

## ABSTRACT

Title of dissertation:      VARIABLE SELECTION PROPERTIES  
   OF L1 PENALIZED REGRESSION IN  
   GENERALIZED LINEAR MODELS

Chon Sam, Doctor of Philosophy, 2008

Dissertation directed by:   Professor Paul J. Smith

A hierarchical Bayesian formulation in Generalized Linear Models (GLMs) is proposed in this dissertation. Under this Bayesian framework, empirical and fully Bayes variable selection procedures related to Least Absolute Selection and Shrinkage Operator (LASSO) are developed. By specifying a double exponential prior for the covariate coefficients and prior probabilities for each candidate model, the posterior distribution of candidate model given data is closely related to LASSO, which shrinks some coefficient estimates to zero, thereby performing variable selection. Various variable selection criteria, empirical Bayes (CML) and fully Bayes under the conjugate prior (FBC\_Conj), with flat prior (FBC\_Flat) a special case, are given explicitly for linear, logistic and Poisson models. Our priors are data dependent, so we are performing a version of objective Bayes analysis.

Consistency of  $L_p$  penalized estimators in GLMs is established under regularity conditions. We also derive the limiting distribution of  $\sqrt{n}$  times the estimation error for  $L_p$  penalized estimators in GLMs.

Simulation studies and data analysis results of the Bayesian criteria mentioned

above are carried out. They are also compared to the popular information criteria, Cp, AIC and BIC.

The simulations yield the following findings. The Bayesian criteria behave very differently in linear, Poisson and logistic models. For logistic models, the performance of CML is very impressive, but it seldom does any variable selection in Poisson cases. The CML performance in the linear case is somewhere in between. In the presence of a predictor coefficient nearly zero and some significant predictors, CML picks out the significant predictors most of the time in the logistic case and fairly often in the linear case, while FBC\_Conj tends to select the significant predictors equally well in all linear, Poisson and logistic models. The behavior of fully Bayes criteria depends strongly on their chosen priors for the Poisson and logistic cases, but not in the linear case. From the simulation studies, the Bayesian criteria are generally more likely than Cp and AIC to choose correct predictors.

**Keywords:** Variable Selection; Generalized Linear Models; Hierarchical Bayes Formulation; Least Absolute Shrinkage and Selection Operator (LASSO); Information criteria;  $L_p$  penalty; Asymptotic theory

Variable Selection Properties of L1 Penalized  
Regression in Generalized Linear Models

by

Chon Sam

Dissertation submitted to the Faculty of the Graduate School of the  
University of Maryland, College Park in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy  
2008

Advisory Committee:  
Professor Paul J. Smith, Chair/Advisor  
Professor Francis Alt  
Professor Benjamin Kedem  
Professor Partha Lahiri  
Professor Eric Slud

© Copyright by  
Chon Sam  
2008

## DEDICATION

To my beloved mother, father and sister, Carol.

## ACKNOWLEDGMENTS

I wish to express my deep appreciation to my advisor, Professor Paul Smith, for his enlightening guidance and enormous support throughout the path of my research and my graduate studies. His broad knowledge of statistics and mathematics, as well as his hands-on approach to theory, have been a great resource to me. I am especially grateful for his encouragement, professional advice and patience throughout our many hours of discussions. Not only has he taught me how to conduct rigorous research, but his positive attitude and his way of thinking when facing obstacles will definitely have a key influence on my future work. My gratitude to Professor Smith goes beyond words.

I would also like to thank Professor Benjamin Kedem who opened the doors to my graduate study in Maryland. Professor Kedem kindly provided me with financial support in the form of a teaching assistantship which enabled me to focus more on my studies. Additionally, I have learned a lot from Professor Kedem's rich experience in statistical analysis and his Times Series RITs.

I am very grateful to Professor Eric Slud who taught me a couple statistics courses. Professor Slud has always been encouraging and supportive of my research. I have immensely benefited from the Estimating Equations RIT offered by him and other statistics faculty members. Moreover, I am very thankful to Professor Slud for his constructive and detailed comments and suggestions, as well as all of his help in advancing this research.

I would also like to acknowledge Professor Francis Alt for serving on my dis-

sertation committee. I am indebted to Professor Alt for his invaluable comments on preparations of an earlier version of the abstract of this dissertation.

I owe my gratitude to Professor Partha Lahiri for serving on my dissertation committee. He was especially helpful by directing me to appropriate literature related to my research. I thank him for his valuable remarks on my dissertation.

I would like to thank the faculty and staff in the Mathematics Department. Professor Konstantina Trivisa and Mrs. Alverda McCoy deserve a special mention. I thank them for providing me with the financial support over the summer and my last semester which gave me the much needed time to devote to the completion of this dissertation. Working with them has been a joy.

I would like to extend my sincere thanks to the Faculty of Science and Technology at the University of Macau for encouraging me to pursue graduate studies in the United States.

I would like to acknowledge my buddies from UNC-Chapel Hill for their continued support and helping to make the transition back to the United States smooth and painless. Thanks are also due to my friends in the Mathematics Department of UMCP for their warm friendship and for providing technical support for my computer. In particular, I must thank Zhiwei, Lu and Ru. I truly appreciate the constant encouragement and wisdom from Zhiwei, the intellectual conversations with Lu, and the generous emotional support from Ru. I especially enjoyed her homemade gourmet foods.

Finally, I wish to thank my parents and my sister for strengthening me with their unwavering love and their never-ending support. As the intensiveness of this

research led to many years of disruption to our family life, I am especially grateful for their understanding. They, especially my sister, are the driving force of my studies. Making them proud is one thing I have enjoyed doing over these years. Thanks also go to Andy Ip who always believes in me.

While it is impossible to list all the people who have provided support in my graduate studies, I would like to offer all of them - those mentioned above and those I was unable to name specifically my sincere thanks. I would not have succeeded and have come this far without them. My graduate school experience has been challenging but I have overcome the hardships with determination and perseverance, and it will be something that I will cherish for the rest of my life.



# TABLE OF CONTENTS

<b>1</b>	<b>Introduction and Overview</b>	<b>1</b>
<b>2</b>	<b>Literature Review</b>	<b>3</b>
2.1	Generalized Linear Models . . . . .	3
2.2	Model Selection . . . . .	4
2.3	Variable Selection Methods in Linear Models . . . . .	5
2.3.1	Automatic Selection Procedures . . . . .	5
2.3.2	Information Criteria . . . . .	6
2.4	Regression with $L_\nu$ Penalty . . . . .	7
2.4.1	Numerical Package in computing LASSO Estimates . . . . .	8
2.4.2	Asymptotics for Penalized Regression Estimators . . . . .	9
2.5	Bayesian Model Selection . . . . .	11
2.5.1	Hierarchical Bayesian Formulation . . . . .	12
<b>3</b>	<b>LASSO Model Selection</b>	<b>15</b>
3.1	Hierarchical Bayesian Formulation for Logistic Regression . . . . .	15
3.2	Analysis of Posterior Probability . . . . .	20
3.2.1	Regular Class . . . . .	24
3.2.2	Nonregular Class . . . . .	25
3.3	Connection to LASSO . . . . .	28
<b>4</b>	<b>Bayesian Model Selection Criteria</b>	<b>31</b>
4.1	Empirical Bayes Criterion for Logistic Model . . . . .	31
4.2	Fully Bayes Criterion . . . . .	34
4.2.1	Restricted Region . . . . .	34
4.2.2	Flat priors . . . . .	37
4.2.3	Conjugate Priors . . . . .	42
4.3	Poisson Models . . . . .	46
4.4	Linear Models . . . . .	49
4.5	Implementation of the Bayesian Criteria . . . . .	52
<b>5</b>	<b>Asymptotic Results for LASSO-type Estimators in GLM</b>	<b>53</b>
<b>6</b>	<b>Simulation Studies and Data Analysis</b>	<b>69</b>
6.1	Simulation Studies . . . . .	69
6.1.1	Measures of Various Performance Criteria . . . . .	71

6.2	Linear Models . . . . .	73
6.3	Poisson Models . . . . .	84
6.4	Logistic Models . . . . .	87
6.5	Summary of Simulation Results . . . . .	90
6.6	South Africa Heart Disease Data Analysis . . . . .	91
<b>7</b>	<b>Summary and Future Research</b>	<b>105</b>
7.1	Summary . . . . .	105
7.2	Future Research . . . . .	107
7.2.1	Priors Specification . . . . .	107
7.2.2	Bayesian Model Averaging . . . . .	108
7.2.3	Model Selections in GLMs with Noncanonical Link . . . . .	108

## LIST OF TABLES

6.1	Simulation Results for Linear Models . . . . .	76
6.2	Simulation Results for Linear Models (continued) . . . . .	77
6.3	Simulation Results for Linear Models (continued) . . . . .	78
6.4	Simulation Results for Poisson Models . . . . .	85
6.5	Simulation Results for Logistic Models . . . . .	88
6.6	Bootstrap results for South Africa Heart Disease data using AIC. The coefficient estimates ( $\hat{\beta}^b$ ) are obtained by AIC. . . . .	94
6.7	Bootstrap results for South Africa Heart Disease data using BIC. The coefficient estimates ( $\hat{\beta}^b$ ) are obtained by BIC. . . . .	95
6.8	Bootstrap results for South Africa Heart Disease data using CML. The coefficient estimates ( $\hat{\beta}^b$ ) are obtained by CML. . . . .	96
6.9	Bootstrap results for South Africa Heart Disease data using FBC_Flat. The coefficient estimates ( $\hat{\beta}^b$ ) are obtained by FBC_Flat. . . . .	97
6.10	Bootstrap results for South Africa Heart Disease data using FBC_Conj. The coefficient estimates ( $\hat{\beta}^b$ ) are obtained by FBC_Conj. . . . .	98

## LIST OF FIGURES

4.1	Restricted Region $R'$ . . . . .	38
6.1	Histograms of model size for Model I from linear models based on 200 Monte Carlo replications. . . . .	75
6.2	Histograms of model size for Model II from linear models based on 200 Monte Carlo replications. . . . .	79
6.3	Histograms of model size for Model III from linear models based on 200 Monte Carlo replications. . . . .	79
6.4	Histograms of model size for Model IV from linear models based on 200 Monte Carlo replications. . . . .	80
6.5	Histograms of model size for Model V from linear models based on 200 Monte Carlo replications. . . . .	80
6.6	Histograms of model size for Model VI from linear models based on 200 Monte Carlo replications. . . . .	81
6.7	Histograms of model size for Model VII from linear models based on 200 Monte Carlo replications. . . . .	81
6.8	Histograms of model size for Poisson Model I based on 200 Monte Carlo replications. . . . .	86
6.9	Histograms of model size for Poisson Model II based on 200 Monte Carlo replications. . . . .	86
6.10	Histograms of model size for Logistic Model I based on 200 Monte Carlo replications. . . . .	89
6.11	Histograms of model size for Logistic Model II based on 200 Monte Carlo replications. . . . .	89

6.12	Solution path for South Africa Heart Disease data by <code>glmnet</code> (except the <code>famhist</code> feature). The horizontal axis is the $\lambda$ scaled by the maximum $\lambda$ from all the steps of the solution path provided by <code>glmnet</code> and the vertical axis represents the values of the estimated coefficients. The vertical lines are the various models chosen by the three Bayesian criteria, as well as AIC and BIC. . . . .	93
6.13	The bootstrap distribution of the coefficient estimates chosen by AIC. The red vertical bars represent $\hat{\beta}$ selected by AIC from the whole data and the blue thick bars are the frequencies of zero coefficients. . . .	99
6.14	The bootstrap distribution of the coefficient estimates chosen by BIC. The red vertical bars represent $\hat{\beta}$ selected by BIC from the whole data and the blue thick bars are the frequencies of zero coefficients. . . .	100
6.15	The bootstrap distribution of the coefficient estimates chosen by CML. The red vertical bars represent $\hat{\beta}$ selected by CML from the whole data and the blue thick bars are the frequencies of zero coefficients. .	101
6.16	The bootstrap distribution of the coefficient estimates chosen by <code>FBC_Flat</code> . The red vertical bars represent $\hat{\beta}$ selected by <code>FBC_Flat</code> from the whole data and the blue thick bars are the frequencies of zero coefficients. .	102
6.17	The bootstrap distribution of the coefficient estimates chosen by <code>FBC_Conj</code> . The red vertical bars represent $\hat{\beta}$ selected by <code>FBC_Conj</code> from the whole data and the blue thick bars are the frequencies of zero coefficients. . . . .	103

# Chapter 1

## Introduction and Overview

Consider the variable selection problem, where there are  $n$  observations of a dependent variable  $\mathbf{Y} = (y_1, y_2, \dots, y_n)^T$  and a set of  $p$  potential explanatory variables or predictors, namely,  $X_1, X_2, \dots, X_p$ . Some of these predictors are redundant or irrelevant, and therefore, the problem is to identify a subset of the predictors that best describes the underlying relationship revealed by the data, in order to provide estimation accuracy and enhance model interpretability.

Variable selection is very common in all disciplines. In the case of normal linear regression, we have

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1.1}$$

where  $\mathbf{X}$  is a  $n \times (p + 1)$  matrix,  $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T$  and  $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$ . The variable selection problem focuses on identifying the subset of nonzero  $\beta_j$ .

The common variable selection methods for linear models (roughly in chronological order) are Mallows'  $C_p$  [26], the Akaike information criterion ( $AIC$ ) [1], the Bayesian information criterion ( $BIC$ ) [33], the risk inflation criterion ( $RIC$ ) [11], the Least Absolute Shrinkage and Selection Operator (LASSO) [35], the minimum

description length (*MDL*) [17], the least angles regression (LAR) and the forward stagewise regression [8].

However, in applications when the dependent variable is categorical or discrete, instead of linear models one should use Generalized Linear Models (GLM). For GLM, supposing that the dependent variable follows an exponential family, we have

$$g(E(\mathbf{Y})) = \mathbf{X}\boldsymbol{\beta},$$

where  $g(\cdot)$  is the link function. The key feature is that the mean of  $\mathbf{Y}$  is a (nonlinear) transformation of a linear combination of predictors. Although the linear model is a special case of the GLM, the existing variable selection methods in linear models do not carry through to GLM automatically. Specifically, the variable selection problem in GLM is that the underlying mechanism of  $\mathbf{Y}$  and the data can be described by selecting some predictors such that the transformed mean,  $g(E(Y))$ , is a linear combination of the predictors in the subset.

The dissertation is organized as below: Chapter 2 summarizes various variable selection criteria in GLMs. A hierarchical framework is built in Chapter 3 which is directly related to Least Absolute Shrinkage and Selection Operator (LASSO). An empirical and a fully Bayes variable selection procedures are developed for linear, logistic and Poisson models in Chapter 4. Chapter 5 gives some asymptotic properties for Bridge estimators in GLMs. Some simulation studies and data analysis results of the performance of the Bayesian criteria derived in Chapter 4 are presented in Chapter 6. Finally, some conclusions and future research are provided in Chapter 7.

## Chapter 2

### Literature Review

## 2.1 Generalized Linear Models

A generalized linear model can be characterized by three components, which are the distribution of the response variable, the link function and the predictors. The response variable,  $\mathbf{Y}$  consists of independent measurements that ought to come from an exponential family distribution, of the form

$$f(\mathbf{Y}|\boldsymbol{\theta}, \phi) = \prod_{i=1}^n \exp \left\{ \frac{y_i \theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \right\}, \quad (2.1)$$

where  $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_n)^T$  and  $\phi$  are unknown parameters that may depend on the predictors  $X_0, X_1, \dots, X_p$  and  $\phi$  is called the dispersion parameter.

The parameters of the distribution are related to the predictors in a special way. The connection is achieved by taking a transformation of the mean through the link function and expressing it in terms of the linear predictors. That is,

$$E(y_i) = \mu_i = b'(\theta_i);$$

$$\text{Var}(y_i) = \phi b''(\theta_i);$$



and

$$\eta_i \equiv g(\mu_i) = \mathbf{x}_i \boldsymbol{\beta},$$

where  $g(\cdot)$  is the link function and  $\mathbf{x}_i$ ,  $i = 1, \dots, n$ , is the  $i$ th row of the design matrix  $\mathbf{X}$ . The dimension of  $\mathbf{X}$  is  $n \times (p + 1)$ , because the matrix always includes a 0th column of ones to accommodate the intercept. The link function that transforms the mean to the natural parameter,  $\boldsymbol{\theta}$ , is called the canonical link. For the canonical link, we have

$$\boldsymbol{\eta} \equiv \boldsymbol{\theta} = g(\boldsymbol{\mu}) = (b')^{-1}(\boldsymbol{\mu}),$$

where  $\boldsymbol{\eta} = (\eta_1, \eta_2, \dots, \eta_n)^T$  and  $\boldsymbol{\mu}$  is the mean of  $\mathbf{Y}$ .

With the help of the link function, the transformed mean  $g(E(\mathbf{Y}))$  can now be modeled by the linear predictors. That is,

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta},$$

where  $\boldsymbol{\eta} = g(\boldsymbol{\mu})$  as mentioned above. Refer to McCullagh and Nelder [27] and Kedem and Fokianos [19] for examples of GLM.

## 2.2 Model Selection

From the section about GLM above, one can see that the underlying problem is indeed how to choose the predictors from a large set of potentially available explanatory variables, in order to attain accurate inference and to obtain good predictions. Let the binary vector,  $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2, \dots, \gamma_p)^T$  index each candidate model, where  $\gamma_i$  takes value either 1 or 0,  $i = 1, 2, \dots, n$ , depending on whether or

not the  $i$ th predictor is included in the model, and let  $|\gamma|$  be the size of the candidate model, with  $|\gamma| = \sum_{i=1}^p \gamma_i$ . The variable selection problem in GLM can be described as follows: one attempts to identify the vector  $\gamma$ , such that

$$\eta = X_{\gamma}\beta_{\gamma}.$$

## 2.3 Variable Selection Methods in Linear Models

For linear models, there are two types of variable selection methods that are commonly used in practice, automatic selection procedures and information-based criteria.

### 2.3.1 Automatic Selection Procedures

The automatic selection procedures are data-driven and include forward selection, backward elimination and stepwise selection procedures.

The forward selection procedure starts from the null model. Then it performs a test to find the significant variables, by checking if the p-value of the variables falls below some pre-set threshold. Among the significant variables, the procedure chooses the most significant one and adds it to the model. One then refits the data using this one variable model and searches for the next variable to enter, and so on. This process continues until none of the remaining variables are significant.

Unlike the forward selection procedure, the backward elimination procedure begins from the full model including all predictors. At each step, each variable is tested for elimination from the model, by comparing the p-value of the variable to

the pre-defined level. From the variables whose p-values are above the chosen level, the least significant one is deleted. With this reduced model, one may refit the data and search for the next least significant variable to exclude. This procedure stops when all remaining variables are statistically significant.

Stepwise selection is a mixture of the forward and backward procedures. This procedure allows dropping or adding variables at the various steps. It initially uses a forward selection procedure. But after each selection, the procedure employs a backward approach by deleting variables if they later appear to be insignificant. After refitting the data with the new model and repeatedly applying the stepwise rule, the process terminates when all currently included variables satisfying a retention criterion and no additional variables satisfy an inclusion criterion. These criteria are chosen to avoid an endless loop.

### **2.3.2 Information Criteria**

Information criteria are model selection methods that penalize the loglikelihood for complexity of the model, where complexity depends on the number of explanatory variables in the model. The most well known ones are the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). AIC, closely related to Mallows'  $C_p$  [26], tries to minimize the Kullback-Leibler divergence between the true distribution and the estimate from a candidate model, whereas BIC favors a model with the highest asymptotic posterior model probability. The goal is to select a model by minimizing the information criteria and obtain estimates of  $\beta$ .

Akaike [1] proposed AIC, which is

$$\text{AIC} = -2 \log L(\mathbf{X}, \hat{\boldsymbol{\theta}}) + 2m,$$

where  $L$  is the likelihood function,  $\hat{\boldsymbol{\theta}}$  is the maximum likelihood estimator of the parameter vector and  $m$  is the complexity variable. Schwarz [33] took a Bayesian approach and derived BIC, that is,

$$\text{BIC} = -2 \log L(\mathbf{X}, \hat{\boldsymbol{\theta}}) + m \log n,$$

where  $L$ ,  $\hat{\boldsymbol{\theta}}$  and  $m$  are as defined previously.

Both the AIC and BIC criteria take the form of loglikelihoods with a deterministic penalty. The difference is that when the true model is among the candidate models, BIC selects the true model with probability approaching unity as  $n$  goes to infinity. This property is called *consistency* and is shared by the minimum description length (MDL) method, originating from coding theory and discussed by Hansen and Yu [17]. However, AIC is not consistent. Instead, if the true model is not in any of the candidate models, AIC asymptotically chooses the model which has the minimum average squared error. See Shao [34] and references contained in his paper.

## 2.4 Regression with $L_\nu$ Penalty

An alternative approach is to estimate  $\boldsymbol{\beta}$  by minimizing the penalized loglikelihood criterion of the form,

$$-\log L(\mathbf{X}, \boldsymbol{\xi}) + \lambda_n \sum_{i=1}^p |\xi_i|^\nu, \tag{2.2}$$

where  $\boldsymbol{\xi}$  is any point in the parameter space,  $\lambda_n > 0$  is the tradeoff parameter between the likelihood and the penalty, and  $\nu > 0$ . Estimators of  $\boldsymbol{\beta}$  obtained in this way are called Bridge estimators by Frank and Friedman [12].

Note that the information criteria are the limiting cases when  $\nu \rightarrow 0$  because

$$\lim_{\nu \downarrow 0} \sum_{i=1}^p |\xi_i|^\nu = \sum_{i=1}^p I(\xi_i \neq 0).$$

For linear models, in the case of  $\nu = 2$ , the method is called ridge regression. Moreover,  $\nu = 1$  refers to Least Absolute Shrinkage and Selection Operator (LASSO) proposed by Tibshirani [35]. LASSO estimates some of the  $\beta_i$  exactly at zero and produces a sparse representation of  $\boldsymbol{\beta}$ . Researchers recognize this attractive feature of LASSO and use it for automatic model selection.

Besides the information criteria, LASSO and ridge regression, there are also other penalties for the regression function, such as the Risk Inflation Criterion (RIC) of Foster and George [11], the penalty functions of Fan and Li [10], and Least Angle Regression (LAR) and stagewise regression developed by Efron et al [8].

### 2.4.1 Numerical Package in computing LASSO Estimates

The LASSO estimates vary as the tradeoff parameter or regularization parameter,  $\lambda_n$ , moves from zero to infinity. Hence, for each  $\lambda_n$ , the nonzero coefficients from the LASSO estimation correspond to selected variables.

The LASSO estimates depend heavily on the regularization parameter. Through the algorithm developed by Osborne, Presnell, and Turlach [28] for the linear case and extended by Lokhorst [25] to include GLM, LASSO estimates are obtained for

a pre-specified set of  $\lambda_n$ . The algorithm is available in R as the package `lasso2`.

However, in order to select the best model, one would like the regularization parameters to run through the whole path from zero to infinity to see where the minimum of some designated criterion occurs. This is made feasible by the efficient `lars` algorithm provided by Efron et al [8], which includes LASSO as one of its options along with LAR and Stagewise Regression for the linear case, under the `lars` package in R. Park and Hastie [29] developed the R package, `glmpath`, to handle GLM problems. A new R package, `glmnet`, was recently developed by Friedman, Hastie and Tibshirani [13]. The `glmnet` software provides fast algorithms via cyclical coordinate descent method for fitting linear, multinomial and logistic models with elastic-net penalties, which is a weighted combination of  $L_1$  and  $L_2$  penalties. All three packages, `lars`, `glmpath` and `glmnet`, allow one to find solutions of  $L_1$  penalized regression problems for the entire path of  $\lambda_n$ .

## 2.4.2 Asymptotics for Penalized Regression Estimators

For linear models, Knight and Fu [21] develop the asymptotics for the Bridge Estimators. They prove that under regularity conditions on the design and on the order of magnitude of  $\lambda_n$ , the estimator is consistent. Supposed that  $Y$  is centered, the covariates are centered and scaled with unit standard deviation. Then the  $L_1$  penalized least squares problem can be written as:

$$\sum_{i=1}^n (Y_i - \mathbf{x}_i \boldsymbol{\xi})^2 + \lambda_n \sum_{j=1}^p |\xi_j|^\nu = \min! \tag{2.3}$$

There are two regularity conditions on the design:

**Condition LC1**  $C_n = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^T \mathbf{x}_i \rightarrow C,$

where  $\mathbf{x}_i$  is a row vector which represents the  $i$ th row of the design matrix  $\mathbf{X}$  and  $C$  is a nonnegative definite matrix and

**Condition LC2**  $\frac{1}{n} \max_{1 \leq i \leq n} \mathbf{x}_i \mathbf{x}_i^T \rightarrow 0.$

Since the covariates are scaled, the diagonal elements of  $C_n$  and  $C$  are all identically equal to 1.

According to Knight and Fu [21], define the random function

$$Z_n(\boldsymbol{\xi}) = \frac{1}{n} \sum_{j=1}^n (Y_j - \mathbf{x}_j \boldsymbol{\xi})^2 + \frac{\lambda_n}{n} \sum_{i=1}^p |\xi_i|^\nu. \quad (2.4)$$

The minimum of (2.4) occurs when  $\boldsymbol{\xi} = \hat{\boldsymbol{\beta}}_n$ .

In general, for  $\nu > 0$ , Knight and Fu prove that  $\hat{\boldsymbol{\beta}}_n$  is consistent if  $\lambda_n = o(n)$ .

More specifically, they establish the following result.

**Theorem 2.1 (Knight & Fu, 2000 [21])** *If  $C$  in Condition LC1 is nonsingular and  $\lambda_n/n \rightarrow \lambda_0 \geq 0$ , then  $\hat{\boldsymbol{\beta}}_n \rightarrow_p \operatorname{argmin}(Z)$  where*

$$Z(\boldsymbol{\xi}) = (\boldsymbol{\xi} - \boldsymbol{\beta})^T C (\boldsymbol{\xi} - \boldsymbol{\beta}) + \lambda_0 \sum_{i=1}^p |\xi_i|^\nu.$$

*Thus if  $\lambda_n = o(n)$ ,  $\operatorname{argmin}(Z) = \boldsymbol{\beta}$  and so  $\hat{\boldsymbol{\beta}}_n$  is consistent.*

In fact, for  $\nu \geq 1$ , they also derive the limiting distribution of  $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$  if  $\lambda_n = O(\sqrt{n})$  and prove its  $\sqrt{n}$ -consistency if  $\lambda_n = o(\sqrt{n})$ .

**Theorem 2.2 (Knight & Fu, 2000 [21])** *Suppose that  $\nu \geq 1$ . If  $\lambda_n/\sqrt{n} \rightarrow \lambda_0 \geq 0$  and  $C$  is nonsingular, then*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \rightarrow_d \operatorname{argmin}(V),$$

where if  $\nu > 1$ ,

$$V(\mathbf{u}) = -2\mathbf{u}^T \mathbf{W} + \mathbf{u}^T C \mathbf{u} + \nu \lambda_0 \sum_{j=1}^p u_j \operatorname{sgn}(\beta_j) |\beta_j|^{\nu-1},$$

if  $\nu = 1$ ,

$$V(\mathbf{u}) = -2\mathbf{u}^T \mathbf{W} + \mathbf{u}^T C \mathbf{u} + \lambda_0 \sum_{j=1}^p [u_j \operatorname{sgn}(\beta_j) I(\beta_j \neq 0) + |u_j| I(\beta_j = 0)],$$

and  $\mathbf{W}$  has a  $N(\mathbf{0}, \sigma^2 \mathbf{C})$  distribution.

When  $\nu \leq 1$ , they need to assume a different rate of growth of  $\lambda_n$  to get a limiting distribution.

**Theorem 2.3 (Knight & Fu, 2000 [21])** *Suppose that  $\nu \leq 1$ . If  $\lambda_n/n^{\nu/2} \rightarrow \lambda_0 \geq 0$  and  $C$  is nonsingular, then*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \rightarrow_d \operatorname{argmin}(V),$$

where

$$V(\mathbf{u}) = -2\mathbf{u}^T \mathbf{W} + \mathbf{u}^T C \mathbf{u} + \lambda_0 \sum_{j=1}^p |u_j|^\nu I(\beta_j = 0),$$

and  $\mathbf{W}$  has a  $N(\mathbf{0}, \sigma^2 \mathbf{C})$  distribution.

The consistency results for these estimators will be generalized to GLM in Chapter 5.

## 2.5 Bayesian Model Selection

Some researchers have attempted to solve the variable selection problem using a Bayesian approach (Raftery and Richardson [32], Raftery [30], Clyde [6], George



and Foster [14] and Dellaportas, Forster and Ntzoufras [7]). In particular, George and Foster [14] showed that for linear models, the criteria  $C_p$ ,  $AIC$  and  $BIC$  correspond to selection of the model with maximum posterior probability under a particular class of priors in a hierarchical Bayesian formulation.

### 2.5.1 Hierarchical Bayesian Formulation

The hierarchical Bayesian formulation first assigns a prior distribution  $\pi(\boldsymbol{\gamma}|\boldsymbol{\psi}_1)$  on the model space, where  $\boldsymbol{\gamma}$  is the binary vector that represents each candidate model and  $\boldsymbol{\psi}_1$  is the hyperparameter vector associated with the prior of  $\boldsymbol{\gamma}$ . For each candidate model, the Bayesian formulation further puts a prior distribution  $P(\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma}, \boldsymbol{\psi}_2)$  on the model specific coefficient vector  $\boldsymbol{\beta}_\gamma$ , where  $\boldsymbol{\psi}_2$  is the hyperparameter vector from the prior of  $\boldsymbol{\beta}_\gamma$ . Bayesians obtain a posterior distribution  $\pi(\boldsymbol{\gamma}|\mathbf{Y}, \boldsymbol{\psi}_1, \boldsymbol{\psi}_2)$  by updating the prior distribution over the model space with the data  $\mathbf{Y}$ :

$$\pi(\boldsymbol{\gamma}|\mathbf{Y}, \boldsymbol{\psi}_1, \boldsymbol{\psi}_2) = \frac{P(\mathbf{Y}|\boldsymbol{\gamma}, \boldsymbol{\psi}_2)\pi(\boldsymbol{\gamma}|\boldsymbol{\psi}_1)}{\sum_{\boldsymbol{\gamma}} P(\mathbf{Y}|\boldsymbol{\gamma}, \boldsymbol{\psi}_2)\pi(\boldsymbol{\gamma}|\boldsymbol{\psi}_1)}, \quad (2.5)$$

where

$$P(\mathbf{Y}|\boldsymbol{\gamma}, \boldsymbol{\psi}_2) = \int P(\mathbf{Y}|\boldsymbol{\beta}_\gamma, \boldsymbol{\gamma})P(\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma}, \boldsymbol{\psi}_2) d\boldsymbol{\beta}_\gamma \quad (2.6)$$

is the marginal distribution of  $\mathbf{Y}$  after integrating out  $\boldsymbol{\beta}_\gamma$  with respect to the prior distribution  $P(\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma}, \boldsymbol{\psi}_2)$ .

There are two ways to handle the hyperparameters  $\boldsymbol{\psi}_1$  and  $\boldsymbol{\psi}_2$ , namely, empirical Bayes and fully Bayes. Empirical Bayes estimates the hyperparameters through the data and plugs them into the posterior distribution to obtain  $\pi(\boldsymbol{\gamma}|\mathbf{Y}, \hat{\boldsymbol{\psi}}_1, \hat{\boldsymbol{\psi}}_2)$ .

It chooses the model with the maximum posterior probability. Fully Bayes imposes a prior on  $\boldsymbol{\psi}_1$  and  $\boldsymbol{\psi}_2$  and follows the standard Bayesian procedure to integrate out the hyperparameters. The resulting posterior distribution  $\pi(\boldsymbol{\gamma}|\mathbf{Y})$  is again used as a variable selection criterion to find the model with the largest posterior probability,

$$\begin{aligned}
\pi(\boldsymbol{\gamma}|\mathbf{Y}) &= \int \int_D \pi(\boldsymbol{\gamma}|\mathbf{Y}, \boldsymbol{\psi}_1, \boldsymbol{\psi}_2) P(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2|\mathbf{Y}) d\boldsymbol{\psi}_1 d\boldsymbol{\psi}_2 \\
&= \int \int_D \frac{P(\mathbf{Y}|\boldsymbol{\gamma}, \boldsymbol{\psi}_2)\pi(\boldsymbol{\gamma}|\boldsymbol{\psi}_1)}{P(\mathbf{Y}|\boldsymbol{\psi}_1, \boldsymbol{\psi}_2)} \frac{P(\mathbf{Y}|\boldsymbol{\psi}_1, \boldsymbol{\psi}_2)\pi(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2)}{P(\mathbf{Y})} d\boldsymbol{\psi}_1 d\boldsymbol{\psi}_2 \\
&= \int \int_D \frac{P(\mathbf{Y}|\boldsymbol{\gamma}, \boldsymbol{\psi}_2)\pi(\boldsymbol{\gamma}|\boldsymbol{\psi}_1)}{P(\mathbf{Y})} \pi(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2) d\boldsymbol{\psi}_1 d\boldsymbol{\psi}_2, \tag{2.7}
\end{aligned}$$

where  $P(\mathbf{Y}|\boldsymbol{\gamma}, \boldsymbol{\psi}_2)$  is given in (2.6) and  $D$  is the hyperparameter space of  $\boldsymbol{\psi}_1$  and  $\boldsymbol{\psi}_2$ .

This hierarchical Bayesian formulation is conceptually attractive as it is able to incorporate various selection criteria, such as *AIC* and *BIC*, and put them in a unified framework. George and Foster [14] first proposed the Empirical Bayes approach for normal linear models using an independence prior for the models so that each predictor is in the model independent from the other predictors with the same inclusion probability  $q$ . They also imposed a conjugate prior for the model coefficients, and estimated the hyperparameters using either a marginal maximum likelihood criterion or a conditional maximum likelihood (CML) criterion.

Using the same priors as George and Foster, Wang and George [36] extended the empirical Bayes method to GLM. Wang and George also developed a fully Bayes approach to allow superimposing a prior distribution on the hyperparameters. By maximizing the posterior distribution, they derived a fully Bayes criterion for variable selection for GLM.

Yuan and Lin [38] also took the empirical Bayes approach for linear models, but they formulated the hierarchical Bayes paradigm in a different way. By specifying a double exponential prior for the model coefficients and giving the following priors with a determinant factor for the models,

$$\pi(\boldsymbol{\gamma}) \propto q^{|\boldsymbol{\gamma}|} (1 - q)^{p - |\boldsymbol{\gamma}|} \sqrt{\det(\mathbf{X}_{\boldsymbol{\gamma}}^T \mathbf{X}_{\boldsymbol{\gamma}})}, \quad (2.8)$$

Yuan and Lin established a variable selection criterion for linear model which is equivalent to minimizing the  $L_1$  penalized likelihood. By using the LARS algorithm [8] in  $R$ , they showed that they can compute their empirical Bayes criterion efficiently and therefore perform variable selection.

## Chapter 3

### LASSO Model Selection

In this chapter, a hierarchical Bayes formulation is carried out for logistic regression as an illustration of extensions to GLM. By specifying a special prior for the covariate coefficients, the posterior distribution is closely related to LASSO and thus allows one to do variable selection in GLM problems.

### 3.1 Hierarchical Bayesian Formulation for Logistic Regression

Yuan and Lin [38] formulated a hierarchical setup for linear regression. We extend their approach to formulate a hierarchical structure for logistic data. This extension accounts for the fact that  $\text{Var}(\mathbf{Y})$  depends on  $\boldsymbol{\beta}$ . Suppose  $\mathbf{Y} = (y_1, y_2, \dots, y_n)^T$  is the observation vector,  $\mathbf{X}$  is an  $n \times (p + 1)$  design matrix with  $\mathbf{x}_i$  representing the  $i$ th row and the binary vector  $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2, \dots, \gamma_p)^T$  index of each model is as defined in Section 2.2. The zero-th column of  $\mathbf{X}$  is a column of ones. Ignoring the intercept term, the size of the candidate model  $|\boldsymbol{\gamma}|$  is defined to be  $|\boldsymbol{\gamma}| = \sum_{i=1}^p \gamma_i$ .

Moreover,  $\mathbf{X}_\gamma$  denotes the columns of  $\mathbf{X}$  that are in the model and  $\mathbf{x}_{i\gamma}$  is the  $i$ th row of  $\mathbf{X}_\gamma$ .

Assume that  $y_i, i = 1, 2, \dots, n$ , are independent and each can take values of either 1 or 0. The success probability of  $y_i$  is

$$P(y_i = 1 | \mathbf{x}_{i\gamma}) = \frac{\exp(\mathbf{x}_{i\gamma}\boldsymbol{\beta}_\gamma)}{1 + \exp(\mathbf{x}_{i\gamma}\boldsymbol{\beta}_\gamma)},$$

while

$$P(y_i = 0 | \mathbf{x}_{i\gamma}) = \frac{1}{1 + \exp(\mathbf{x}_{i\gamma}\boldsymbol{\beta}_\gamma)}$$

is the probability of failure. Since  $y_i$  follows a Bernoulli distribution, its mean is

$$\mu_{i\gamma} = 1 \times P(y_i = 1 | \mathbf{x}_{i\gamma}) + 0 \times P(y_i = 0 | \mathbf{x}_{i\gamma}) = P(y_i = 1 | \mathbf{x}_{i\gamma})$$

Using the canonical link function, which is the logit, one may express the transformed mean as a linear combination of the predictors.

$$\text{logit} \mu_{i\gamma} = \log \left( \frac{P(y_i = 1 | \mathbf{x}_{i\gamma})}{P(y_i = 0 | \mathbf{x}_{i\gamma})} \right) = \mathbf{x}_{i\gamma}\boldsymbol{\beta}_\gamma.$$

The density function of  $y_i | \boldsymbol{\beta}_\gamma, \boldsymbol{\gamma}$  is

$$\begin{aligned} f(y_i | \boldsymbol{\beta}_\gamma, \boldsymbol{\gamma}) &= (P(y_i = 1 | \mathbf{x}_{i\gamma}))^{y_i} (P(y_i = 0 | \mathbf{x}_{i\gamma}))^{1-y_i} \\ &= (\exp(\mathbf{x}_{i\gamma}\boldsymbol{\beta}_\gamma))^{y_i} \frac{1}{1 + \exp(\mathbf{x}_{i\gamma}\boldsymbol{\beta}_\gamma)}. \end{aligned} \quad (3.1)$$

Given the model  $\boldsymbol{\gamma}$ ,

$$\beta_0 | \boldsymbol{\gamma} \sim N(0, \sigma_0^2). \quad (3.2)$$

The quantity  $\sigma_0^2$  is large to create a vague prior. The intercept is not subject to selection. Since  $\gamma_j = 0$  means that the  $j$ th predictor is not in the model,  $\beta_j | \gamma_j = 0$

is degenerate at 0. If  $\gamma_j = 1$ ,  $\beta_j$  has a double exponential prior distribution with hyperparameter  $\tau$ . Furthermore, the  $\beta_j$  are conditionally independent given  $\gamma$ .

Therefore,

$$P(\beta_j|\gamma_j, \tau) = \begin{cases} 0 & \text{if } \gamma_j = 0 \\ \tau/2 \exp(-\tau|\beta_j|) & \text{if } \gamma_j = 1 \end{cases} \quad (3.3)$$

where  $j = 1, 2, \dots, p$ . This double exponential prior will enable the method to set certain  $\beta_j$  equal to zero.

For  $\gamma$ , instead of the widely used independence prior which assumes that each predictor enters the model independently with common probability  $q$ , for computational simplicity, assume that the prior of  $\gamma$  is proportional to the independence prior times a function of sample quantities; that is,

$$\pi(\gamma|q) = q^{|\gamma|}(1-q)^{p-|\gamma|} \sqrt{\det(\mathbf{A} + \mathbf{H})} \exp\left(-\frac{1}{2}(\mathbf{e} - \mathbf{t})^T(\mathbf{A} + \mathbf{H})^{-1}(\mathbf{e} - \mathbf{t})\right) \quad (3.4)$$

where  $\mathbf{A}$  is a  $(|\gamma| + 1) \times (|\gamma| + 1)$  matrix, with

$$\mathbf{A} = \sum_{i=1}^n \left( \mathbf{x}_{i\gamma}^T \frac{\exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_{\gamma}^*)}{(1 + \exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_{\gamma}^*))^2} \mathbf{x}_{i\gamma} \right), \quad (3.5)$$

$\mathbf{H}$  is a  $(|\gamma| + 1) \times (|\gamma| + 1)$  matrix, with

$$\mathbf{H} = \begin{pmatrix} 1/(2\sigma_0^2) & \mathbf{0}_{1 \times |\gamma|} \\ \mathbf{0}_{|\gamma| \times 1} & \mathbf{0}_{|\gamma| \times |\gamma|} \end{pmatrix}, \quad (3.6)$$

$\mathbf{e}$  and  $\mathbf{t}$  are both length  $|\gamma| + 1$  vectors, with

$$\mathbf{e}^T = \sum_{i=1}^n \left( y_i - \frac{\exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_{\gamma}^*)}{1 + \exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_{\gamma}^*)} \right) \mathbf{x}_{i\gamma} = (e_0, e_1, \dots, e_{|\gamma|}), \quad (3.7)$$

and

$$\mathbf{t}^T = \left( \frac{\beta_0^*}{\sigma_0^2}, 0, \dots, 0 \right), \quad (3.8)$$

where

$$\boldsymbol{\beta}_\gamma^* = \operatorname{argmin}_{\boldsymbol{\beta}_\gamma} \left( \sum_{i=1}^n (\log(1 + \exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_\gamma)) - y_i \mathbf{x}_{i\gamma} \boldsymbol{\beta}_\gamma) + \frac{\beta_0^2}{2\sigma_0^2} + \lambda \sum_{\substack{\gamma_j=1 \\ j \neq 0}} |\beta_j| \right).$$

The distributions (3.1), (3.2), (3.3) and (3.4) comprise a hierarchical Bayesian formulation with some hyperparameters  $\tau, q$  and  $\sigma_0$ . In this section, we assume that  $\sigma_0$  is fixed and known, but later in Section 3.3, we assume  $\sigma_0$  approaches infinity. The remaining parameters  $\tau$  and  $q$  can be obtained by empirical Bayes, which will be discussed in Section 4.1. From the fully Bayes point of view, one can put hyperpriors on the parameters and this will be further investigated in Section 4.1.

Our priors involve the observed  $y_i$  explicitly in  $\mathbf{e}$  (3.7) and through the data dependent quantity  $\boldsymbol{\beta}_\gamma^*$ , which is part of  $\mathbf{A}, \mathbf{H}$  and  $\mathbf{e}$ . This is a version of objective Bayes (Berger and Pericchi [4], Berger [5]).

Putting (3.1), (3.2), (3.3) and (3.4) together, one may write the joint distribution  $P(\boldsymbol{\gamma}, \boldsymbol{\beta}, \mathbf{Y})$  as

$$\begin{aligned} P(\boldsymbol{\gamma}, \boldsymbol{\beta}, \mathbf{Y}) &\propto \left( \prod_{i=1}^n \exp(y_i \mathbf{x}_{i\gamma} \boldsymbol{\beta}_\gamma) \frac{1}{1 + \exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_\gamma)} \right) \frac{1}{\sigma_0 \sqrt{2\pi}} \exp\left(-\frac{\beta_0^2}{2\sigma_0^2}\right) \left(\frac{\tau}{2}\right)^{|\boldsymbol{\gamma}|} \\ &\quad \times \exp\left(-\tau \sum_{\substack{\gamma_j=1 \\ j \neq 0}} |\beta_j|\right) \left(\frac{q}{1-q}\right)^{|\boldsymbol{\gamma}|} (1-q)^p \sqrt{\det(\mathbf{A} + \mathbf{H})} \\ &\quad \times \exp\left(-\frac{1}{2}(\mathbf{e} - \mathbf{t})^T (\mathbf{A} + \mathbf{H})^{-1} (\mathbf{e} - \mathbf{t})\right) \end{aligned}$$

$$\begin{aligned}
&= (1-q)^p \frac{1}{\sigma_0 \sqrt{2\pi}} \left( \frac{q}{1-q} \cdot \frac{\tau}{2} \cdot \sqrt{2\pi} \right)^{|\gamma|} \frac{\sqrt{\det(\mathbf{A} + \mathbf{H})}}{(\sqrt{2\pi})^{|\gamma|}} \\
&\quad \times \exp \left( -\frac{1}{2} (\mathbf{e} - \mathbf{t})^T (\mathbf{A} + \mathbf{H})^{-1} (\mathbf{e} - \mathbf{t}) \right) \\
&\quad \times \exp \left[ - \left( \sum_{i=1}^n (\log(1 + \exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_\gamma)) - y_i \mathbf{x}_{i\gamma} \boldsymbol{\beta}_\gamma) + \frac{\beta_0^2}{2\sigma_0^2} + \tau \sum_{\substack{\gamma_j=1 \\ j \neq 0}} |\beta_j| \right) \right].
\end{aligned} \tag{3.9}$$

One may obtain the conditional distribution of  $P(\boldsymbol{\gamma}, \boldsymbol{\beta} | \mathbf{Y})$ , which is

$$P(\boldsymbol{\gamma}, \boldsymbol{\beta} | \mathbf{Y}) = \frac{P(\boldsymbol{\gamma}, \boldsymbol{\beta}, \mathbf{Y})}{\sum_{\boldsymbol{\gamma}} \int_{\boldsymbol{\beta}} P(\boldsymbol{\gamma}, \boldsymbol{\beta}, \mathbf{Y}) d\boldsymbol{\beta} P(\boldsymbol{\gamma})}.$$

After reparameterizing  $(\tau, q)$  in terms of  $(\lambda, k)$ , we obtain

$$\begin{aligned}
&P(\boldsymbol{\gamma} | \mathbf{Y}) \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} P(\boldsymbol{\gamma}, \boldsymbol{\beta} | \mathbf{Y}) d\boldsymbol{\beta}_\gamma \\
&= G(\mathbf{Y}) k^{|\gamma|} \frac{1}{\sigma_0} \frac{\sqrt{\det(\mathbf{A} + \mathbf{H})}}{(\sqrt{2\pi})^{|\gamma|+1}} \exp \left( -\frac{1}{2} (\mathbf{e} - \mathbf{t})^T (\mathbf{A} + \mathbf{H})^{-1} (\mathbf{e} - \mathbf{t}) \right) \\
&\quad \times \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp \left[ - \left( \sum_{i=1}^n (\log(1 + \exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_\gamma)) - y_i \mathbf{x}_{i\gamma} \boldsymbol{\beta}_\gamma) \right. \right. \\
&\quad \left. \left. + \frac{\beta_0^2}{2\sigma_0^2} + \lambda \sum_{\substack{\gamma_j=1 \\ j \neq 0}} |\beta_j| \right) \right] d\boldsymbol{\beta}_\gamma, \tag{3.10}
\end{aligned}$$

where

$$k = \left( \frac{q}{1-q} \cdot \frac{\tau}{2} \cdot \sqrt{2\pi} \right),$$

$\lambda = \tau$ , and  $G(\mathbf{Y})$  is a function of  $\mathbf{Y}$  not depending on  $\boldsymbol{\gamma}$ .



## 3.2 Analysis of Posterior Probability

For the purpose of variable selection, one would like to evaluate the posterior probability and select the model with maximum posterior probability. This involves calculating the high dimensional integrals in (3.10) which do not have a closed form solution. The integration can only be done by approximation using analytical or numerical methods.

The candidate models are divided into two classes: regular and nonregular as defined in Yuan and Lin [38]:

**Definition 3.1 (Yuan and Lin, 2005 [38])** *For a dataset  $(\mathbf{X}, \mathbf{Y})$  and a given regularization parameter  $\lambda$ ,*

- (1) *a model  $\gamma$  is called regular if and only if  $\beta_{\gamma}^*$  does not contain 0's or  $|\gamma| = 0$  and*
- (2) *a model  $\gamma$  is called nonregular if  $\beta_{\gamma}^*$  contains at least one zero component.*

By means of Taylor expansion and Laplace approximation, we give an expression for the posterior probability for the regular class in Section 3.2.1. Then we show that the posterior probability for the nonregular class is dominated by its regular class counterpart in Section 3.2.2. That is, if  $\gamma$  is regular, we can find a regular model  $\gamma^*$  with  $P(\gamma|\mathbf{Y}) \leq P(\gamma^*|\mathbf{Y})$ .

Let

$$\beta_{\gamma}^* = \operatorname{argmin}_{\beta_{\gamma}} \left( \sum_{i=1}^n (\log(1 + \exp(\mathbf{x}_{i\gamma}\beta_{\gamma})) - y_i \mathbf{x}_{i\gamma}\beta_{\gamma}) + \frac{\beta_0^2}{2\sigma_0^2} + \lambda \sum_{\substack{\gamma_j=1 \\ j \neq 0}} |\beta_j| \right).$$

Define  $\boldsymbol{\beta}_\gamma = \boldsymbol{\beta}_\gamma^* + \mathbf{u}$ . The posterior probability  $P(\boldsymbol{\gamma}|\mathbf{Y})$  becomes

$$\begin{aligned}
& P(\boldsymbol{\gamma}|\mathbf{Y}) \\
&= G(\mathbf{Y})k^{|\boldsymbol{\gamma}|} \frac{1}{\sigma_0} \frac{\sqrt{\det(\mathbf{A} + \mathbf{H})}}{(\sqrt{2\pi})^{|\boldsymbol{\gamma}|+1}} \exp\left(-\frac{1}{2}(\mathbf{e} - \mathbf{t})^T(\mathbf{A} + \mathbf{H})^{-1}(\mathbf{e} - \mathbf{t})\right) \\
&\quad \times \left\{ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left[-\left(\sum_{i=1}^n (\log(1 + \exp(\mathbf{x}_{i\boldsymbol{\gamma}}(\boldsymbol{\beta}_\gamma^* + \mathbf{u})) - y_i \mathbf{x}_{i\boldsymbol{\gamma}}(\boldsymbol{\beta}_\gamma^* + \mathbf{u}))\right.\right.\right. \\
&\quad \left.\left.\left. - \log(1 + \exp(\mathbf{x}_{i\boldsymbol{\gamma}}\boldsymbol{\beta}_\gamma^*)) + y_i \mathbf{x}_{i\boldsymbol{\gamma}}\boldsymbol{\beta}_\gamma^*\right) + \frac{(\beta_0^* + u_0)^2}{2\sigma_0^2} - \frac{\beta_0^{*2}}{2\sigma_0^2}\right.\right. \\
&\quad \left.\left.\left. + \lambda \sum_{\substack{\gamma_j=1 \\ j \neq 0}} (|\beta_j^* + u_j| - |\beta_j^*|)\right)\right] d\mathbf{u}\right\} \\
&\quad \times \exp\left[-\min_{\boldsymbol{\beta}_\gamma} \left(\sum_{i=1}^n (\log(1 + \exp(\mathbf{x}_{i\boldsymbol{\gamma}}\boldsymbol{\beta}_\gamma)) - y_i \mathbf{x}_{i\boldsymbol{\gamma}}\boldsymbol{\beta}_\gamma) + \frac{\beta_0^2}{2\sigma_0^2} + \lambda \sum_{\substack{\gamma_j=1 \\ j \neq 0}} |\beta_j|)\right)\right].
\end{aligned} \tag{3.11}$$

The integral in (3.11) are approximated as follows, using Taylor expansions of the logarithm of the integrand and Laplace's method.

$$\begin{aligned}
& \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left[-\left(\sum_{i=1}^n (\log(1 + \exp(\mathbf{x}_{i\boldsymbol{\gamma}}(\boldsymbol{\beta}_\gamma^* + \mathbf{u})) - y_i \mathbf{x}_{i\boldsymbol{\gamma}}(\boldsymbol{\beta}_\gamma^* + \mathbf{u}))\right.\right. \\
&\quad \left.\left. - \log(1 + \exp(\mathbf{x}_{i\boldsymbol{\gamma}}\boldsymbol{\beta}_\gamma^*)) + y_i \mathbf{x}_{i\boldsymbol{\gamma}}\boldsymbol{\beta}_\gamma^*\right) + \frac{(\beta_0^* + u_0)^2}{2\sigma_0^2} - \frac{\beta_0^{*2}}{2\sigma_0^2}\right. \\
&\quad \left.\left. + \lambda \sum_{\substack{\gamma_j=1 \\ j \neq 0}} (|\beta_j^* + u_j| - |\beta_j^*|)\right)\right] d\mathbf{u} \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left[-\left(\sum_{i=1}^n (\log(1 + \exp(\mathbf{x}_{i\boldsymbol{\gamma}}(\boldsymbol{\beta}_\gamma^* + \mathbf{u})) - y_i \mathbf{x}_{i\boldsymbol{\gamma}}(\boldsymbol{\beta}_\gamma^* + \mathbf{u}))\right.\right. \\
&\quad \left.\left. - \log(1 + \exp(\mathbf{x}_{i\boldsymbol{\gamma}}\boldsymbol{\beta}_\gamma^*)) + y_i \mathbf{x}_{i\boldsymbol{\gamma}}\boldsymbol{\beta}_\gamma^*\right) + \frac{1}{2\sigma_0^2}(\beta_0^{*2} + 2\beta_0^*u_0 + u_0^2) - \frac{\beta_0^{*2}}{2\sigma_0^2}\right. \\
&\quad \left.\left. + \lambda \sum_{\substack{\gamma_j=1 \\ j \neq 0}} (|\beta_j^* + u_j| - |\beta_j^*|)\right)\right] d\mathbf{u}
\end{aligned}$$

$$\begin{aligned}
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp \left[ - \left( \sum_{i=1}^n (\log(1 + \exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_{\gamma}^*))) + \frac{\exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_{\gamma}^*)}{1 + \exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_{\gamma}^*)} \mathbf{x}_{i\gamma} \mathbf{u} \right. \right. \\
&\quad \left. \left. + \frac{1}{2} \mathbf{u}^T \mathbf{x}_{i\gamma}^T \frac{\exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_{\gamma}^*)}{(1 + \exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_{\gamma}^*))^2} \mathbf{x}_{i\gamma} \mathbf{u} + R(\mathbf{u}) - \log(1 + \exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_{\gamma}^*)) \right. \right. \\
&\quad \left. \left. - y_i \mathbf{x}_{i\gamma} \mathbf{u} + \frac{1}{2\sigma_0^2} (2\beta_0^* u_0 + u_0^2) + \lambda \sum_{\substack{\gamma_j=1 \\ j \neq 0}} (|\beta_j^* + u_j| - |\beta_j^*|) \right) \right] d\mathbf{u} \quad (3.12) \\
&\approx \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp \left[ - \left( \frac{1}{2} \mathbf{u}^T \left( \sum_{i=1}^n \left( \mathbf{x}_{i\gamma}^T \frac{\exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_{\gamma}^*)}{(1 + \exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_{\gamma}^*))^2} \mathbf{x}_{i\gamma} \right) \right) \mathbf{u} \right. \right. \\
&\quad \left. \left. - \left( \sum_{i=1}^n \left( y_i - \frac{\exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_{\gamma}^*)}{1 + \exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_{\gamma}^*)} \right) \mathbf{x}_{i\gamma} \right) \mathbf{u} + \frac{1}{2\sigma_0^2} (2\beta_0^* u_0 + u_0^2) \right. \right. \\
&\quad \left. \left. + \lambda \sum_{\substack{\gamma_j=1 \\ j \neq 0}} (|\beta_j^* + u_j| - |\beta_j^*|) \right) \right] d\mathbf{u} \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp \left[ - \left( \frac{1}{2} \mathbf{u}^T (\mathbf{A} + \mathbf{H}) \mathbf{u} - (\mathbf{e} - \mathbf{t})^T \mathbf{u} \right. \right. \\
&\quad \left. \left. + \lambda \sum_{\substack{\gamma_j=1 \\ j \neq 0}} (|\beta_j^* + u_j| - |\beta_j^*|) \right) \right] d\mathbf{u} \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp \left[ - \left( \frac{1}{2} \mathbf{u}^T (\mathbf{A} + \mathbf{H}) \mathbf{u} - (\mathbf{e} - \mathbf{t})^T (\mathbf{A} + \mathbf{H})^{-1} (\mathbf{A} + \mathbf{H}) \mathbf{u} \right. \right. \\
&\quad \left. \left. \pm \frac{1}{2} (\mathbf{e} - \mathbf{t})^T (\mathbf{A} + \mathbf{H})^{-1} (\mathbf{A} + \mathbf{H}) (\mathbf{A} + \mathbf{H})^{-1} (\mathbf{e} - \mathbf{t}) \right. \right. \\
&\quad \left. \left. + \lambda \sum_{\substack{\gamma_j=1 \\ j \neq 0}} (|\beta_j^* + u_j| - |\beta_j^*|) \right) \right] d\mathbf{u} \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp \left[ - \left( \frac{1}{2} \mathbf{u}^T \boldsymbol{\Psi}^{-1} \mathbf{u} - \mathbf{m}^T \boldsymbol{\Psi}^{-1} \mathbf{u} + \frac{1}{2} \mathbf{m}^T \boldsymbol{\Psi}^{-1} \mathbf{m} \right. \right. \\
&\quad \left. \left. - \frac{1}{2} (\mathbf{e} - \mathbf{t})^T (\mathbf{A} + \mathbf{H})^{-1} (\mathbf{A} + \mathbf{H}) (\mathbf{A} + \mathbf{H})^{-1} (\mathbf{e} - \mathbf{t}) \right. \right. \\
&\quad \left. \left. + \lambda \sum_{\substack{\gamma_j=1 \\ j \neq 0}} (|\beta_j^* + u_j| - |\beta_j^*|) \right) \right] d\mathbf{u}
\end{aligned}$$

$$\begin{aligned}
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp \left[ - \left( \frac{1}{2} (\mathbf{u} - \mathbf{m})^T \boldsymbol{\Psi}^{-1} (\mathbf{u} - \mathbf{m}) \right. \right. \\
&\quad \left. \left. + \lambda \sum_{\substack{\gamma_j=1 \\ j \neq 0}} (|\beta_j^* + u_j| - |\beta_j^*|) \right) \right] d\mathbf{u} \\
&\quad \times \exp \left( \frac{1}{2} (\mathbf{e} - \mathbf{t})^T (\mathbf{A} + \mathbf{H})^{-1} (\mathbf{e} - \mathbf{t}) \right), \tag{3.13}
\end{aligned}$$

where  $R(\mathbf{u})$  is the Taylor series remainder term in (3.12) and the following approximate equality comes from dropping  $R(\mathbf{u})$ . We write,

$$\boldsymbol{\Psi}^{-1} = \mathbf{A} + \mathbf{H},$$

and

$$\mathbf{m} = (\mathbf{A} + \mathbf{H})^{-1} (\mathbf{e} - \mathbf{t}).$$

The remainder  $R(\mathbf{u}) = o(\|\mathbf{u}\|^2)$  as  $\mathbf{u} \rightarrow \mathbf{0}$ , according to the multidimensional Taylor theorem. Therefore there exist  $c, \delta > 0$ , such that  $R(\mathbf{u}) < c\|\mathbf{u}\|^2$  when  $\|\mathbf{u}\| < \delta$ . Moreover if  $\|\mathbf{u}\| < \delta$ ,  $\exp[\lambda \sum_{\gamma_j=1, j \neq 0} (|\beta_j^* + u_j| - |\beta_j^*|)]$  lies in an interval  $(1 - \eta, 1 + \eta)$  where  $\eta$  is small. By modifying the proof of Proposition 4.7.1 of Lange [22], we conclude that the ratio of (3.12) and (3.13) converges to 1 as  $\|\mathbf{A} + \mathbf{H}\| \rightarrow \infty$ . From (3.5) it can be seen that  $\mathbf{A}$  is a sum of nonnegative matrices, so under mild conditions on the  $\mathbf{x}_i$ ,  $\|\mathbf{A}\| \rightarrow \infty$  as  $n \rightarrow \infty$ . Therefore our Laplace approximation (3.12)/(3.13)  $\approx 1$  holds.

Define

$$\begin{aligned}
f(\mathbf{u}) &= \frac{1}{2} (\mathbf{u} - \mathbf{m})^T (\mathbf{A} + \mathbf{H}) (\mathbf{u} - \mathbf{m}) \\
&\quad - \frac{1}{2} (\mathbf{e} - \mathbf{t})^T (\mathbf{A} + \mathbf{H})^{-1} (\mathbf{e} - \mathbf{t}) + \lambda \sum_{\substack{\gamma_j=1 \\ j \neq 0}} (|\beta_j^* + u_j| - |\beta_j^*|)
\end{aligned}$$

Note that by the definition of  $\boldsymbol{\beta}_\gamma^*$ ,  $f(\mathbf{u})$  is minimized at  $\mathbf{u} = \mathbf{0}$ . Observe that  $\mathbf{m}$ ,  $\mathbf{A} + \mathbf{H}$  and the second term of  $f(\mathbf{u})$  are constant with respect to  $\mathbf{u}$ .

### 3.2.1 Regular Class

Because  $\boldsymbol{\beta}_\gamma^*$  does not contain zeros,  $f(\mathbf{u})$  is differentiable in a neighborhood of  $\mathbf{u} = \mathbf{0}$ , and

$$\frac{\partial^2 f(\mathbf{u})}{\partial \mathbf{u} \partial \mathbf{u}^T} \Big|_{\mathbf{u}=\mathbf{0}} = \left( \sum_{i=1}^n \left( \mathbf{x}_{i\gamma}^T \frac{\exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_\gamma^*)}{(1 + \exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_\gamma^*))^2} \mathbf{x}_{i\gamma} \right) \right) + \begin{pmatrix} \frac{1}{2\sigma_0^2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} = \mathbf{A} + \mathbf{H}. \quad (3.14)$$

Using Taylor expansion and the Laplace approximation (3.12)/(3.13)  $\approx 1$ , we obtain

$$\begin{aligned} & \frac{\sqrt{\det(\mathbf{A} + \mathbf{H})}}{(\sqrt{2\pi})^{|\gamma|+1}} \exp\left(-\frac{1}{2}(\mathbf{e} - \mathbf{t})^T (\mathbf{A} + \mathbf{H})^{-1} (\mathbf{e} - \mathbf{t})\right) \\ & \times \left\{ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp \left[ - \left( \sum_{i=1}^n (\log(1 + \exp(\mathbf{x}_{i\gamma}(\boldsymbol{\beta}_\gamma^* + \mathbf{u})) - y_i \mathbf{x}_{i\gamma}(\boldsymbol{\beta}_\gamma^* + \mathbf{u})) \right. \right. \right. \\ & \quad \left. \left. \left. - \log(1 + \exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_\gamma^*)) + y_i \mathbf{x}_{i\gamma} \boldsymbol{\beta}_\gamma^*) + \frac{(\beta_0^* + u_0)^2}{2\sigma_0^2} - \frac{\beta_0^{*2}}{2\sigma_0^2} \right. \right. \right. \\ & \quad \left. \left. \left. + \lambda \sum_{\substack{\gamma_j=1 \\ j \neq 0}} (|\beta_j^* + u_j| - |\beta_j^*|) \right) \right] d\mathbf{u} \right\} \\ & \approx 1, \end{aligned} \quad (3.15)$$

Substituting (3.15) into (3.11), the posterior probability is asymptotically

$$\begin{aligned} & P(\boldsymbol{\gamma} | \mathbf{Y}) \\ & = G(\mathbf{Y}) k^{|\boldsymbol{\gamma}|} \frac{1}{\sigma_0} \\ & \quad \times \exp \left[ - \min_{\boldsymbol{\beta}_\gamma} \left( \sum_{i=1}^n (\log(1 + \exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_\gamma)) - y_i \mathbf{x}_{i\gamma} \boldsymbol{\beta}_\gamma) + \frac{\beta_0^2}{2\sigma_0^2} + \lambda \sum_{\substack{\gamma_j=1 \\ j \neq 0}} |\beta_j| \right) \right]. \end{aligned} \quad (3.16)$$

With high probability, in large samples with  $n \rightarrow \infty$ , the difference between  $\beta_\gamma$  and  $\beta_\gamma^*$  which is  $\mathbf{u}$ , is very small. Hence, the approximation above (3.15) holds. (The statement is proved as Theorem 5.3 in Chapter 5.)

### 3.2.2 Nonregular Class

The posterior probability of the nonregular models cannot be obtained in the same way since  $f(\mathbf{u})$  is not differentiable at  $\mathbf{u} = \mathbf{0}$ . As discussed in this section, we can show that the posterior probability of a nonregular model is always dominated by that of a regular submodel. If one wants to do variable selection, one may simply search through the models in the regular model class, which results in a parsimonious model compared to the nonregular models.

Consider a model  $\gamma = (\gamma_0, \gamma_1, \gamma_2, \dots, \gamma_p)^T$ , that is, a  $(p+1)$ -dimensional vector taking the form of  $(1, \dots, 1, 0, \dots, 0)$ , where the first  $|\gamma| + 1$  components are 1's, and  $|\gamma| = \sum_{i=1}^p \gamma_i$ . The corresponding  $\beta_\gamma = (\beta_0, \beta_1, \dots, \beta_s, 0, \dots, 0)$ , where  $s < |\gamma|$ . Let  $\gamma^* = (\gamma_0, \gamma_1, \dots, \gamma_s, 0, \dots, 0)$ , with  $\gamma_0^* = \gamma_1^* = \dots = \gamma_s^* = 1$  and  $\gamma_j^* = 0$ ,  $j > s$ . We wish to show that  $P(\gamma|\mathbf{Y}) \leq P(\gamma^*|\mathbf{Y})$ .

Since  $f(\mathbf{u})$  is minimized at  $\mathbf{u} = \mathbf{0}$ , the derivative of  $f(\mathbf{u})$  evaluated at  $\mathbf{u} = \mathbf{0}$ , is

$$\left. \frac{\partial f}{\partial u_j} \right|_{\mathbf{u}=\mathbf{0}} = 0, \quad \forall j \leq s,$$

so that, after substituting for  $\mathbf{A}$ ,  $\mathbf{H}$ ,  $\mathbf{e}$  and  $\mathbf{t}$ ,

$$\begin{cases} \sum_{i=1}^n \left( y_i - \frac{\exp(\mathbf{x}_i \gamma \beta_\gamma^*)}{1 + \exp(\mathbf{x}_i \gamma \beta_\gamma^*)} \right) \mathbf{x}_{ij} - (\beta_0^* / \sigma_0^2) = \lambda \operatorname{sgn}(\beta_{\gamma_0}^*), & j = 0, \\ \sum_{i=1}^n \left( y_i - \frac{\exp(\mathbf{x}_i \gamma \beta_\gamma^*)}{1 + \exp(\mathbf{x}_i \gamma \beta_\gamma^*)} \right) \mathbf{x}_{ij} = \lambda \operatorname{sgn}(\beta_{\gamma_j}^*), & 0 < j \leq s. \end{cases}$$

For  $s < j \leq |\gamma| + 1$ ,  $\beta_{\gamma j}^* = 0$ . Therefore,

$$\left. \frac{\partial f}{\partial u_j} \right|_{u_j=0^+; u_l=0, \forall l \neq 0} = - \left( \sum_{i=1}^n \left( y_i - \frac{\exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_{\gamma}^*)}{1 + \exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_{\gamma}^*)} \right) \mathbf{x}_{ij} - \frac{\beta_0^*}{\sigma_0^2} \right) + \lambda \geq 0,$$

so that

$$\lambda \geq \sum_{i=1}^n \left( y_i - \frac{\exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_{\gamma}^*)}{1 + \exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_{\gamma}^*)} \right) \mathbf{x}_{ij} - \frac{\beta_0^*}{\sigma_0^2},$$

and

$$\left. \frac{\partial f}{\partial u_j} \right|_{u_j=0^-; u_l=0, \forall l \neq 0} = - \left( \sum_{i=1}^n \left( y_i - \frac{\exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_{\gamma}^*)}{1 + \exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_{\gamma}^*)} \right) \mathbf{x}_{ij} - \frac{\beta_0^*}{\sigma_0^2} \right) - \lambda \leq 0,$$

so that

$$\lambda \geq - \left( \sum_{i=1}^n \left( y_i - \frac{\exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_{\gamma}^*)}{1 + \exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_{\gamma}^*)} \right) \mathbf{x}_{ij} - \frac{\beta_0^*}{\sigma_0^2} \right),$$

and  $\lambda > 0$ . Therefore,

$$\lambda \geq \left| \sum_{i=1}^n \left( y_i - \frac{\exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_{\gamma}^*)}{1 + \exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_{\gamma}^*)} \right) \mathbf{x}_{ij} - \frac{\beta_0^*}{\sigma_0^2} \right|, \quad (3.17)$$

for  $j = 0$ . Similarly,

$$\lambda \geq \left| \sum_{i=1}^n \left( y_i - \frac{\exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_{\gamma}^*)}{1 + \exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_{\gamma}^*)} \right) \mathbf{x}_{ij} \right|, \quad (3.18)$$

if  $0 < j \leq s$ , and

$$\beta_{\gamma j}^* = 0, \quad (3.19)$$

if  $s < j \leq |\gamma| + 1$ .

The expression below, which is part of the formula for  $P(\gamma|\mathbf{Y})$  in (3.11) can

be bounded using Laplace's approximation and (3.17), (3.18) and (3.19):

$$\begin{aligned}
& \frac{\sqrt{\det(\mathbf{A} + \mathbf{H})}}{(\sqrt{2\pi})^{|\gamma|+1}} \exp\left(-\frac{1}{2}(\mathbf{e} - \mathbf{t})^T(\mathbf{A} + \mathbf{H})^{-1}(\mathbf{e} - \mathbf{t})\right) \\
& \times \left\{ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left[-\left(\sum_{i=1}^n (\log(1 + \exp(\mathbf{x}_{i\gamma}(\boldsymbol{\beta}_{\gamma}^* + \mathbf{u})) - y_i \mathbf{x}_{i\gamma}(\boldsymbol{\beta}_{\gamma}^* + \mathbf{u}))\right.\right.\right. \\
& \quad \left.\left.\left. - \log(1 + \exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_{\gamma}^*)) + y_i \mathbf{x}_{i\gamma} \boldsymbol{\beta}_{\gamma}^*) + \frac{(\beta_0^* + u_0)^2}{2\sigma_0^2} - \frac{\beta_0^{*2}}{2\sigma_0^2}\right.\right.\right. \\
& \quad \left.\left.\left. + \lambda \sum_{\substack{\gamma_j=1 \\ j \neq 0}} (|\beta_j^* + u_j| - |\beta_j^*|)\right)\right] d\mathbf{u} \right\} \\
& < \frac{\sqrt{\det(\mathbf{A} + \mathbf{H})}}{(\sqrt{2\pi})^{|\gamma|+1}} \exp\left(-\frac{1}{2}(\mathbf{e} - \mathbf{t})^T(\mathbf{A} + \mathbf{H})^{-1}(\mathbf{e} - \mathbf{t})\right) \\
& \quad \times \left\{ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left[-\left(\frac{1}{2}\mathbf{u}^T(\mathbf{A} + \mathbf{H})\mathbf{u}\right)\right] d\mathbf{u} \right\} \\
& < \frac{\sqrt{\det(\mathbf{A} + \mathbf{H})}}{(\sqrt{2\pi})^{|\gamma|+1}} \left\{ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left[-\left(\frac{1}{2}\mathbf{u}^T(\mathbf{A} + \mathbf{H})\mathbf{u}\right)\right] d\mathbf{u} \right\} \\
& = 1. \tag{3.20}
\end{aligned}$$

Hence, asymptotically

$$P(\boldsymbol{\gamma}|\mathbf{Y})$$

$$\begin{aligned}
& < G(\mathbf{Y}) k^{|\gamma|} \frac{1}{\sigma_0} \\
& \quad \times \exp\left[-\min\left(\sum_{i=1}^n (\log(1 + \exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_{\gamma})) - y_i \mathbf{x}_{i\gamma} \boldsymbol{\beta}_{\gamma}) + \frac{\beta_0^2}{2\sigma_0^2} + \lambda \sum_{\substack{\gamma_j=1 \\ j \neq 0}} |\beta_j|)\right)\right]. \tag{3.21}
\end{aligned}$$

Since  $\beta_{\gamma_j}^* = \beta_{\gamma^*_j}^*$ , for any  $j \leq s$ , comparing the posterior probability for the nonregular model (3.21) and the posterior probability for the regular model (3.16), asymptotically, the ratio is

$$\frac{P(\boldsymbol{\gamma}|\mathbf{Y})}{P(\boldsymbol{\gamma}^*|\mathbf{Y})} \leq k^{|\gamma|-s}. \tag{3.22}$$



In fact, we have established  $P(\boldsymbol{\gamma}|\mathbf{Y})/P(\boldsymbol{\gamma}^*|\mathbf{Y}) < k^{|\boldsymbol{\gamma}|-s}(1+\epsilon)$ . The factor  $(1+\epsilon)$  arises by accounting for the error in the Laplace approximations in the numerator and denominator. The size of these errors comes from bounding  $\exp(R(\mathbf{u}))$ . However, if  $n$  is sufficiently large,  $\mathbf{u}$  is very close to zero with high probability, and therefore  $\epsilon$  can be made arbitrarily small.

The fact that  $P(\boldsymbol{\gamma}|\mathbf{Y}) \leq P(\boldsymbol{\gamma}^*|\mathbf{Y})$  if  $k \leq 1$  implies that, in order to locate the model with the maximum posterior probability, one only needs to concentrate on the models in the regular class and does not need to consider the models in the nonregular class. One may set  $k = 1$  to achieve this. This constrains our choice of prior distributions.

### 3.3 Connection to LASSO

In this section, the connection of the hierarchical structure to LASSO is elucidated. Before that, let us first discuss  $\boldsymbol{\beta}_{\boldsymbol{\gamma}^*}$ , the minimizer of the criterion embedded in the posterior probability  $P(\boldsymbol{\gamma}|\mathbf{Y})$ , in a limiting sense.

Supposed that  $\kappa_0 = 1/\sigma_0$ . Define

$$\begin{aligned} h_{\kappa_0}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) &= h_{\kappa_0}(\beta_0, \boldsymbol{\beta}_{[-1]}) \\ &= \sum_{i=1}^n (\log(1 + \exp(\mathbf{x}_{i\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})) - y_i \mathbf{x}_{i\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}) + \frac{\beta_0^2}{2\sigma_0^2} + \lambda \sum_{\substack{\gamma_j=1 \\ j \neq 0}} |\beta_j| \\ &= \sum_{i=1}^n (\log(1 + \exp(\mathbf{x}_{i\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})) - y_i \mathbf{x}_{i\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}) + \frac{\beta_0^2}{2} \kappa_0^2 + \lambda \sum_{\substack{\gamma_j=1 \\ j \neq 0}} |\beta_j|, \end{aligned}$$

where  $\boldsymbol{\beta}_{\boldsymbol{\gamma}[-1]}$  is the vector  $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$  deleting the zero-th component (the intercept).

For each fixed  $\kappa_0$ , recall that  $(\beta_0^*(\kappa_0), \boldsymbol{\beta}_{\boldsymbol{\gamma}[-1]}^*(\kappa_0))$  is the minimizer of

$h_{\kappa_0}(\beta_0, \boldsymbol{\beta}_{\gamma[-1]})$ . Assuming  $\gamma$  is a model in the regular class, the minimizer  $(\beta_0^*(\kappa_0), \boldsymbol{\beta}_{\gamma[-1]}^*(\kappa_0))$  is obtained by solving the system of equations

$$\frac{\partial h_{\kappa_0}(\boldsymbol{\beta}_{\gamma})}{\partial \boldsymbol{\beta}_{\gamma}} = \begin{bmatrix} \frac{\partial}{\partial \beta_0} \\ \frac{\partial}{\partial \boldsymbol{\beta}_{\gamma[-1]}} \end{bmatrix} h_{\kappa_0}(\beta_0, \boldsymbol{\beta}_{\gamma[-1]}) = \mathbf{0}.$$

Since  $\gamma$  is a regular model,  $\partial h_{\kappa_0}(\boldsymbol{\beta}_{\gamma})/\partial \boldsymbol{\beta}_{\gamma}$  is differentiable for each fixed  $\kappa_0$ . By the Implicit Function Theorem,  $(\beta_0^*(\kappa_0), \boldsymbol{\beta}_{\gamma[-1]}^*(\kappa_0))$ , is a continuously differentiable function of  $\kappa_0$ . Let  $\kappa_0 \rightarrow 0$ ,

$$(\beta_0^*(\kappa_0), \boldsymbol{\beta}_{\gamma[-1]}^*(\kappa_0)) \rightarrow (\check{\beta}_0^*(\kappa_0), \check{\boldsymbol{\beta}}_{\gamma[-1]}^*(\kappa_0)), \quad (3.23)$$

and let  $(\check{\beta}_0^*(\kappa_0), \check{\boldsymbol{\beta}}_{\gamma[-1]}^*(\kappa_0))$  be the minimizer of the limiting criterion,

$$\tilde{h}(\boldsymbol{\gamma}) \equiv \operatorname{argmin}_{\boldsymbol{\beta}_{\gamma}} \left( \sum_{i=1}^n (\log(1 + \exp(\mathbf{x}_{i\boldsymbol{\gamma}} \boldsymbol{\beta}_{\gamma})) - y_i \mathbf{x}_{i\boldsymbol{\gamma}} \boldsymbol{\beta}_{\gamma}) + \lambda \sum_{\substack{\gamma_j=1 \\ j \neq 0}} |\beta_j| \right). \quad (3.24)$$

Note that  $\tilde{h}(\boldsymbol{\gamma})$  is of the form of a LASSO-type criterion, where the first part is like the loglikelihood of the logistic regression model and the second part is the penalty component with regularization parameter  $\lambda$ .

In general, according to (3.22), if  $k \leq 1$ , one may confine the search for the highest posterior probability to the regular class and skip the entire nonregular class. Recall that for each fixed  $\sigma_0$ , the posterior probability for the regular model asymptotically is

$$P(\boldsymbol{\gamma}|\mathbf{Y}) \approx G(\mathbf{Y}) \frac{1}{\sigma_0} \exp(-h(\boldsymbol{\gamma})).$$

As a result, one should target the regular model which minimizes  $h(\boldsymbol{\gamma})$ . Also, as  $\sigma_0 \rightarrow \infty$ , which is equivalent to  $\kappa_0 \rightarrow 0$ , one just has to focus on minimizing  $\tilde{h}(\boldsymbol{\gamma})$ .

Fortunately, by the proposition below there is no need to go through each individual model in the regular class to determine the minimizer of  $\tilde{h}(\boldsymbol{\gamma})$  :

**Proposition 3.1** *Let  $\hat{\boldsymbol{\beta}}$  minimize  $\sum_{i=1}^n [\log(1 + \exp(\mathbf{x}_i \boldsymbol{\beta})) - y_i \mathbf{x}_i \boldsymbol{\beta}] + \lambda \sum_{j=1}^p |\beta_j|$ , and let model  $\hat{\boldsymbol{\gamma}}$  be such that  $\hat{\gamma}_j = I(\hat{\beta}_j \neq 0)$ , where  $I(\cdot)$  is the indicator function. Then  $\hat{\boldsymbol{\gamma}}$  is the regular model that minimizes  $\tilde{h}(\boldsymbol{\gamma})$ .*

**Proof:** Note that if  $\boldsymbol{\gamma}_1$  is a submodel of  $\boldsymbol{\gamma}_2$ ,  $\tilde{h}(\boldsymbol{\gamma}_1) \geq \tilde{h}(\boldsymbol{\gamma}_2)$  due to the fact that  $\tilde{h}$  is a decreasing function of each of the components of  $\boldsymbol{\gamma}$ .

By the definition of  $\boldsymbol{\gamma}$  in Section 2.2,  $\boldsymbol{\gamma} = \mathbf{1}_{p+1}$  represents the full model. Therefore, for any regular model  $\boldsymbol{\gamma}$ ,

$$\tilde{h}(\hat{\boldsymbol{\gamma}}) = \tilde{h}(\mathbf{1}) \leq \tilde{h}(\boldsymbol{\gamma})$$

since  $\tilde{h}(\hat{\boldsymbol{\gamma}})$  is a regular model.  $\square$

Observe that  $\tilde{h}(\mathbf{1})$  is exactly the LASSO criterion and not surprisingly,  $\tilde{h}(\hat{\boldsymbol{\gamma}})$  is also the model produced by LASSO. Through Proposition 3.1, our hierarchical Bayes formulations (3.1), (3.2), (3.3) and (3.4) are connected to LASSO with  $k = 1$  and  $\kappa_0 \rightarrow 0$ . With the available R package `glmnet` of Friedman, Hastie and Tibshirani [13], or the R package `glmplath` of Park and Hastie [29], one can compute the LASSO estimate for the entire  $\lambda$  path from  $\lambda = 0$  to  $\lambda \rightarrow \infty$ .

## Chapter 4

### Bayesian Model Selection Criteria

The hierarchical Bayes formulation in the last chapter requires the specification for the values of the hyperparameters. Two approaches will be taken to deal with this problem. Section 4.1 explores the empirical Bayes method and Section 4.2 discussed the fully Bayes approach, both for logistic case. These Bayesian variable selection criteria are presented in Section 4.3 for Poisson regression and in Section 4.4 for linear regression.

#### 4.1 Empirical Bayes Criterion for Logistic Model

Taking  $k = 1$  and assuming  $\sigma_0$  is fixed, one would like to estimate  $\lambda$  in the posterior probability  $P(\boldsymbol{\gamma}|\mathbf{Y})$ , which turns out to be the regularization parameter in LASSO. The empirical Bayes method advocates estimating the parameter from the data. It selects the  $\lambda$  that maximizes the marginal density

$$f(\mathbf{Y}|\lambda) = \sum_{\boldsymbol{\gamma}} \int_{\boldsymbol{\beta}_{\boldsymbol{\gamma}}} P(\mathbf{Y}, \boldsymbol{\gamma}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}) d\boldsymbol{\beta}_{\boldsymbol{\gamma}}. \quad (4.1)$$

(Recall that  $P(\mathbf{Y}, \boldsymbol{\gamma}, \boldsymbol{\beta}_{\boldsymbol{\gamma}})$  defined in (3.9) involves the parameter  $\lambda$ .)

If the number of variables is relatively small, this maximization problem can be solved easily. However, as more variables are introduced, it is increasingly more difficult to calculate this maximizer numerically. Consider an individual term in (4.1), that is, the conditional density of  $\mathbf{Y}$  given a model  $\gamma$ :

$$\begin{aligned} f(\mathbf{Y}|\gamma, \lambda) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\frac{\lambda}{2}\right)^{|\gamma|} \exp \left[ - \left( \sum_{i=1}^n (\log(1 + \exp(\mathbf{x}_{i\gamma}\boldsymbol{\beta}_\gamma)) - y_i \mathbf{x}_{i\gamma}\boldsymbol{\beta}_\gamma) \right. \right. \\ &\quad \left. \left. + \lambda \sum_{\substack{\gamma_j=1 \\ j \neq 0}} |\beta_j| \right) \right] \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{\beta_0^2}{2\sigma_0^2}\right) d\boldsymbol{\beta}_\gamma. \end{aligned}$$

Given a particular tuning parameter  $\lambda$ , let the selected model be  $\hat{\gamma}_\lambda$ . Instead of maximizing the marginal density  $f(\mathbf{Y}|\gamma, \lambda)$  (4.1), one can maximize  $f(\mathbf{Y}|\hat{\gamma}_\lambda, \lambda)$ , the largest component of  $f(\mathbf{Y}|\gamma, \lambda)$ , and select  $\lambda$  this way. George and Foster [14] used a similar approach for linear models with Gaussian priors.

From Section 3.2.2, it is shown that  $\hat{\gamma}$  is regular. Then, for each fixed  $\sigma_0$ ,  $f(\mathbf{Y}|\hat{\gamma}_\lambda, \lambda)$  can be approximated as  $n \rightarrow \infty$  by (3.11), (3.15) and Proposition 3.1 to obtain:

$$\begin{aligned} f(\mathbf{Y}|\hat{\gamma}_\lambda, \lambda) &\approx \left(\frac{\lambda}{2}\right)^{|\hat{\gamma}_\lambda|} \frac{1}{\sigma_0} (\sqrt{2\pi})^{|\hat{\gamma}_\lambda|} (\det(\mathbf{A} + \mathbf{H}))^{-\frac{1}{2}} \exp\left(\frac{1}{2}(\mathbf{e} - \mathbf{t})^T (\mathbf{A} + \mathbf{H})^{-1} (\mathbf{e} - \mathbf{t})\right) \\ &\quad \times \exp \left[ - \min_{\boldsymbol{\beta}_\gamma} \left( \sum_{i=1}^n (\log(1 + \exp(\mathbf{x}_{i\hat{\gamma}_\lambda}\boldsymbol{\beta}_{\hat{\gamma}_\lambda})) - y_i \mathbf{x}_{i\hat{\gamma}_\lambda}\boldsymbol{\beta}_{\hat{\gamma}_\lambda}) + \frac{\beta_0^2}{2\sigma_0^2} + \lambda \sum_{\substack{i \in \hat{\gamma}_\lambda \\ i \neq 0}} |\beta_j| \right) \right] \\ &= \left(\frac{\lambda}{2}\right)^{|\hat{\gamma}_\lambda|} \frac{1}{\sigma_0} (\sqrt{2\pi})^{|\hat{\gamma}_\lambda|} (\det(\mathbf{A} + \mathbf{H}))^{-\frac{1}{2}} \exp\left(\frac{1}{2}(\mathbf{e} - \mathbf{t})^T (\mathbf{A} + \mathbf{H})^{-1} (\mathbf{e} - \mathbf{t})\right) \\ &\quad \times \exp \left[ - \min_{\boldsymbol{\beta}} \left( \sum_{i=1}^n (\log(1 + \exp(\mathbf{x}_i\boldsymbol{\beta})) - y_i \mathbf{x}_i\boldsymbol{\beta}) + \frac{\beta_0^2}{2\sigma_0^2} + \lambda \sum_{i=1}^p |\beta_j| \right) \right]. \quad (4.2) \end{aligned}$$

Therefore, maximizing  $f(\mathbf{Y}|\hat{\gamma}_\lambda, \lambda)$  is approximately equivalent to minimizing the negative logarithm of (4.2). The terms not involving  $\lambda$  can be dropped because they do not affect the minimization. This minimization criterion is named the CML criterion as in George and Foster [14], and it consists of three terms. For each fixed  $\sigma_0$ , the CML criterion by

$$\begin{aligned}
& \text{CML}_{\kappa_0}(\lambda) \\
&= -|\hat{\gamma}_\lambda| \log\left(\frac{\lambda}{2}\right) - |\hat{\gamma}_\lambda| \log \sqrt{2\pi} + \frac{1}{2} \log(\det(\mathbf{A} + \mathbf{H})) \\
&\quad - \left(\frac{1}{2}(\mathbf{e} - \mathbf{t})^T(\mathbf{A} + \mathbf{H})^{-1}(\mathbf{e} - \mathbf{t})\right) \\
&\quad + \min_{\boldsymbol{\beta}} \left( \sum_{i=1}^n (\log(1 + \exp(\mathbf{x}_i \boldsymbol{\beta})) - y_i \mathbf{x}_i \boldsymbol{\beta}) + \frac{\beta_0^2}{2\sigma_0^2} + \lambda \sum_{i=1}^p |\beta_j| \right) \\
&= -|\hat{\gamma}_\lambda| \log\left(\frac{\lambda}{2}\right) - |\hat{\gamma}_\lambda| \log \sqrt{2\pi} + \frac{1}{2} \log(\det(\mathbf{A} + \mathbf{H})) \\
&\quad - \left(\frac{1}{2}(\mathbf{e} - \mathbf{t})^T(\mathbf{A} + \mathbf{H})^{-1}(\mathbf{e} - \mathbf{t})\right) \\
&\quad + \min_{\boldsymbol{\beta}} \left( \sum_{i=1}^n (\log(1 + \exp(\mathbf{x}_i \boldsymbol{\beta})) - y_i \mathbf{x}_i \boldsymbol{\beta}) + \frac{\beta_0^2}{2} \kappa_0^2 + \lambda \sum_{i=1}^p |\beta_j| \right), \quad (4.3)
\end{aligned}$$

where  $\kappa_0 = 1/\sigma_0$  was defined in Chapter 3. As  $\kappa_0 \rightarrow 0$  or equivalently  $\sigma_0 \rightarrow \infty$ , the minimizer of  $\text{CML}_{\kappa_0}$  converges to the minimizer of CML by the Implicit Function Theorem and (3.23), where by definition

$$\begin{aligned}
\text{CML}(\lambda) &= -|\hat{\gamma}_\lambda| \log\left(\frac{\lambda}{2}\right) - |\hat{\gamma}_\lambda| \log \sqrt{2\pi} + \frac{1}{2} \log(\det(\mathbf{A} + \mathbf{H})) \\
&\quad - \left(\frac{1}{2}(\mathbf{e} - \mathbf{t})^T(\mathbf{A} + \mathbf{H})^{-1}(\mathbf{e} - \mathbf{t})\right) \\
&\quad + \min_{\boldsymbol{\beta}} \left( \sum_{i=1}^n (\log(1 + \exp(\mathbf{x}_i \boldsymbol{\beta})) - y_i \mathbf{x}_i \boldsymbol{\beta}) + \lambda \sum_{i=1}^p |\beta_j| \right). \quad (4.4)
\end{aligned}$$

## 4.2 Fully Bayes Criterion

The fully Bayes approach deals with the hyperparameters,  $\tau$  and  $q$ , by supplying them with a prior distribution. To implement the Bayesian procedure, one needs to integrate out the hyperparameters to obtain the posterior distribution.

Consider the same hierarchical formulations (3.1), (3.2), (3.3) and (3.4), but now assume multiplicative priors on  $\tau$  and  $q$  on a restricted region, such that

$$\pi(\tau, q) = \pi(\tau)\pi(q).$$

The model is once more divided into two classes: regular and nonregular. Flat priors will be explored in Section 4.2.1, and conjugate prior distributions for  $\tau$  and  $q$  will be investigated in Section 4.2.3.

### 4.2.1 Restricted Region

Building upon the hierarchical formulations (3.1), (3.2) and (3.3), the marginal distribution of  $\mathbf{Y}$  given  $\gamma$  and  $\tau$  for the regular case is, by (3.13)

$$\begin{aligned} P(\mathbf{Y}|\gamma, \tau) &= \int_{-\infty}^{\infty} \left( \frac{1}{\sigma_0 \sqrt{2\pi}} \right) \left( \frac{\tau}{2} \right)^{|\gamma|} \\ &\quad \times \exp \left[ - \left( \sum_{i=1}^n (\log(1 + \exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_\gamma)) - y_i \mathbf{x}_{i\gamma} \boldsymbol{\beta}_\gamma) + \frac{\beta_0^2}{2\sigma_0^2} + \tau \sum_{\substack{\gamma_j=1 \\ j \neq 0}} |\beta_j| \right) \right] d\boldsymbol{\beta}_\gamma \end{aligned}$$

$$\begin{aligned}
&\approx \left(\frac{1}{\sigma_0}\right) \left(\frac{\tau}{2}\right)^{|\gamma|} (\sqrt{2\pi})^{|\gamma|} (\det(\mathbf{A} + \mathbf{H}))^{-\frac{1}{2}} \exp\left(\frac{1}{2}(\mathbf{e} - \mathbf{t})^T (\mathbf{A} + \mathbf{H})^{-1} (\mathbf{e} - \mathbf{t})\right) \\
&\quad \times \exp\left[-\min_{\boldsymbol{\beta}_\gamma} \left(\sum_{i=1}^n (\log(1 + \exp(\mathbf{x}_{i\gamma}\boldsymbol{\beta}_\gamma)) - y_i \mathbf{x}_{i\gamma}\boldsymbol{\beta}_\gamma) + \frac{\beta_0^2}{2\sigma_0^2} + \tau \sum_{\substack{\gamma_j=1 \\ j \neq 0}} |\beta_j|)\right)\right],
\end{aligned} \tag{4.5}$$

while  $P(\mathbf{Y}|\boldsymbol{\gamma}, \tau)$  for the nonregular case is, using (3.13) and (3.20),

$$\begin{aligned}
&P(\mathbf{Y}|\boldsymbol{\gamma}, \tau) \\
&= \int_{-\infty}^{\infty} \left(\frac{1}{\sigma_0\sqrt{2\pi}}\right) \left(\frac{\tau}{2}\right)^{|\gamma|} \\
&\quad \times \exp\left[-\left(\sum_{i=1}^n (\log(1 + \exp(\mathbf{x}_{i\gamma}\boldsymbol{\beta}_\gamma)) - y_i \mathbf{x}_{i\gamma}\boldsymbol{\beta}_\gamma) + \frac{\beta_0^2}{2\sigma_0^2} + \tau \sum_{\substack{\gamma_j=1 \\ j \neq 0}} (|\beta_j|)\right)\right] d\boldsymbol{\beta}_\gamma \\
&< \left(\frac{1}{\sigma_0}\right) \left(\frac{\tau}{2}\right)^{|\gamma|} (\sqrt{2\pi})^{|\gamma|} (\det(\mathbf{A} + \mathbf{H}))^{-\frac{1}{2}} \\
&\quad \times \exp\left[-\min_{\boldsymbol{\beta}_\gamma} \left(\sum_{i=1}^n (\log(1 + \exp(\mathbf{x}_{i\gamma}\boldsymbol{\beta}_\gamma)) - y_i \mathbf{x}_{i\gamma}\boldsymbol{\beta}_\gamma) + \frac{\beta_0^2}{2\sigma_0^2} + \tau \sum_{\substack{\gamma_j=1 \\ j \neq 0}} (|\beta_j|)\right)\right],
\end{aligned} \tag{4.6}$$

where  $\mathbf{A}$ ,  $\mathbf{H}$ ,  $\mathbf{e}$  and  $\mathbf{t}$  are defined in (3.5), (3.6), (3.7) and (3.8). Incorporating (3.4), which is the prior distribution for  $\boldsymbol{\gamma}$ , one may express the posterior distribution of  $\boldsymbol{\gamma}$  given  $\mathbf{Y}$ ,  $\tau$  and  $q$  by

$$\pi(\boldsymbol{\gamma}|\mathbf{Y}, \tau, q) \propto \pi(\boldsymbol{\gamma}|q) P(\mathbf{Y}|\boldsymbol{\gamma}, \tau).$$



For the regular case, from (3.4) and (4.5),

$$\begin{aligned}
& \pi(\boldsymbol{\gamma}|\mathbf{Y}, \tau, q) \\
& \propto q^{|\boldsymbol{\gamma}|}(1-q)^{p-|\boldsymbol{\gamma}|} \left(\frac{1}{\sigma_0}\right) \left(\frac{\tau}{2}\right)^{|\boldsymbol{\gamma}|} (\sqrt{2\pi})^{|\boldsymbol{\gamma}|} \\
& \quad \times \exp \left[ - \min_{\boldsymbol{\beta}_\gamma} \left( \sum_{i=1}^n (\log(1 + \exp(\mathbf{x}_{i\boldsymbol{\gamma}}\boldsymbol{\beta}_\gamma)) - y_i \mathbf{x}_{i\boldsymbol{\gamma}}\boldsymbol{\beta}_\gamma) + \frac{\beta_0^2}{2\sigma_0^2} + \tau \sum_{\substack{\gamma_j=1 \\ j \neq 0}} (|\beta_j|) \right) \right].
\end{aligned} \tag{4.7}$$

Similarly, for the nonregular case using (3.4), (4.6) and (3.20),

$$\begin{aligned}
& \pi(\boldsymbol{\gamma}|\mathbf{Y}, \tau, q) \\
& \propto q^{|\boldsymbol{\gamma}|}(1-q)^{p-|\boldsymbol{\gamma}|} \left(\frac{1}{\sigma_0}\right) \left(\frac{\tau}{2}\right)^{|\boldsymbol{\gamma}|} (\sqrt{2\pi})^{|\boldsymbol{\gamma}|} \\
& \quad \times \exp \left( \frac{1}{2}(\mathbf{e} - \mathbf{t})^T (\mathbf{A} + \mathbf{H})^{-1} (\mathbf{e} - \mathbf{t}) \right) \\
& \quad \times \exp \left[ - \min_{\boldsymbol{\beta}_\gamma} \left( \sum_{i=1}^n (\log(1 + \exp(\mathbf{x}_{i\boldsymbol{\gamma}}\boldsymbol{\beta}_\gamma)) - y_i \mathbf{x}_{i\boldsymbol{\gamma}}\boldsymbol{\beta}_\gamma) + \frac{\beta_0^2}{2\sigma_0^2} + \tau \sum_{\substack{\gamma_j=1 \\ j \neq 0}} (|\beta_j|) \right) \right] \\
& < q^{|\boldsymbol{\gamma}|}(1-q)^{p-|\boldsymbol{\gamma}|} \left(\frac{1}{\sigma_0}\right) \left(\frac{\tau}{2}\right)^{|\boldsymbol{\gamma}|} (\sqrt{2\pi})^{|\boldsymbol{\gamma}|} \\
& \quad \times \exp \left[ - \min_{\boldsymbol{\beta}_\gamma} \left( \sum_{i=1}^n (\log(1 + \exp(\mathbf{x}_{i\boldsymbol{\gamma}}\boldsymbol{\beta}_\gamma)) - y_i \mathbf{x}_{i\boldsymbol{\gamma}}\boldsymbol{\beta}_\gamma) + \frac{\beta_0^2}{2\sigma_0^2} + \tau \sum_{\substack{\gamma_j=1 \\ j \neq 0}} (|\beta_j|) \right) \right].
\end{aligned} \tag{4.8}$$

Following the notation in Section 3.2.2, let  $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2, \dots, \gamma_p)^T$ , where the first  $|\boldsymbol{\gamma}| + 1$  components are 1's and  $|\boldsymbol{\gamma}| = \sum_{i=1}^p \gamma_i$ . Only the first  $s + 1$  elements of those  $|\boldsymbol{\gamma}| + 1$  components of the vector  $\boldsymbol{\beta}_{\boldsymbol{\gamma}^*}$  are nonzero. Recall that  $\boldsymbol{\gamma}^*$  is a submodel of  $\boldsymbol{\gamma}$  with its first  $s + 1$  components equal to one.

Since  $\beta_{\gamma,i}^* = \beta_{\gamma^*,i}^*$  for any  $i \leq s$ ,

$$\begin{aligned} \frac{\pi(\boldsymbol{\gamma}|\mathbf{Y}, \tau, q)}{\pi(\boldsymbol{\gamma}^*|\mathbf{Y}, \tau, q)} &\leq q^{|\boldsymbol{\gamma}|-s}(1-q)^{-(|\boldsymbol{\gamma}|-s)} \left(\frac{\tau}{2}\right)^{|\boldsymbol{\gamma}|-s} (\sqrt{2\pi})^{|\boldsymbol{\gamma}|-s} \\ &\leq \left(\frac{q}{1-q} \frac{\tau}{2} \sqrt{2\pi}\right)^{|\boldsymbol{\gamma}|-s} = k^{|\boldsymbol{\gamma}|-s} \end{aligned}$$

It is clear that if  $q(1-q)^{-1}(\tau/2)\sqrt{2\pi} \leq 1$ , the posterior probability of the nonregular case is smaller or equal to the posterior probability of a regular case. We impose this condition on any prior distribution for  $(\tau, q)$ . However, the area under the region  $R$  is unbounded, where

$$R = \left\{ (\tau, q) : \left( \frac{1-q}{q} \sqrt{\frac{2}{\pi}} \right) \geq \tau, \tau > 0, 0 < q < 1 \right\}.$$

Instead, the prior distribution of  $(\tau, q)$  is restricted to the bounded region  $R'$ , where  $r$  is a fixed value, and

$$R' = \left\{ (\tau, q) : \left( \frac{1-q}{q} \sqrt{\frac{2}{\pi}} \right) \geq r, \tau \leq r, 0 < q < 1 \right\}. \quad (4.9)$$

Forcing  $(\tau, q) \in R'$  assures that the posterior will be proper. (See Figure 4.1.) Therefore, when searching for the best model with the highest posterior probability, one should concentrate on the models in the regular class with the restricted region  $R'$ .

## 4.2.2 Flat priors

Since the model with the maximum posterior probability is contained in the regular class, we should focus only on the regular model for the rest of this chapter. Assuming that  $\tau$  and  $q$  both have a uniform prior over the restricted region  $R'$  to

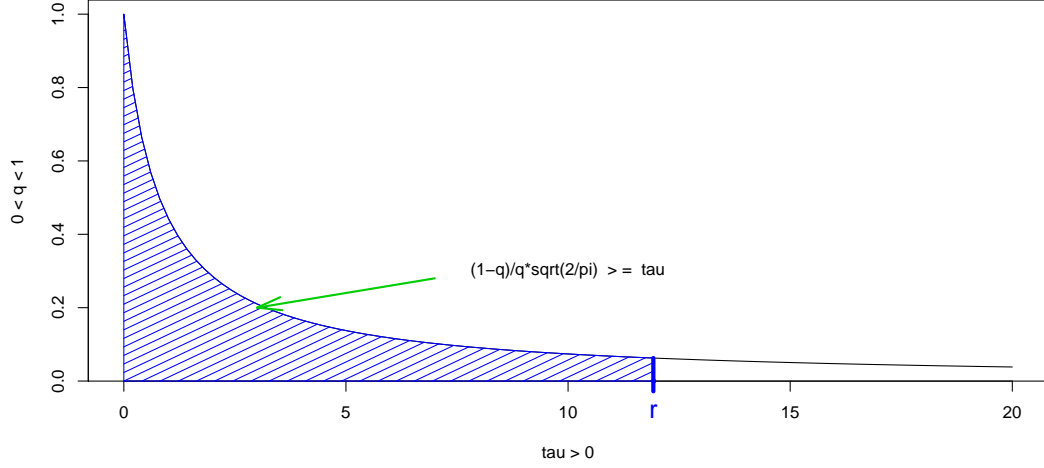


Figure 4.1: Restricted Region  $R'$

reflect not much prior information on the hyperparameters and independent priors.

Then for each fixed  $\sigma_0$ , the posterior distribution becomes

$$\begin{aligned}
& \pi_{\sigma_0}(\boldsymbol{\gamma} | \mathbf{Y}) \\
& \propto \left( \frac{1}{\sigma_0} \right) \left( \frac{1}{2} \right)^{|\boldsymbol{\gamma}|} (\sqrt{2\pi})^{|\boldsymbol{\gamma}|} \\
& \quad \times \exp \left[ - \min_{\boldsymbol{\beta}_\gamma} \left( \sum_{i=1}^n (\log(1 + \exp(\mathbf{x}_i \boldsymbol{\beta}_\gamma)) - y_i \mathbf{x}_i \boldsymbol{\beta}_\gamma) + \frac{\beta_0^2}{2\sigma_0^2} + \tau \sum_{\substack{\gamma_j=1 \\ j \neq 0}} (|\beta_j|) \right) \right] \\
& \quad \times \int_{R'} \int \tau^{|\boldsymbol{\gamma}|} q^{|\boldsymbol{\gamma}|} (1-q)^{p-|\boldsymbol{\gamma}|} d\tau dq \\
& = \left( \frac{1}{\sigma_0} \right) \left( \frac{1}{2} \right)^{|\boldsymbol{\gamma}|} (\sqrt{2\pi})^{|\boldsymbol{\gamma}|} \\
& \quad \times \exp \left[ - \min_{\boldsymbol{\beta}} \left( \sum_{i=1}^n (\log(1 + \exp(\mathbf{x}_i \boldsymbol{\beta})) - y_i \mathbf{x}_i \boldsymbol{\beta}) + \frac{\beta_0^2}{2\sigma_0^2} + \tau \sum_{i=1}^p |\beta_j| \right) \right] \\
& \quad \times \int_{R'} \int \tau^{|\boldsymbol{\gamma}|} q^{|\boldsymbol{\gamma}|} (1-q)^{p-|\boldsymbol{\gamma}|} d\tau dq. \tag{4.10}
\end{aligned}$$

Maximizing the posterior probability is equivalent to maximizing the terms involving  $\gamma$  since the terms independent of  $\gamma$  do not contribute to the maximization. Therefore, for each fixed  $\sigma_0$ , the posterior probability satisfies

$$\begin{aligned} \pi_{\sigma_0}(\boldsymbol{\gamma}|\mathbf{Y}) \propto & \left(\frac{1}{2}\right)^{|\boldsymbol{\gamma}|} (\sqrt{2\pi})^{|\boldsymbol{\gamma}|} \exp \left[ -\min_{\boldsymbol{\beta}} \left( \sum_{i=1}^n (\log(1 + \exp(\mathbf{x}_i\boldsymbol{\beta})) - y_i\mathbf{x}_i\boldsymbol{\beta}) + \frac{\beta_0^2}{2\sigma_0^2} \right. \right. \\ & \left. \left. + \tau \sum_{i=1}^p |\beta_j| \right) \right] \times \int_{R'} \int \tau^{|\boldsymbol{\gamma}|} q^{|\boldsymbol{\gamma}|} (1-q)^{p-|\boldsymbol{\gamma}|} d\tau dq. \end{aligned} \quad (4.11)$$

By the Implicit Function Theorem and (3.23), as  $\sigma_0 \rightarrow \infty$ , the maximizer of  $\pi_{\sigma_0}(\boldsymbol{\gamma}|\mathbf{Y})$  converges to the maximizer of  $\pi(\boldsymbol{\gamma}|\mathbf{Y})$ , where

$$\begin{aligned} \pi(\boldsymbol{\gamma}|\mathbf{Y}) \propto & \left(\frac{1}{2}\right)^{|\boldsymbol{\gamma}|} (\sqrt{2\pi})^{|\boldsymbol{\gamma}|} \exp \left[ -\min_{\boldsymbol{\beta}} \left( \sum_{i=1}^n (\log(1 + \exp(\mathbf{x}_i\boldsymbol{\beta})) - y_i\mathbf{x}_i\boldsymbol{\beta}) \right. \right. \\ & \left. \left. + \tau \sum_{i=1}^p |\beta_j| \right) \right] \times \int_{R'} \int \tau^{|\boldsymbol{\gamma}|} q^{|\boldsymbol{\gamma}|} (1-q)^{p-|\boldsymbol{\gamma}|} d\tau dq. \end{aligned} \quad (4.12)$$

Partition the restricted region  $R'$  into  $R_1$  and  $R_2$ , where  $R_1$  is the rectangular region bounded by the axes, the horizontal line  $\tau = r$  and the vertical line  $q = 1/(1 + r\sqrt{\pi/2})$ , and  $R_2$  is the region bounded by the  $q$ -axis, the vertical line  $q = 1/(1 + r\sqrt{\pi/2})$  and the curve  $\tau = ((1-q)/q)\sqrt{2/\pi}$ . The integral in (4.12) depends on whether  $\sum_{j=1}^p |\check{\beta}_j^*|$  is zero or not and  $\check{\beta}_j^*$  is defined in (3.23).

(i) If  $\sum_{j=1}^p |\check{\beta}_j^*| = 0$ ,

$$\begin{aligned} \pi(\boldsymbol{\gamma}|\mathbf{Y}) \propto & \left(\frac{1}{2}\right)^{|\boldsymbol{\gamma}|} (\sqrt{2\pi})^{|\boldsymbol{\gamma}|} \exp \left[ -\sum_{i=1}^n \left( \log(1 + \exp(\mathbf{x}_i\check{\boldsymbol{\beta}}^*)) - y_i\mathbf{x}_i\check{\boldsymbol{\beta}}^* \right) \right] \\ & \times \left[ \int_{R_1} \int \tau^{|\boldsymbol{\gamma}|} q^{|\boldsymbol{\gamma}|} (1-q)^{p-|\boldsymbol{\gamma}|} d\tau dq + \int_{R_2} \int \tau^{|\boldsymbol{\gamma}|} q^{|\boldsymbol{\gamma}|} (1-q)^{p-|\boldsymbol{\gamma}|} d\tau dq \right], \end{aligned} \quad (4.13)$$

where  $\check{\boldsymbol{\beta}}^*$  is defined in (3.23). From simple calculations, the first integral in (4.13)

is

$$\begin{aligned}
& \int_{R_1} \int \tau^{|\gamma|} q^{|\gamma|} (1-q)^{p-|\gamma|} d\tau dq \\
&= \int_0^{(1+r\sqrt{\frac{\pi}{2}})^{-1}} \int_0^r \tau^{|\gamma|} q^{|\gamma|} (1-q)^{p-|\gamma|} d\tau dq \\
&= \int_0^{(1+r\sqrt{\frac{\pi}{2}})^{-1}} q^{|\gamma|} (1-q)^{p-|\gamma|} \left. \frac{\tau^{|\gamma|+1}}{|\gamma|+1} \right|_0^r dq \\
&= \frac{r^{|\gamma|+1}}{|\gamma|+1} \int_0^{(1+r\sqrt{\frac{\pi}{2}})^{-1}} q^{|\gamma|} (1-q)^{p-|\gamma|} dq \\
&= \frac{r^{|\gamma|+1}}{|\gamma|+1} \frac{\Gamma(|\gamma|+1)\Gamma(p-|\gamma|+1)}{\Gamma(p+2)} B_0 \left( \frac{1}{1+r\sqrt{\frac{\pi}{2}}} \right), \tag{4.14}
\end{aligned}$$

where  $B_0(\cdot)$  is the CDF of the beta distribution with parameters  $\alpha = |\gamma| + 1$ ,  $\beta = p - |\gamma| + 1$ . The second integral is

$$\begin{aligned}
& \int_{R_2} \int \tau^{|\gamma|} q^{|\gamma|} (1-q)^{p-|\gamma|} d\tau dq \\
&= \int_{(1+r\sqrt{\frac{\pi}{2}})^{-1}}^1 \int_0^{\frac{1-q}{q}\sqrt{\frac{2}{\pi}}} \tau^{|\gamma|} q^{|\gamma|} (1-q)^{p-|\gamma|} d\tau dq \\
&= \int_{(1+r\sqrt{\frac{\pi}{2}})^{-1}}^1 q^{|\gamma|} (1-q)^{p-|\gamma|} \left. \frac{\tau^{|\gamma|+1}}{|\gamma|+1} \right|_0^{\frac{1-q}{q}\sqrt{\frac{2}{\pi}}} dq \\
&= \frac{1}{|\gamma|+1} \int_{(1+r\sqrt{\frac{\pi}{2}})^{-1}}^1 q^{|\gamma|} (1-q)^{p-|\gamma|} \left( \frac{1-q}{q} \sqrt{\frac{2}{\pi}} \right)^{|\gamma|+1} dq \\
&= \left( \sqrt{\frac{2}{\pi}} \right)^{|\gamma|+1} \frac{1}{|\gamma|+1} \int_{(1+r\sqrt{\frac{\pi}{2}})^{-1}}^1 q^{-1} (1-q)^{p+1} dq. \tag{4.15}
\end{aligned}$$

Plugging (4.14) and (4.15) into (4.13), for  $\sum_{j=1}^p |\check{\beta}_j^*| = 0$ , the logarithm of the

posterior distribution satisfies

$$\begin{aligned}
\log(\pi(\boldsymbol{\gamma}|\mathbf{Y})) &= \frac{|\boldsymbol{\gamma}|}{2} \log\left(\frac{\pi}{2}\right) - \sum_{i=1}^n \left( \log(1 + \exp(\mathbf{x}_i \check{\boldsymbol{\beta}}^*)) - y_i \mathbf{x}_i \check{\boldsymbol{\beta}}^* \right) - \log(|\boldsymbol{\gamma}| + 1) \\
&\quad + \log \left[ r^{|\boldsymbol{\gamma}|+1} \frac{\Gamma(|\boldsymbol{\gamma}| + 1) \Gamma(p - |\boldsymbol{\gamma}| + 1)}{\Gamma(p + 2)} B_0 \left( \frac{1}{1 + r \sqrt{\frac{\pi}{2}}} \right) \right. \\
&\quad \left. + \left( \sqrt{\frac{2}{\pi}} \right)^{|\boldsymbol{\gamma}|+1} \int_{(1+r\sqrt{\frac{\pi}{2}})^{-1}}^1 q^{-1} (1-q)^{p+1} dq \right] + \text{Const.}
\end{aligned} \tag{4.16}$$

(ii) If  $\sum_{j=1}^p |\check{\beta}_j^*| > 0$ ,

$$\begin{aligned}
\pi(\boldsymbol{\gamma}|\mathbf{Y}) &\propto \left(\frac{1}{2}\right)^{|\boldsymbol{\gamma}|} (\sqrt{2\pi})^{|\boldsymbol{\gamma}|} \exp \left[ - \sum_{i=1}^n \left( \log(1 + \exp(\mathbf{x}_i \check{\boldsymbol{\beta}}^*)) - y_i \mathbf{x}_i \check{\boldsymbol{\beta}}^* \right) \right] \\
&\quad \times \left[ \int_{R_1} \int \tau^{|\boldsymbol{\gamma}|} \exp \left( -\tau \sum_{j=1}^p |\check{\beta}_j^*| \right) q^{|\boldsymbol{\gamma}|} (1-q)^{p-|\boldsymbol{\gamma}|} d\tau dq \right. \\
&\quad \left. + \int_{R_2} \int \tau^{|\boldsymbol{\gamma}|} \exp \left( -\tau \sum_{j=1}^p |\check{\beta}_j^*| \right) q^{|\boldsymbol{\gamma}|} (1-q)^{p-|\boldsymbol{\gamma}|} d\tau dq \right],
\end{aligned} \tag{4.17}$$

where  $\check{\boldsymbol{\beta}}^*$  is defined in (3.23). The first integral in (4.17) is

$$\begin{aligned}
&\int_{R_1} \int \tau^{|\boldsymbol{\gamma}|} \exp \left( -\tau \sum_{j=1}^p |\check{\beta}_j^*| \right) q^{|\boldsymbol{\gamma}|} (1-q)^{p-|\boldsymbol{\gamma}|} d\tau dq \\
&= \int_0^{(1+r\sqrt{\frac{\pi}{2}})^{-1}} \int_0^r \tau^{|\boldsymbol{\gamma}|} \exp \left( -\tau \sum_{j=1}^p |\check{\beta}_j^*| \right) q^{|\boldsymbol{\gamma}|} (1-q)^{p-|\boldsymbol{\gamma}|} d\tau dq \\
&= \frac{\Gamma(|\boldsymbol{\gamma}| + 1)}{\left( \sum_{j=1}^p |\check{\beta}_j^*| \right)^{|\boldsymbol{\gamma}|+1}} G_0(r) \int_0^{(1+r\sqrt{\frac{\pi}{2}})^{-1}} q^{|\boldsymbol{\gamma}|} (1-q)^{p-|\boldsymbol{\gamma}|} dq \\
&= \frac{\Gamma(|\boldsymbol{\gamma}| + 1)}{\left( \sum_{j=1}^p |\check{\beta}_j^*| \right)^{|\boldsymbol{\gamma}|+1}} G_0(r) \frac{\Gamma(|\boldsymbol{\gamma}| + 1) \Gamma(p - |\boldsymbol{\gamma}| + 1)}{\Gamma(p + 2)} B_0 \left( \frac{1}{1 + r \sqrt{\frac{\pi}{2}}} \right),
\end{aligned} \tag{4.18}$$

where  $G_0(\cdot)$  is the CDF of the gamma distribution with parameters  $\alpha = |\boldsymbol{\gamma}| + 1$ ,  $\lambda =$

$\sum_{j=1}^p |\check{\beta}_j^*|$ . The second integral in (4.17) is

$$\begin{aligned}
& \int_{R_2} \int \tau^{|\gamma|} \exp\left(-\tau \sum_{j=1}^p |\check{\beta}_j^*|\right) q^{|\gamma|} (1-q)^{p-|\gamma|} d\tau dq \\
&= \int_{(1+r\sqrt{\frac{\pi}{2}})^{-1}}^1 \int_0^{\frac{1-q}{q}\sqrt{\frac{2}{\pi}}} \tau^{|\gamma|} \exp\left(-\tau \sum_{j=1}^p |\check{\beta}_j^*|\right) q^{|\gamma|} (1-q)^{p-|\gamma|} d\tau dq \\
&= \frac{\Gamma(|\gamma|+1)}{\left(\sum_{j=1}^p |\check{\beta}_j^*|\right)^{|\gamma|+1}} \int_{(1+r\sqrt{\frac{\pi}{2}})^{-1}}^1 q^{|\gamma|} (1-q)^{p-|\gamma|} G_0\left(\frac{1-q}{q}\sqrt{\frac{2}{\pi}}\right) dq. \tag{4.19}
\end{aligned}$$

After substituting (4.18) and (4.19) into (4.17), we obtain

$$\begin{aligned}
\log(\pi(\boldsymbol{\gamma}|\mathbf{Y})) &= \frac{|\gamma|}{2} \log\left(\frac{\pi}{2}\right) - \sum_{i=1}^n \left(\log(1 + \exp(\mathbf{x}_i \check{\boldsymbol{\beta}}^*)) - y_i \mathbf{x}_i \check{\boldsymbol{\beta}}^*\right) \\
&\quad + \log \Gamma(|\gamma|+1) - (|\gamma|+1) \log\left(\sum_{j=1}^p |\check{\beta}_j^*|\right) \\
&\quad + \log \left[ G_0(r) \frac{\Gamma(|\gamma|+1)\Gamma(p-|\gamma|+1)}{\Gamma(p+2)} B_0\left(\frac{1}{1+r\sqrt{\frac{\pi}{2}}}\right) \right. \\
&\quad \left. + \int_{(1+r\sqrt{\frac{\pi}{2}})^{-1}}^1 q^{|\gamma|} (1-q)^{p-|\gamma|} G_0\left(\frac{1-q}{q}\sqrt{\frac{2}{\pi}}\right) dq \right] + \text{Const.} \tag{4.20}
\end{aligned}$$

### 4.2.3 Conjugate Priors

Under restricted region  $R'$ , consider the conjugate priors

$$\tau \sim \text{Gamma}\left(a, \frac{1}{b}\right), \quad \tau > 0, \quad q \sim \text{Beta}(\alpha, \beta), \quad 0 < q < 1.$$

Similar to the arguments in Section 4.2.1, focusing on the relevant terms and applying the Implicit Function Theorem, the posterior probability is proportional to

$$\begin{aligned}
& \pi(\boldsymbol{\gamma}|\mathbf{Y}) \\
& \propto \left(\frac{1}{2}\right)^{|\gamma|} (\sqrt{2\pi})^{|\gamma|} \exp\left[-\min_{\boldsymbol{\beta}} \left(\sum_{i=1}^n (\log(1 + \exp(\mathbf{x}_i \boldsymbol{\beta})) - y_i \mathbf{x}_i \boldsymbol{\beta}) + \tau \sum_{i=1}^p |\beta_j|\right)\right] \\
& \quad \times \int_{R'} \int \tau^{|\gamma|+a-1} \exp\left(-\frac{\tau}{b}\right) q^{|\gamma|+\alpha-1} (1-q)^{p-|\gamma|+\beta-1} d\tau dq. \tag{4.21}
\end{aligned}$$

As in Section 4.2.2, the restricted region is decomposed into  $R_1$  and  $R_2$ . The evaluation of the integral in (4.21) relies on whether the sum of  $\left(\sum_{j=1}^p |\check{\beta}_j^*| + 1/b\right)$  is zero or not.

(i) If  $\left(\sum_{j=1}^p |\check{\beta}_j^*| + 1/b\right) = 0$ ,

$$\begin{aligned} \pi(\boldsymbol{\gamma}|\mathbf{Y}) &\propto \left(\frac{1}{2}\right)^{|\boldsymbol{\gamma}|} (\sqrt{2\pi})^{|\boldsymbol{\gamma}|} \exp\left[-\sum_{i=1}^n \left(\log(1 + \exp(\mathbf{x}_i \check{\boldsymbol{\beta}}^*)) - y_i \mathbf{x}_i \check{\boldsymbol{\beta}}^*\right)\right] \\ &\quad \times \left[ \int_{R_1} \int \tau^{|\boldsymbol{\gamma}|+a-1} q^{|\boldsymbol{\gamma}|+\alpha-1} (1-q)^{p-|\boldsymbol{\gamma}|+\beta-1} d\tau dq \right. \\ &\quad \left. + \int_{R_2} \int \tau^{|\boldsymbol{\gamma}|+a-1} q^{|\boldsymbol{\gamma}|+\alpha-1} (1-q)^{p-|\boldsymbol{\gamma}|+\beta-1} d\tau dq \right]. \end{aligned} \quad (4.22)$$

The first integral in (4.22) is

$$\begin{aligned} &\int_{R_1} \int \tau^{|\boldsymbol{\gamma}|+a-1} q^{|\boldsymbol{\gamma}|+\alpha-1} (1-q)^{p-|\boldsymbol{\gamma}|+\beta-1} d\tau dq \\ &= \int_0^{(1+r\sqrt{\frac{\pi}{2}})^{-1}} \int_0^r \tau^{|\boldsymbol{\gamma}|+a-1} q^{|\boldsymbol{\gamma}|+\alpha-1} (1-q)^{p-|\boldsymbol{\gamma}|+\beta-1} d\tau dq \\ &= \frac{r^{|\boldsymbol{\gamma}|+a}}{|\boldsymbol{\gamma}|+a} \int_0^{(1+r\sqrt{\frac{\pi}{2}})^{-1}} q^{|\boldsymbol{\gamma}|+\alpha-1} (1-q)^{p-|\boldsymbol{\gamma}|+\beta-1} dq \\ &= \frac{r^{|\boldsymbol{\gamma}|+a}}{|\boldsymbol{\gamma}|+a} \frac{\Gamma(\alpha + |\boldsymbol{\gamma}|)\Gamma(p - |\boldsymbol{\gamma}| + \beta)}{\Gamma(p + \alpha + \beta)} B\left(\frac{1}{1 + r\sqrt{\frac{\pi}{2}}}\right), \end{aligned} \quad (4.23)$$

where  $B(\cdot)$  is the CDF of the beta distribution with parameters  $\alpha = |\boldsymbol{\gamma}| + \alpha$ ,  $\beta = p - |\boldsymbol{\gamma}| + \beta$ . The second integral is

$$\begin{aligned} &\int_{R_2} \int \tau^{|\boldsymbol{\gamma}|+a-1} q^{|\boldsymbol{\gamma}|+\alpha-1} (1-q)^{p-|\boldsymbol{\gamma}|+\beta-1} d\tau dq \\ &= \int_{(1+r\sqrt{\frac{\pi}{2}})^{-1}}^1 \int_0^{\frac{1-q}{q}\sqrt{\frac{2}{\pi}}} \tau^{|\boldsymbol{\gamma}|+a-1} q^{|\boldsymbol{\gamma}|+\alpha-1} (1-q)^{p-|\boldsymbol{\gamma}|+\beta-1} d\tau dq \\ &= \frac{1}{|\boldsymbol{\gamma}|+a} \int_{(1+r\sqrt{\frac{\pi}{2}})^{-1}}^1 q^{|\boldsymbol{\gamma}|+\alpha-1} (1-q)^{p-|\boldsymbol{\gamma}|+\beta-1} \left(\frac{1-q}{q}\sqrt{\frac{2}{\pi}}\right)^{|\boldsymbol{\gamma}|+a} dq \end{aligned} \quad (4.24)$$

$$= \left(\sqrt{\frac{2}{\pi}}\right)^{|\boldsymbol{\gamma}|+a} \frac{1}{|\boldsymbol{\gamma}|+a} \int_{(1+r\sqrt{\frac{\pi}{2}})^{-1}}^1 q^{\alpha-a-1} (1-q)^{p+a+\beta-1} dq. \quad (4.25)$$



Maximizing the posterior probability is the same as maximizing the logarithm of the posterior probability, so that,

$$\begin{aligned}
\log(\pi(\boldsymbol{\gamma}|\mathbf{Y})) &= \frac{|\boldsymbol{\gamma}|}{2} \log\left(\frac{\pi}{2}\right) - \sum_{i=1}^n \left( \log(1 + \exp(\mathbf{x}_i \check{\boldsymbol{\beta}}^*)) - y_i \mathbf{x}_i \check{\boldsymbol{\beta}}^* \right) - \log(|\boldsymbol{\gamma}| + a) \\
&+ \log \left[ r^{|\boldsymbol{\gamma}|+a} \frac{\Gamma(\alpha + |\boldsymbol{\gamma}|) \Gamma(p - |\boldsymbol{\gamma}| + \beta)}{\Gamma(p + \alpha + \beta)} B\left(\frac{1}{1 + r\sqrt{\frac{\pi}{2}}}\right) \right. \\
&\left. + \left(\sqrt{\frac{2}{\pi}}\right)^{|\boldsymbol{\gamma}|+a} \int_{(1+r\sqrt{\frac{\pi}{2}})^{-1}}^1 q^{\alpha-a-1} (1-q)^{p+a+\beta-1} dq \right] + \text{Const.}
\end{aligned} \tag{4.26}$$

(ii) If  $\left(\sum_{j=1}^p |\check{\beta}_j^*| + 1/b\right) > 0$ ,

$$\begin{aligned}
\pi(\boldsymbol{\gamma}|\mathbf{Y}) &\propto \left(\frac{1}{2}\right)^{|\boldsymbol{\gamma}|} (\sqrt{2\pi})^{|\boldsymbol{\gamma}|} \exp \left[ - \sum_{i=1}^n \left( \log(1 + \exp(\mathbf{x}_i \check{\boldsymbol{\beta}}^*)) - y_i \mathbf{x}_i \check{\boldsymbol{\beta}}^* \right) \right] \\
&\times \left[ \int_{R_1} \int \tau^{|\boldsymbol{\gamma}|+a-1} \exp \left[ -\tau \left( \sum_{j=1}^p |\check{\beta}_j^*| + \frac{1}{b} \right) \right] q^{|\boldsymbol{\gamma}|+a-1} (1-q)^{p-|\boldsymbol{\gamma}|+\beta-1} d\tau dq \right. \\
&\left. + \int_{R_2} \int \tau^{|\boldsymbol{\gamma}|+a-1} \exp \left[ -\tau \left( \sum_{j=1}^p |\check{\beta}_j^*| + \frac{1}{b} \right) \right] q^{|\boldsymbol{\gamma}|+a-1} (1-q)^{p-|\boldsymbol{\gamma}|+\beta-1} d\tau dq \right].
\end{aligned} \tag{4.27}$$

The first integral in (4.27) is

$$\begin{aligned}
&\int_{R_1} \int \tau^{|\boldsymbol{\gamma}|+a-1} \exp \left[ -\tau \left( \sum_{j=1}^p |\check{\beta}_j^*| + \frac{1}{b} \right) \right] q^{|\boldsymbol{\gamma}|+a-1} (1-q)^{p-|\boldsymbol{\gamma}|+\beta-1} d\tau dq \\
&= \int_0^{(1+r\sqrt{\frac{\pi}{2}})^{-1}} \int_0^r \tau^{|\boldsymbol{\gamma}|+a-1} \exp \left[ -\tau \left( \sum_{j=1}^p |\check{\beta}_j^*| + \frac{1}{b} \right) \right] \\
&\quad \times q^{|\boldsymbol{\gamma}|+a-1} (1-q)^{p-|\boldsymbol{\gamma}|+\beta-1} d\tau dq \\
&= \frac{\Gamma(|\boldsymbol{\gamma}| + a) G_1(r)}{\left(\sum_{j=1}^p |\check{\beta}_j^*| + \frac{1}{b}\right)^{|\boldsymbol{\gamma}|+a}} \int_0^{(1+r\sqrt{\frac{\pi}{2}})^{-1}} q^{|\boldsymbol{\gamma}|+a-1} (1-q)^{p-|\boldsymbol{\gamma}|+\beta-1} dq \\
&= \frac{\Gamma(|\boldsymbol{\gamma}| + a) G_1(r)}{\left(\sum_{j=1}^p |\check{\beta}_j^*| + \frac{1}{b}\right)^{|\boldsymbol{\gamma}|+a}} \frac{\Gamma(\alpha + |\boldsymbol{\gamma}|) \Gamma(p - |\boldsymbol{\gamma}| + \beta)}{\Gamma(p + \alpha + \beta)} B\left(\frac{1}{1 + r\sqrt{\frac{\pi}{2}}}\right), \tag{4.28}
\end{aligned}$$

where  $G_1(\cdot)$  is the CDF of the gamma distribution with parameters  $\alpha = |\boldsymbol{\gamma}| + a$ ,  $\lambda = \sum_{j=1}^p |\check{\beta}_j^*| + 1/b$ . The second integral is

$$\begin{aligned}
& \int_{R_2} \int \tau^{|\boldsymbol{\gamma}|+a-1} \exp \left[ -\tau \left( \sum_{j=1}^p |\check{\beta}_j^*| + \frac{1}{b} \right) \right] q^{|\boldsymbol{\gamma}|+\alpha-1} (1-q)^{p-|\boldsymbol{\gamma}|+\beta-1} d\tau dq \\
&= \int_{(1+r\sqrt{\frac{\pi}{2}})^{-1}}^1 \int_0^{\frac{1-q}{q}\sqrt{\frac{2}{\pi}}} \tau^{|\boldsymbol{\gamma}|+a-1} \exp \left[ -\tau \left( \sum_{j=1}^p |\check{\beta}_j^*| + \frac{1}{b} \right) \right] \\
&\quad \times q^{|\boldsymbol{\gamma}|+\alpha-1} (1-q)^{p-|\boldsymbol{\gamma}|+\beta-1} d\tau dq \\
&= \frac{\Gamma(|\boldsymbol{\gamma}| + a)}{\left( \sum_{j=1}^p |\check{\beta}_j^*| + \frac{1}{b} \right)^{|\boldsymbol{\gamma}|+a}} \int_{(1+r\sqrt{\frac{\pi}{2}})^{-1}}^1 q^{|\boldsymbol{\gamma}|+\alpha-1} (1-q)^{p-|\boldsymbol{\gamma}|+\beta-1} G_1 \left( \frac{1-q}{q} \sqrt{\frac{2}{\pi}} \right) dq.
\end{aligned} \tag{4.29}$$

The highest posterior probability can be obtained by maximizing  $\log(\pi(\boldsymbol{\gamma}|\mathbf{Y}))$ , where

$$\begin{aligned}
& \log(\pi(\boldsymbol{\gamma}|\mathbf{Y})) \\
&= \frac{|\boldsymbol{\gamma}|}{2} \log \left( \frac{\pi}{2} \right) - \sum_{i=1}^n \left( \log(1 + \exp(\mathbf{x}_i \check{\boldsymbol{\beta}}^*)) - y_i \mathbf{x}_i \check{\boldsymbol{\beta}}^* \right) \\
&\quad + \log \Gamma(|\boldsymbol{\gamma}| + a) - (|\boldsymbol{\gamma}| + a) \log \left( \sum_{j=1}^p |\check{\beta}_j^*| + \frac{1}{b} \right) \\
&\quad + \log \left[ G_1(r) \frac{\Gamma(\alpha + |\boldsymbol{\gamma}|) \Gamma(p - |\boldsymbol{\gamma}| + \beta)}{\Gamma(p + \alpha + \beta)} B \left( \frac{1}{1 + r\sqrt{\frac{\pi}{2}}} \right) \right. \\
&\quad \left. + \int_{(1+r\sqrt{\frac{\pi}{2}})^{-1}}^1 q^{|\boldsymbol{\gamma}|+\alpha-1} (1-q)^{p-|\boldsymbol{\gamma}|+\beta-1} G_1 \left( \frac{1-q}{q} \sqrt{\frac{2}{\pi}} \right) dq \right] + \text{Const}
\end{aligned} \tag{4.30}$$

Note that (4.16) and (4.20) are special cases of (4.26) and (4.30) with  $a = 1$ ,  $b = +\infty$ ,  $\alpha = 1$  and  $\beta = 1$ .

### 4.3 Poisson Models

In this section, the dependent variable  $\mathbf{Y}$  follows a Poisson distribution. Using the methods described in Chapter 3 and 4 with priors similar to those in Section 3.1, we derive the empirical Bayes and fully Bayes criteria for Poisson models and summarize the results below.

Assume that  $y_i, i = 1, 2, \dots, n$ , are conditionally independent given  $\mu_{i\gamma}, i = 1, 2, \dots, n$ , and that  $y_i|\mu_{i\gamma}$  follows a Poisson distribution,

$$y_i|\mu_{i\gamma} \sim \frac{\exp(-\mu_{i\gamma})}{y_i!} \mu_{i\gamma}^{y_i},$$

where

$$\log \mu_{i\gamma} = \mathbf{x}_{i\gamma} \boldsymbol{\beta}_\gamma.$$

The density of  $y_i|\boldsymbol{\beta}_\gamma, \gamma$  can be written as

$$f(y_i|\boldsymbol{\beta}_\gamma, \gamma) = \frac{1}{y_i!} \exp(-\exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_\gamma)) \exp(y_i \mathbf{x}_{i\gamma} \boldsymbol{\beta}_\gamma). \quad (4.31)$$

Taking the priors (3.2), (3.3) and (3.4) and calculating parallel to the derivation of the logistic case yields the empirical Bayes criterion:

$$\begin{aligned} \text{CML}(\lambda) &= -|\hat{\gamma}_\lambda| \log \left( \frac{\lambda}{2} \right) - |\hat{\gamma}_\lambda| \log \sqrt{2\pi} + \frac{1}{2} \log(\det(\mathbf{A} + \mathbf{H})) \\ &\quad - \left( \frac{1}{2} (\mathbf{e} - \mathbf{t})^T (\mathbf{A} + \mathbf{H})^{-1} (\mathbf{e} - \mathbf{t}) \right) \\ &\quad + \min_{\boldsymbol{\beta}} \left( \sum_{i=1}^n (\exp(\mathbf{x}_i \boldsymbol{\beta}) - y_i \mathbf{x}_i \boldsymbol{\beta}) + \lambda \sum_{j=1}^p |\beta_j| \right), \end{aligned} \quad (4.32)$$

where

$$\mathbf{A} = \sum_{i=1}^n (\mathbf{x}_{i\gamma}^T \exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_\gamma^*) \mathbf{x}_{i\gamma}), \quad (4.33)$$

$$\mathbf{e}^T = \sum_{i=1}^n (y_i - \exp(\mathbf{x}_{i\gamma} \boldsymbol{\beta}_{\gamma^*})) \mathbf{x}_{i\gamma}, \quad (4.34)$$

and

$$\check{\boldsymbol{\beta}}^* = \operatorname{argmin}_{\boldsymbol{\beta}} \left( \sum_{i=1}^n (\exp(\mathbf{x}_i \boldsymbol{\beta}) - y_i \mathbf{x}_i \boldsymbol{\beta}) + \lambda \sum_{j=1}^p |\beta_j| \right).$$

By analysis analogous to Section 4.2.2 and Section 4.2.3 and hyperpriors specified in those two sections, the fully Bayes criteria are as follows:

### I. Flat priors

(i) If  $\sum_{j=1}^p |\check{\beta}_j^*| = 0$ ,

$$\begin{aligned} \log(\pi(\boldsymbol{\gamma}|\mathbf{Y})) &= \frac{|\boldsymbol{\gamma}|}{2} \log\left(\frac{\pi}{2}\right) - \sum_{i=1}^n \left( \exp(\mathbf{x}_i \check{\boldsymbol{\beta}}^*) - y_i \mathbf{x}_i \check{\boldsymbol{\beta}}^* \right) - \log(|\boldsymbol{\gamma}| + 1) \\ &+ \log \left[ r^{|\boldsymbol{\gamma}|+1} \frac{\Gamma(|\boldsymbol{\gamma}| + 1) \Gamma(p - |\boldsymbol{\gamma}| + 1)}{\Gamma(p + 2)} B_0 \left( \frac{1}{1 + r \sqrt{\frac{\pi}{2}}} \right) \right. \\ &\quad \left. + \left( \sqrt{\frac{2}{\pi}} \right)^{|\boldsymbol{\gamma}|+1} \int_{(1+r\sqrt{\frac{\pi}{2}})^{-1}}^1 q^{-1} (1-q)^{p+1} dq \right] + \text{Const}, \end{aligned} \quad (4.35)$$

where  $B_0(\cdot)$  is the CDF of the beta distribution with parameters  $\alpha = |\boldsymbol{\gamma}| + 1$ ,  $\beta = p - |\boldsymbol{\gamma}| + 1$ .

(ii) If  $\sum_{j=1}^p |\check{\beta}_j^*| > 0$ ,

$$\begin{aligned} \log(\pi(\boldsymbol{\gamma}|\mathbf{Y})) &= \frac{|\boldsymbol{\gamma}|}{2} \log\left(\frac{\pi}{2}\right) - \sum_{i=1}^n \left( \exp(\mathbf{x}_i \check{\boldsymbol{\beta}}^*) - y_i \mathbf{x}_i \check{\boldsymbol{\beta}}^* \right) \\ &+ \log \Gamma(|\boldsymbol{\gamma}| + 1) - (|\boldsymbol{\gamma}| + 1) \log \left( \sum_{j=1}^p |\check{\beta}_j^*| \right) \\ &+ \log \left[ G_0(r) \frac{\Gamma(|\boldsymbol{\gamma}| + 1) \Gamma(p - |\boldsymbol{\gamma}| + 1)}{\Gamma(p + 2)} B_0 \left( \frac{1}{1 + r \sqrt{\frac{\pi}{2}}} \right) \right. \\ &\quad \left. + \int_{(1+r\sqrt{\frac{\pi}{2}})^{-1}}^1 q^{|\boldsymbol{\gamma}|} (1-q)^{p-|\boldsymbol{\gamma}|} G_0 \left( \frac{1-q}{q} \sqrt{\frac{2}{\pi}} \right) dq \right] + \text{Const}, \end{aligned} \quad (4.36)$$

where  $G_0(\cdot)$  is the CDF of the gamma distribution with parameters  $\alpha = |\gamma| + 1$ ,  $\lambda = \sum_{j=1}^p |\check{\beta}_j^*|$ .

## II. Conjugate priors

(i) If  $\left(\sum_{j=1}^p |\check{\beta}_j^*| + 1/b\right) = 0$ ,

$$\begin{aligned} \log(\pi(\gamma|\mathbf{Y})) &= \frac{|\gamma|}{2} \log\left(\frac{\pi}{2}\right) - \sum_{i=1}^n (\exp(\mathbf{x}_i \boldsymbol{\beta}^*) - y_i \mathbf{x}_i \boldsymbol{\beta}^*) - \log(|\gamma| + a) \\ &+ \log \left[ r^{|\gamma|+a} \frac{\Gamma(\alpha + |\gamma|) \Gamma(p - |\gamma| + \beta)}{\Gamma(p + \alpha + \beta)} B\left(\frac{1}{1 + r\sqrt{\frac{\pi}{2}}}\right) \right. \\ &\quad \left. + \left(\sqrt{\frac{2}{\pi}}\right)^{|\gamma|+a} \int_{(1+r\sqrt{\frac{\pi}{2}})^{-1}}^1 q^{\alpha-a-1} (1-q)^{p+a+\beta-1} dq \right] + \text{Const}, \end{aligned} \quad (4.37)$$

where  $B(\cdot)$  is the CDF of the beta distribution with parameters  $\alpha = |\gamma| + \alpha$ ,  $\beta = p - |\gamma| + \beta$ .

(ii) If  $\left(\sum_{j=1}^p |\check{\beta}_j^*| + 1/b\right) > 0$ ,

$$\begin{aligned} \log(\pi(\gamma|\mathbf{Y})) &= \frac{|\gamma|}{2} \log\left(\frac{\pi}{2}\right) - \sum_{i=1}^n (\exp(\mathbf{x}_i \check{\boldsymbol{\beta}}^*) - y_i \mathbf{x}_i \check{\boldsymbol{\beta}}^*) \\ &+ \log \Gamma(|\gamma| + a) - (|\gamma| + a) \log\left(\sum_{j=1}^p |\check{\beta}_j^*| + \frac{1}{b}\right) \\ &+ \log \left[ G_1(r) \frac{\Gamma(\alpha + |\gamma|) \Gamma(p - |\gamma| + \beta)}{\Gamma(p + \alpha + \beta)} B\left(\frac{1}{1 + r\sqrt{\frac{\pi}{2}}}\right) \right. \\ &\quad \left. + \int_{(1+r\sqrt{\frac{\pi}{2}})^{-1}}^1 q^{|\gamma|+\alpha-1} (1-q)^{p-|\gamma|+\beta-1} G_1\left(\frac{1-q}{q} \sqrt{\frac{2}{\pi}}\right) dq \right] + \text{Const}, \end{aligned} \quad (4.38)$$

where  $G_1(\cdot)$  is the CDF of the gamma distribution with parameters  $\alpha = |\gamma| + a$ ,  $\lambda = \sum_{j=1}^p |\check{\beta}_j^*| + 1/b$ .

## 4.4 Linear Models

Yuan and Lin [38] gave the empirical Bayes criterion for linear models. Building upon their priors, we specify hyperpriors for the hyperparameters and develop the fully Bayes criterion in this section.

Consider a linear model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where  $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$ . Each predictor is centered and scaled so that its sample mean is 0 and sample standard deviation is 1.

Assigning a double exponential prior (3.3) for  $\beta_j$  and using the following prior for  $\boldsymbol{\gamma}$ ,

$$\pi(\boldsymbol{\gamma}|q) = q^{|\boldsymbol{\gamma}|} (1-q)^{p-|\boldsymbol{\gamma}|} \sqrt{\det(\mathbf{X}_{\boldsymbol{\gamma}}^T \mathbf{X}_{\boldsymbol{\gamma}})}. \quad (4.39)$$

Yuan and Lin [38] gave the empirical Bayes criterion:

$$\begin{aligned} \text{CML}(\lambda) = & -(n + |\hat{\boldsymbol{\gamma}}_{\lambda}|) \left[ \log \left( \frac{\min_{\boldsymbol{\beta}} (\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^p |\beta_j|)}{n + |\hat{\boldsymbol{\gamma}}_{\lambda}|} \right) + 1 \right] \\ & + \log(\det(\mathbf{X}_{\hat{\boldsymbol{\gamma}}_{\lambda}}^T \mathbf{X}_{\hat{\boldsymbol{\gamma}}_{\lambda}})) - 2|\hat{\boldsymbol{\gamma}}_{\lambda}| \ln(\sqrt{2\pi}\lambda/4), \end{aligned} \quad (4.40)$$

In order to obtain a proper posterior distribution, define the restricted region as

$$R' = \left\{ (\tau, q) : \left( \frac{1-q}{q} \sqrt{\frac{2}{\pi\sigma^2}} \right) \geq \tau, \tau \leq r, 0 < q < 1 \right\}, \quad (4.41)$$

where  $r$  is a fixed value and  $\sigma^2$  is known.

We will write

$$\boldsymbol{\beta}^* = \min_{\boldsymbol{\beta}} (\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^p |\beta_j|)$$

throughout this section. Under the restricted region  $R'$ , we employ the flat priors in Section 4.2.2 and conjugate priors in Section 4.2.3. The fully Bayes criteria are as follows:

### I. Flat priors

(i) If  $\sum_{j=1}^p |\check{\beta}_j^*| = 0$ ,

$$\begin{aligned} \log(\pi(\boldsymbol{\gamma}|\mathbf{Y})) &= -(n - |\boldsymbol{\gamma}|) \log(\sqrt{2\pi\sigma^2}) - |\boldsymbol{\gamma}| \log 2 - \frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^*\|^2}{2\sigma^2} - \log(|\boldsymbol{\gamma}| + 1) \\ &+ \log \left[ r^{|\boldsymbol{\gamma}|+1} \frac{\Gamma(|\boldsymbol{\gamma}| + 1)\Gamma(p - |\boldsymbol{\gamma}| + 1)}{\Gamma(p + 2)} B_0 \left( \frac{1}{1 + r\sqrt{\pi\sigma^2/2}} \right) \right. \\ &\left. + \left( \sqrt{\frac{2}{\pi\sigma^2}} \right)^{|\boldsymbol{\gamma}|+1} \int_{(1+r\sqrt{\pi\sigma^2/2})^{-1}}^1 q^{-1}(1-q)^{p+1} dq \right] + \text{Const}, \end{aligned} \quad (4.42)$$

where  $B_0(\cdot)$  is the CDF of the beta distribution with parameters  $\alpha = |\boldsymbol{\gamma}| + 1$ ,  $\beta = p - |\boldsymbol{\gamma}| + 1$ .

(ii) If  $\sum_{j=1}^p |\check{\beta}_j^*| > 0$ ,

$$\begin{aligned} \log(\pi(\boldsymbol{\gamma}|\mathbf{Y})) &= -(n - |\boldsymbol{\gamma}|) \log(\sqrt{2\pi\sigma^2}) - |\boldsymbol{\gamma}| \log 2 - \frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^*\|^2}{2\sigma^2} \\ &+ \log \Gamma(|\boldsymbol{\gamma}| + 1) - (|\boldsymbol{\gamma}| + 1) \log \left( \sum_{j=1}^p |\check{\beta}_j^*| \right) \\ &+ \log \left[ G_0(r) \frac{\Gamma(|\boldsymbol{\gamma}| + 1)\Gamma(p - |\boldsymbol{\gamma}| + 1)}{\Gamma(p + 2)} B_0 \left( \frac{1}{1 + r\sqrt{\pi\sigma^2/2}} \right) \right. \\ &\left. + \int_{(1+r\sqrt{\pi\sigma^2/2})^{-1}}^1 q^{|\boldsymbol{\gamma}|}(1-q)^{p-|\boldsymbol{\gamma}|} G_0 \left( \frac{1-q}{q} \sqrt{\frac{2}{\pi\sigma^2}} \right) dq \right] + \text{Const}, \end{aligned} \quad (4.43)$$

where  $G_0(\cdot)$  is the CDF of the gamma distribution with parameters  $\alpha = |\boldsymbol{\gamma}| + 1$ ,  $\lambda = \sum_{j=1}^p |\check{\beta}_j^*|$ .

## II. Conjugate priors

(i) If  $\left(\sum_{j=1}^p |\beta_j^*| + 1/b\right) = 0$ ,

$$\begin{aligned}
& \log(\pi(\boldsymbol{\gamma}|\mathbf{Y})) \\
&= -(n - |\boldsymbol{\gamma}|) \log(\sqrt{2\pi\sigma^2}) - |\boldsymbol{\gamma}| \log 2 - \frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^*\|^2}{2\sigma^2} - \log(|\boldsymbol{\gamma}| + a) \\
&\quad + \log \left[ r^{|\boldsymbol{\gamma}|+a} \frac{\Gamma(\alpha + |\boldsymbol{\gamma}|)\Gamma(p - |\boldsymbol{\gamma}| + \beta)}{\Gamma(p + \alpha + \beta)} B\left(\frac{1}{1 + r\sqrt{\pi\sigma^2/2}}\right) \right. \\
&\quad \left. + \left(\sqrt{\frac{2}{\pi\sigma^2}}\right)^{|\boldsymbol{\gamma}|+a} \int_{(1+r\sqrt{\pi\sigma^2/2})^{-1}}^1 q^{\alpha-a-1} (1-q)^{p+a+\beta-1} dq \right] + \text{Const},
\end{aligned} \tag{4.44}$$

where  $B(\cdot)$  is the CDF of the beta distribution with parameters  $\alpha = |\boldsymbol{\gamma}| + \alpha$ ,  $\beta = p - |\boldsymbol{\gamma}| + \beta$ .

(ii) If  $\left(\sum_{j=1}^p |\beta_j^*| + 1/b\right) > 0$ ,

$$\begin{aligned}
& \log(\pi(\boldsymbol{\gamma}|\mathbf{Y})) \\
&= -(n - |\boldsymbol{\gamma}|) \log(\sqrt{2\pi\sigma^2}) - |\boldsymbol{\gamma}| \log 2 - \frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^*\|^2}{2\sigma^2} \\
&\quad + \log \Gamma(|\boldsymbol{\gamma}| + a) - (|\boldsymbol{\gamma}| + a) \log \left( \sum_{j=1}^p |\beta_j^*| + \frac{1}{b} \right) \\
&\quad + \log \left[ G_1(r) \frac{\Gamma(\alpha + |\boldsymbol{\gamma}|)\Gamma(p - |\boldsymbol{\gamma}| + \beta)}{\Gamma(p + \alpha + \beta)} B\left(\frac{1}{1 + r\sqrt{\pi\sigma^2/2}}\right) \right. \\
&\quad \left. + \int_{(1+r\sqrt{\pi\sigma^2/2})^{-1}}^1 q^{|\boldsymbol{\gamma}|+\alpha-1} (1-q)^{p-|\boldsymbol{\gamma}|+\beta-1} G_1\left(\frac{1-q}{q} \sqrt{\frac{2}{\pi\sigma^2}}\right) dq \right] + \text{Const},
\end{aligned} \tag{4.45}$$

where  $G_1(\cdot)$  is the CDF of the gamma distribution with parameters  $\alpha = |\boldsymbol{\gamma}| + a$ ,  $\lambda = \sum_{j=1}^p |\check{\beta}_j^*| + 1/b$ .



## 4.5 Implementation of the Bayesian Criteria

For each of the models considered in this Chapter, the CML and fully Bayes criteria depend on  $\check{\beta}^*$ , where  $\check{\beta}^*$  minimizes (3.24), the penalized negative loglikelihood for a given penalty  $\lambda \sum_{i=1}^p |\beta_j|$ . Available R packages (`lars` [8], `glmnet` [29], `glmnet` [13]) permit one to solve the LASSO problem for all values of  $\lambda$ . Therefore the CML and fully Bayes criteria can be computed for each  $\lambda$  and hence the optimum  $\lambda$  can be determined. This means that model selection using CML or fully Bayes criteria can be implemented.

The R packages mentioned above already produce optimum  $\lambda$  values for the  $C_p$ , AIC and BIC criteria. Therefore our criteria can be applied directly to solve  $L_1$  penalized likelihood problems. We will apply our criteria to both simulated and real- world data in Chapter 6.

## Chapter 5

### Asymptotic Results for LASSO-type Estimators in GLM

In this Chapter, extension of Theorems 2.1, 2.2 and 2.3 will be presented when the regularization parameter  $\lambda$  is deterministic. The following regularity conditions are imposed throughout this chapter.

**Condition C1**  $C_n(\boldsymbol{\xi}) = \sum_{i=1}^n \mathbf{x}_i^T \phi b''(\mathbf{x}_i \boldsymbol{\xi}) \mathbf{x}_i / n \xrightarrow{\text{uniformly}} C(\boldsymbol{\xi}),$

where  $\mathbf{x}_i$  is a row vector which represents the  $i$ th row of the design matrix  $\mathbf{X}$  and  $C = C(\boldsymbol{\xi})$  is a nonnegative definite matrix. The convergence is uniform over  $\boldsymbol{\xi} \in K$ , a compact and convex set containing  $\boldsymbol{\beta}$ .

**Condition C2**  $\max_{1 \leq i \leq n} \mathbf{x}_i \mathbf{x}_i^T / n \rightarrow 0.$

Assume  $C$  is nonsingular throughout this chapter. Write the likelihood function as  $L(\mathbf{Y}, \boldsymbol{\beta}) = \prod_{i=1}^n l_i(y_i, \boldsymbol{\beta})$ . Define the random function

$$\begin{aligned} Z_n(\boldsymbol{\xi}) &= -\frac{1}{n} \sum_{i=1}^n \log \left[ \frac{l_i(y_i, \boldsymbol{\xi})}{l_i(y_i, \boldsymbol{\beta})} \right] + \frac{\lambda_n}{n} \sum_{j=1}^p (|\xi_j|^\nu - |\beta_j|^\nu) \\ &= -\frac{1}{\phi} \sum_{i=1}^n \frac{1}{n} [y_i \mathbf{x}_i \boldsymbol{\xi} - b(\mathbf{x}_i \boldsymbol{\xi}) - y_i \mathbf{x}_i \boldsymbol{\beta} + b(\mathbf{x}_i \boldsymbol{\beta})] \\ &\quad + \frac{\lambda_n}{n} \sum_{j=1}^p (|\xi_j|^\nu - |\beta_j|^\nu), \end{aligned} \tag{5.1}$$

which is minimized at  $\boldsymbol{\xi} = \hat{\boldsymbol{\beta}}_n$ .

We are interested in the minimizer of  $Z_n(\boldsymbol{\xi})$  and will show that the minimizer of  $Z_n(\boldsymbol{\xi})$  converges to the minimizer of some limiting function  $Z(\boldsymbol{\xi})$ . The following two theorems are used to establish the required convergence:

**Theorem 5.1 (Andersen & Gill, 1982 [2])** *Let  $E$  be an open convex subset of  $\mathbb{R}^p$  and let  $F_1, F_2, \dots$ , be a sequence of random concave functions on  $E$  such that  $\forall x \in E, F_n(x) \rightarrow_p f(x)$  as  $n \rightarrow \infty$  where  $f$  is some real function on  $E$ . Then  $f$  is also concave and for all compact  $A \subset E$ ,*

$$\sup_{x \in A} |F_n(x) - f(x)| \rightarrow 0 \text{ as } n \rightarrow \infty.$$

**Definition 5.1 (Kim & Pollard, 1990 [20])** *Let  $(\Omega, \mathcal{A}, \mathbb{P})$  be a probability space. The outer expectation of a bounded, real function  $f$  on  $\Omega$  is defined by*

$$\mathbb{P}^* f = \inf\{\mathbb{P} g : f \leq g \text{ and } g \text{ is integrable}\}.$$

**Definition 5.2 (Kim & Pollard, 1990 [20])** *Let  $(\mathcal{X}, \rho)$  be a metric space and  $\mathcal{U}(\mathcal{X})$  be the set of bounded, uniformly continuous, real functions on  $\mathcal{X}$ . For maps  $X_n$  from  $\Omega$  into  $\mathcal{X}$  and a probability measure  $Q$  on the Borel  $\sigma$ -field of  $\mathcal{X}$ , define the convergence in distribution  $X_n \rightsquigarrow Q$  to mean:*

- (i)  $Q$  has separable support;
- (ii)  $\mathbb{P}^* h(X_n) \rightarrow Q h$  for each  $h$  in  $\mathcal{U}(\mathcal{X})$ .

**Theorem 5.2 (Kim & Pollard, 1990 [20])** *Let  $B_{\text{loc}}(\mathbb{R}^d)$  be the space of all locally bounded real functions on  $\mathbb{R}^d$ , equipped with the topology of uniform conver-*

gence on compacta. Let  $\{Z_n\}$  be random maps into  $B_{\text{loc}}(\mathbb{R}^d)$  and  $\{t_n\}$  be random maps into  $\mathbb{R}^d$  such that:

(i)  $Z_n \rightsquigarrow Q$  for a Borel measure  $Q$  concentrated on  $C_{\text{max}}(\mathbb{R}^d)$ , where  $C_{\text{max}}(\mathbb{R}^d)$  is the separable set of continuous functions  $x(\cdot)$  in  $B_{\text{loc}}(\mathbb{R}^d)$  such that  $x(t) \rightarrow -\infty$  as  $|t| \rightarrow \infty$  and  $x(\cdot)$  attains a unique maximum in  $\mathbb{R}^d$ ;

(ii)  $t_n = O_p(1)$ ;

(iii)  $Z_n(t_n) \geq \sup_t Z_n(t) - \alpha_n$  for random variables  $\{\alpha_n\}$  of order  $o_p(1)$ .

Then  $t_n \rightsquigarrow \text{argmax}(Z)$  for a  $Z$  with distribution  $Q$ .

**Lemma 5.1** Let  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$  be independent and identically distributed with joint cumulative distribution function  $H(\mathbf{x})$ . Let  $K \subset \mathbb{R}^p$  be a compact set and let  $\boldsymbol{\beta}$  belong to the interior of  $K$ . Define

$$F_n(\boldsymbol{\xi}) = \frac{1}{n} \sum_{i=1}^n [b(\mathbf{x}_i \boldsymbol{\xi}) - b(\mathbf{x}_i \boldsymbol{\beta}) - b'(\mathbf{x}_i \boldsymbol{\beta}) \mathbf{x}_i (\boldsymbol{\xi} - \boldsymbol{\beta})].$$

and assume that  $F_n(\boldsymbol{\xi}) \rightarrow F(\boldsymbol{\xi})$  uniformly for  $\boldsymbol{\xi} \in K$ . Then if  $\lambda_n/n \rightarrow \lambda_0 \in (0, \infty)$ , for  $Z_n(\boldsymbol{\xi})$  in (5.1),

$$\sup_{\boldsymbol{\xi} \in K} |Z_n(\boldsymbol{\xi}) - Z(\boldsymbol{\xi})| \xrightarrow{p} 0 \tag{5.2}$$

where

$$Z(\boldsymbol{\xi}) = \lim \left\{ \frac{1}{\phi} \sum_{i=1}^n \frac{1}{n} [b(\mathbf{x}_i \boldsymbol{\xi}) - b(\mathbf{x}_i \boldsymbol{\beta}) - b'(\mathbf{x}_i \boldsymbol{\beta}) \mathbf{x}_i (\boldsymbol{\xi} - \boldsymbol{\beta})] \right\} + \lambda_0 \sum_{j=1}^p (|\xi_j|^\nu - |\beta_j|^\nu).$$

**Proof:** Consider the function

$$\begin{aligned}
Z_n(\boldsymbol{\xi}) &= -\frac{1}{\phi} \sum_{i=1}^n \frac{1}{n} [y_i \mathbf{x}_i \boldsymbol{\xi} - b(\mathbf{x}_i \boldsymbol{\xi}) - y_i \mathbf{x}_i \boldsymbol{\beta} + b(\mathbf{x}_i \boldsymbol{\beta})] \\
&\quad + \frac{\lambda_n}{n} \sum_{j=1}^p (|\xi_j|^\nu - |\beta_j|^\nu) \\
&= -\frac{1}{n} \sum_{i=1}^n \left[ \frac{1}{\phi} (y_i - b'(\mathbf{x}_i \boldsymbol{\beta})) \mathbf{x}_i (\boldsymbol{\xi} - \boldsymbol{\beta}) \right] \\
&\quad + \frac{1}{\phi} \sum_{i=1}^n \frac{1}{n} [b(\mathbf{x}_i \boldsymbol{\xi}) - b(\mathbf{x}_i \boldsymbol{\beta}) - b'(\mathbf{x}_i \boldsymbol{\beta}) \mathbf{x}_i (\boldsymbol{\xi} - \boldsymbol{\beta})] \\
&\quad + \frac{\lambda_n}{n} \sum_{j=1}^p (|\xi_j|^\nu - |\beta_j|^\nu). \tag{5.3}
\end{aligned}$$

Note that

$$\begin{aligned}
&\text{Var} \left[ \frac{1}{n\phi} \sum_{i=1}^n (y_i - b'(\mathbf{x}_i \boldsymbol{\beta})) \mathbf{x}_i (\boldsymbol{\xi} - \boldsymbol{\beta}) \right] \\
&= \frac{1}{n^2 \phi^2} \sum_{i=1}^n (\boldsymbol{\xi} - \boldsymbol{\beta})^T \mathbf{x}_i^T (E(y_i - b'(\mathbf{x}_i \boldsymbol{\beta}))^2) \mathbf{x}_i (\boldsymbol{\xi} - \boldsymbol{\beta}) \\
&= \frac{1}{n\phi^2} (\boldsymbol{\xi} - \boldsymbol{\beta})^T \left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^T \phi b''(\mathbf{x}_i \boldsymbol{\beta}) \mathbf{x}_i \right) (\boldsymbol{\xi} - \boldsymbol{\beta}) \\
&\rightarrow 0, \tag{5.4}
\end{aligned}$$

since  $(1/n) \sum_{i=1}^n \mathbf{x}_i^T \phi b''(\mathbf{x}_i \boldsymbol{\beta}) \mathbf{x}_i$  converges uniformly to a finite nonnegative matrix

$C(\boldsymbol{\beta})$  by Condition C1. From (5.4) and the compactness of  $K$ , we have

$$\sup_{\boldsymbol{\xi} \in K} \left\{ \text{Var} \left[ \frac{1}{n\phi} \sum_{i=1}^n (y_i - b'(\mathbf{x}_i \boldsymbol{\beta})) \mathbf{x}_i (\boldsymbol{\xi} - \boldsymbol{\beta}) \right] \right\} \rightarrow 0. \tag{5.5}$$

Using Chebyshev's inequality on the first term of  $Z_n(\boldsymbol{\xi})$ , when  $n \rightarrow \infty$ ,

$$\begin{aligned}
&P \left[ \left| \frac{1}{n\phi} \sum_{i=1}^n (y_i - b'(\mathbf{x}_i \boldsymbol{\beta})) \mathbf{x}_i (\boldsymbol{\xi} - \boldsymbol{\beta}) \right| > \epsilon \right] \\
&\leq \frac{1}{\epsilon^2} \text{Var} \left[ \frac{1}{n\phi} \sum_{i=1}^n (y_i - b'(\mathbf{x}_i \boldsymbol{\beta})) \mathbf{x}_i (\boldsymbol{\xi} - \boldsymbol{\beta}) \right] \\
&\rightarrow 0, \tag{5.6}
\end{aligned}$$

uniformly over  $\boldsymbol{\xi} \in K$ . Then for each  $\boldsymbol{\xi} \in K$ ,

$$\frac{1}{n\phi} \sum_{i=1}^n (y_i - b'(\mathbf{x}_i\boldsymbol{\beta}))\mathbf{x}_i(\boldsymbol{\xi} - \boldsymbol{\beta}) \rightarrow 0. \quad (5.7)$$

Supposed that  $T(\boldsymbol{\xi}) = |\sum_{i=1}^n (y_i - b'(\mathbf{x}_i\boldsymbol{\beta}))\mathbf{x}_i(\boldsymbol{\xi} - \boldsymbol{\beta})/(n\phi)|$ . If  $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2 \in K$  and  $\omega \in [0, 1]$ , then

$$\begin{aligned} & T(\omega\boldsymbol{\xi}_1 + (1 - \omega)\boldsymbol{\xi}_2) \\ &= \left| \frac{1}{n\phi} \sum_{i=1}^n (y_i - b'(\mathbf{x}_i\boldsymbol{\beta}))\mathbf{x}_i(\omega\boldsymbol{\xi}_1 + (1 - \omega)\boldsymbol{\xi}_2 - \boldsymbol{\beta}) \right| \\ &= \left| \frac{1}{n\phi} \sum_{i=1}^n (y_i - b'(\mathbf{x}_i\boldsymbol{\beta}))\mathbf{x}_i(\omega\boldsymbol{\xi}_1 + (1 - \omega)\boldsymbol{\xi}_2 - (\omega + 1 - \omega)\boldsymbol{\beta}) \right| \\ &= \left| \frac{\omega}{n\phi} \sum_{i=1}^n (y_i - b'(\mathbf{x}_i\boldsymbol{\beta}))\mathbf{x}_i(\boldsymbol{\xi}_1 - \boldsymbol{\beta}) + \frac{1 - \omega}{n\phi} \sum_{i=1}^n (y_i - b'(\mathbf{x}_i\boldsymbol{\beta}))\mathbf{x}_i(\boldsymbol{\xi}_2 - \boldsymbol{\beta}) \right| \\ &\leq \omega \left| \frac{1}{n\phi} \sum_{i=1}^n (y_i - b'(\mathbf{x}_i\boldsymbol{\beta}))\mathbf{x}_i(\boldsymbol{\xi}_1 - \boldsymbol{\beta}) \right| + (1 - \omega) \left| \frac{1}{n\phi} \sum_{i=1}^n (y_i - b'(\mathbf{x}_i\boldsymbol{\beta}))\mathbf{x}_i(\boldsymbol{\xi}_2 - \boldsymbol{\beta}) \right| \\ &= \omega T(\boldsymbol{\xi}_1) + (1 - \omega) T(\boldsymbol{\xi}_2). \end{aligned}$$

Hence,  $|\sum_{i=1}^n (y_i - b'(\mathbf{x}_i\boldsymbol{\beta}))\mathbf{x}_i(\boldsymbol{\xi} - \boldsymbol{\beta})/(n\phi)|$  is a convex function. Also, the function converges pointwise to 0 by (5.7). From the results of Theorem II.1 from Andersen and Gill [2], we have

$$\sup_{\boldsymbol{\xi} \in K} \left| \frac{1}{n\phi} \sum_{i=1}^n (y_i - b'(\mathbf{x}_i\boldsymbol{\beta}))\mathbf{x}_i(\boldsymbol{\xi} - \boldsymbol{\beta}) - 0 \right| \rightarrow_p 0, \quad (5.8)$$

as  $n \rightarrow \infty$ .

Observe that

$$\begin{aligned}
& |Z_n(\boldsymbol{\xi}) - Z(\boldsymbol{\xi})| \\
& \leq \left| \frac{1}{\phi} \sum_{i=1}^n \frac{1}{n} [b(\mathbf{x}_i \boldsymbol{\xi}) - b(\mathbf{x}_i \boldsymbol{\beta}) - b'(\mathbf{x}_i \boldsymbol{\beta}) \mathbf{x}_i (\boldsymbol{\xi} - \boldsymbol{\beta})] \right. \\
& \quad \left. - \lim \left\{ \frac{1}{\phi} \sum_{i=1}^n \frac{1}{n} [b(\mathbf{x}_i \boldsymbol{\xi}) - b(\mathbf{x}_i \boldsymbol{\beta}) - b'(\mathbf{x}_i \boldsymbol{\beta}) \mathbf{x}_i (\boldsymbol{\xi} - \boldsymbol{\beta})] \right\} \right| \\
& \quad + \left| \frac{1}{n \phi} \sum_{i=1}^n (y_i - b'(\mathbf{x}_i \boldsymbol{\beta})) \mathbf{x}_i (\boldsymbol{\xi} - \boldsymbol{\beta}) \right| \\
& \quad + \left| \left( \frac{\lambda_n}{n} - \lambda_0 \right) \sum_{j=1}^p (|\xi_j|^\nu - |\beta_j|^\nu) \right|. \tag{5.9}
\end{aligned}$$

The supremum of the last term

$$\sup_{\boldsymbol{\xi} \in K} \left\{ \left| \left( \frac{\lambda_n}{n} - \lambda_0 \right) \sum_{j=1}^p (|\xi_j|^\nu - |\beta_j|^\nu) \right| \right\} \rightarrow_p 0 \tag{5.10}$$

because  $\boldsymbol{\xi}$  is bounded which follows from  $\boldsymbol{\xi} \in K$  and the function  $\sum_{j=1}^p |\xi_j|^\nu$  is bounded. Then,

$$\begin{aligned}
& \sup_{\boldsymbol{\xi} \in K} |Z_n(\boldsymbol{\xi}) - Z(\boldsymbol{\xi})| \\
& \leq \sup_{\boldsymbol{\xi} \in K} \left| \frac{1}{\phi} \sum_{i=1}^n \frac{1}{n} [b(\mathbf{x}_i \boldsymbol{\xi}) - b(\mathbf{x}_i \boldsymbol{\beta}) - b'(\mathbf{x}_i \boldsymbol{\beta}) \mathbf{x}_i (\boldsymbol{\xi} - \boldsymbol{\beta})] \right. \\
& \quad \left. - \lim \left\{ \frac{1}{\phi} \sum_{i=1}^n \frac{1}{n} [b(\mathbf{x}_i \boldsymbol{\xi}) - b(\mathbf{x}_i \boldsymbol{\beta}) - b'(\mathbf{x}_i \boldsymbol{\beta}) \mathbf{x}_i (\boldsymbol{\xi} - \boldsymbol{\beta})] \right\} \right| \\
& \quad + \sup_{\boldsymbol{\xi} \in K} \left| \frac{1}{n \phi} \sum_{i=1}^n (y_i - b'(\mathbf{x}_i \boldsymbol{\beta})) \mathbf{x}_i (\boldsymbol{\xi} - \boldsymbol{\beta}) \right| \\
& \quad + \sup_{\boldsymbol{\xi} \in K} \left| \left( \frac{\lambda_n}{n} - \lambda_0 \right) \sum_{j=1}^p (|\xi_j|^\nu - |\beta_j|^\nu) \right|. \\
& \rightarrow_p 0. \tag{5.11}
\end{aligned}$$

by the hypothesized uniform convergence of  $F_n(\boldsymbol{\xi})$ , (5.8) and (5.10).  $\square$

**Remark 5.1** It can be seen from the proof of Lemma 5.1 that

$$Z(\boldsymbol{\xi}) = \lim \left\{ \frac{1}{\phi} \sum_{i=1}^n \frac{1}{n} [b(\mathbf{x}_i \boldsymbol{\xi}) - b(\mathbf{x}_i \boldsymbol{\beta}) - y_i \mathbf{x}_i (\boldsymbol{\xi} - \boldsymbol{\beta})] \right\} + \lambda_0 \sum_{j=1}^p (|\xi_j|^\nu - |\beta_j|^\nu).$$

**Remark 5.2** The hypothesis about the convergence of  $F_n(\boldsymbol{\xi})$  can be modified by adding conditions on the  $\mathbf{x}_i$  or on  $H$ . As an example, we state the following corollary.

**Corollary 5.1** *If the  $\mathbf{x}_i$  are uniformly bounded, then the conclusion of Lemma 5.1 holds.*

**Proof:** Boundedness of the  $\mathbf{x}_i$ , the strong law of large numbers and compactness of  $K$  guarantee the uniform convergence of  $F_n(\boldsymbol{\xi})$  to

$$F_n(\boldsymbol{\xi}) = E_H [b(\mathbf{x}_i \boldsymbol{\xi}) - b(\mathbf{x}_i \boldsymbol{\beta}) - b'(\mathbf{x}_i \boldsymbol{\beta}) \mathbf{x}_i (\boldsymbol{\xi} - \boldsymbol{\beta})]. \quad \square$$

Theorem 5.3 below shows that the Bridge estimators  $\hat{\boldsymbol{\beta}}_n$  are consistent if  $\lambda_n$  is of order  $o(n)$ .

**Theorem 5.3** *If  $C(\boldsymbol{\beta})$  defined in Condition C1 is nonsingular and  $\lambda_n/n \rightarrow \lambda_0 \geq 0$ , then  $\hat{\boldsymbol{\beta}}_n \rightarrow_p \operatorname{argmin}(Z)$  where*

$$Z(\boldsymbol{\xi}) = \lim \left\{ \frac{1}{\phi} \sum_{i=1}^n \frac{1}{n} [b(\mathbf{x}_i \boldsymbol{\xi}) - b(\mathbf{x}_i \boldsymbol{\beta}) - y_i \mathbf{x}_i (\boldsymbol{\xi} - \boldsymbol{\beta})] \right\} + \lambda_0 \sum_{j=1}^p (|\xi_j|^\nu - |\beta_j|^\nu).$$

*Thus if  $\lambda_n = o(n)$ ,  $\operatorname{argmin}(Z) = \boldsymbol{\beta}$  and so  $\hat{\boldsymbol{\beta}}_n$  is consistent.*

**Proof:** Define  $Z_n$  as in (5.1). We need the following two conditions:

$$\sup_{\boldsymbol{\xi} \in K} |Z_n(\boldsymbol{\xi}) - Z(\boldsymbol{\xi})| \rightarrow_p 0 \quad (5.12)$$

for any compact set  $K$  and

$$\hat{\boldsymbol{\beta}}_n = O_p(1). \quad (5.13)$$



Under (5.12) and (5.13), we have

$$\operatorname{argmin}(Z_n) \rightarrow_p \operatorname{argmin}(Z).$$

For  $\nu \geq 1$ ,  $Z_n$  is convex, and therefore (5.2) and (5.13) are true from the uniform convergence of  $Z_n(\boldsymbol{\xi})$  to  $Z(\boldsymbol{\xi})$  and by applying Theorems II.1 from Andersen and Gill [2]. For  $\nu < 1$ ,  $Z_n$  is not convex, but (5.2) follows from Lemma 5.1. Note that

$$Z_n(\boldsymbol{\xi}) \geq -\frac{1}{n} \sum_{i=1}^n \log \left[ \frac{l(y_i, \boldsymbol{\xi})}{l(y_i, \boldsymbol{\beta})} \right] = Z_n^{(0)}(\boldsymbol{\xi}),$$

for all  $\boldsymbol{\xi}$ . Since  $\operatorname{argmin}(Z_n^{(0)}) = O_p(1)$ , it follows that  $\operatorname{argmin}(Z_n) = O_p(1)$ .  $\square$

The limiting distribution of the Bridge estimators can be obtained when  $\lambda_n$  grows slowly. Theorem 5.4 shows that the Bridge Estimator is  $\sqrt{n}$ -consistent when  $\lambda_n = o(\sqrt{n})$  for  $\nu \geq 1$ , whereas Theorem 5.5 proves that the rate of growth should be  $\lambda_n = o(n^{\nu/2})$  when  $\nu < 1$ . Before stating the theorems, let us consider the following lemma:

**Lemma 5.2** *If  $y$  follows an exponential family distribution which has the following probability density function:*

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

*then the moment generating function of  $\exp(y/\phi)$  is*

$$E \left[ \exp \left( \frac{ty}{\phi} \right) \right] = \exp \left[ \frac{1}{\phi} (b(t + \theta) - b(\theta)) \right].$$

**Proof:**

$$\begin{aligned}
& E \left[ \exp \left( \frac{ty}{\phi} \right) \right] \\
&= \int \exp \left( \frac{ty}{\phi} \right) \exp \left[ \frac{y\theta - b(\theta)}{\phi} \right] \exp(c(y, \phi)) dy \\
&= \int \exp \left[ \frac{1}{\phi} (b(t + \theta) - b(\theta)) \right] \exp \left[ \frac{1}{\phi} ((t + \theta)y - b(t + \theta)) \right] \exp(c(y, \phi)) dy \\
&= \exp \left[ \frac{1}{\phi} (b(t + \theta) - b(\theta)) \right]. \quad \square
\end{aligned} \tag{5.14}$$

**Lemma 5.3** Suppose that  $\mathbf{Y}$ ,  $\mathbf{u}$  are vectors with length  $n$  and  $p + 1$  respectively.

If  $\mathbf{Y}$  follows an exponential family distribution (2.1), then the moment generating

function of  $\exp [(1/(\sqrt{n}\phi)) \sum_{i=1}^n (y_i - b'(\mathbf{x}_i\boldsymbol{\beta})) \mathbf{x}_i\mathbf{u}]$  satisfies

$$\begin{aligned}
M_n(t) &= E \left\{ \exp \left[ \frac{t}{\sqrt{n}\phi} \sum_{i=1}^n (y_i - b'(\mathbf{x}_i\boldsymbol{\beta})) \mathbf{x}_i\mathbf{u} \right] \right\} \\
&\rightarrow \exp \left[ \frac{t^2}{2\phi^2} \mathbf{u}^T \mathbf{C}(\boldsymbol{\beta}) \mathbf{u} \right].
\end{aligned}$$

Thus, as  $n \rightarrow \infty$ ,  $\exp \left[ \frac{1}{\sqrt{n}\phi} \sum_{i=1}^n (y_i - b'(\mathbf{x}_i\boldsymbol{\beta})) \mathbf{x}_i\mathbf{u} \right] \rightarrow_d \mathbf{u}^T \mathbf{W}$ , where  $\mathbf{W}$  has a  $N(\mathbf{0}, \mathbf{C}(\boldsymbol{\beta})/\phi^2)$  distribution.

**Proof:** Observe that by Taylor expansion,

$$\sum_{i=1}^n b \left( \mathbf{x}_i \left( \boldsymbol{\beta} + \frac{t\mathbf{u}}{\sqrt{n}} \right) \right) = \sum_{i=1}^n \left[ b(\mathbf{x}_i\boldsymbol{\beta}) + \frac{t}{\sqrt{n}} b'(\mathbf{x}_i\boldsymbol{\beta}) \mathbf{x}_i\mathbf{u} + \frac{t^2}{2n} \mathbf{u}^T \mathbf{x}_i^T b''(\mathbf{x}_i\boldsymbol{\beta}^*) \mathbf{x}_i\mathbf{u} \right], \tag{5.15}$$

where  $\boldsymbol{\beta}^*$  is between  $\boldsymbol{\beta}$  and  $\boldsymbol{\beta} + t/\sqrt{n}$ . By Lemma 5.2,

$$E \left[ \exp \left( \frac{t\mathbf{x}_i\mathbf{u}}{\sqrt{n}\phi} y_i \right) \right] = \exp \left\{ \frac{1}{\phi} \left[ b \left( \mathbf{x}_i\boldsymbol{\beta} + \frac{t}{\sqrt{n}} \mathbf{x}_i\mathbf{u} \right) - b(\mathbf{x}_i\boldsymbol{\beta}) \right] \right\}. \tag{5.16}$$

Then

$$\begin{aligned}
M_n(t) &= \prod_{i=1}^n E \left\{ \exp \left[ \frac{t}{\sqrt{n}\phi} (y_i - b'(\mathbf{x}_i\boldsymbol{\beta})) \mathbf{x}_i \mathbf{u} \right] \right\} \\
&= \prod_{i=1}^n \left\{ \exp \left[ \frac{1}{\phi} \left\{ b \left( \mathbf{x}_i \boldsymbol{\beta} + \frac{t}{\sqrt{n}} \mathbf{x}_i \mathbf{u} \right) - b(\mathbf{x}_i\boldsymbol{\beta}) - \frac{t}{\sqrt{n}\phi} b'(\mathbf{x}_i\boldsymbol{\beta}) \mathbf{x}_i \mathbf{u} \right\} \right] \right\} \\
&= \exp \left[ \frac{1}{\phi} \sum_{i=1}^n \left\{ b \left( \mathbf{x}_i \left( \boldsymbol{\beta} + \frac{t\mathbf{u}}{\sqrt{n}} \right) \right) - b(\mathbf{x}_i\boldsymbol{\beta}) - \frac{t}{\sqrt{n}\phi} b'(\mathbf{x}_i\boldsymbol{\beta}) \mathbf{x}_i \mathbf{u} \right\} \right] \\
&= \exp \left[ \frac{1}{\phi} \sum_{i=1}^n \left\{ b(\mathbf{x}_i\boldsymbol{\beta}) + \frac{t}{\sqrt{n}} b'(\mathbf{x}_i\boldsymbol{\beta}) \mathbf{x}_i \mathbf{u} + \frac{t^2}{2n} \mathbf{u}^T \mathbf{x}_i^T b''(\mathbf{x}_i\boldsymbol{\beta}^*) \mathbf{x}_i \mathbf{u} \right. \right. \\
&\quad \left. \left. - b(\mathbf{x}_i\boldsymbol{\beta}) - \frac{t}{\sqrt{n}\phi} b'(\mathbf{x}_i\boldsymbol{\beta}) \mathbf{x}_i \mathbf{u} \right\} \right] \\
&= \exp \left[ \frac{1}{\phi} \sum_{i=1}^n \left\{ \frac{t^2}{2n} \mathbf{u}^T \mathbf{x}_i^T b''(\mathbf{x}_i\boldsymbol{\beta}) \mathbf{x}_i \mathbf{u} \right. \right. \\
&\quad \left. \left. + \frac{t^2}{2n} \mathbf{u}^T \mathbf{x}_i^T (b''(\mathbf{x}_i\boldsymbol{\beta}^*) - b''(\mathbf{x}_i\boldsymbol{\beta})) \mathbf{x}_i \mathbf{u} \right\} \right] \\
&= \exp \left[ \frac{t^2}{2\phi^2} \mathbf{u}^T \left( \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i^T \phi b''(\mathbf{x}_i\boldsymbol{\beta}) \mathbf{x}_i) \right) \mathbf{u} \right] \\
&\quad \times \exp \left[ \frac{t^2}{2\phi^2} \mathbf{u}^T \left( \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i^T \phi (b''(\mathbf{x}_i\boldsymbol{\beta}^*) - b''(\mathbf{x}_i\boldsymbol{\beta})) \mathbf{x}_i) \right) \mathbf{u} \right] \\
&\longrightarrow \exp \left[ \frac{t^2}{2\phi^2} \mathbf{u}^T C(\boldsymbol{\beta}) \mathbf{u} \right].
\end{aligned}$$

The justification of the last statement is as follows. Since  $b''(\mathbf{x}_i\boldsymbol{\beta}^*) - b''(\mathbf{x}_i\boldsymbol{\beta}) = O(\mathbf{x}_i\boldsymbol{\beta}^*/\sqrt{n})$ , writing  $\|\mathbf{x}\|$  for the  $L_2$  norm of  $\mathbf{x}$ ,

$$\left| \frac{\mathbf{x}_i\boldsymbol{\beta}^*}{\sqrt{n}} \right| \leq \frac{\|\mathbf{x}_i\|^{1/2} \|\boldsymbol{\beta}^*\|^{1/2}}{\sqrt{n}} \leq \underbrace{\left( \max \frac{\|\mathbf{x}_i\|^{1/2}}{\sqrt{n}} \right)}_{\rightarrow 0} \underbrace{\|\boldsymbol{\beta}^*\|^{1/2}}_{O(1)} \rightarrow 0$$

by Condition C2. Therefore

$$\frac{t^2}{2\phi^2} \mathbf{u}^T \left( \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i^T \phi (b''(\mathbf{x}_i\boldsymbol{\beta}^*) - b''(\mathbf{x}_i\boldsymbol{\beta})) \mathbf{x}_i) \right) \mathbf{u} \rightarrow 0,$$

over  $\boldsymbol{\beta}^*$  in the compact and convex set  $K$ , as  $n \rightarrow \infty$ .  $\square$

**Theorem 5.4** Suppose that  $\nu \geq 1$ . If  $\lambda_n/\sqrt{n} \rightarrow \lambda_0 \geq 0$ , and  $C(\boldsymbol{\beta})$  is nonsingular, then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \rightarrow_d \operatorname{argmin}(V),$$

where if  $\nu > 1$ ,

$$V(\mathbf{u}) = -\mathbf{u}^T \mathbf{W} + \frac{1}{2\phi^2} \mathbf{u}^T C(\boldsymbol{\beta}) \mathbf{u} + \nu \lambda_0 \sum_{j=1}^p u_j \operatorname{sgn}(\beta_j) |\beta_j|^{\nu-1},$$

if  $\nu = 1$ ,

$$V(\mathbf{u}) = -\mathbf{u}^T \mathbf{W} + \frac{1}{2\phi^2} \mathbf{u}^T C(\boldsymbol{\beta}) \mathbf{u} + \lambda_0 \sum_{j=1}^p [u_j \operatorname{sgn}(\beta_j) I(\beta_j \neq 0) + |u_j| I(\beta_j = 0)],$$

and  $\mathbf{W}$  has a  $N(\mathbf{0}, \mathbf{C}(\boldsymbol{\beta})/\phi^2)$  distribution.

**Proof:** Define  $V_n(\mathbf{u})$  by

$$V_n(\mathbf{u}) = -\sum_{i=1}^n \log \left[ \frac{l(y_i, \boldsymbol{\beta} + \mathbf{u}/\sqrt{n})}{l(y_i, \boldsymbol{\beta})} \right] + \lambda_n \sum_{j=1}^p \left( \left| \beta_j + \frac{u_j}{\sqrt{n}} \right|^\nu - |\beta_j|^\nu \right), \quad (5.17)$$

where  $\mathbf{u}$  is a length  $p + 1$  vector and observe that  $V_n$  is minimized at  $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ .

Examining  $V_n$  carefully, one finds that

$$\begin{aligned}
V_n(\mathbf{u}) &= -\sum_{i=1}^n \frac{1}{\phi} \left\{ y_i \mathbf{x}_i \left( \boldsymbol{\beta} + \frac{\mathbf{u}}{\sqrt{n}} \right) - b \left( \mathbf{x}_i \left( \boldsymbol{\beta} + \frac{\mathbf{u}}{\sqrt{n}} \right) \right) - y_i \mathbf{x}_i \boldsymbol{\beta} + b(\mathbf{x}_i \boldsymbol{\beta}) \right\} \\
&\quad + \lambda_n \sum_{j=1}^p \left( \left| \beta_j + \frac{u_j}{\sqrt{n}} \right|^\nu - |\beta_j|^\nu \right) \\
&= -\frac{1}{\phi} \sum_{i=1}^n \left\{ y_i \mathbf{x}_i \frac{\mathbf{u}}{\sqrt{n}} - \left[ b(\mathbf{x}_i \boldsymbol{\beta}) + \frac{\mathbf{u}^T \mathbf{x}_i}{\sqrt{n}} b'(\mathbf{x}_i \boldsymbol{\beta}) \right. \right. \\
&\quad \left. \left. + \frac{1}{2n} b'' \left( \mathbf{x}_i \left( \boldsymbol{\beta} + \frac{\tilde{\mathbf{u}}}{\sqrt{n}} \right) \right) \mathbf{u}^T \mathbf{x}_i^T \mathbf{x}_i \mathbf{u} \right] + b(\mathbf{x}_i \boldsymbol{\beta}) \right\} \\
&\quad + \lambda_n \sum_{j=1}^p \left( \left| \beta_j + \frac{u_j}{\sqrt{n}} \right|^\nu - |\beta_j|^\nu \right) \\
&= -\frac{1}{\sqrt{n} \phi} \sum_{i=1}^n (y_i - b'(\mathbf{x}_i \boldsymbol{\beta})) \mathbf{x}_i \mathbf{u} + \frac{1}{2\phi^2} \mathbf{u}^T \left[ \frac{\phi}{n} \sum_{i=1}^n \mathbf{x}_i^T b''(\mathbf{x}_i \boldsymbol{\beta}) \mathbf{x}_i \right] \mathbf{u} \\
&\quad + \frac{1}{2\phi^2} \mathbf{u}^T \left\{ \frac{\phi}{n} \sum_{i=1}^n \mathbf{x}_i^T \left[ b'' \left( \mathbf{x}_i \left( \boldsymbol{\beta} + \frac{\tilde{\mathbf{u}}}{\sqrt{n}} \right) \right) - b''(\mathbf{x}_i \boldsymbol{\beta}) \right] \mathbf{x}_i \right\} \mathbf{u} \\
&\quad + \lambda_n \sum_{j=1}^p \left( \left| \beta_j + \frac{u_j}{\sqrt{n}} \right|^\nu - |\beta_j|^\nu \right), \tag{5.18}
\end{aligned}$$

where  $\tilde{\mathbf{u}}$  is between  $\mathbf{u}$  and  $\boldsymbol{\beta}$ . Similar to the proof of Lemma 5.3, since

$$b'' \left( \mathbf{x}_i \left( \boldsymbol{\beta} + \tilde{\mathbf{u}}/\sqrt{n} \right) \right) - b''(\mathbf{x}_i \boldsymbol{\beta}) = O(\mathbf{x}_i \tilde{\mathbf{u}}/\sqrt{n}),$$

by Condition C2, as  $n \rightarrow \infty$ ,

$$\frac{1}{2\phi^2} \mathbf{u}^T \left\{ \frac{\phi}{n} \sum_{i=1}^n \mathbf{x}_i^T \left[ b'' \left( \mathbf{x}_i \left( \boldsymbol{\beta} + \frac{\tilde{\mathbf{u}}}{\sqrt{n}} \right) \right) - b''(\mathbf{x}_i \boldsymbol{\beta}) \right] \mathbf{x}_i \right\} \mathbf{u} \rightarrow 0,$$

over  $\tilde{\mathbf{u}}$  in the compact and convex set  $K$ .

Hence, by Lemma 5.3 and Central Limit Theorem (CLT),

$$\begin{aligned}
&-\frac{1}{\sqrt{n} \phi} \sum_{i=1}^n (y_i - b'(\mathbf{x}_i \boldsymbol{\beta})) \mathbf{x}_i \mathbf{u} + \frac{1}{2\phi^2} \mathbf{u}^T \left[ \frac{\phi}{n} \sum_{i=1}^n \mathbf{x}_i^T b''(\mathbf{x}_i \boldsymbol{\beta}) \mathbf{x}_i \right] \mathbf{u} \\
&\quad + \frac{1}{2\phi^2} \mathbf{u}^T \left\{ \frac{\phi}{n} \sum_{i=1}^n \mathbf{x}_i^T \left[ b'' \left( \mathbf{x}_i \left( \boldsymbol{\beta} + \frac{\tilde{\mathbf{u}}}{\sqrt{n}} \right) \right) - b''(\mathbf{x}_i \boldsymbol{\beta}) \right] \mathbf{x}_i \right\} \mathbf{u} \\
&\quad \rightarrow_d -\mathbf{u}^T \mathbf{W} + \frac{1}{2\phi^2} \mathbf{u}^T C(\boldsymbol{\beta}) \mathbf{u}.
\end{aligned}$$

When  $\nu > 1$ ,

$$\begin{aligned} \lambda_n \sum_{j=1}^p \left( \left| \beta_j + \frac{u_j}{\sqrt{n}} \right|^\nu - |\beta_j|^\nu \right) &= \lambda_n \sum_{j=1}^p \left| \frac{u_j}{\sqrt{n}} \right|^\nu I(\beta_j = 0) \\ &+ \lambda_n \sum_{j=1}^p \left( \left( \beta_j + \frac{u_j}{\sqrt{n}} \right)^\nu - \beta_j^\nu \right) I(\beta_j > 0) \\ &+ \lambda_n \sum_{j=1}^p \left( \left( -\beta_j - \frac{u_j}{\sqrt{n}} \right)^\nu - (-\beta_j)^\nu \right) I(\beta_j < 0). \end{aligned}$$

In the case of  $\beta_j = 0$ ,

$$\lambda_n \left| \frac{u_j}{\sqrt{n}} \right|^\nu = \frac{\lambda_n}{\sqrt{n}} \sqrt{n} \left| \frac{u_j}{\sqrt{n}} \right|^\nu = \frac{\lambda_n}{\sqrt{n}} \frac{|u_j|^\nu}{(\sqrt{n})^{\nu-1}} \longrightarrow 0,$$

as  $n \rightarrow \infty$ . If  $\beta_j > 0$ , applying the binomial theorem,

$$\begin{aligned} \lambda_n \left( \left( \beta_j + \frac{u_j}{\sqrt{n}} \right)^\nu - \beta_j^\nu \right) &= \lambda_n \left[ \beta_j^\nu + \nu \frac{u_j}{\sqrt{n}} \beta_j^{\nu-1} + o\left(\frac{1}{\sqrt{n}}\right) - \beta_j^\nu \right] \\ &= \frac{\lambda_n}{\sqrt{n}} \left[ \nu u_j \beta_j^{\nu-1} + o\left(\frac{1}{\sqrt{n}}\right) \right] \\ &\longrightarrow \nu \lambda_0 u_j |\beta_j|^{\nu-1}. \end{aligned}$$

Similarly, when  $\beta_j < 0$ ,

$$\begin{aligned} \lambda_n \left( \left( -\beta_j - \frac{u_j}{\sqrt{n}} \right)^\nu - (-\beta_j)^\nu \right) &= \lambda_n \left[ (-\beta_j)^\nu + \nu \left( -\frac{u_j}{\sqrt{n}} \right) (-\beta_j)^{\nu-1} + o\left(\frac{1}{\sqrt{n}}\right) - (-\beta_j)^\nu \right] \\ &= \frac{\lambda_n}{\sqrt{n}} \left[ \nu (-u_j) (-\beta_j)^{\nu-1} + o\left(\frac{1}{\sqrt{n}}\right) \right] \\ &\longrightarrow -\nu \lambda_0 u_j |\beta_j|^{\nu-1}. \end{aligned}$$

Therefore

$$\lambda_n \sum_{j=1}^p \left( \left| \beta_j + \frac{u_j}{\sqrt{n}} \right|^\nu - |\beta_j|^\nu \right) \longrightarrow \nu \lambda_0 \sum_{j=1}^p u_j \operatorname{sgn}(\beta_j) |\beta_j|^{\nu-1}.$$

When  $\nu = 1$ , the penalty component can be decomposed into three parts:

$$\begin{aligned} \lambda_n \sum_{j=1}^p \left( \left| \beta_j + \frac{u_j}{\sqrt{n}} \right| - |\beta_j| \right) &= \lambda_n \sum_{j=1}^p \left( \left( \beta_j + \frac{u_j}{\sqrt{n}} \right) - \beta_j \right) I(\beta_j > 0) \\ &\quad + \lambda_n \sum_{j=1}^p \left( \left( -\beta_j - \frac{u_j}{\sqrt{n}} \right) - (-\beta_j) \right) I(\beta_j < 0) \\ &\quad + \lambda_n \sum_{j=1}^p \left| \frac{u_j}{\sqrt{n}} \right| I(\beta_j = 0). \end{aligned}$$

That is,

$$\lambda_n \sum_{j=1}^p \left( \left| \beta_j + \frac{u_j}{\sqrt{n}} \right| - |\beta_j| \right) \longrightarrow \lambda_0 \sum_{j=1}^p [u_j \operatorname{sgn}(\beta_j) I(\beta_j \neq 0) + |u_j| I(\beta_j = 0)].$$

Hence,  $V_n(\mathbf{u}) \rightarrow_d V(\mathbf{u})$ . Since  $V_n(\mathbf{u})$  is convex and  $V$  has a unique minimum, by Geyer [15], the following result holds:

$$\operatorname{argmin}(V_n) = \sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \rightarrow_d \operatorname{argmin}(V). \quad \square$$

When  $\nu < 1$ , a different rate of growth of  $\lambda_n$  is assumed to get a limiting distribution.

**Theorem 5.5** *Suppose that  $\nu < 1$ . If  $\lambda_n/n^{\nu/2} \rightarrow \lambda_0 \geq 0$  and  $C(\boldsymbol{\beta})$  is nonsingular, then*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \rightarrow_d \operatorname{argmin}(V),$$

where

$$V(\mathbf{u}) = -\mathbf{u}^T \mathbf{W} + \frac{1}{2\phi^2} \mathbf{u}^T C(\boldsymbol{\beta}) \mathbf{u} + \lambda_0 \sum_{j=1}^p |u_j|^\nu I(\beta_j = 0)$$

and  $\mathbf{W}$  has a  $N(\mathbf{0}, \mathbf{C}(\boldsymbol{\beta})/\phi^2)$  distribution.

**Proof:** Following the same idea of the proof of Theorem 5.4, define

$$\begin{aligned}
V_n(\mathbf{u}) &= -\frac{1}{\sqrt{n}\phi} \sum_{i=1}^n (y_i - b'(\mathbf{x}_i\boldsymbol{\beta})) \mathbf{x}_i \mathbf{u} + \frac{1}{2\phi^2} \mathbf{u}^T \left[ \frac{\phi}{n} \sum_{i=1}^n \mathbf{x}_i^T b''(\mathbf{x}_i\boldsymbol{\beta}) \mathbf{x}_i \right] \mathbf{u} \\
&\quad + \frac{1}{2\phi^2} \mathbf{u}^T \left\{ \frac{\phi}{n} \sum_{i=1}^n \mathbf{x}_i^T \left[ b'' \left( \mathbf{x}_i \left( \boldsymbol{\beta} + \frac{\tilde{\mathbf{u}}}{\sqrt{n}} \right) \right) - b''(\mathbf{x}_i\boldsymbol{\beta}) \right] \mathbf{x}_i \right\} \mathbf{u} \\
&\quad + \lambda_n \sum_{j=1}^p \left( \left| \beta_j + \frac{u_j}{\sqrt{n}} \right|^\nu - |\beta_j|^\nu \right). \tag{5.19}
\end{aligned}$$

As in the proof of Theorem 5.4, the sum of the first two terms in  $V_n$  converges in distribution to  $-\mathbf{u}^T \mathbf{W} + \mathbf{u}^T C(\boldsymbol{\beta}) \mathbf{u} / (2\phi^2)$ . The penalty component consists of three pieces depending on the sign of  $\beta_j$ . If  $\beta_j > 0$ , using the binomial theorem,

$$\begin{aligned}
\lambda_n \left( \left( \beta_j + \frac{u_j}{\sqrt{n}} \right)^\nu - \beta_j^\nu \right) &= \frac{\lambda_n}{(\sqrt{n})^\nu} (\sqrt{n})^\nu \left[ \beta_j^\nu + \nu \frac{u_j}{\sqrt{n}} \beta_j^{\nu-1} + o \left( \frac{1}{\sqrt{n}} \right) - \beta_j^\nu \right] \\
&= \frac{\lambda_n}{(\sqrt{n})^\nu} \left[ \frac{1}{n^{(1-\nu)/2}} \nu u_j \beta_j^{\nu-1} + o \left( \frac{1}{n^{(1-\nu)/2}} \right) \right] \\
&\rightarrow \lambda_0 \cdot 0 = 0.
\end{aligned}$$

since  $1/(n^{(1-\nu)/2}) \rightarrow 0$ . Similarly, when  $\beta_j < 0$ ,

$$\lambda_n \left( \left( -\beta_j - \frac{u_j}{\sqrt{n}} \right)^\nu - (-\beta_j)^\nu \right) \rightarrow \lambda_0 \cdot 0 = 0.$$

If  $\beta_j = 0$ ,

$$\lambda_n \left| \frac{u_j}{\sqrt{n}} \right|^\nu = \frac{\lambda_n}{(\sqrt{n})^\nu} (\sqrt{n})^\nu \left| \frac{u_j}{\sqrt{n}} \right|^\nu = \frac{\lambda_n}{\sqrt{n}} |u_j|^\nu \rightarrow \lambda_0 |u_j|^\nu.$$

Therefore,  $\lambda_n \sum_{j=1}^p \left( \left| \beta_j + \frac{u_j}{\sqrt{n}} \right|^\nu - |\beta_j|^\nu \right)$  converges uniformly to

$\lambda_0 \sum_{j=1}^p |u_j|^\nu I(\beta_j = 0)$  over compact and convex sets of  $\mathbf{u}$ . Then,

$$V_n(\cdot) \rightarrow_d V(\cdot)$$

on the space of functions topologized by the uniform convergence on compact and convex sets. Since  $V_n$  is not convex, applying the results of Theorem 2.7 on page



198 from Kim and Pollard [20] ,  $\operatorname{argmin}(V_n) \rightarrow_d \operatorname{argmin}(V)$  if  $\operatorname{argmin}(V_n) = O_p(1)$ .

Note that, for all  $\mathbf{u}$  and  $n$  sufficiently large,

$$\begin{aligned} V_n(\mathbf{u}) &\geq -\sum_{i=1}^n \log \frac{l(y_i, \boldsymbol{\beta} + \mathbf{u}/\sqrt{n})}{l(y_i, \boldsymbol{\beta})} - \lambda_n \sum_{j=1}^p \left| \frac{u_j}{\sqrt{n}} \right|^\nu \\ &\geq g(\mathbf{u}) - (\lambda_0 + \delta) \sum_{j=1}^p |u_j|^\nu \\ &= V_n^{(l)}(\mathbf{u}), \end{aligned}$$

where the function  $g$  is a broken straight line function. The last inequality holds because the loglikelihood function of an exponential family is convex and a convex function can be bounded below by a broken straight line function. Since the first component of  $V_n^{(l)}$  is a linear function of  $\mathbf{u}$ , and the second component is just a fraction of  $\mathbf{u}$ , the first component of  $V_n^{(l)}$  grows faster than the second component. Therefore, we have  $\operatorname{argmin}(V_n^{(l)}) = O_p(1)$  so that  $\operatorname{argmin}(V_n) = O_p(1)$ . Because  $\operatorname{argmin}(V)$  is unique with probability 1,  $\operatorname{argmin}(V_n) \rightarrow_d \operatorname{argmin}(V)$ .  $\square$

## Chapter 6

### Simulation Studies and Data Analysis

In this Chapter we present simulation studies conducted to report the performance of the Bayesian criteria including the empirical Bayes criterion (CML), the fully Bayes criterion with flat prior (FBC.Flat) and fully Bayes criteria with conjugate priors (FBC.Conj) developed in Chapter 4. We also include the popular information criteria Cp, AIC and BIC for the purpose of comparison.

In addition to that, a real dataset on South Africa Heart Disease is analyzed using the proposed Bayesian criteria and compared to the results using AIC and BIC.

#### 6.1 Simulation Studies

Simulation studies are carried out for linear models, as well as for Poisson and logistic models. The Bayesian criteria behave differently according to the various models.

In the linear cases, we look at a variety of settings for the coefficients ( $\beta$ ). We examine the same setups as Tibshirani [35] and Yuan and Lin [38] for the correlated

cases to allow comparison with their results. In the case of independent predictors, we investigate three models which represent

1. some coefficients significantly different from zero and others exactly equal to zero;
2. some coefficients significantly different from zero, but one coefficient nearly zero and the others exactly equal to zero;
3. some coefficients significantly different from zero while the other coefficients are all nearly but not exactly zero.

Note that the last scenario was the case for which Leeb [23] pointed out the peculiar conditional distribution of the coefficients estimates ( $\hat{\beta}$ ) based on the selected model. Leeb showed that the sampling distribution of  $\hat{\beta}$  chosen by a consistent variable selection procedure is not at all asymptotically normal and sometimes may not even have a unimodal distribution. The results of these simulations are reported in Section 6.2.

Our simulations represent prediction problems in which jointly distributed observations  $(\mathbf{x}_i, y_i)$ ,  $i = 1, \dots, n$ , are used to find an accurate predictor of a future  $\mathbf{Y}$  from a future  $\mathbf{x}$ . The model is intended to predict  $\mathbf{Y}$  accurately, rather than to estimate each regression coefficient accurately.

All computations were performed in R. We used existing R packages to compute the entire LASSO path. We developed our own code to evaluate the various Bayesian criteria and to perform model selection, as described in Section 4.5.

For the Poisson and logistic cases, we focus on orthogonal predictors. We examine two models, including a model with some significant predictors and others zero, and a model with one nearly zero predictor and some significant predictors and others exactly zero. These results are reported in Sections 6.3 and 6.4.

### 6.1.1 Measures of Various Performance Criteria

To assess the performance of various variable selection criteria, we use multiple ways for calibration. These include the number of times selecting the correct model (# Correct), the number of times selecting a model containing the correct model (# Contained), the average model size (Model Size), the model error for linear models (Model Error), the prediction error for Poisson models (Pred. Error) and the percentage of prediction accuracy for logistic models (Pred. Acc.).

For linear models, we define model error as follows:

$$\text{Model Error} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \mathbf{V} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}),$$

where  $\mathbf{V} = E(\mathbf{X}^T \mathbf{X})$ . The model error combines possible errors in selecting the correct prediction and sampling variation in  $\hat{\boldsymbol{\beta}}$ . An ‘oracle’ would know the correct model. The oracle’s average model error is

$$\text{Oracle Error} = E_{\text{true}}[\text{Model Error}]$$

and involves only sampling error, since the oracle always chooses the true model. From Anderson [3], we have

- (i) if rows of  $\mathbf{X}$  ( $\mathbf{x}_i$ ) are identical and independent  $N_p(\mathbf{0}, \mathbf{V})$ , then  $(\mathbf{X}^T \mathbf{X}) = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T \sim \text{Wishart}(\mathbf{V}, n)$ ;

$$(ii) E[(\mathbf{X}^T \mathbf{X})^{-1}] = \mathbf{V}/(n - p - 1).$$

Using the facts from Anderson [3] stated above, the oracle error is evaluated as follows:

$$\begin{aligned}
E_{\text{true}}[\text{Model Error}] &= E_{\text{true}}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \mathbf{V} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] \\
&= E_{\text{true}}[\text{tr}(\mathbf{V} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T)] \\
&= \text{tr}[\mathbf{V} E_{\text{true}}\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\}] \\
&= \text{tr}[\mathbf{V} E_{\text{true}}\{E((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T | \mathbf{X})\}] \\
&= \text{tr}[\mathbf{V} E_{\text{true}}(\sigma^2 (\mathbf{X}^T \mathbf{X})^{-1})] \\
&= \sigma^2 \text{tr}[\mathbf{V} \mathbf{V}^{-1} / (n - df_{\text{true}} - 1)] \\
&= \sigma^2 df_{\text{true}} / (n - df_{\text{true}} - 1), \tag{6.1}
\end{aligned}$$

where  $df_{\text{true}}$  is the true model size. The oracle errors are computed in the simulations for linear models.

The prediction error for Poisson models is defined to be

$$\text{Pred. Error} = \sum_{i=1}^n (y_i - \exp(\mathbf{x}_i \hat{\boldsymbol{\beta}}))^2.$$

The oracle prediction error (Oracle Pred. Error) is the expected prediction error under the oracle, which cannot be evaluated analytically. Therefore, we compute the Oracle Pred. Error by Monte Carlo methods with 2000 replications in our simulations for Poisson models.

Finally, we use 0.5 as the cutoff point for logistic models. The data point  $y_i$  is classified to be a success if the estimated success probability is greater than or

equal to 0.5; that is,

$$P(y_i = 1|\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i\hat{\boldsymbol{\beta}})}{1 + \exp(\mathbf{x}_i\hat{\boldsymbol{\beta}})} \geq 0.5.$$

This prediction is then compared with the actual observation and the percentage of prediction accuracy is computed for each sample. The percentage of oracle prediction accuracy (Oracle Pred. Acc.) is defined to be the percentage of expected prediction accuracy under the oracle and this quantity cannot be evaluated analytically. In our simulation studies for logistic models, we compute the percentage of Oracle Pred. Acc. by Monte Carlo methods with 2000 replications.

The measure of number of times selecting the models with the most significant predictors ( $\#$  Significant) is also used when the magnitude of some coefficients in the model are significantly greater than the others. Model size is defined to be the number of nonzero parameters in the model. Histograms of the model size by the various variable selection methods are also presented for the models discussed in the simulation studies.

## 6.2 Linear Models

We consider the following models where the columns of  $\mathbf{X} = [x_{ij}]$  are independent and the  $x_{ij}$  are drawn independently from the standard normal distribution.

A new  $\mathbf{X}$  matrix is generated for each Monte Carlo replication.

Model I. The coefficient vector  $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$  and  $\sigma = 3$ .

Model II. The coefficient vector  $\boldsymbol{\beta} = (3, 1.5, 0, 0.05, 2, 0, 0, 0)^T$  and  $\sigma = 3$ .

Model III. The coefficient vector  $\boldsymbol{\beta} = (3, 1.5, 0.01, 0.01, 2, 0.01, 0.01, 0.01)^T$  and  $\sigma =$

3.

Following the setups in Tibshirani [35] and Yuan and Lin [38], four more models are considered:

Model IV. The coefficient vector  $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$  and  $\sigma = 3$ . The correlation between  $x_{ij}$  and  $x_{ik}$  is  $\rho^{|j-k|}$  with  $\rho = 0.5$ .

Model V. The coefficient vector  $\boldsymbol{\beta} = (0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85)^T$ , same covariance structure as in model IV, and  $\sigma = 3$

Model VI. The coefficient vector  $\boldsymbol{\beta} = (5, 0, 0, 0, 0, 0, 0, 0)^T$  with the same covariance structure as in model IV, and  $\sigma = 2$

Model VII. The coefficient vector  $\boldsymbol{\beta} = (2, \dots, 2, 0, \dots, 0)^T$  where these are 2 blocks of 20 repeated coefficients. The  $\mathbf{X}$ s are correlated in such a way that  $x_{ij} = z_{ij} + w_i$ , where the  $z_{ij}$  and  $w_i$  are independent standard normal random variables, and use  $\sigma = 3$ . Hence,  $\rho(x_{ij}, x_{ik}) = 1/2$  for all  $j \neq k$ .

The simulation studies in this section are carried out in R using `lars` [8]. For Models I to VI, simulated samples of 20 are generated in each of 200 Monte Carlo replications. For Model VII, 200 Monte Carlo data sets are generated, each with sample size 100.

The Cp and BIC criteria are compared with the Bayesian criteria. Various priors are assigned to the fully Bayes criteria including a flat prior and informative priors for  $\tau$  and  $q$  both with small means, to explore the sensitivity of the criterion to the various priors. Under the informative prior (denoted FBC\_Conj), a gamma prior is given to  $\tau$  with hyperparameters  $a = 0.01$ ,  $b = 20$  and  $E(\tau) = 0.2$ , and  $q$  has a beta prior with  $\alpha = 1$ ,  $\beta = 10$  and  $E(q) = 1/11$ . For the flat prior (denoted

FBC.Flat and this notation is used throughout the entire Chapter), both  $\tau$  and  $q$  have an uniform distribution. The restricted region for linear model is defined in (4.41) and  $r = 1$ . Tables 6.1 and 6.3 summarize the simulation results. Figures 6.1 to 6.7 show the histograms of the model size for the various models.

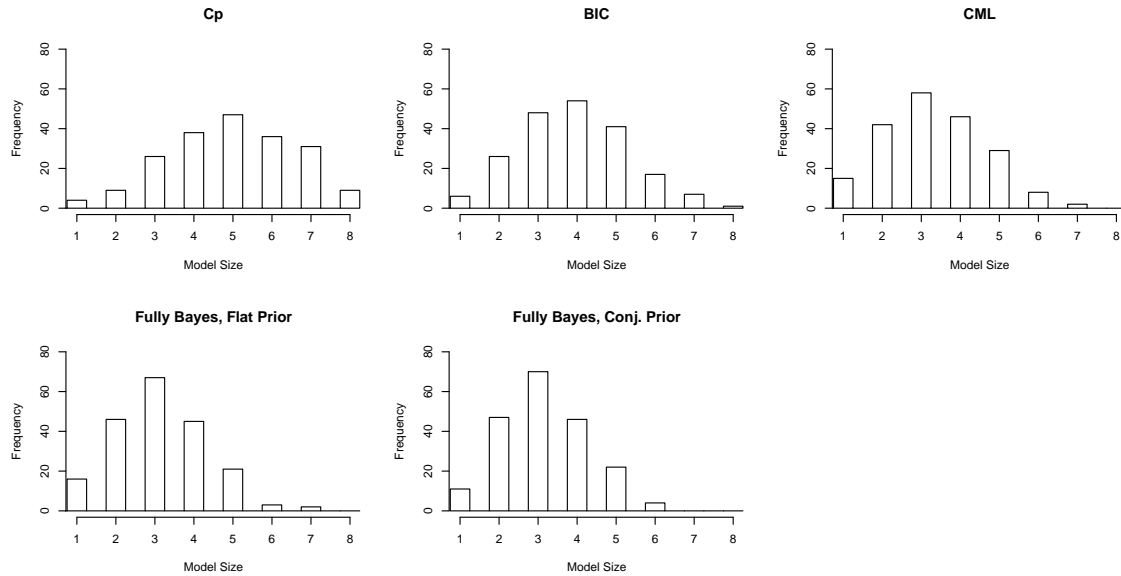


Figure 6.1: Histograms of model size for Model I from linear models based on 200 Monte Carlo replications.

From Table 6.1, the fully Bayes criterion with conjugate prior is more likely than the other criteria to select the correct model. This criterion and BIC produce smaller model errors than the other criteria. Since there are three nonzero predictors, the correct model size should be 3. The average model size for the fully Bayes criteria FBC.Flat and FBC.Conj are 3.13 and 3.17, respectively. CML behaves similarly to the fully Bayes criteria with slightly smaller  $\#$  Correct and model size 3.32. Cp tends to overfit. The histograms (Figure 6.1) shows the mode of the model size is 3



Table 6.1: Simulation Results for Linear Models

	Cp	BIC	CML	FBC_Flat	FBC_Conj
<u>Model I, True Model Size = 3, Oracle Model Error = 1.69</u>					
# Correct	18	34	44	48	50
# Contained	166	143	123	113	116
Model Error*	4.90 (.32)	4.37 (.25)	4.47 (.22)	4.51 (.21)	4.35 (.20)
Model Size*	4.96 (.12)	3.91 (.10)	3.32 (.09)	3.13 (.09)	3.17 (.08)
<u>Model II, True Model Size = 4, Oracle Model Error = 2.25</u>					
# Correct	10	8	10	10	10
# Contained	81	36	31	16	14
# Significant	28	53	60	68	69
Model Error*	4.29 (.24)	3.71 (.19)	3.75 (.20)	3.68 (.20)	3.59 (.18)
Model Size*	4.94 (.12)	3.96 (.10)	3.59 (.10)	3.24 (.08)	3.23 (.07)
<u>Model III, True Model Size = 8, Oracle Model Error = 4.50</u>					
# Correct	10	2	1	0	0
# Contained	10	2	1	0	0
# Significant	22	41	52	70	71
Model Error*	4.42 (.27)	3.72 (.20)	3.63 (.16)	3.48 (.14)	3.49 (.15)
Model Size*	4.78 (.11)	3.90 (.09)	3.56 (.08)	3.32 (.06)	3.29 (.06)

\* Monte Carlo average (Monte Carlo standard error)

Table 6.2: Simulation Results for Linear Models (continued)

	Cp	BIC	CML	FBC_Flat	FBC_Conj
<u>Model IV, True Model Size = 3, Oracle Model Error = 1.69</u>					
# Correct	17	31	37	45	46
# Contained	150	137	129	119	118
Model Error*	4.34 (.29)	3.64 (.19)	3.62 (.17)	3.79 (.17)	3.82 (.18)
Model Size*	4.85 (.11)	4.02 (.09)	3.68 (.07)	3.46 (.08)	3.44 (.07)
<u>Model V, True Model Size = 8, Oracle Model Error = 4.50</u>					
# Correct	8	3	0	0	0
# Contained	8	3	0	0	0
Model Error*	4.38 (.24)	4.15 (.20)	3.97 (.15)	4.54 (.17)	4.55 (.17)
Model Size*	5.74 (.09)	5.35 (.09)	5.04 (.09)	4.59 (.09)	4.57 (.09)

\* Monte Carlo average (Monte Carlo standard error)

Table 6.3: Simulation Results for Linear Models (continued)

	Cp	BIC	CML	FBC_Flat	FBC_Conj
<u>Model VI, True Model Size = 1, Oracle Model Error = 0.25</u>					
# Correct	54	90	0	116	126
# Contained	190	200	200	200	200
Model Error*	1.31 (.11)	0.93 (.07)	0.80 (.05)	0.84 (.06)	0.84 (.06)
Model Size*	3.16 (.14)	2.10 (.09)	2.47 (.06)	1.62 (.06)	1.53 (.06)
<u>Model VII, True Model Size = 20, Oracle Model Error = 2.28</u>					
# Correct	0	0	0	0	0
# Contained	200	199	199	199	199
Model Error*	7.67 (.20)	7.22 (.17)	7.58 (.18)	7.45 (.18)	7.45 (.18)
Model Size*	29.08 (.21)	27.17 (.14)	26.91 (.13)	26.87 (.13)	26.87 (.13)

\* Monte Carlo average (Monte Carlo standard error)

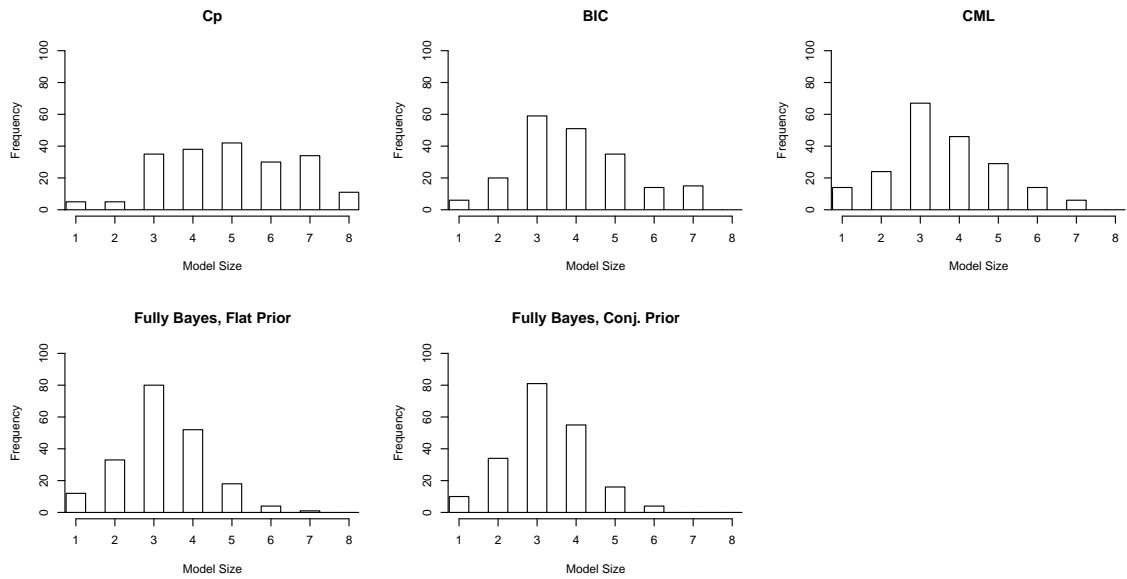


Figure 6.2: Histograms of model size for Model II from linear models based on 200 Monte Carlo replications.

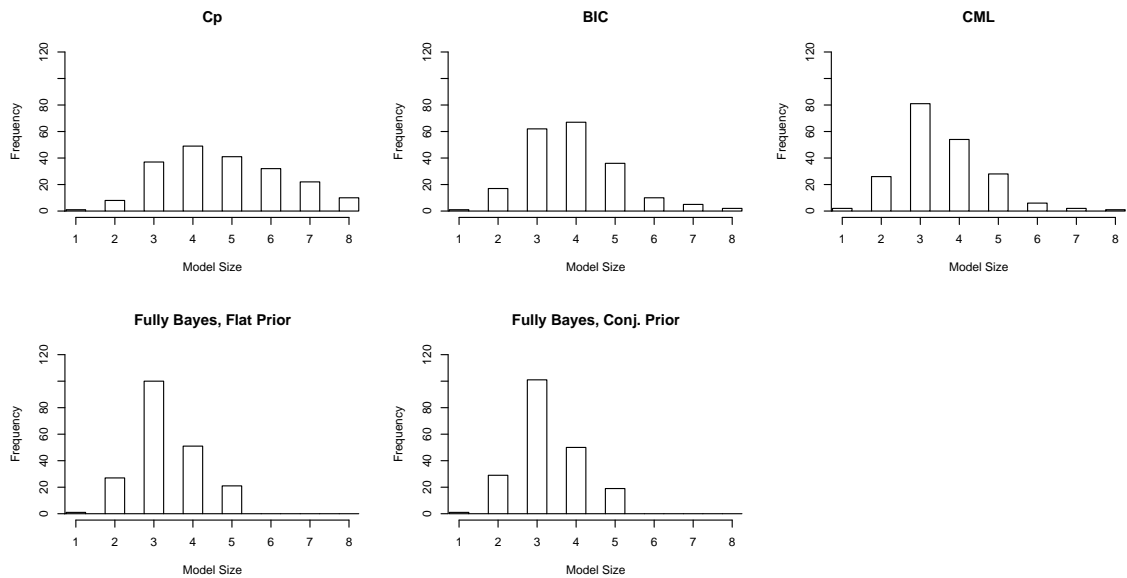


Figure 6.3: Histograms of model size for Model III from linear models based on 200 Monte Carlo replications.

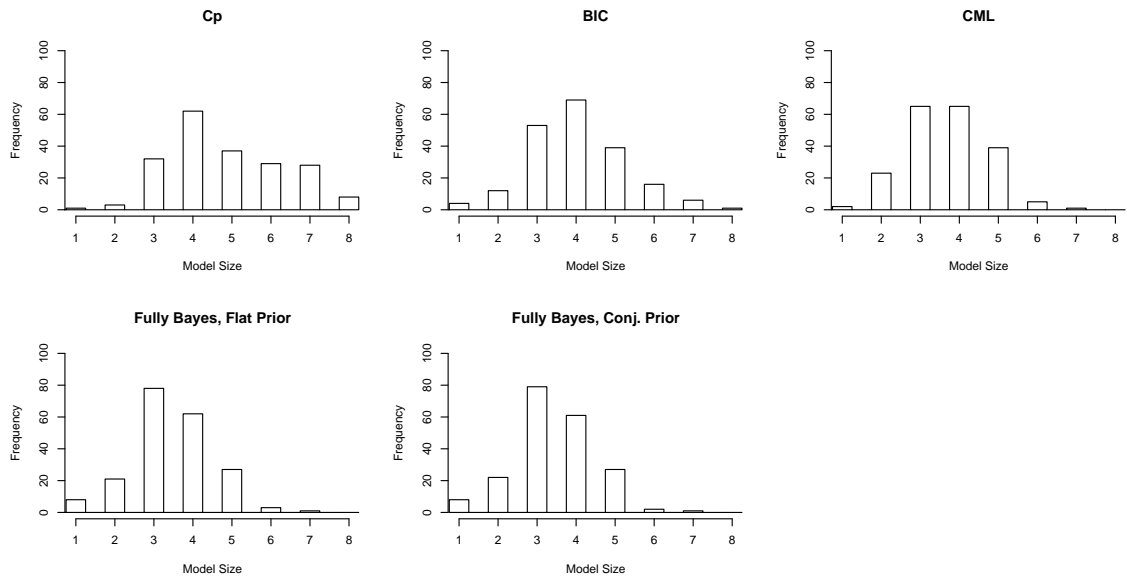


Figure 6.4: Histograms of model size for Model IV from linear models based on 200 Monte Carlo replications.

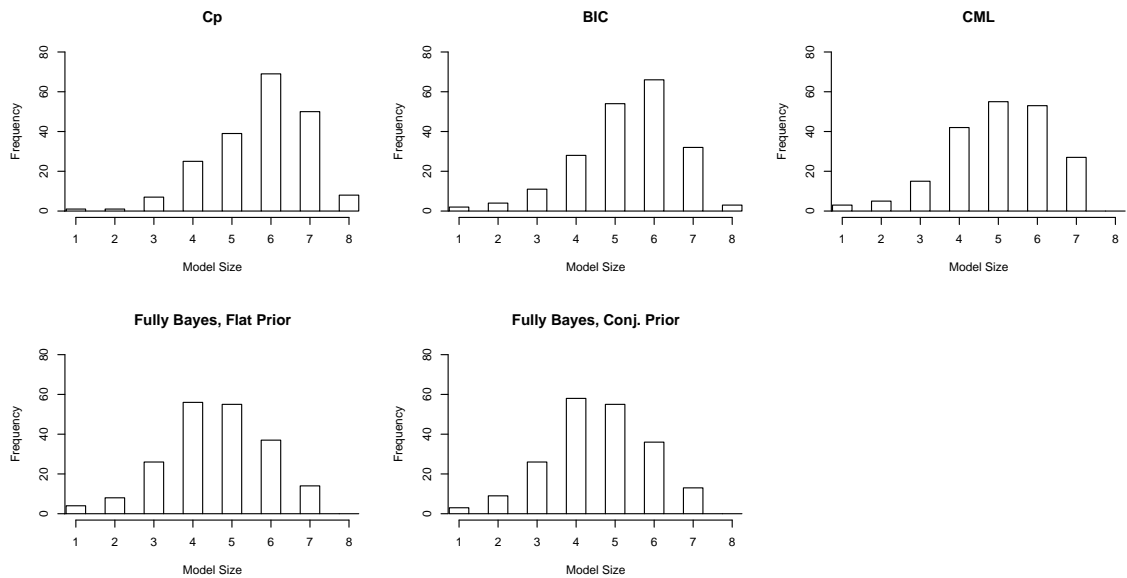


Figure 6.5: Histograms of model size for Model V from linear models based on 200 Monte Carlo replications.

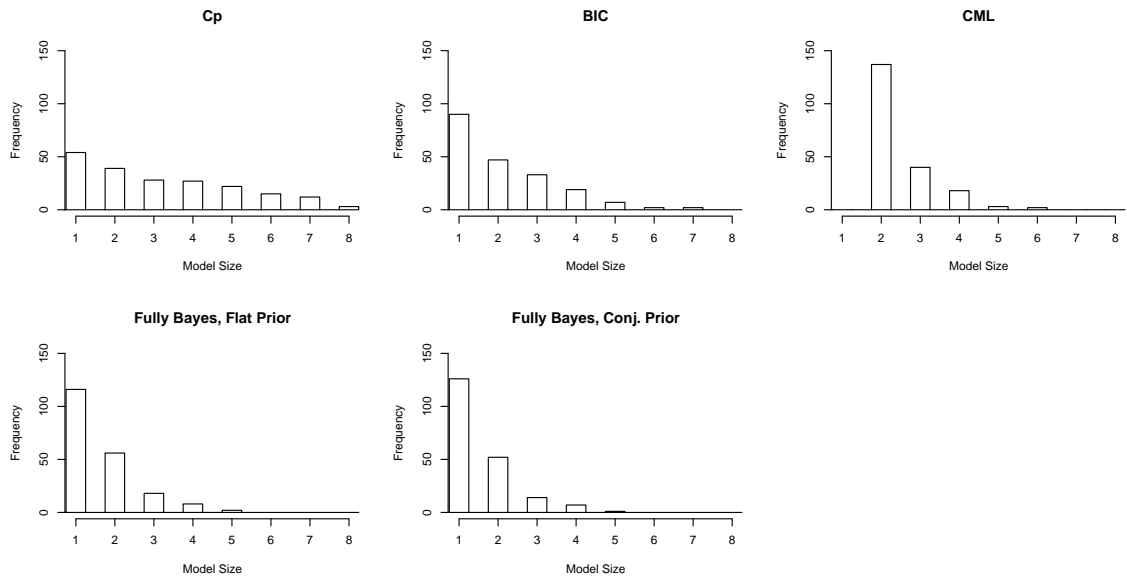


Figure 6.6: Histograms of model size for Model VI from linear models based on 200 Monte Carlo replications.

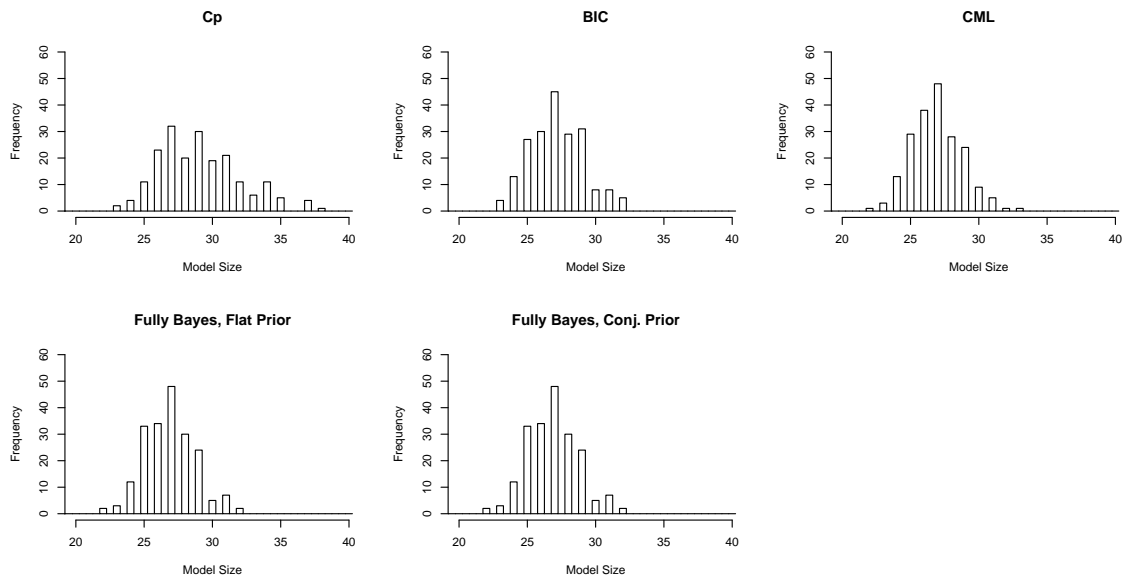


Figure 6.7: Histograms of model size for Model VII from linear models based on 200 Monte Carlo replications.

for CML, FBC\_Flat and FBC\_Conj.

Model II is similar to Model I with a nearly zero predictor added to the model. Since the original three predictors ( $X_1$ ,  $X_2$ ,  $X_5$ ) are much more significant than  $X_4$  in magnitude, the measure  $\#$  Significant is used to see if any of these criteria can detect the significant predictors. Table 6.1 shows that none of the criteria performs well in terms of number of times selecting the correct model. However, all three Bayesian criteria are able to pick up the significant predictors more often than Cp and BIC. The average model size and the histograms (Figure 6.2) show that the Bayesian criteria tend to pick models with three significant predictors while Cp has a tendency to overfit.

Model III is the full model with some of the coefficients very close to zero, while keeping the coefficients of ( $X_1$ ,  $X_2$ ,  $X_5$ ) as in Models I and II. As expected from what Leeb [23] mentioned in his paper, all variable selection criteria perform poorly, even a consistent procedure like BIC. However, the Bayesian criteria detect the significant predictors more frequent by than Cp and BIC. The fully Bayes criteria include the significant predictors 35% of the time and they also outperform other criteria in terms of model error. Figure 6.3 shows that the fully Bayes criteria and CML are likely to pick models with size 3.

Model IV is similar to Model I but with correlated  $\mathbf{X}$ s. The Bayesian criteria outperform the other criteria in terms of number of times picking the correct model and model size. CML performs fairly compared to other criteria, but it does the best for model error. The histograms (Figure 6.4) show that there is a peak for the Bayesian criteria at 3 while the peak for Cp and BIC occurs at 4.

Model V is a full model with small coefficients. All the criteria perform poorly for the measures used as they all try to do some variable selections. The Bayesian criteria tend to pick smaller models than  $C_p$  and BIC. CML performs slightly better than the others in terms of model error.

Model VI is the case with a single significant predictor. The fully Bayes criteria performs substantially better than the other criteria in terms of number of times selecting the correct model and the model size. On the other hand, CML always chooses the correct predictor, but it also adds in an extra unnecessary predictor so the model size is usually 2. Figure 6.6 describes the distribution of the model size for this model.

Finally, Model VII contains 20 nonzero parameters and none of the criteria behaves satisfactorily. While BIC does slightly better than the other criteria in terms of model error, CML, FBC\_Flat and FBC\_Conj are superior in terms of model size. Figure 6.7 shows the distribution of the nonzero parameters are concentrated on the upper 20s for all the criteria.

The results from Models IV to V are similar to the results obtained by Tibshirani [35] and Yuan and Lin [38]. The Bayesian criteria perform variable selection to some extent. For the simulated models discussed above, the fully Bayes criteria usually outperform the other criteria in terms of number of times selecting the correct models. In the linear case, we can see that the behavior of the fully Bayes criteria is not very sensitive to the priors. The results from FBC\_Flat are very similar to the results from FBC\_Conj. CML, FBC\_Flat and FBC\_Conj tend to select the significant predictors more often than  $C_p$  and BIC.  $C_p$  has a tendency of overfit.



## 6.3 Poisson Models

We consider the following models where the columns of  $\mathbf{X} = [x_{ij}]$  are orthogonal and the  $x_{ij}$  are drawn independently from the uniform distribution ranging from  $-1/2$  to  $1/2$ . A new  $\mathbf{X}$  matrix is generated for each Monte Carlo replication. The intercept  $\beta_0 = 0.5$ .

Model I. The coefficient vector  $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$  and  $\sigma_0 = 10$ .

Model II. The coefficient vector  $\boldsymbol{\beta} = (3, 1.5, 0, 0.05, 2, 0, 0, 0)^T$  and  $\sigma_0 = 10$ .

Two hundred datasets were generated, each with sample size 20. The Bayesian criteria are compared with the AIC and BIC criteria. The conjugate prior (FBC\_Conj) for  $\tau$  is a gamma distribution with  $a = 2$ ,  $b = 20$ , and  $q$  has a beta prior with  $\alpha = 2$  and  $\beta = 20$ . The restricted region is defined in (4.9) and  $r = 1$ . The simulation studies in this Section are carried out in R using `glm` [29]. The simulation results are presented in Table 6.4.

When the significant predictor coefficients are largely different from zero as in Model I, FBC\_Conj selects the correct model twice as often as than FBC\_Flat and BIC, and almost four times more often than AIC. FBC\_Conj also outperforms the other criteria in terms of average model size. Histograms of the model size (Figure 6.8) also confirm that the model size of FBC\_Conj is concentrated near 4 (three nonzero parameters and the intercept). However, it does not do very well in terms of prediction error. FBC\_Flat behaves similarly to BIC. CML does not do any variable selection at all as it is inclined to select the nearly full model. Both AIC and BIC overfit.

Table 6.4: Simulation Results for Poisson Models

	AIC	BIC	CML	FBC_Flat	FBC_Conj
<u>Model I, True Model Size = 4, Oracle Pred. Error = 39.97</u>					
# Correct	25	41	0	46	94
# Contained	195	193	180	193	177
Pred. Error*	33.93 (1.27)	41.43 (1.48)	73.89 (11.11)	38.65 (1.33)	60.00 (2.17)
Model Size*	6.20 (.10)	5.58 (.09)	7.55 (.12)	5.65 (.10)	4.53 (.07)
<u>Model II, True Model Size = 5, Oracle Pred. Error = 34.49</u>					
# Correct	6	9	0	5	15
# Contained	104	73	168	75	39
# Significant	15	34	0	39	71
Pred. Error*	34.03 (1.37)	41.47 (1.59)	35.80 (4.25)	38.00 (1.42)	52.42 (1.89)
Model Size*	6.50 (.10)	5.76 (.09)	8.02 (.07)	5.85 (.10)	4.91 (.07)

\* Monte Carlo average (Monte Carlo standard error)

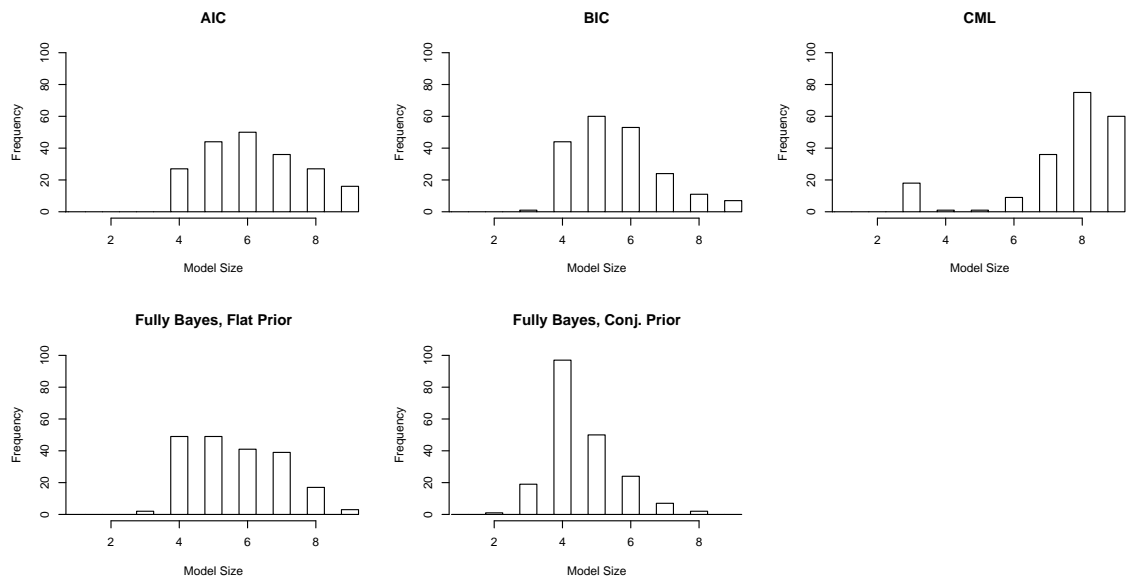


Figure 6.8: Histograms of model size for Poisson Model I based on 200 Monte Carlo replications.

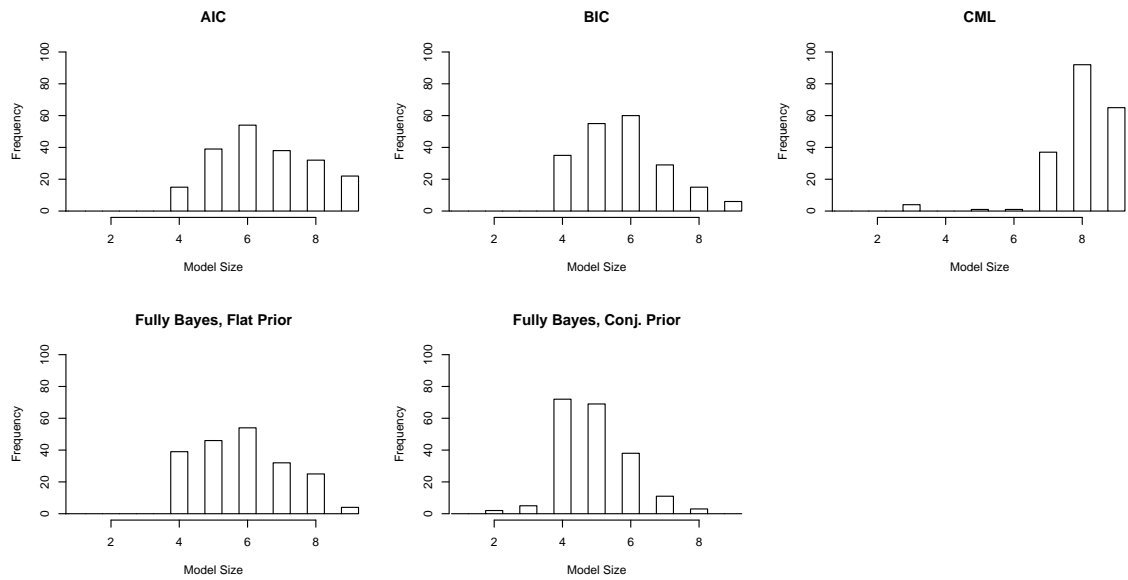


Figure 6.9: Histograms of model size for Poisson Model II based on 200 Monte Carlo replications.

In Model II, there is a predictor much smaller in magnitude than all the other predictors. All the criteria behave poorly in terms of number of times picking the correct model, with FBC\_Conj the best. In addition to that, # Significant shows that the fully Bayes criteria seem to pick out the significant predictors more often than AIC, BIC and CML. The average model size for FBC\_Conj is 4.91, which is closer to the true model size 5 with four nonzero parameters and an intercept compared to other criteria. The results from FBC\_Flat are comparable to the ones from BIC. From the histograms (Figure 6.9), CML again likes to select a bigger model.

The CML criterion does not seem to perform any variable selection in the Poisson models. The performance of FBC\_Flat seems to be comparable to the performance of BIC for the simulated models. FBC\_Conj performs better than other criteria in terms of number of times selecting the correct model and the model size, and it tends to recognize the significant predictors. From the simulation results, the fully Bayes criteria behave differently depending on their priors.

## 6.4 Logistic Models

We consider the following models where the columns of  $\mathbf{X} = [x_{ij}]$  are orthogonal and the  $x_{ij}$  are drawn independently from the standard normal distribution. A new  $\mathbf{X}$  matrix is generated for each Monte Carlo replication. The intercept  $\beta_0 = 0.5$ .

Model I. The coefficients  $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$  and  $\sigma_0 = 10$ .

Model II. The coefficients  $\boldsymbol{\beta} = (3, 1.5, 0, 0.05, 2, 0, 0, 0)^T$  and  $\sigma_0 = 10$ .

Two hundred datasets were generated, each with sample size 100. The Bayesian criteria were compared to the AIC and BIC criteria. The conjugate prior (FBC\_conj) for  $\tau$  is a gamma distribution with  $a = 1.2$ ,  $b = 2$ , and  $q$  has a beta prior with  $\alpha = 1.2$  and  $\beta = 20$ . The restricted region is defined in (4.9) and  $r = 1$ . The simulation study in this section is carried out in R using `glmnet` [13]. Table 6.5 summarizes the simulation results.

Table 6.5: Simulation Results for Logistic Models

	AIC	BIC	CML	FBC_Flat	FBC_Conj
<u>Model I, True Model Size = 4, Oracle Pred. Acc. (%) = 87.71</u>					
# Correct	25	118	179	49	100
# Contained	200	200	199	200	200
Pred. Acc.* (%)	88.76 (.24)	88.04 (.25)	87.41 (.23)	88.46 (.23)	88.07 (.23)
Model Size*	6.20 (.10)	4.68 (.07)	4.10 (.02)	5.41 (.08)	4.77 (.07)
<u>Model II, True Model Size = 5, Oracle Pred. Acc. (%) = 87.97</u>					
# Correct	8	8	2	8	7
# Contained	200	200	195	200	200
# Significant	46	134	188	72	129
Pred. Acc.* (%)	88.23 (.24)	87.23 (.24)	87.04 (.24)	87.93 (.23)	87.46 (.23)
Model Size *	5.77 (.10)	4.43 (.05)	4.01 (.02)	5.05 (.07)	4.48 (.05)

\* Monte Carlo average (Monte Carlo standard error)

From Table 6.5, the prediction accuracy for all criteria sits around 88%, but

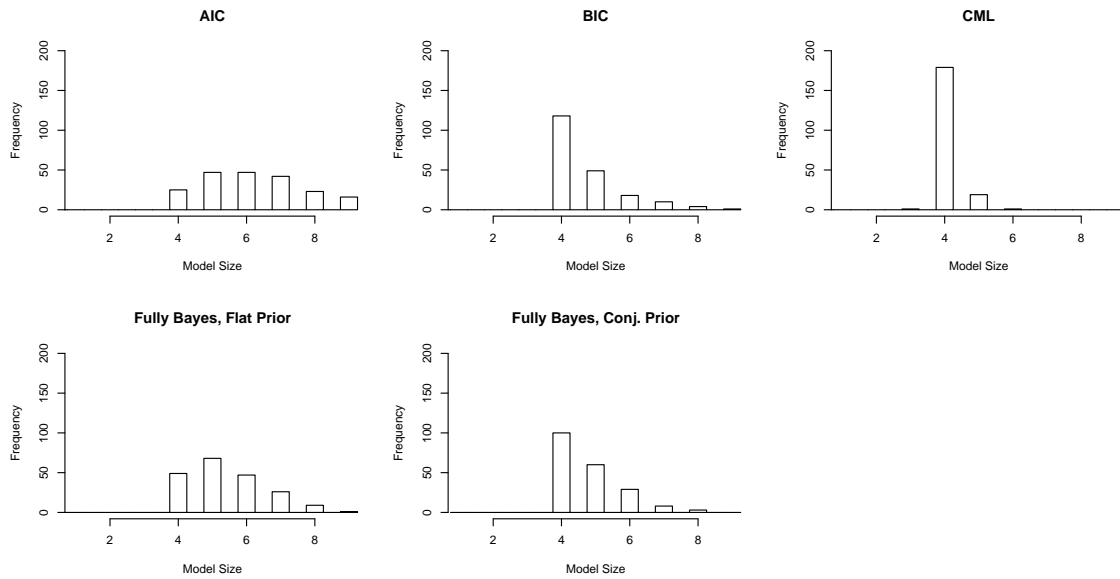


Figure 6.10: Histograms of model size for Logistic Model I based on 200 Monte Carlo replications.

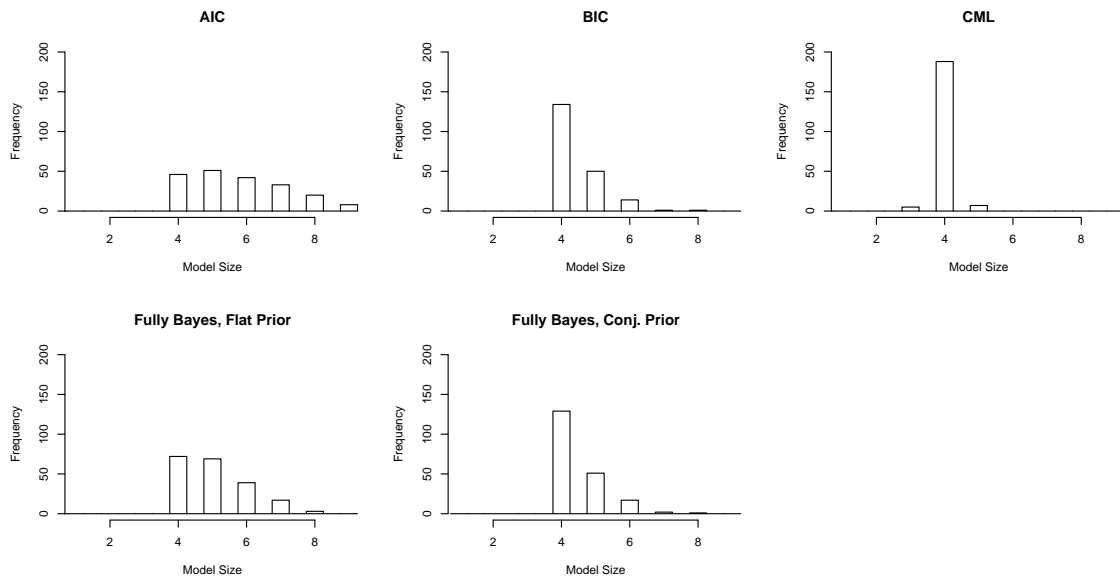


Figure 6.11: Histograms of model size for Logistic Model II based on 200 Monte Carlo replications.

CML performs exceptionally well for Model I. It selects the correct model 90% of times and the model size is very close to the true model size 4 including three nonzero coefficients and the intercept. FBC\_Flat tends to choose a bigger model and it does not do as well as FBC\_Conj. FBC\_Conj behaves comparably to BIC. The histograms in Figure 6.10 clearly shows that the model size of CML highly concentrates around 4.

For Model II, none of the criteria is able to detect the small coefficient of  $X_4$  so none of them do well in terms of picking the correct model. However, CML is very efficient in selecting the significant predictors, which is shown in the measure of % Significant and in the histograms (Figure 6.11). Similar to Model I, the behavior of FBC\_Conj is comparable to that of BIC as shown in Table 6.5 and in Figure 6.11. FBC\_Flat tends to overfit but it is able to detect the significant predictors more often than AIC. The prediction accuracy for all criteria falls around 87%.

In the logistic case, CML performs remarkably well when the predictors are very significant as in Models I and II. FBC\_Conj and BIC behave similarly for the simulated models. FBC\_Flat tend to select larger models but not as big as AIC. From the simulation results, it is obvious that the behavior of fully Bayes criterion is sensitive to their priors.

## 6.5 Summary of Simulation Results

The Bayesian criteria behavior differs from linear, Poisson and logistic models. While CML performs exceptionally well in logistic cases and is terrible in Poisson

cases, CML performs variable selection to some extent in linear cases. The fully Bayes criteria performance are sensitive to their chosen priors for Poisson and logistic models, but it does not seem to be true in the linear cases. The behavior of FBC\_Conj and BIC is similar in the logistic cases, while the behavior of FBC\_Flat is more comparable to BIC in the Poisson cases.

## 6.6 South Africa Heart Disease Data Analysis

This heart disease dataset, used previously in Hastie, Tibshirani and Friedman [16] and Park and Hastie [29], consists of nine feature attributes from the medical record of 432 males in a heart disease high-risk region of the Western Cape, South Africa. The responses are binary with the value one indicating the presence of coronary heart disease. The nine feature attributes are systolic blood pressure (sbp), tobacco (cumulative tobacco usage in kilograms), low density lipoprotein cholesterol (ldl), adiposity, family history of heart disease (famhist) with the value one indicating the presence of family history of heart disease, type-A behavior (typea), obesity, current alcohol consumption (alcohol) and age at onset (age).

Logistic regression is used for the heart disease data since the response variable is binary. The three Bayesian criteria (CML, FBC\_Flat and FBC\_Conj) and two information criteria (AIC and BIC) are applied to select the explanatory variables (features) related to heart disease. FBC\_Conj employs the same priors for  $\tau$  and  $q$  as in the logistic models in Section 6.4 and the restricted region is defined in (4.9) with  $r = 10$ . We use  $\sigma_0 = 15$  for this dataset. The analysis is performed in R using



`glmnet` [13] and the results are shown on the second column of Tables 6.6 to 6.10. CML is clearly the most aggressive criterion as it tends to select the least number of predictors including tobacco, famhist and age. These features are shared by all the other criteria. BIC selects two extra feature variables which are ldl and typea. The model chosen by FBC\_Conj differs from BIC only by the feature sbp. Finally, both FBC\_Flat and AIC favor the six-feature model including obesity and all the predictors that FBC\_Conj selects.

Using `glmnet` by Friedman, Hastie and Tibshirani [13], the entire solution path is plotted in Figure 6.12 except famhist for graphical display purpose. However, it is always included in the model as indicated by all the above criteria. The horizontal axis is the lambda scaled by the maximum lambda from all the steps of the solution path provided by `glmnet` and the vertical axis represents the values of the estimated coefficients. The vertical line are the various models chosen by the three Bayesian criteria, as well as AIC and BIC.

Bootstrap analysis (Efron and Tibshirani [9]) is used to validate the coefficient estimates selected by the above variable selection criteria. We generate 1000 bootstrap samples. For each sample, we fit a logistic regression path by `glmnet` and record the coefficient estimates chosen by the above mentioned variable selection criteria. The estimated bootstrap means ( $\text{Mean}(\hat{\beta}^b)$ ) and standard errors ( $\text{SE}(\hat{\beta}^b)$ ) are then computed from the 1000 bootstrap samples for each selection criterion. Tables 6.6 to 6.10 summarize the results for the variable selection criteria AIC, BIC, CML, FBC\_Flat and FBC\_Conj respectively. The second column represents the coefficient estimate ( $\hat{\beta}^b$ ) for each feature obtained by the particular selection criterion.

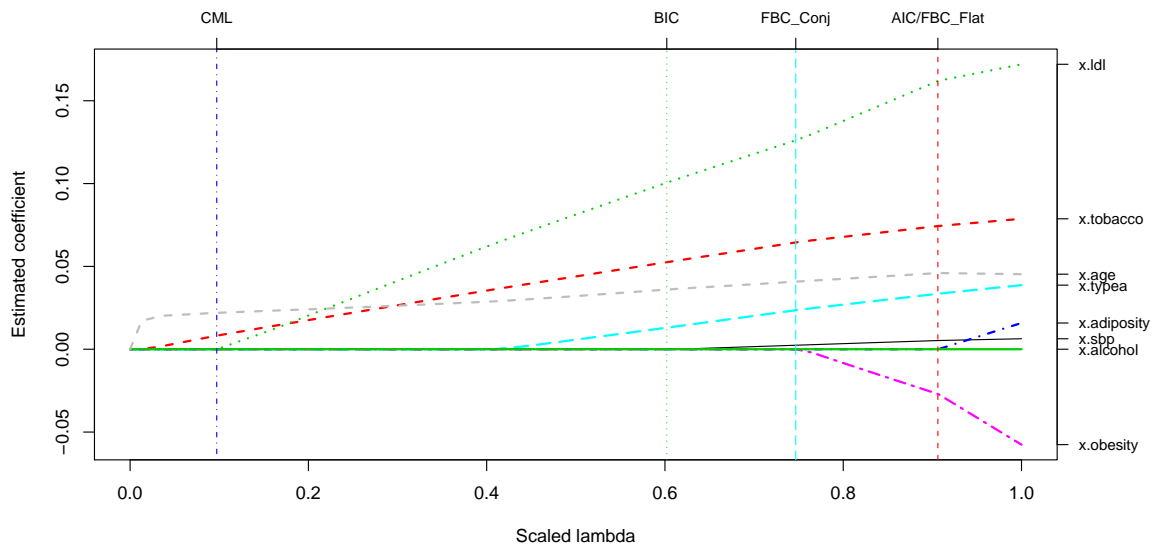


Figure 6.12: Solution path for South Africa Heart Disease data by glmnet (except the famhist feature). The horizontal axis is the lambda scaled by the maximum lambda from all the steps of the solution path provided by glmnet and the vertical axis represents the values of the estimated coefficients. The vertical lines are the various models chosen by the three Bayesian criteria, as well as AIC and BIC.

Mean( $\hat{\beta}^b$ ) and SE( $\hat{\beta}^b$ ) are given on the third and fourth column and the last column is the number of times that the coefficient is nonzero ( $\#$  Nonzero) out of the 1000 bootstrap samples. For the coefficients of the predictors estimated at zero, fewer than 40% of the bootstrap estimates are nonzeros. The only exception is ldl by CML with 49% of the bootstrap samples nonzero. However, the  $t$  statistics shows this predictor is not significant.

Table 6.6: Bootstrap results for South Africa Heart Disease data using AIC. The coefficient estimates ( $\hat{\beta}^b$ ) are obtained by AIC.

Feature	$\hat{\beta}$	Mean( $\hat{\beta}^b$ )	SE( $\hat{\beta}^b$ )	$\#$ Nonzero
sbp	0.0053	0.0050	0.0051	661
tobacco	0.0744	0.0734	0.0284	991
ldl	0.1619	0.1530	0.0622	987
adiposity	0	0.0099	0.0224	316
famhist	0.8573	0.8361	0.2423	999
typea	0.0335	0.0324	0.0141	969
obesity	-0.0270	-0.0359	0.0465	530
alcohol	0	0.0002	0.0037	411
age	0.0460	0.0434	0.0106	1000

The estimated bootstrap means (Mean( $\hat{\beta}^b$ )) and standard errors (SE( $\hat{\beta}^b$ )) are based on 1000 bootstrap samples. The last column are the number of times that the coefficient is nonzero ( $\#$  Nonzero) out of the 1000 samples.

Figures 6.13 to 6.17 show the bootstrap distribution of the coefficient estimates

Table 6.7: Bootstrap results for South Africa Heart Disease data using BIC. The coefficient estimates ( $\hat{\beta}^b$ ) are obtained by BIC.

Feature	$\hat{\beta}$	Mean( $\hat{\beta}^b$ )	SE( $\hat{\beta}^b$ )	# Nonzero
sbp	0	0.0023	0.0041	284
tobacco	0.0526	0.0559	0.0285	922
ldl	0.1005	0.1079	0.0639	868
adiposity	0	0.0015	0.0066	96
famhist	0.5986	0.6325	0.2813	940
typea	0.0130	0.0176	0.0153	663
obesity	0	-0.0047	0.0176	85
alcohol	0	0.0002	0.0017	63
age	0.0360	0.0369	0.0099	999

The estimated bootstrap means (Mean( $\hat{\beta}^b$ )) and standard errors (SE( $\hat{\beta}^b$ )) are based on 1000 bootstrap samples. The last column are the number of times that the coefficient is nonzero (# Nonzero) out of the 1000 samples.

Table 6.8: Bootstrap results for South Africa Heart Disease data using CML. The coefficient estimates ( $\hat{\beta}^b$ ) are obtained by CML.

Feature	$\hat{\beta}$	Mean( $\hat{\beta}^b$ )	SE( $\hat{\beta}^b$ )	# Nonzero
sbp	0	0	0.0005	9
tobacco	0.0082	0.0223	0.0246	559
ldl	0	0.0337	0.0445	488
adiposity	0	0.0002	0.0020	14
famhist	0.0991	0.2381	0.2108	842
typea	0	0.0014	0.0048	100
obesity	0	-0.0001	0.0021	3
alcohol	0	0	0	0
age	0.0220	0.0239	0.0081	993

The estimated bootstrap means (Mean( $\hat{\beta}^b$ )) and standard errors (SE( $\hat{\beta}^b$ )) are based on 1000 bootstrap samples. The last column are the number of times that the coefficient is nonzero (# Nonzero) out of the 1000 samples.

Table 6.9: Bootstrap results for South Africa Heart Disease data using FBC\_Flat.

The coefficient estimates ( $\hat{\beta}^b$ ) are obtained by FBC\_Flat.

Feature	$\hat{\beta}$	Mean( $\hat{\beta}^b$ )	SE( $\hat{\beta}^b$ )	# Nonzero
sbp	0.0053	0.0045	0.0049	619
tobacco	0.0744	0.0712	0.0277	990
ldl	0.1619	0.1468	0.0604	985
adiposity	0	0.0071	0.0183	254
famhist	0.8573	0.8092	0.2410	999
typea	0.0335	0.0302	0.0138	961
obesity	-0.0270	-0.0278	0.0411	448
alcohol	0	0.0003	0.0033	341
age	0.0460	0.0429	0.0104	1000

The estimated bootstrap means (Mean( $\hat{\beta}^b$ )) and standard errors (SE( $\hat{\beta}^b$ )) are based on 1000 bootstrap samples. The last column are the number of times that the coefficient is nonzero (# Nonzero) out of the 1000 samples.

Table 6.10: Bootstrap results for South Africa Heart Disease data using FBC\_Conj.

The coefficient estimates ( $\hat{\beta}^b$ ) are obtained by FBC\_Conj.

Feature	$\hat{\beta}$	Mean( $\hat{\beta}^b$ )	SE( $\hat{\beta}^b$ )	# Nonzero
sbp	0.0024	0.0043	0.0048	590
tobacco	0.0645	0.0698	0.0273	989
ldl	0.1261	0.1431	0.0597	982
adiposity	0	0.0059	0.0167	222
famhist	0.7359	0.7935	0.2396	999
typea	0.0235	0.0290	0.0140	947
obesity	0	-0.0239	0.0379	403
alcohol	0	0.0003	0.0031	305
age	0.0408	0.0425	0.0103	1000

The estimated bootstrap means (Mean( $\hat{\beta}^b$ )) and standard errors (SE( $\hat{\beta}^b$ )) are based on 1000 bootstrap samples. The last column are the number of times that the coefficient is nonzero (# Nonzero) out of the 1000 samples.

chosen by AIC, BIC, CML, FBC\_Flat and FBC\_Conj respectively assuming that the original data are randomly re-sampled with replacement from the population. The red vertical bars represent  $\hat{\beta}$  selected by the particular selection criterion from the whole data and the blue thick bars are the frequencies of zero coefficients. As indicated in the histograms for each selection criterion, the red bar situates near the center of the bootstrap distribution for predictors whose coefficient estimates are nonzero. For the coefficients of the predictors estimated at zero, the histograms have a peak exactly at zero.

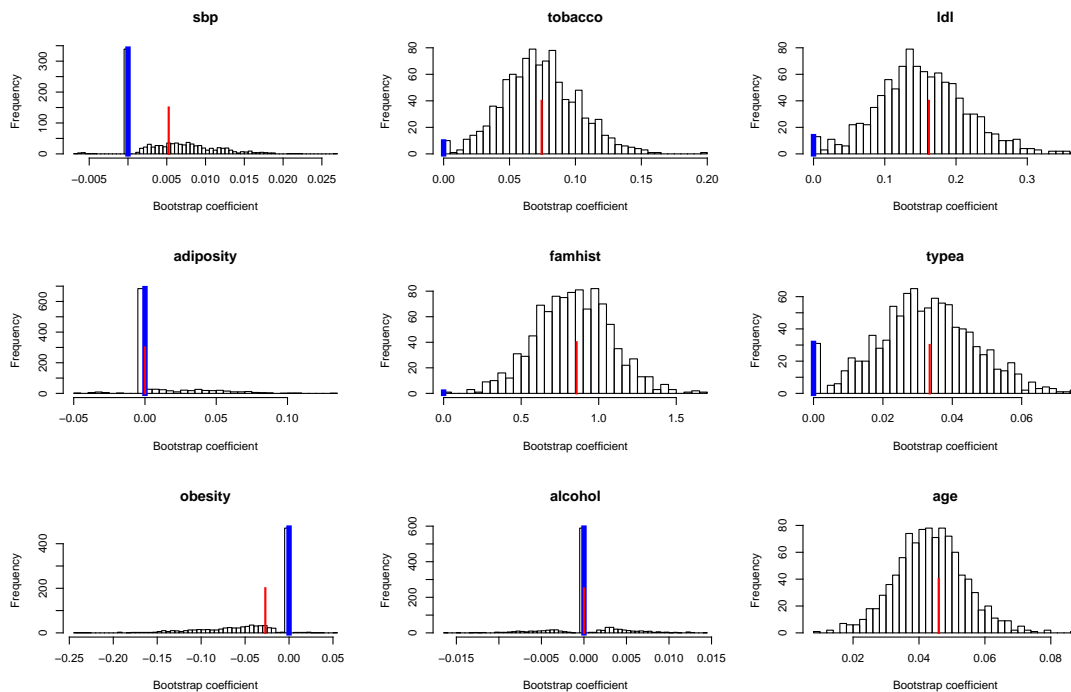


Figure 6.13: The bootstrap distribution of the coefficient estimates chosen by AIC. The red vertical bars represent  $\hat{\beta}$  selected by AIC from the whole data and the blue thick bars are the frequencies of zero coefficients.

The bootstrap results confirm the coefficient estimates selected by various



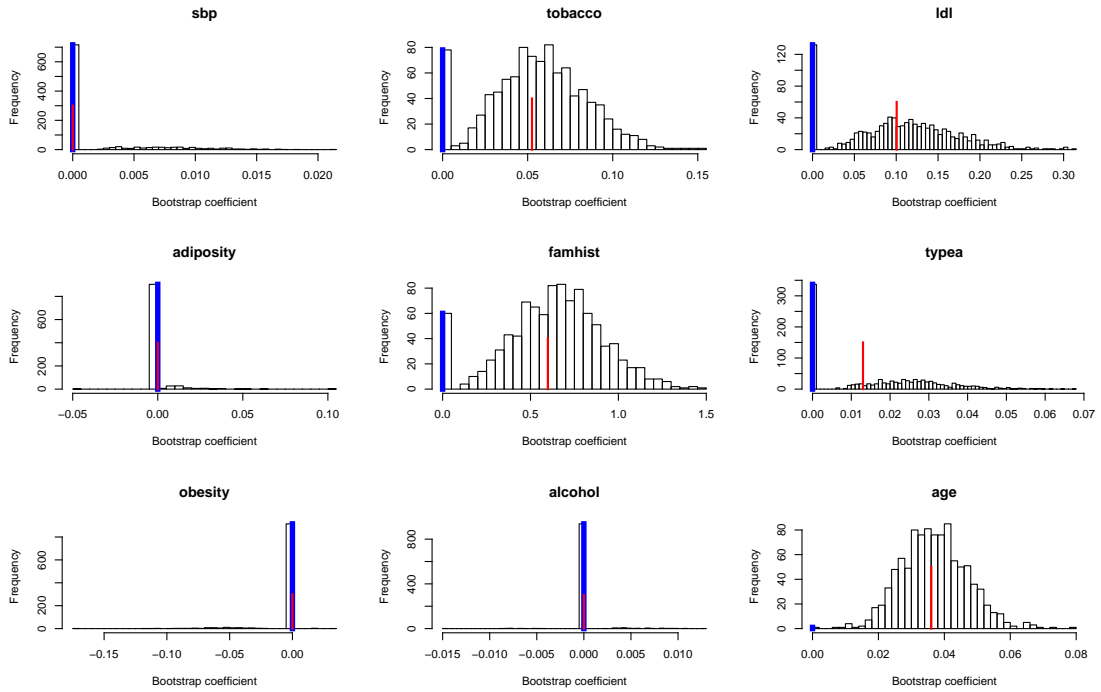


Figure 6.14: The bootstrap distribution of the coefficient estimates chosen by BIC. The red vertical bars represent  $\hat{\beta}$  selected by BIC from the whole data and the blue thick bars are the frequencies of zero coefficients.

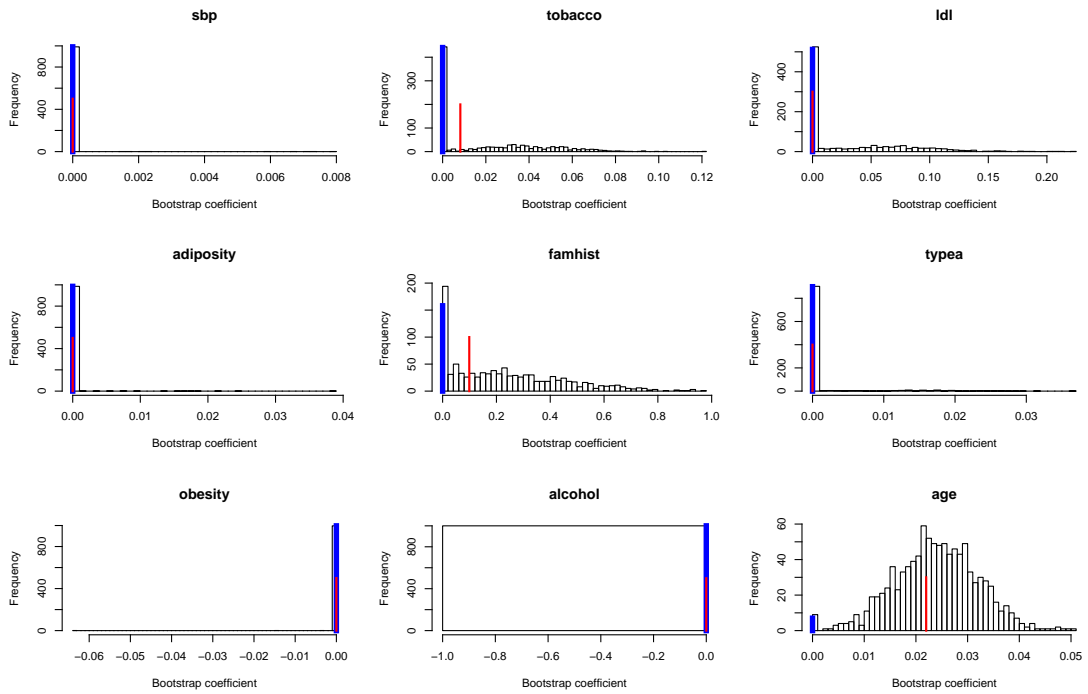


Figure 6.15: The bootstrap distribution of the coefficient estimates chosen by CML.

The red vertical bars represent  $\hat{\beta}$  selected by CML from the whole data and the blue thick bars are the frequencies of zero coefficients.

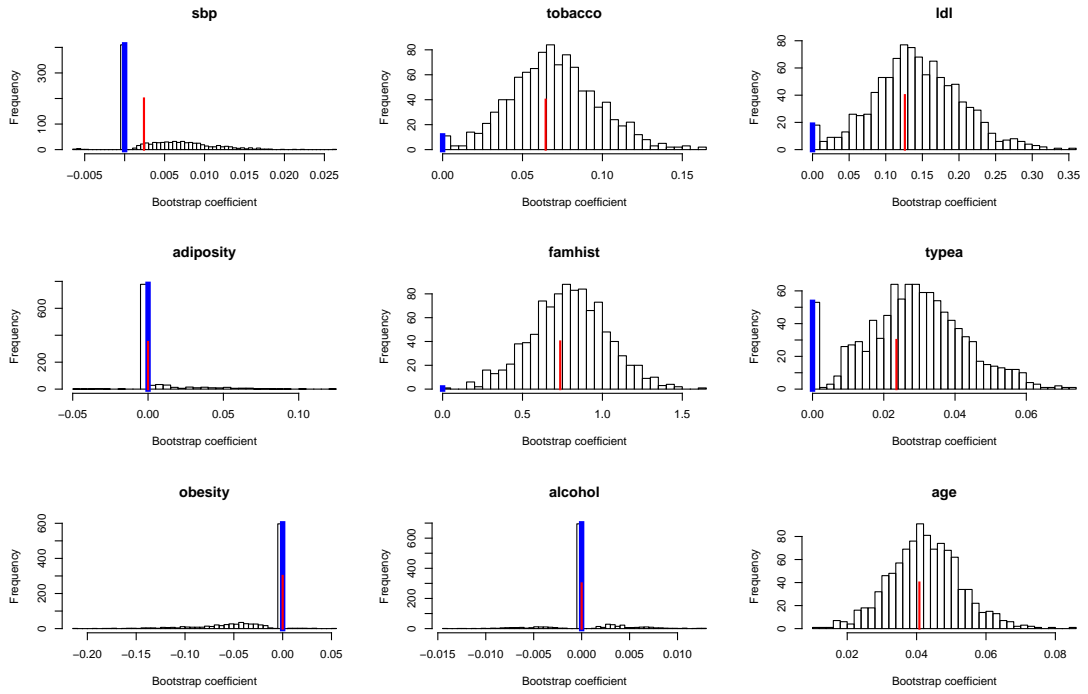


Figure 6.16: The bootstrap distribution of the coefficient estimates chosen by FBC.Flat. The red vertical bars represent  $\hat{\beta}$  selected by FBC.Flat from the whole data and the blue thick bars are the frequencies of zero coefficients.

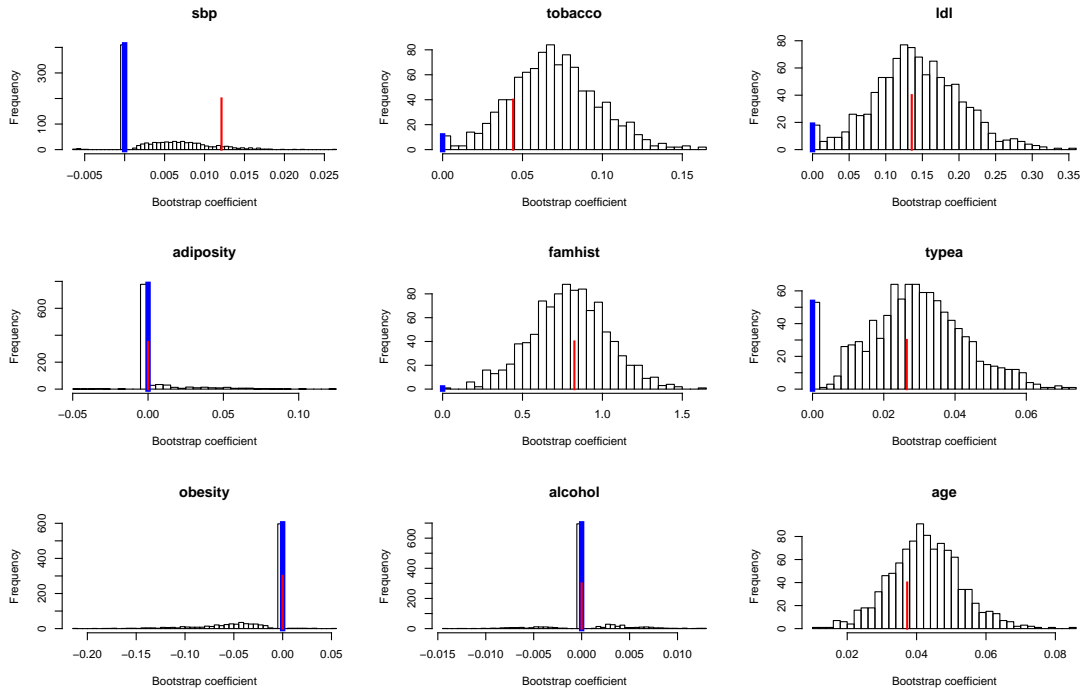


Figure 6.17: The bootstrap distribution of the coefficient estimates chosen by FBC\_Conj. The red vertical bars represent  $\hat{\beta}$  selected by FBC\_Conj from the whole data and the blue thick bars are the frequencies of zero coefficients.

variable selection criteria. The FBC.Flat and AIC tend to pick a bigger model compared to other criterion. Even though the behavior of FBC.Conj is similar to the BIC's in the simulation studies for the models with independent columns of  $\mathbf{X}$ s (Section 6.4), it is not the case for this correlated heart disease data. The model selected by FBC.Conj has one additional predictor, sbp, as opposed to the one chosen by BIC.

In an unpublished preliminary version of Park and Hastie [29], the authors also performed a bootstrap analysis on the same heart disease dataset using `glmpath`. They employed BIC as the selection criterion to choose the model. Their BIC results obtained by `glmpath` are different from my BIC results using `glmnet`. The `glmpath` and `glmnet` algorithms choose the value of the regularization parameter  $\lambda$  differently and consequently select a different model. However, the model picked by my FBC.Conj is the same as that chosen by Park and Hastie's BIC criterion. Their estimated coefficients and my estimated coefficients are very close, allowing the numerical discrepancy between `glmpath` and `glmnet`. Furthermore, the histograms obtained by my FBC.Conj and Park and Hastie's BIC possess the same shape and resemble each other.

## Chapter 7

### Summary and Future Research

#### 7.1 Summary

In this dissertation, we have proposed a hierarchical Bayesian formulation in GLMs and developed two Bayesian variable selection procedures related to LASSO. By specifying a double exponential prior for the covariate coefficients and a special prior for each candidate model, we have shown that the posterior distribution of the candidate model given the data is closely related to LASSO which has the property of shrinking some coefficient estimates to zero, and hence allows one to perform variable selection.

Since the selected model will be the one with maximum posterior probability, using a logistic regression as an illustration for the GLM, we evaluated the posterior probability both for the regular and nonregular classes. We have also shown that the posterior probability for the nonregular class is dominated by its regular class counterpart. Therefore, the search for the model with maximum posterior probability can be strictly confined in the regular class.

We derived an empirical Bayes (CML) and a fully Bayes criterion under the Bayesian formulation for variable selection in GLMs. The fully Bayes criterion, FBC\_Conj, employs a conjugate gamma prior for  $\tau$  with hyperparameters  $a$  and  $b$  and a conjugate beta prior for  $q$  with hyperparameters  $\alpha$  and  $\beta$ . The fully Bayes criterion flat prior, FBC\_Flat, is a special case of FBC\_Conj with  $a = 1$ ,  $b = +\infty$ ,  $\alpha = 1$  and  $\beta = 1$ . We have devised the CML, FBC\_Flat and FBC\_Conj criteria for logistic and Poisson models, as well as the fully Bayes criterion for the linear model.

The asymptotic behavior of the Bridge estimators in GLMs has been explored. We have proved that under regularity conditions on the design, if  $\lambda_n/n$  goes to zero, the estimator is consistent. Furthermore, we are able to characterize the estimation error. We also derive the limiting distribution of  $\sqrt{n}$  times the estimation error for all Bridge estimators in GLMs.

The performance of the Bayesian variable selection criteria has been studied by simulations and a real data analysis. The Bayesian variable selection criteria are also compared to the popular information criteria, Cp, AIC and BIC and they behave very differently in linear, Poisson and logistic models. For logistic models, the performance of CML is very impressive but it almost does not do any variable selection in Poisson cases. The CML performance in linear case is somewhere in between. In the presence of a predictor coefficient nearly zero and some significant predictors, CML picks the significant predictors most of the time in the logistic case and fairly often in linear case, while FBC\_Conj tends to select the significant predictors equally well in all linear, Poisson and logistic models. The behavior of fully Bayes criteria depends strongly on their chosen priors for Poisson and logistic

cases. However, such a distinction is not obvious in linear case. From the simulation studies, the Bayesian criteria are generally more likely than  $C_p$  and AIC to choose correct predictors. The real data analysis also showed that CML tends to pick a smaller model in logistic case and this agrees with our simulation results.

## 7.2 Future Research

There are a few issues to be investigated as extensions of my work:

### 7.2.1 Priors Specification

The double-exponential prior for the coefficients is chosen because the resulting posterior distribution of candidate model given data is closely related to LASSO. However, the special priors for candidate models are selected mainly due to computational and analytical convenience. The simulation and real data analysis results show that the fully Bayes criteria are sensitive to their prior. This prompts us to think if there are other ways to specify the priors and hyperpriors so that the behavior of these Bayesian criteria will have better model selection performance, such that they will penalize heavily for bigger models. Moreover, we would like to explore the possibility of finding priors that are flexible enough to incorporate the expert information such as the preference that certain variables must be included in the model, as well as the adjustment of the priors if some variables are highly correlated to each other.



## 7.2.2 Bayesian Model Averaging

In practice, model selection is usually based on some criteria and hence a single representable model that captures the essential information of the data is chosen. One criticism is that the selected model may be unstable as it does not properly account for model uncertainty. The Bayesian criteria from my work take model uncertainty into account as there is a prior distribution for the candidate models. Another approach to handle the model uncertainty is Bayesian Model Averaging where each candidate model is associated with a prior probability and prior distributions are chosen for the parameters within each model. Using the Bayesian mechanism, the posterior probabilities of the models can be computed and the result is a weighted combination of some models. Refer to Raftery, Madigan and Hoeting [31] and Hoeting, Madigan and Raftery [18] for details in Bayesian model averaging. We would like to investigate application of Bayesian model averaging as extension of this work to increase model stability and prediction accuracy.

## 7.2.3 Model Selections in GLMs with Noncanonical Link

For the Bayesian variable selection criteria in GLMs proposed in this dissertation, only canonical link is considered. A natural extension is to explore the use of noncanonical link in the GLMs. Noncanonical link is common in practice but it also requires more technicalities in computing the posterior distribution. In conjunction to that, will the connection of posterior distribution to LASSO preserved to allow model selection? These are the issues waiting to be addressed.

## BIBLIOGRAPHY

- [1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *Proceedings of International Symposium on Information Theory*, Ed. B. N. Petrov and F. Csaki, 267-281, Budapest: Akademia Kiado.
- [2] Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study, *Annals of Statistics*, **10**, 1100-1120.
- [3] Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*, second edition, New York: Wiley.
- [4] Berger, J. O. and Pericchi, L. R. (2001). Objective Bayesian Methods for Model Selection: Introduction and Comparison. *Lecture Notes-Monograph Series, Model Selection*, **38**, 135-207.
- [5] Berger, J. O. (2006). The Case for Objective Bayesian Analysis. *Bayesian Analysis*, **1**, 385-402.
- [6] Clyde, M. A. (1999). Bayesian model averaging and model search strategies. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), *Bayesian Statistics*, **6** (pp. 157-185). Oxford: University Press.
- [7] Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2002). On bayesian model and variable selection using mcmc. *Statistics and Computing*, **12**, 27-36.
- [8] Efron, B., Johnston, I., Hastie, T. and Tibshirani, R. (2004). Least angle regression, *Annals of Statistics*, **32**, 407-499.
- [9] Efron, B., and Tibshirani, R. (1993). *An introduction to the Bootstrap*, Boca Raton: Chapman and Hall.
- [10] Fan, J. and Li, R. (2001). Variable selection via noncave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*, **96**, 1348-1360.

- [11] Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression, *Annals of Statistics*, **22**, 1947-1975.
- [12] Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools (with discussion), *Technometrics*, **35**, 109-148.
- [13] Friedman, J., Hastie T. and Tibshirani, R. (2008) Regularized Paths for Generalized Linear Models via Coordinate Descent Manuscript, Department of Statistics, Stanford University. Available at <http://www-stat.stanford.edu/hastie/Papers/glmnet.pdf>.
- [14] George, E. I. and Foster, D. P. (2000). Calibration and empirical Bayes variable selection, *Biometrika*, **87**, 731-747.
- [15] Geyer, C. J. (1996). On the asymptotics of convex stochastic optimization. Unpublished manuscript.
- [16] Hastie, T., Tibshirani R. and Friedman J. (2002). *The Elements of Statistical Learning; Data Mining, Inference, and Prediction*, New York: Springer-Verlag.
- [17] Hansen, M. and Yu, B. (2001). Model selection and the principle of minimum description length, *Journal of the American Statistical Association*, **96**, 746-774.
- [18] Hoeting, J., Madigan, D., and Raftery, A. E. (1999). Bayesian model averaging: A tutorial (with discussion), *Statistical Science*, **14**, 382-417. Corrected version available at <http://www.stat.washington.edu/www/research/online/hoeting1999.pdf>.
- [19] Kedem, B. and Fokianos, K. (2002). *Regression Models for Time Series Analysis*, New York: Wiley.
- [20] Kim, J. and Pollard, D (1990). Cube root asymptotics, *Annals of Statistics*, **18**, 191-219.
- [21] Knight, K. and Fu, W (2000). Asymptotics for LASSO-type estimators *Annals of Statistics*, **28**, 1356-1378.
- [22] Lange, K. (1999). *Numerical Analysis for Statisticians*, New York: Springer-Verlag.
- [23] Leeb, H. and Pötscher (2006). Can one estimate the conditional distribution of post-model-selection estimators? *Annals of Statistics*, **34**, 2554-2591.

- [24] Leeb, H. and Pötscher (2005). Model selection and inference: facts and fiction *Econometric Theory*, **21**, 21-59.
- [25] Lokhorst, J. (1999). The LASSO and Generalised Linear Models. Honors Project, Department of Statistics, University of Adelaide. Available as file Doc/justin.lokhorst.ps.gz from <http://www.maths.uwa.edu.au/berwin/software/lasso.html>.
- [26] Mallows, C. L. (1973). Some comments on  $C_p$ , *Technometrics*, **15**, 661-675.
- [27] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, London: Chapman and Hall.
- [28] Osborne, M.R., Presnell, B. and Turlach, B.A. (2000). A new approach to variable selection in least squares problems, *IMA Journal of Numerical Analysis* **20**, 389-403.
- [29] Park, M. Y. and Hastie, T. (2007).  $L_1$  Regularization path algorithm for generalized linear model, *Journal of the Royal Statistical Society: Series B*, **69**, 659-677.
- [30] Raftery, A. E. (1996). Approximate bayes factors and accounting for model uncertainty in generalized linear models, *Biometrika*, **83**, 251-266.
- [31] Raftery, A. E., Madigan, D. M., and Hoeting, J. (1997). Model selection and accounting for model uncertainty in linear regression models, *Journal of the American Statistical Association*, **92**, 179-191.
- [32] Raftery, A. E. and Richardson, S. (1993). Model selection for generalized linear models via glib, with application to epidemiology. In D. A. Berry and D. K. Stangl (Eds.), *Bayesian Biostatistics*. New York: Marcel Dekker.
- [33] Schwarz, G. (1978). Estimating the dimension of a model, *Annals of Statistics*, **6**, 461-464.
- [34] Shao, J. (1997). An asymptotic theory for linear model selection (with Discussion), *Statistica Sinica*, **7**, 221-242.
- [35] Tibshirani, R. J. (1996). Regression shrinkage and selection via the LASSO, *Journal of the Royal Statistical Society: Series B*, **58**, 267-288.
- [36] Wang, X. and George, E. I. (2007). Adaptive Bayesian criteria in variable selection for generalized linear models, *Statistica Sinica*, **17**, 667-690.

- [37] Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation, *Biometrika*, **92**, 937-950.
- [38] Yuan, M. and Lin, Y. (2005). Efficient empirical Bayes variable selection and estimation in linear models, *Journal of the American Statistical Association*, **100**, 1215-1225.