < 0$ is much less than for the prior $\beta_{12} \sim \text{Normal}(0, 10^2)$. Therefore, for $\beta_{12} \sim \text{Normal}(0, 1000^2)$, since most of the posterior mass is in the area $\beta_{02} > 0, \beta_{12} > 0$, this leads to posterior credible interval with both bounds near 1. For $\beta_{12} \sim \text{Normal}(0, 10^2)$, since there is more posterior mass in the area of $\beta_{02} > 0, \beta_{12} < 0$, this leads to a posterior credible interval that becomes $(0, 1)$ immediately after the last observation.

Neither of these prior distributions lead to results that are consistent with our scientific expectations, which suggests that our "non-informative" priors have impacts that we did not anticipate. Consequently, we want to investigate ways of specifying a more

informative prior to help with the prediction.

4.2 Fixed Reliability Prior

Based on experience and intuition, we often have information about what we expect to see for the system reliability. We now develop two prior elicitation methodologies: Fixed Reliability and Fixed Time prior distributions. Both of the proposed methods use an "overlap probability," either on time, or predictive probability, as a guide to investigate the sensitivity of the prior.

4.2.1 Prior Elicitation

We want to construct a prior distribution for the regression coefficients of component i, (β_{0i}, β_{1i}). We choose two reliability values and specify a time interval in which we think those reliabilities will occur. For example, suppose that it is believed that the component should be 90% reliable sometime between $t = 200$ and $t = 300$, and that it should drop to 50% reliable sometime around $t = 500$. We can then coerce this information into a prior distribution.

In general, suppose we believe that reliability p_{1i} will occur in the interval (r_1, r_2) and that reliability p_{2i}, with $p_{2i} < p_{1i}$, will occur in (r_3, r_4). We view these intervals as having a width of two standard deviations. We might choose to elicit them by asking for a range of time where the reliability is "likely to occur." We then specify prior distributions for the true (t^*_{1i}, t^*_{2i}), which are the parameters that represent the actual time that the reliability hits p_{1i} and p_{2i}. Based on the prior information, we specify

independent Normal priors for $t^*_{1i}$and t^*_{2i}. So that,

$$\begin{aligned} t^*_{1i} &\sim \text{Normal}(\mu_1, \sigma_1^2) \\ t^*_{2i} &\sim \text{Normal}(\mu_2, \sigma_2^2) \end{aligned} \tag{4.5}$$

where

$$\begin{aligned} \mu_1 &= \frac{r_1 + r_2}{2}, \\ \mu_2 &= \frac{r_3 + r_4}{2}, \\ \sigma_1 &= \frac{r_1 + r_2}{2} - r_1, \\ \sigma_2 &= \frac{r_3 + r_4}{2} - r_3 \end{aligned}$$

Since we can write

$$\begin{aligned} t^*_{1i} &= \frac{c_{1i} - \beta_{0i}}{\beta_{1i}} \\ t^*_{2i} &= \frac{c_{2i} - \beta_{0i}}{\beta_{1i}} \end{aligned} \tag{4.6}$$

where $c_{1i} = \log(\frac{p_{1i}}{1-p_{1i}})$, $c_{2i} = \log(\frac{p_{2i}}{1-p_{2i}})$, the regression coefficients β_{0i}, β_{1i} are a known transformation of t^*_{1i}, t^*_{2i}. Then the joint prior distribution of β_{0i}, β_{1i} is derived as

$$f(\beta_{0i}, \beta_{1i}) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(\frac{c_{1i}-\beta_{0i}}{\beta_{12}} - \mu_1)^2}{2\sigma_1^2}} \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(\frac{c_{2i}-\beta_{0i}}{\beta_{1i}} - \mu_2)^2}{2\sigma_2^2}} \left| \frac{1}{\beta_{1i}^3}(c_{2i} - c_{1i}) \right| . \tag{4.7}$$

4.2.2 Sensitivity Analysis

If we know r_1, r_2, r_3, r_4, then we can specify the prior for β_{0i}, β_{1i} by Equation (4.7). However, we are more interested in the situation that where we may not be completely confident in our specification of r_1, r_2, r_3, r_4. In this section, we examine how changes in the prior distribution for (t^*_{1i}, t^*_{2i}) affect prediction and posterior credible intervals.

Instead of focusing on r_1, r_2, r_3, r_4, we use "overlap probability" for the following analysis. Typically, we expect reliability to decrease as time increases. The "overlap probability" is the probability that reliability stays the same or increases as time increases.

In general, we expect that reliability is monotonically decreasing as a function of time. Since $p_{1i} > p_{2i}$, this would imply that $t^*_{1i} < t^*_{2i}$, although we do not enforce this constraint in the prior, as we allow for the possibility of reliability to increase with time. Define the "overlap probability" P_r as

$$P_r = P_r(t^*_{2i} - t^*_{1i} < 0 | p_{1i} > p_{2i}) \ , \tag{4.8}$$

which is the probability that a later time is associated with a higher reliability. Since t^*_{1i} and t^*_{2i} are random variables, we have $z = t^*_{1i} - t^*_{2i} \sim \text{Normal}(\mu_2 - \mu_1, \sigma_1^2 + \sigma_2^2)$ and

$$\begin{aligned} P_r = P(t^*_{2i} - t^*_{1i} < 0) &= P(z < 0) \\ &= \Phi\left(\frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right) \end{aligned} \tag{4.9}$$

If we assume that $\sigma_1 = \sigma_2 = \sigma$, we have

$$P(t^*_{2i} - t^*_{1i} < 0) = \Phi(\mu_1 - \mu_2/\sqrt{2}\sigma) = P_r \tag{4.10}$$

and

$$\sigma = -\frac{\mu_1 - \mu_2}{\sqrt{2}\Phi^{-1}(P_r)} \ . \tag{4.11}$$

For prior elicitation, suppose we fix μ_1 and μ_2. From Equation (4.11), we find a one-to-one transformation between σ and the overlap probability P_r. In addition, there is also one-to-one transformation between σ (or P_r) and r_1, r_2, r_3, r_4 based on the prior information framework, with

$$\begin{aligned} r_1 &= \mu_1 - \sigma, r_2 = \mu_1 + \sigma \\ r_3 &= \mu_2 - \sigma, r_4 = \mu_2 + \sigma \end{aligned} \tag{4.12}$$

We can then consider the overlap probability, P_r, as the prior probability of reliability increasing with time, and view the sensitivity of our results.

4.3 Fixed Time Prior

4.3.1 Prior Elicitation

Before examining the use of the Fixed Reliability prior, we will consider another specification of a prior distribution for the logistic regression parameters where for any two times, we specify a range of possible reliabilities. Suppose that at time t_{1i}, we believe p_{1i} is in the range (r_1, r_2) and at t_{2i} where $t_{1i} < t_{2i}$, we believe p_{2i} in (r_3, r_4). We will treat these intervals as containing two standard deviations, and we might choose to elicit them by asking for a range of reliabilities "likely to occur" at the given times.

We now use independent beta prior distributions for p_{1i} and p_{2i},

$$p_{1i} \sim \text{Beta}(a, b) \tag{4.13}$$

$$p_{2i} \sim \text{Beta}(c, d) \tag{4.14}$$

To choose the parameters of the beta distributions, we use the center of the range as the mean and the width of the range as two standard deviations. Then,

$$\begin{aligned} \frac{a}{a+b} &= \frac{r_1 + r_2}{2} = m_1 \\ \frac{c}{c+d} &= \frac{r_3 + r_4}{2} = m_2 \\ \frac{ab}{(a+b)^2(a+b+1)} &= (\frac{r_1 + r_2}{2} - r_1)^2 = \sigma_1^2 \\ \frac{cd}{(c+d)^2(c+d+1)} &= (\frac{r_3 + r_4}{2} - r_3)^2 = \sigma_2^2 \end{aligned} \tag{4.15}$$

Since β_{0i} and β_{1i} can be written as functions of p_{1i} and p_{2i}, the prior distribution of (β_{0i}, β_{1i}) is as

$$f(\beta_{0i}, \beta_{1i}) = \frac{e^{(\beta_{0i}+t_{1i}\beta_{1i})a}(1 + e^{\beta_{0i}+t_{1i}\beta_{1i}})^{-a-b}}{B(a, b)} \cdot \frac{e^{(\beta_{0i}+t_{2i}\beta_{1i})c}(1 + e^{\beta_{0i}+t_{2i}\beta_{1i}})^{-c-d}}{B(c, d)} \cdot (t_{2i} - t_{1i}) \tag{4.16}$$

where $B(\cdot, \cdot)$ is the beta function.

4.3.2 Sensitivity Analysis

We are interested in understanding how predictions and posterior credible intervals change using the Fixed Time prior, and, in particular, how these changes relate to the choice of r_1, r_2, r_3, r_4. We again use"overlap probability" to examine the performance of

different choice of parameters.

Suppose that we fix the prior means at m_1 and m_2 and vary the standard deviations $\sigma_1 = \sigma_2 = \sigma$. Using Equation (4.15), we can express the unknown a, b, c, d as a function of σ:

$$\begin{aligned} a &= m_1\{\frac{m_1(1-m_1)}{\sigma^2} - 1\} \\ b &= (1-m_1)\{\frac{m_1(1-m_1)}{\sigma^2} - 1\} \end{aligned} \tag{4.17}$$

and similarly for c and d.

We again define the overlap probability as the probability that reliability is lower at an earlier time. In this case, we have

$$\begin{aligned} P_r = P(p_{2i} > p_{1i} | t_{1i} < t_{2i}) &= E[P(p_{2i} > p_{1i} \mid p_{1i})] \\ &= \int_0^1 \{\int_{p_{1i}}^1 \frac{1}{B(c,d)} p_{2i}^{c-1}(1-p_{2i})^{d-1} dp_{2i}\} \frac{1}{B(a,b)} p_{1i}^{a-1}(1-p_{1i})^{b-1} dp_{1i} \end{aligned} \tag{4.18}$$

After substitution, we now have that σ is the only unknown on the right hand side of Equation (4.18), and for fixed P_r this can be solved. Thus, the overlap probability is a one-to-one function of σ once again.

As with the prior described in Section 4.3.1, we will also vary the overlap probability to examine the effect on the corresponding posterior distribution. For this prior, however, we have to consider that there are constraints on the prior parameters a, b, c, d. In particular, the first constraint is that a, b, c, d all must be positive to be a valid distribution. A second constraint that we will impose on the beta prior is that it is weather monotone or unimodal. We thus rule out the case of the U-shaped Beta distribution in that reliability would be most peaked at both zero and one. As we increase P_r, it will automatically

lead to all of a, b, c, d less than zero numerically. Therefore, there exists a maximum P_r we could reach.

Since the relation between a, b, c, d and the overlap probability is monotonic, instead of changing σ, we can find the extreme value for a, b, c, d that induces the maximum overlap probability and then express σ_1, σ_2 in term of a, b, c, d. This method provides a more flexible way to induce a prior for the unknown parameters since we can loosen the assumption that $\sigma_1 = \sigma_2$.

Consider the prior for p_1 as an example. A large σ_1 would suggest a potential large overlap probability. As a is a monotonically decreasing function of σ_1, we want the smallest a satisfying the two constraints. If $m_1 \geq \frac{1}{2}$, we have $a \geq b$ since $b = \frac{1-m_1}{m_1}a$. The lower bound of a we can have in order to satisfy condition (2) is 1. Summarizing,

$$\begin{aligned} a = 1, b = \frac{1-m_1}{m_1} \text{ if } m_1 \geq \frac{1}{2} \\ a = \frac{m_1}{1-m_1}, b = 1 \text{ if } m_1 \leq \frac{1}{2} \end{aligned} \tag{4.19}$$

Similar results hold for c and d. Using these relationships, we can compute the maximum overlap probability for fixed prior means m_1 and m_2.

4.4 Simulation Results and Recommendations

In this section, we consider the results of a sensitivity analysis for the proposed priors in terms of the overlap probability. We use a 95% central credible interval as the criterion to check the performance of the prior.

Figures (4.4) and (4.5) use the Fixed Reliability prior. With the initial prior information that $p_1 = 0.9$ occurs in a range of times centered at $t_1^* = 200$ and $p_2 = 0.5$

occurs in a range of times centered at $t_1^* = 300$, we consider overlap probabilities $P_r \in (0.001, 0.01, 0.05, 0.3, 0.4, 0.45)$. Each row in the figures corresponds to one value of P_r. In the figures, the first column displays the prior contour, the second column is the posterior contour, and the last column is the posterior credible interval.

From the third column of plot, we can observe that the width of the central 95% credible interval past $t = 154$, the final observed failure, gets larger as the overlap probability increases. The posterior credible interval is "reasonable" until the overlap probability of $P_r = 0.05$. As the overlap probability reaches $P_r = 0.30$, the posterior credible interval becomes $(0, 1)$ immediately after past the last observation at $t = 154$. As P_r continues to increase, the posterior mean and the credible intervals begin to converge towards large reliability values.

To interpret the behaviors of the posterior credible intervals, consider the prior and posterior distributions. Notice that, first, Up to $P_r = 0.05$, most of the prior mass is concentrated in the second and fourth quadrants, with $\beta_0 > 0, \beta_1 < 0$ or $\beta_0 < 0, \beta_1 > 0$. Since the data does not support a negative β_0, which implies many failures at early times, the posterior distribution is concentrated in the area with $\beta_0 > 0, \beta_1 < 0$. Second, as the overlap probability increases, both the prior and posterior of β_0, β_1 is less correlated, and there is more variation between β_0, β_1. Until $P_r = 0.05$, the correlation is moderate. When the overlap probability increases to 0.30, we observe that there is mass in the first quadrant with $\beta_0 > 0, \beta_1 > 0$ and larger variance. This behavior of prior distribution results in some mass in $\beta_0 > 0, \beta_1 > 0$ in the posterior contour, which leads to posterior credible intervals approaching $(0, 1)$.

Figure 4.6 shows the prior and posterior contours and posterior credible interval from the Fixed Time prior. The initial prior information is that at time $t_1 = 200$, the reliability p_1 is centered at $m_1 = 0.9$; at time $t_2 = 300$, the reliability p_2 is centered at $m_2 = 0.5$.

We vary the overlap probability, with $P_r \in \{0.001, 0.05, 0.1\}$ to perform the sensitivity analysis. $P_r = 0.1$ is the maximum overlap we can reach under the constraints. The plots shows that up to the maximum overlap probability of 0.1, the prediction result are still reasonable, with no steep drop in the lower confidence interval, as observed in the use of naive non-informative prior. The prior and posterior contours confirm this behavior: Prior mass is concentrated in $\beta_0 > 0, \beta_1 < 0$, as is the posterior distribution; Until we consider the maximum overlap probability, there is no mass in the first quadrant for both the prior and posterior contours.

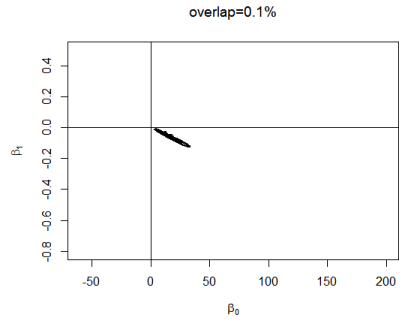
The sensitivity analysis and prior specification depends heavily on the given information. In order to more fully understand the prior specification, we also explore prior elicitation using different initial guesses. We either change the guess of the mean reliability m_1, m_2, or we change the initial time t_1, t_2. Figure 4.7 shows the situation where we fix time $t_1 = 200, t_3 = 300$, but vary the belief on the corresponding mean reliability, with $(m_1, m_2) \in \{(0.9, 0.75), (0.75, 0.5), (0.75, 0.25), (0.5, 0.25)\}$.

Each figure shows results for the maximum overlap probability each scenario can reach. In the Figure 4.7(a)(b) specification, the posterior credible interval is "reasonable," with the lower credible interval bound not dropping sharply immediately after $t = 154$. However, in panels (c) and (d), it appears that the initial prior specification is not reasonable, as the resulting posterior credible intervals quickly move toward $(0, 1)$.

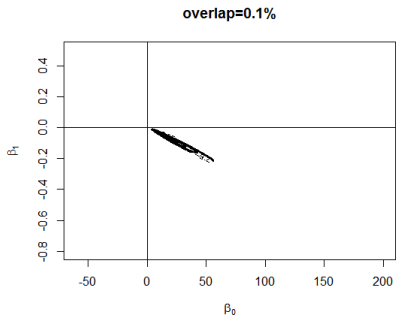
In Figure 4.8, we also change the initial prior belief on time, with $t_1 = 170, t_2 = 250$. This specification suggests that reliability starts to decrease at an earlier time. While the results are consistent with the prior specification, the posterior credible intervals seem to widen too quickly since there have been no observed failures.

Our sensitivity analysis suggests that the proposed prior elicitation might be more useful than naive non-informative priors. For the Fixed Reliability prior, we would sug-

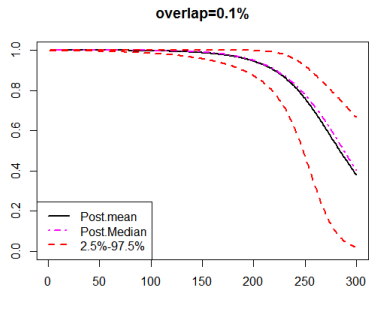
gest $P_r = 0.05$ as the maximum, which seems to lead to a reasonable posterior results. For the fixed time prior, we seem to have robust reasonable results across the range of parameters, which recommends its use.



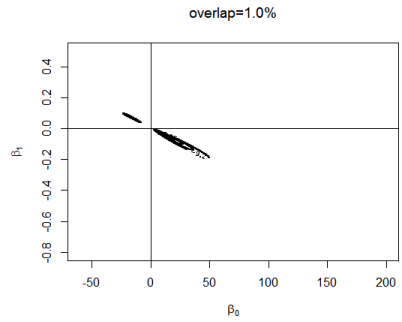
(a) overlap prob of 0.001



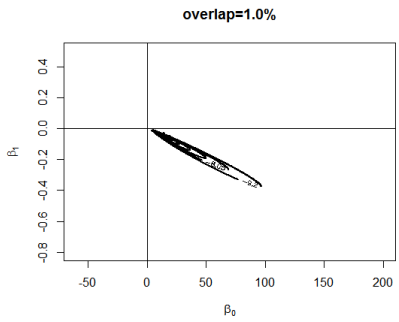
(b) overlap prob of 0.001



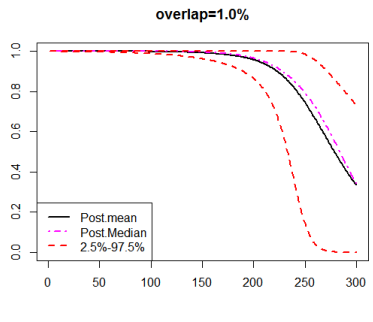
(c) overlap prob of 0.001



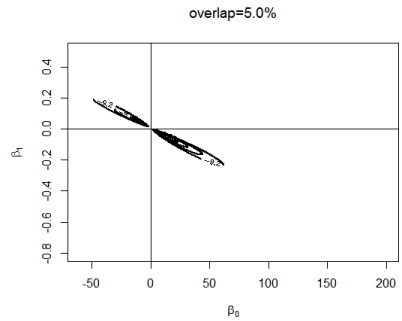
(d) overlap prob of 0.01



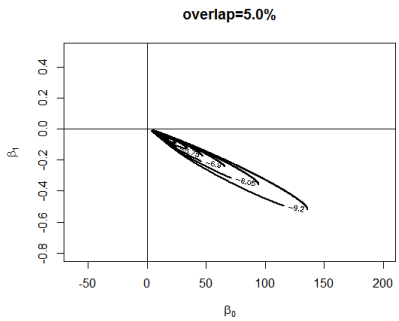
(e) overlap prob of 0.01



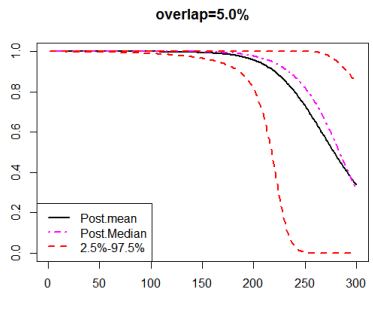
(f) overlap prob of 0.01



(g) overlap prob of 0.05

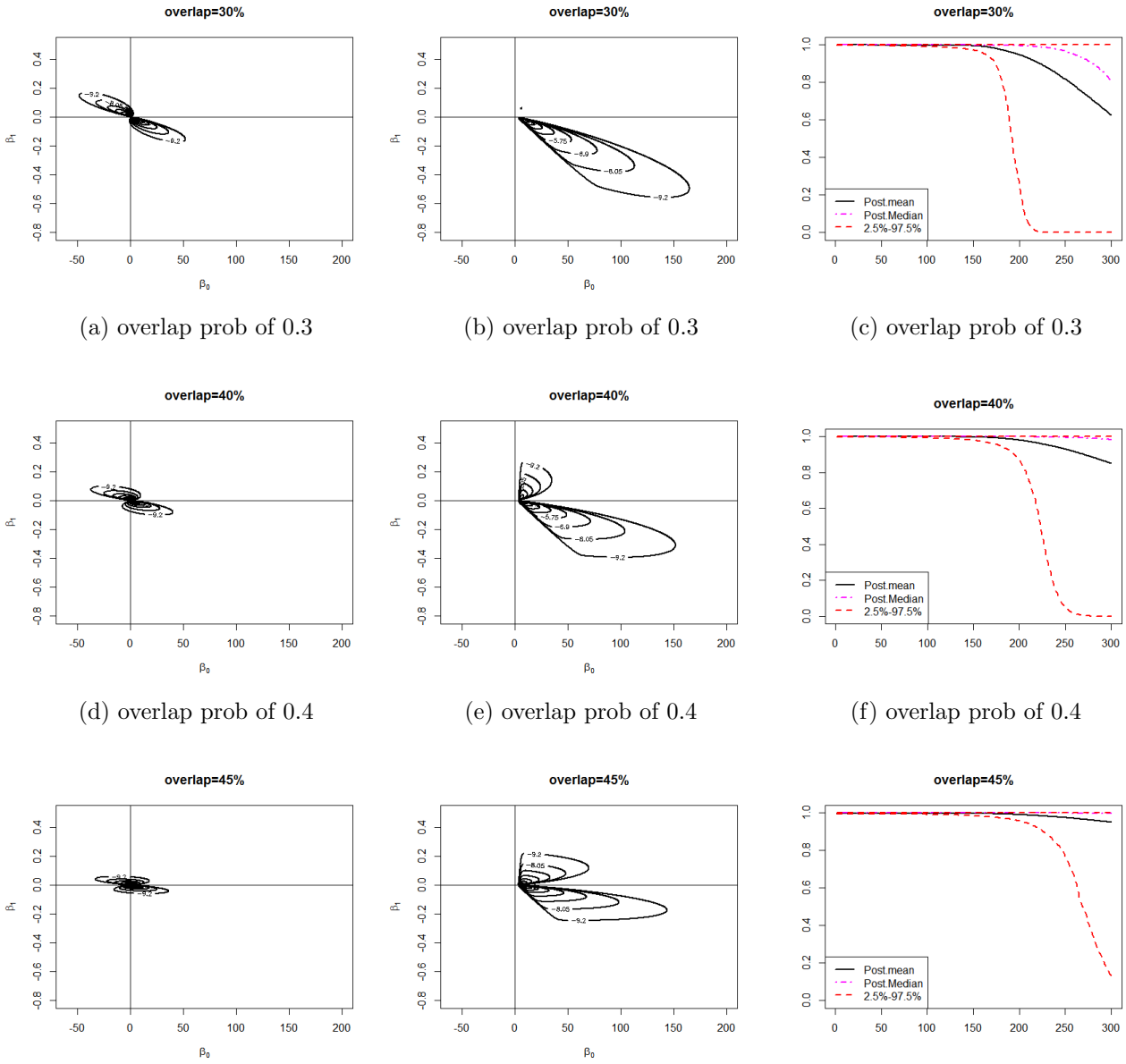


(h) overlap prob of 0.05



(i) overlap prob of 0.05

Figure 4.4: Component 2 posterior confidence interval under different situation. Fixed predictive probability with random time.

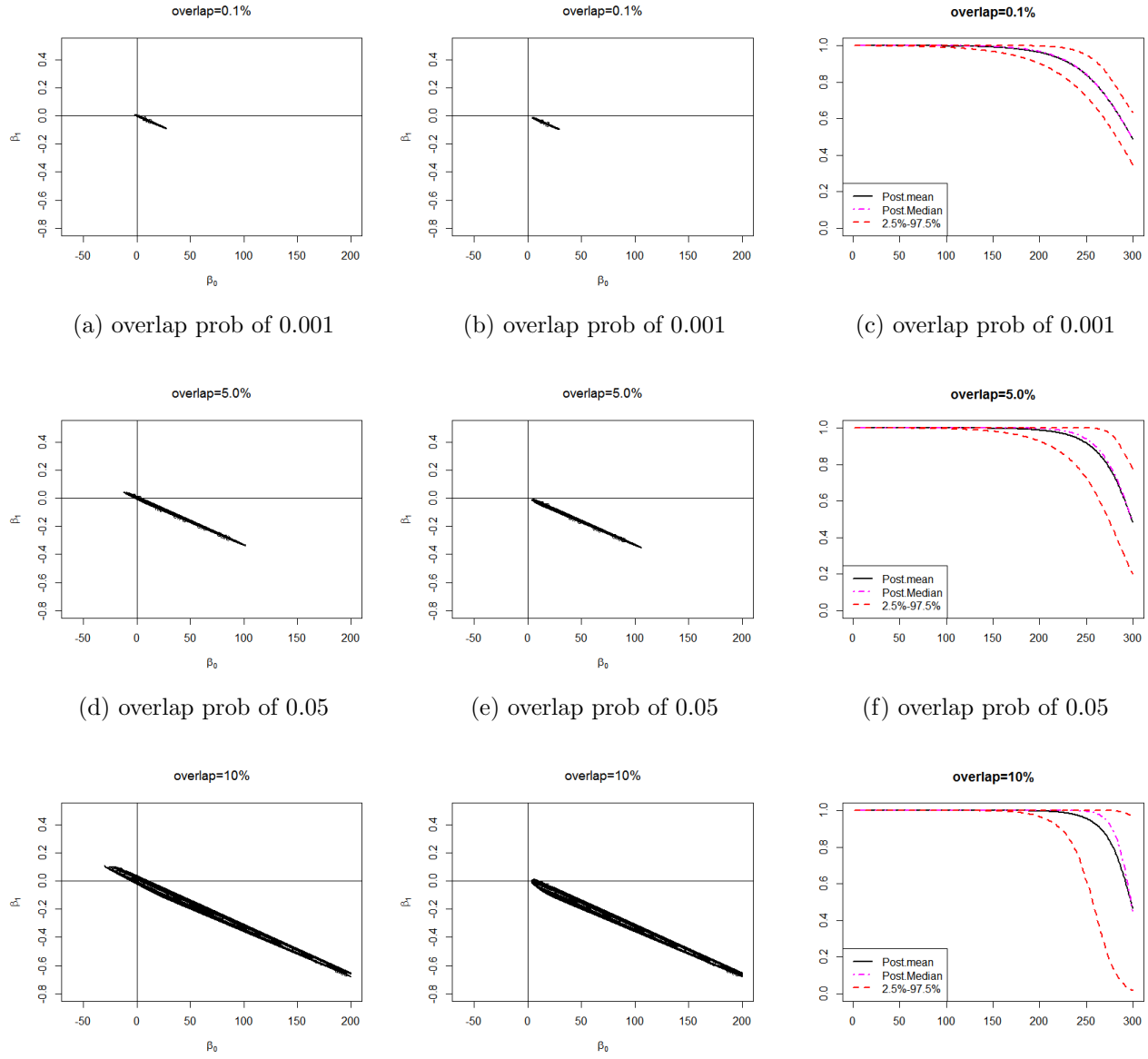


(a) overlap prob of 0.3 (b) overlap prob of 0.3 (c) overlap prob of 0.3

(d) overlap prob of 0.4 (e) overlap prob of 0.4 (f) overlap prob of 0.4

(g) overlap prob of 0.45 (h) overlap prob of 0.45 (i) overlap prob of 0.45

Figure 4.5: Component 2 posterior confidence interval under different situation. Fixed predictive probability with random time.

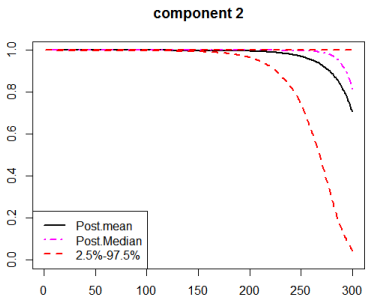


(a) overlap prob of 0.001 (b) overlap prob of 0.001 (c) overlap prob of 0.001

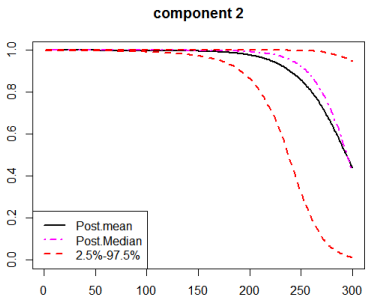
(d) overlap prob of 0.05 (e) overlap prob of 0.05 (f) overlap prob of 0.05

(g) overlap prob of 0.1 (h) overlap prob of 0.1 (i) overlap prob of 0.1

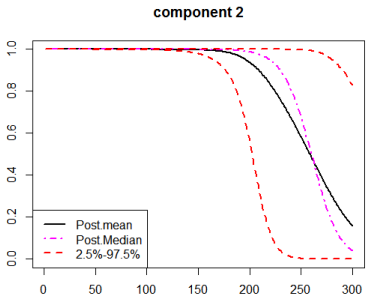
Figure 4.6: Component 2 posterior confidence interval under different situation. Fixed time with random predictive probability.



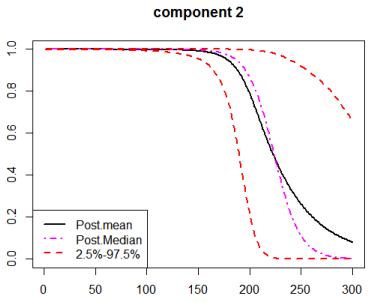
(a) $p_1 = 0.9, p_2 = 0.75$



(b) $p_1 = 0.75, p_2 = 0.5$

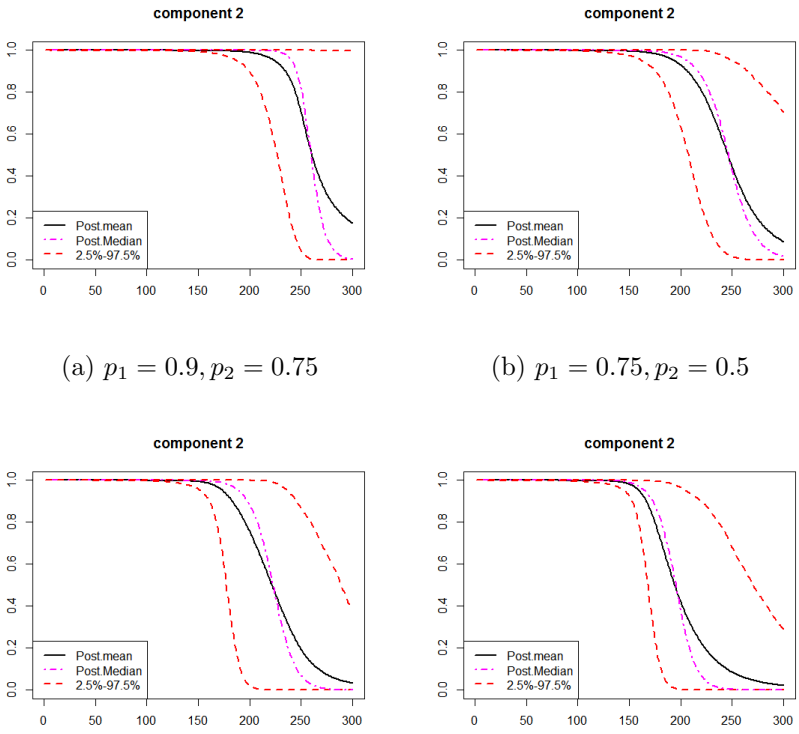


(c) $p_1 = 0.75, p_2 = 0.25$



(d) $p_1 = 0.5, p_2 = 0.25$

Figure 4.7: Fixed time with random predictive probability. We fixed $t_1 = 200, t_2 = 300$ and vary the predictive probability. Each figure plot the extreme situation which gives the most overlap.



(a) $p_1 = 0.9, p_2 = 0.75$ (b) $p_1 = 0.75, p_2 = 0.5$

(c) $p_1 = 0.75, p_2 = 0.25$ (d) $p_1 = 0.5, p_2 = 0.25$

Figure 4.8: Fixed time with random predictive probability. We fixed $t_1 = 170, t_2 = 250$ and vary the predictive probability. Each figure plot the extreme situation which gives the most overlap.

Chapter 5

Contributions and Future Work

This dissertation considers three problems. The first (Chapter 2) focuses on Bayesian logistic regression variable selection. We extend the Joint Credible Region approach [5] to the logistic regression model. We propose the use of the Normal-Gamma prior to construct the Joint Credible Region and develop methodology to tune the hyper-parameters. The prior and hyperparameters assist with the variable selection. In addition, we explore various scenarios by simulation to highlight the influence of correlation between important predictors on the selection results. We demonstrate that the method has superiority over existing logistic regression variable selection methods: LASSO, Forward Selection and Screening, in terms of overall performance and selection properties. We provide recommendations for the use of the new approach.

The second problem (Chapter 3) is an application of logistic regression variable selection. We extend the JCR approach to reliability variable selection. We perform variable selection by simultaneous modeling system-level and component-level data. Instead of focusing on the full likelihood function, as in previous literature, we use imputation to improve the MCMC efficiency. We treat unobserved component data as missing and

simultaneously update the "missing components" based on the system's status. We also consider selection criteria. Instead of AIC/BIC, we propose an approximate AIC/BIC, which improves efficiency and saves computation time. The new method is clearly superior to variable selection by screening.

In the third problem (Chapter 4), we develop two prior distributions to help with the prediction of system reliability at a future time point. We examined the situation where data are unbalanced and the default prior fails to make reasonable predictions. We provide a statistical framework to construct the new prior distributions and demonstrate that they are able to make reasonable prediction for the reliability for at least a few additional time points. We also examine the sensitivity the prediction results to the prior distributions.

For future research, we could consider the following ideas. For logistic regression variable selection, we use a single mode elliptical contour to construct the Joint Credible Region, as in the linear regression model [5]. However, we could consider other possibilities. One possibility to construct the Joint Credible Region is two-mode elliptical contours, which may require heavy computation. For shrinkage purposes, we use the Normal-Gamma prior. One possible extension is to consider a generalized beta mixture prior (Normal-Gamma-Gamma) prior [4] In addition, we assume the predictors are not correlated when developing the prior distribution. Further modifications could be considered to place a correlated prior on the regression coefficients. Finally, we could generalize the approach to Generalized Linear Models and explore the suitability of the method.

For reliability variable selection, we use the Normal-Gamma prior. We assume the predictors within components are uncorrelated and independent across the components. In future work, we could consider the situation where some of components are dependent on other components, and for example, they might share the same predictors.

For reliability assessment, we only consider a very simple case, with only five components in series. For future work, we could extend the framework to more complicated systems that may require the use of minimum cut sets to describe the system. In addition, we only consider one covariate, which is "time." We could extend the work to multiple covariates, perhaps performing variable selection using Joint Credible Region approach before making predictions.

REFERENCES

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[2] F. Allella, E. Chiodo, D. Lauria, and M. Pagano. Negative loggamma distribution for data uncertainty modelling in reliability analysis of complex systems methodology and robustness. *International Journal of Quality & Reliability Management*, 18(3):307–323, 2001.

[3] Flavio Allella, Elio Chiodo, and Davide Lauria. Optimal reliability allocation under uncertain conditions, with application to hybrid electric vehicle design. *International Journal of Quality & Reliability Management*, 22(6):626–641, 2005.

[4] Artin Armagan, David B. Dunson, and Merlise Clyde. Generalized beta mixtures of gaussians. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, pages 523–531, USA, 2011. Curran Associates Inc.

[5] Howard D. Bondell and Brian J. Reich. Consistent high-dimensional bayesian variable selection via penalized credible regions. *Journal of the American Statistical Association*, 107(500):1610–1624, 2012.

[6] Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

[7] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

[8] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[9] J Friedman, T Hastie, and R Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1–22, 2010.

[10] Edward I. George and Robert E. McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

[11] Steven N Goodman. Toward evidence-based medical statistics. 2: The bayes factor. *Annals of Internal Medicine*, 130(12):1005–1013, 1999.

[12] Jim E. Griffin and Philip J. Brown. Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188, 2010.

[13] M. Hamada, H. Martz, C. S. Reese, T. Graves, V. Johnson, and A. Wilson. A fully Bayesian approach for combining multilevel failure information in fault tree quantification and corresponding optimal resource allocation. *Reliability Engineering and Systems Safety*, 86(3):297–305, 2004.

[14] M S Hamada, A Wilson, C S Reese, and H F Martz. *Bayesian Reliability*. Springer, 2008.

[15] Timothy E. Hanson, Adam J. Branscum, and Wesley O. Johnson. Informative g -priors for logistic regression. *Bayesian Analysis*, 9(3):597–612, 2014.

[16] A. Hoyland and M. Rausand. *System Reliability Theory : Models and Statistical Methods*. New York : J. Wiley & Sons, 2008.

[17] V Johnson, T Graves, M Hamada, and C. S Reese. A hierarchical model for estimating the reliability of complex systems. In J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West, editors, *Bayesian Statistics 7*, pages 199–213. Oxford:Oxford University Press, 2003.

[18] R. E Kass and A. E Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773795, 1995.

[19] David V. Mastran. Incorporating component and system test data into the same assessment: A Bayesian approach. *Operations Research*, 24(3):491–499, 1976.

[20] David V. Mastran and Nozer D. Singpurwalla. A Bayesian estimation of the reliability of coherent structures. *Operations Research*, 26(4):663–672, 1978.

[21] Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[22] J. B. Parker. Bayesian prior distributions for multi-component systems. *Naval Research Logistics Quarterly*, 19(3):509–515, 1972.

[23] Nicholas G. Polson and James G. Scott. Shrink globally, act locally: Sparse Bayesian regularization and prediction. In J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West, editors, *Bayesian Statistics 9*. Oxford Univ. Press, 2010.

[24] Nicholas G. Polson, James G. Scott, and Jesse Windle. Bayesian inference for logistic models using plyagamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349, 2013.

[25] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[26] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58:267–288, 1996.

[27] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society (Series B)*, 68:49–67, 2006.

[28] Dietmar Zellner, Frieder Keller, and Gnter E. Zellner. Variable selection in logistic regression models. *Communications in Statistics - Simulation and Computation*, 33(3):787–805, 2004.

[29] R. Zoh, A. Wilson, S. VanderWiel, and E. Lawrence. Using the negative log-gamma distribution for system reliability assessment. *Journal of Risk and Reliability*, 0(0):1748006X17692154, 2017.

[30] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

[31] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[32] Hui Zou and Runze Li. One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, 36(4):1509–1533, 2008.