Bayesian Empirical Likelihood for Linear Regression and Penalized Regression

by

Adel Bedoui

(Under the Direction of Nicole A. Lazar)

Abstract

The likelihood function plays an essential role in statistical analysis. It helps to estimate a set of parameters of interest. To make inferences, usually one must specify a parametric model given data, which is a challenging task because it requires specification of a correct distribution, and this parametric model may be prone to bias that arises either from the estimation of a parameter or an incorrect specification of the probability distribution. Non-parametric approaches are used as a remedy to overcome the misspecification of the model but can be computationally costly. In this dissertation, we proposed an alternative approach based on Bayesian empirical likelihood for linear regression and penalized regression. This method is semi-parametric because it combines a nonparametric and a parametric model. The advantage of this approach is that it does not require the assumption of a parametric model nor the linearity of estimators; that is, we avoided problems with model misspecification. By using a Hamiltonian Monte Carlo, we averted the problem of convergence and the daunting task of finding an adequate proposal density in the Metropolis-Hastings method. Additionally, we showed that the maximum empirical likelihood estimator is consistent. Moreover, the resulting posterior density under the Bayesian empirical likelihood framework lacks a closed

form, which makes it difficult to obtain the exact distribution. For this purpose, we derived the asymptotic distribution of the regression parameters in the linear regression along with Bayesian credible intervals.

INDEX WORDS: Empirical likelihood, Bayesian statistics, Hamiltonian Monte Carlo, Penalized regression, Linear model, Asymptotic distribution

Bayesian Empirical Likelihood for Linear Regression and Penalized Regression

by

Adel Bedoui

B.S., University of Abdelmalek Essaadi , 2004

B.S., The Ohio State University , 2010

M.S., University of Texas at El Paso, 2013

A Dissertation Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2017

Bayesian Empirical Likelihood for Linear Regression and Penalized Regression

by

Adel Bedoui

Approved:

Professor:	Nicole A. Lazar
Committee:	Shuyang Bai
	Lynne Billard
	Cheolwoo Park
	Lynne Seymour

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
August 2017

Bayesian Empirical Likelihood for Linear Regression and Penalized Regression

Adel Bedoui

July 14, 2017

DEDICATION

To my parents

Acknowledgments

My warmest and unreserved thanks go to Dr. Nicole Lazar for her guidance and support. Her vast knowledge and her faith in me were vital for the completion of my Ph.D. I appreciate her contributions, time, ideas, and constructive criticism to improve my research. I am indebted for my adviser's constant assistance, encouragement, guidance and the tremendous support she provided throughout my doctoral study, especially during my illness.

I would also like to extend my appreciation to members of my committee: Dr. Shuyang Bai, Dr. Lynne Billard, Dr. Gauri Datta, Dr. Lynne Seymour, and Dr. Cheolwoo Park for generously offering their time, support, and guidance throughout the preparation and review of this dissertation. I would like to thank Dr. Gauri Datta and Dr. Lynne Seymour for their valuable knowledge. They introduced me to the Hamiltonian Monte Carlo method, which is a crucial piece in the completion of this dissertation. I thank the Department of Statistics for the funding it provided through my Doctoral program. I am grateful to all the professors, graduate students, and staff in the Department of Statistics.

A very special thanks to Dr. Kim Gilbert, Dr. Cheolwoo Park, and Dr. T.N. Sriram for their support during my sick leave period.

I will forever be thankful to my former adviser at The University of Texas at El Paso,

Dr. Ori Rosen. Professor Ori has been helpful in providing advice many times during my graduate school career. He was the reason why I decided to go to pursue a Ph.D. degree.

I cannot forget to thank my friends for their support and guidance. Last but not the least, I would like to thank my parents, my brothers, and sister for their unconditional support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Empirical likelihood (EL) is a nonparametric method first introduced by Owen (1988, 1990), although it can be considered as an extension of calibration estimation in survey sampling (Hartley and Rao, 1968; Deville and Sarndal, 1992). It is an estimation method inspired by maximum likelihood but without assuming a parametric model for the data. Hence, we avert the problem of model misspecification. One of the advantages of the EL approach is its flexibility to incorporate constraints and prior information (Kuk and Mak, 1989; Chen and Qin, 1993; Owen, 2001). Qin and Lawless (1994) extended the work of Owen by linking moment conditions and developing methods of combining information about parameters. In some settings, Owen (1988, 1990, 2001) showed that EL inherits properties of a parametric model. For instance, the limiting distribution of the likelihood ratio test based on EL for a univariate mean is χ^2. This parallels the result (Wilks, 1938) for parametric likelihood ratio tests. Another feature of EL is that it admits a Bartlett correction (DiCiccio et al., 1991); that is, the coverage error in EL can be reduced to n^{-2}. Baggerly (1998) showed that empirical likelihood is the only member of the Cressie-Read power divergence family to be Bartlett-correctable. Also, one can obtain data-determined confidence intervals through the Wilks statistics, which does not require the estimation of variance (Owen, 2001).

Empirical likelihood has also been extended to linear models, correlation models, ANOVA and variance modeling (Owen, 1991, 2001), generalized linear models (Kolaczyk, 1994), Bayesian settings (Lazar, 2003), weighted empirical likelihood (Wu, 2004), exponentially tilted empirical likelihood (Schennach, 2007), covariance estimation (Chaudhuri et al., 2007), generalized linear models incorporating population level information (Chaudhuri et al., 2008), Bayesian empirical likelihood for small area estimation (Chaudhuri and Ghosh, 2011), and Bayesian empirical likelihood for quantile regression (Yang and He, 2012). To sum up, EL is considered a powerful approach, as an alternative to likelihood, compared to other methods because without specifying a model to given data, it retains many desirables properties of likelihood.

Next, we introduce the essential concepts of EL as well as its most important properties. Most of the results can be found in the book *Empirical Likelihood* (Owen, 2001).

1.1 Empirical Likelihood

Suppose $X_1, \cdots, X_n \in \mathbb{R}$ are iid random variables generated from an unknown distribution. The cumulative distribution (CDF) is $F_{X_i}(x_i) = P(X_i \leq x_i)$, where $x_i \in \mathbb{R}$. Denote $P(X_i < x_i)$ by $F(x_i^-)$ and $P(X_i \leq x_i)$ by $F(x_i)$, so we can write $w_i = P(X_i = x_i) = F(x_i) - F(x_i^-)$. Let the notation $\mathbb{1}_{\delta_x}$ represent the indicator function of event δ_x that takes the value 1 if the assertion δ_x is true and 0 otherwise.

Definition 1.1.1. The empirical cumulative distribution function (ECDF) of $X_1, \cdots, X_n$ is

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq x},$$

where $x \in \mathbb{R}$. Using similar notation, one can define the nonparametric likelihood of the CDF F given $X_1, \cdots, X_n \in \mathbb{R}$ with common CDF F as follows:

$$L(F) = \prod_{i=1}^{n} \big(F(X_i) - F(X_i^-)\big) = \prod_{i=1}^{n} w_i.$$

The basic idea of Owen (1988) is to construct a multinomial distribution $F(w_1, \cdots, w_n)$ that places probability w_i on each observation. For each $i = 1, \cdots, n$, the probability must be non-negative and

$$\sum_{i=1}^{n} w_i = 1. \tag{1.1}$$

For inference, we explore the empirical likelihood ratio

$$R(F) = \frac{L(F)}{L(F_n)} = \prod_{i=1}^{n} n w_i,$$

with constraints $w_i \geq 0$ and $\sum_{i=1}^{n} w_i = 1$. By definition, $L(F) = 0$ if the distribution is continuous.

An important property of the ECDF is that it is the nonparametric maximum likelihood estimator (NPMLE) of the distribution F, that is, if $X_1, \cdots, X_n \in \mathbb{R}$ are independent random variables with a common CDF F, where F_n is their ECDF, then for any CDF F and for $F \neq F_n$

$$L(F) < L(F_n).$$

The NPMLE maintains the invariance property of maximum likelihood for functions. Given a function T that depends on data with common CDF F and a parameter of interest θ, such that $\theta = T(F)$, the NPMLE of θ is $\hat{\theta} = T(F_n)$.

Suppose $X_1, \cdots, X_n$ are iid random variables generated from a distribution F. We are interested in estimating a p-dimensional parameter $\boldsymbol{\theta} = T(F)$. Godambe (1960) provided a method of estimating equations that specifies how $\boldsymbol{\theta}$ should be determined:

$$E\{\mathbf{g}(X_i, \boldsymbol{\theta})\} = \mathbf{0},$$

where $\mathbf{g}$ is a $r \times 1$ vector-valued function and $r \geq p$.

Example 1.1.1. Let $\{(V_i, Y_i),\ i = 1, \cdots, n\}$ be a random sample. We are interested in estimating the correlation $\rho = \dfrac{\text{cov}(\boldsymbol{V}, \boldsymbol{Y})}{\sigma_v \sigma_y}$. To estimate ρ, we need to estimate $\mu_v = E(\boldsymbol{V})$, $\mu_y = E(\boldsymbol{Y})$, $\sigma_v = \sqrt{\text{Var}(\boldsymbol{V})}$, $\sigma_y = \sqrt{\text{Var}(\boldsymbol{Y})}$, and $\sigma_{vy} = E(\boldsymbol{VY})$. Let $\boldsymbol{\theta}^T = (\mu_v,\ \mu_y,\ \sigma_v^2,\ \sigma_y^2,\ \sigma_{vy})$. The set of estimating equations for estimating ρ is as follows:

$$
\begin{aligned}
g_1(\boldsymbol{V}, \boldsymbol{Y}, \boldsymbol{\theta}) &= \boldsymbol{V} - \mu_v, \\
g_2(\boldsymbol{V}, \boldsymbol{Y}, \boldsymbol{\theta}) &= \boldsymbol{Y} - \mu_y, \\
g_3(\boldsymbol{V}, \boldsymbol{Y}, \boldsymbol{\theta}) &= (\boldsymbol{V} - \mu_v)^2 - \sigma_v^2, \\
g_4(\boldsymbol{V}, \boldsymbol{Y}, \boldsymbol{\theta}) &= (\boldsymbol{Y} - \mu_y)^2 - \sigma_y^2, \\
g_5(\boldsymbol{V}, \boldsymbol{Y}, \boldsymbol{\theta}) &= (\boldsymbol{VY} - \mu_v \mu_y) - \sigma_{vy}.
\end{aligned}
$$

Let $\mathbf{g} = (g_1,\ g_2,\ g_3,\ g_4,\ g_5)^T$. Then the solution to the estimating equation

$$
\frac{1}{n} \sum_{i=1}^{n} \mathbf{g}(V_i,\ Y_i,\ \boldsymbol{\theta}) = \mathbf{0}
$$

is an estimator $\hat{\rho} = \dfrac{\hat{\sigma}_{vy}}{\hat{\sigma}_v \hat{\sigma}_y}$ for ρ.

Qin and Lawless (1994) linked estimating equations and empirical likelihood by using information about a parameter of interest regarding functions. Suppose that we have data $x_1, \cdots, x_n$ from some unknown distribution F. We are interested in inference concerning some function of F of p-dimension, $\boldsymbol{\theta}(F)$. When $\boldsymbol{\theta}(F)$ can be determined by an estimating equation $\boldsymbol{g}(x_i, \boldsymbol{\theta})$, the empirical likelihood ratio function is defined by

$$
\begin{aligned}
R(\boldsymbol{\theta}) &= \sup\Big\{ R(F) |\ w_i \geq 0,\ \sum_{i=1}^{n} w_i = 1, \sum_{i=1}^{n} w_i \boldsymbol{g}(X_i, \boldsymbol{\theta}) = 0 \Big\} \\
&= \sup\Big\{ \prod_{i=1}^{n} n w_i |\ w_i \geq 0,\ \sum_{i=1}^{n} w_i = 1, \sum_{i=1}^{n} w_i \boldsymbol{g}(X_i, \boldsymbol{\theta}) = 0 \Big\},
\end{aligned}
\tag{1.2}
$$

Equation (1.2) maximizes the empirical likelihood ratio subject to constraints $\sum_{i=1}^{n} w_i = 1$ and $\sum_{i=1}^{n} w_i \boldsymbol{g}(X_i, \boldsymbol{\theta}) = 0$. To maximize the constrained equation in (1.2), we use the Lagrange multiplier approach (Rockafellar, 1993). The Lagrangian equation to our maximization problem is

$$l(\boldsymbol{\theta},\ w_1, \cdots,\ w_n, \boldsymbol{\lambda},\ \gamma) = \ n\log(n) + \ \sum_{i=1}^{n} \log(w_i) \ - \ n\boldsymbol{\lambda}^T \sum_{i=1}^{n} w_i \boldsymbol{g}(X_i, \boldsymbol{\theta}) \ - \ \gamma\Big\{\sum_{i=1}^{n} w_i - 1\Big\} \tag{1.3}$$

where $\boldsymbol{\lambda}$ and γ are Lagrange multipliers. Taking the derivative of (1.3) with respect to w_i and setting it equal to zero, we obtain

$$\begin{aligned} &\frac{1}{w_i} - \gamma - n\boldsymbol{\lambda}^T \boldsymbol{g}(X_i, \boldsymbol{\theta}) = 0, \text{ for each } i = 1, \cdots, n \\ &w_i = \frac{1}{\gamma + n\boldsymbol{\lambda}^T \boldsymbol{g}(X_i, \boldsymbol{\theta})}. \end{aligned} \tag{1.4}$$

To estimate γ, we multiply (1.4) by $\sum_{i=1}^{n} w_i$, which fulfills the condition in (1.1)

$$\begin{aligned} &\sum_{i=1}^{n} w_i \Big(\frac{1}{w_i} - \gamma - n\boldsymbol{\lambda}^T \boldsymbol{g}(X_i, \boldsymbol{\theta})\Big) = 0 \\ &n - \gamma \sum_{i=1}^{n} w_i - n \sum_{i=1}^{n} w_i \boldsymbol{\lambda}^T \boldsymbol{g}(X_i, \boldsymbol{\theta}) = 0 \\ &\hat{\gamma} = n. \end{aligned} \tag{1.5}$$

We find that the maximum empirical likelihood estimator of the weight is simply

$$\hat{w}_i = \frac{1}{n(1 + \boldsymbol{\lambda}^T \boldsymbol{g}(X_i, \hat{\boldsymbol{\theta}}))}.$$

The result is an equation in terms of $\boldsymbol{\theta}$ where $\boldsymbol{\lambda} \in R^{p+1}$ is a function of $\boldsymbol{\theta}$ that solves

$$\sum_{i=1}^{n} \frac{\boldsymbol{g}(X_i, \boldsymbol{\theta})}{1 + \boldsymbol{\lambda}^T \boldsymbol{g}(X_i, \boldsymbol{\theta})} = \mathbf{0}.$$

1.2 Empirical Likelihood for Univariate Mean

In this section, we present a particular example of the general approach from Qin and Lawless (1994) with a specific univariate estimating equation. Let the population mean, $\mu \in \mathbb{R}$, be our parameter of interest. Suppose $X_1, \cdots, X_n$ are independent random variables with common CDF F and $E(X_i) = \mu$. The population mean can be estimated by $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$. The estimating equation has the following form:

$$g(x_i, \mu) = x_i - \mu.$$

By linking the above estimating equation and EL, the empirical likelihood ratio for the univariate mean is defined as follows:

$$R(\mu) = \sup\Big\{ \prod_{i=1}^{n} n w_i \mid w_i \ \geq \ 0, \ \sum_{i=1}^{n} w_i = 1, \ \sum_{i=1}^{n} w_i (x_i \ - \ \mu) \ = 0 \Big\}.$$

Maximizing $\prod_{i=1}^{n} n w_i$ is equivalent to maximizing $\sum_{i=1}^{n} \log(n w_i)$ under the two constraints $\sum_{i=1}^{n} w_i = 1$ and $\sum_{i=1}^{n} w_i (x_i - \mu) = 0$. We solve the optimization problem by using the Lagrange approach

$$l(\mu, \ \lambda, \ \gamma) \ = \ \sum_{i=1}^{n} \log(n w_i) \ - \ \lambda \sum_{i=1}^{n} w_i (x_i - \mu) \ - \ \gamma (1 \ - \ \sum_{i=1}^{n} w_i) \tag{1.6}$$

where λ and γ are the Lagrange multipliers. The first-order conditions for the maximization of (1.6) with respect to w_i, γ, and λ are

$$\begin{aligned} \frac{1}{w_i} &= \gamma + n\lambda(x_i - \mu) \\ \sum_{i=1}^{n} w_i &= 1 \\ \sum_{i=1}^{n} w_i(x_i - \mu) &= 0. \end{aligned} \tag{1.7}$$

Now, multiplying the first equation in (1.7) by w_i, summing over i, and using the second and third equations, we find that $\hat{\gamma} = n$ and

$$\hat{w_i} = n^{-1}\Big\{1 + \lambda(x_i - \mu)\Big\}^{-1}.$$

Substituting the values of $\hat{\gamma}$ and $\hat{w_i}$ into equation (1.6) we obtain

$$l(\mu, \lambda) = -\sum_{i=1}^{n} \log(1 + \lambda(x_i - \mu)) \tag{1.8}$$

where the value of λ is the solution of

$$m(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \frac{x_i - \mu}{1 + \lambda(x_i - \mu)} = 0. \tag{1.9}$$

For this example, the estimator of the mean is the same as the sample mean. Owen (1990) proved that when $X_1, \cdots, X_n$ are iid with finite mean μ and finite variance, then $-2\log(R(\mu)) \xrightarrow{d} \chi_1^2$.

Proof: The Lagrange multiplier λ is the solution to equation (1.9). Note that $m(0) = \bar{x} - \mu$. Let denote $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$. The Taylor expansion of m in the neighborhood of 0 gives

$$\begin{aligned} 0 = m(\lambda) &= m(0) + \lambda m^{'}(0) + o_P(n^{-1/2}) \\ &= \bar{x} - \mu - \lambda\hat{\sigma}^2 + o_P(n^{-1/2}). \end{aligned}$$

Thus,

$$\lambda \;=\; \frac{\bar{x} \;-\; \mu}{\hat{\sigma}^2} \;+\; o_P(n^{-1/2}) \;=\; O_P(n^{-1/2}).$$

Recall that the Taylor expansion of $\log(1 \;+\; x)$ is $x \;-\; \frac{x^2}{2} \;+\; O(x^3)$. From equation (1.8) we have

$$\begin{aligned}
-2\log(R(\mu)) \;&=\; 2\sum_{i=1}^{n}\log(1 \;+\; \lambda(x_i \;-\; \mu)) \\
&=\; 2n\lambda(\bar{x} \;-\; \mu) \;-\; n\lambda^2\hat{\sigma}^2 \;+\; o_P(1) \\
&=\; \frac{2n\lambda(\bar{x} \;-\; \mu)^2}{\hat{\sigma}^2} \;-\; \frac{n(\bar{x} \;-\; \mu)^2}{\hat{\sigma}^2} \;+\; o_P(1) \\
&=\; \frac{n\lambda(\bar{x} \;-\; \mu)^2}{\hat{\sigma}^2} \;+\; o_P(1) \\
&\xrightarrow{d}\; \chi_1^2
\end{aligned}$$

■

This is an exciting result because it parallels the Wilk's test result provided that $Var(X_i) \in (0, \infty)$. As such, it permits the construction of a rejection region, which can be used to build tests and confidence intervals for the functionals of interest. We reject the value of μ at level α when $-2\log\left(R(\mu)\right) > \chi_1^2$. The unrejected values of μ form a $100(1-\alpha)\%$ confidence region.

1.3 Fundamentals of Bayesian Inference

In this section, we present the fundamentals of Bayesian inference. Recall that in the frequentist setting, the data are repeatable random samples where the underlying parameters remain constant. However, in the Bayesian framework, the data are observed from the realized sample, i.e., fixed, where the parameters are random variables. The core of the Bayesian analysis is Bayes' theorem, which gives a coherent mathematical framework for updating our

belief in light of new data. Let $X = (X_1, \cdots, X_n)$ be independent random variables generated by a family of parametric models $\Pi = \{\pi(x|\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$, where Θ is a parameter space of possible values of $\boldsymbol{\theta}$, $\Theta \subset \mathbb{R}^p$, where p is known. In addition, we assume that the form of the density $\pi(x|\boldsymbol{\theta})$ is known but $\boldsymbol{\theta}$ is unknown. In addition to the model (likelihood), we specify a prior distribution for $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta})$. By Bayes' rule, the posterior density (distribution for $\boldsymbol{\theta}$ given the data X) is :

$$\pi(\boldsymbol{\theta}|\ X) \ = \ \frac{\pi(\boldsymbol{\theta}, X)}{\pi(X)} \ = \ \frac{\pi(\boldsymbol{\theta})\pi(X|\boldsymbol{\theta})}{\pi(X)} \tag{1.10}$$

where

$$\pi(X) \ = \ \begin{cases} \int\limits_{\Theta} \pi(\boldsymbol{\theta})\pi(X|\boldsymbol{\theta}) & \text{if } X \text{ is continuous,} \\ \sum\limits_{\Theta} \pi(\boldsymbol{\theta})\pi(X|\boldsymbol{\theta}) & \text{if } X \text{ is discrete.} \end{cases}$$

The term $\pi(X)$ is known as the marginal of X and can be omitted in equation (1.10) yielding the unnormalized posterior density

$$\pi(\boldsymbol{\theta}|X) \ \propto \ \pi(X|\boldsymbol{\theta})\pi(X).$$

Often, the analytic derivation of the posterior distribution, where algebra starts to bury the statistical science, is not easy, making the Bayesian inference a ponderous task. Fortunately, the development of powerful computers has made Bayesian analytics more tractable and the implementation of Markov Chain Monte Carlo (MCMC) approaches feasible. MCMC methods are a class of algorithms for sampling from a posterior distribution based on constructing a Markov chain. It is a general method based on drawing values of a parameter, θ, from approximate distributions and then correcting those draws to better approximate the target posterior distribution, $\pi(\theta|x)$ (Gelman, 2006). The sampling is done sequentially, with the distribution of sampled draws depending on the last value drawn; hence, the draws form a

Markov chain. The Key to the method's success, however, is not the Markov property but rather that the approximate distributions are improved at each step in the simulation, in the sense of converging to the target distribution (Gelman et al., 2013). Next, we define the Markov chain and Monte Carlo. A Markov chain M is a discrete time stochastic process $\{M_0, M_1, \cdots\}$ with the property that the distribution of M_t, given all previous values in the process $\{M_0, M_1, \cdots, M_{t-1}\}$, only depends on M_{t-1}. That is,

$$P(M_t \in A | M_0, M_1, \cdots, M_{t-1}) = P(M_t \in A | M_{t-1}) \text{ for any set } A.$$

For the distribution of M_t to converge, the chain needs to satisfy three properties:

- *Irreducibility*: A Markov chain is irreducible if the chain can reach any state from any other state with positive probability and in a finite amount of time.
- *Aperiodicity*: The chain is aperiodic if it does not get trapped in cycles. That is, the chain does not oscillate between sets of states in a regular periodic fashion.
- *Positive recurrent*: For any state x_i, the expected number of steps required for the chain to return x_i is finite.

The term "Monte Carlo" was first used by Ulam and Von Neumann (Cooper et al., 1989). It is a method of approximating an expectation by the sample mean of a function of simulated random variables. Markov chain and Monte Carlo can be combined to solve some delicate problems in areas such as Bayesian inference, molecular computational biology, bioinformatics, etc. The idea is to construct a Markov chain that converges to the desired distribution after many iterations. That is, MCMC allows us to estimate any statistic by ergodic averages.

$$E[h(t)]_\pi \approx \frac{1}{s} \sum_{i=1}^{s} h(t^{(i)})$$

where π is the posterior distribution of interest, $E[h(t)]$ is the desired expectation and $h(t^{(i)})$ is the i^{th} simulated sample from π. One might refer to these papers that fired the initial shots in the MCMC revolution that came to statistics (Metropolis et al., 1953; Hastings, 1970; Geman and Geman, 1984).

1.4 MCMC Methods

Many methods have been created for sampling from the posterior distribution. In this section, we introduce two MCMC algorithms: Gibbs sampling and the Metropolis-Hastings algorithm.

Gibbs Sampling Algorithm

Gibbs sampling is one of the MCMC algorithms that is suitable to generate samples from the posterior distribution. The algorithm was named after the physicist Josiah Willard Gibbs and described by brothers Stuart and Donald Geman in 1984. To produce samples using the Gibbs method, we sweep through each variable to sample from its conditional distribution with the remaining variables fixed to their current values (Lynch, 2007). For example, let X_1, X_2, and X_3 be random variables and set their initial values to $x_1^{(0)}$,$x_2^{(0)}$, and $x_3^{(0)}$. At iteration i, we sample $x_1^{(i)}$, $x_2^{(i)}$, and $x_3^{(i)}$ from $\pi(X_1 = x_1|X_2 = x_2^{(i-1)}, X_3 = x_3^{(i-1)})$, $\pi(X_2 = x_2|X_1 = x_1^{(i)}, X_3 = x_3^{(i-1)})$, and $\pi(X_3 = x_3|X_1 = x_1^{(i)}, X_2 = x_2^{(i)})$, respectively. This process continues until the chain converges. Algorithm 1 summarizes the Gibbs sampler process.

Metropolis-Hastings Algorithm (MH)

The Gibbs sampler is useful when the posterior density has a standard distribution. However, there are cases for which the posterior density is not of a known form. The Metropolis-

Algorithm 1: Gibbs sampler algorithm.

Initialize $x^{(0)} \sim \pi(x)$, $\pi(x)$ is a proposal distribution.

for iteration i=1,2,... **do**

$x_1 \sim \pi(X_1 = x_1 | X_2 = x_2^{(i-1)}, X_3 = x_3^{(i-1)}, \cdots, X_S = x_S^{(i-1)})$

$x_2 \sim \pi(X_2 = x_2 | X_1 = x_1^{(i)}, X_3 = x_3^{(i-1)}, \cdots, X_S = x_S^{(i-1)})$

$\vdots$

$x_S \sim \pi(X_S = x_S | X_1 = x_1^{(i)}, X_3 = x_3^{(i)}, \cdots, X_{S-1} = x_{S-1}^{(i)})$

end for

Hastings algorithm (Metropolis et al., 1953), named after Nicholas Metropolis, is a method to produce samples from the posterior distribution for which direct sampling is difficult. Suppose we have a density Q that can generate candidate observations. We also refer to Q as the jumping or proposal density. When the process is in state $\boldsymbol{\theta}$, we propose jumping to point $\boldsymbol{\theta}^*$ with the candidate value drawn according to Q. We evaluate the proposed state by calculating the acceptance probability of moving from our current value $\boldsymbol{\theta}$ to the proposed value $\boldsymbol{\theta}^*$

$$\alpha\left(\boldsymbol{\theta}^* | \boldsymbol{\theta}\right) = \min\left\{1,\ \frac{Q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)\pi(\boldsymbol{\theta}^*)}{Q(\boldsymbol{\theta}^*|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}\right\}.$$

$\frac{Q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)\pi(\boldsymbol{\theta}^*)}{Q(\boldsymbol{\theta}^*|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}$ is the ratio of the target density for the proposed value versus the current value multiplied by the ratio of the proposal density values. The algorithm consists of three main components. First, generate a candidate sample from the proposal density. Second, compute the acceptance probability α. Third, accept the candidate sample with probability α or reject it with probability $1 - \alpha$. Algorithm 2 details the general Metropolis-Hastings algorithm.

Algorithm 2: Metropolis-Hastings algorithm.

Initialize $x^{(0)}$.
for iteration i=1,2,... **do**
  Propose: $x^{cand} \sim Q(x^{(i)}|x^{(i-1)})$
  Acceptance Probability:

$$\alpha(x^{cand}|x^{(i-1)}) = \min\left\{1, \frac{Q(x^{(i-1)}|x^{cand})\pi(x^{cand})}{Q(x^{cand}|x^{(i-1)})\pi(x^{(i-1)})}\right\}$$

  $u \sim \text{uniform}(u; 0, 1)$
  if $u < \alpha$ **then**
    Accept the proposal: $x^{(i)} \leftarrow x^{cand}$
  else
    Reject the proposal: $x^{(i)} \leftarrow x^{(i-1)}$
  end if
end for

1.5 Bayesian Empirical Likelihood

Using EL under the Bayesian framework has captured the attention of many researchers. Lazar (2003) discussed the validity of using EL as an alternative to the likelihood function by exploring the characteristics of Bayesian inference with profile EL ratio in place of the data density. She provided simulation via Monte Carlo and further discussion to assess the validity and the appropriateness of the resulting posterior by using the method proposed by Monahan and Boos (1992). Grendár and Judge (2009) showed that Bayesian empirical likelihood (BEL) and Bayesian maximum a posteriori (MAP) estimators are consistent under misspecification of the model. They also demonstrated that the point estimators obtained by empirical likelihood and Bayesian MAP are asymptotically equivalent. Rao and Wu (2010) applied BEL to survey sampling; Chaudhuri and Ghosh (2011) to small area estimation; Yang and He (2012) to quantile regression; and Mengersen et al. (2013) to approximate Bayesian computation.

The BEL scheme is as follows. Let $X_1, \cdots, X_n$ be independent random variables with an

unknown distribution $F_{\boldsymbol{\theta}} \in \mathcal{F}_{\boldsymbol{\theta}}$ depending on a parameter $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p) \in \Theta \subseteq \mathbb{R}^p$. $\mathcal{F}_{\boldsymbol{\theta}}$ is a family of distributions described by $\boldsymbol{\theta}$. By placing a prior distribution $\pi(\boldsymbol{\theta})$ on $\boldsymbol{\theta}$, the posterior empirical likelihood density is

$$\pi(\boldsymbol{\theta}|\ X) = \frac{R(\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int\limits_{\Theta} R(\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} \ \propto \ R(\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \tag{1.11}$$

where $R(\boldsymbol{\theta})$ is the profile empirical likelihood ratio.

1.6 Scope of Dissertation

In this dissertation, we propose a Bayesian approach based on empirical likelihood for linear regression, ridge regression, and least absolute shrinkage and selection operator (lasso) regression. In the Bayesian framework, we replace the likelihood function with the profile empirical likelihood ratio. This method is semi-parametric because it combines EL and prior, which are a non-parametric and a parametric model, respectively. All Bayesian and frequentist methods assume a statistical model to data. In contrast, the empirical likelihood approach does not require the assumption of a parametric model. Hence, we avoid problems with model misspecification.

The ridge and lasso regressions impose l_2 and l_1 penalties, respectively, on the parameters of the linear regression. We begin by deriving the profile empirical likelihood ratio for the linear regression. Then, we derive BEL for ridge and lasso regression where we introduce the penalty in the form of a hyperparameter. The ridge and lasso estimates can be interpreted as Bayesian posterior mean estimates when the regression parameters have independent Normal and Laplace priors, respectively. The implementation of Gibbs sampling and Metropolis-Hastings, under the BEL approach, has limitations and poses challenges. Thus, we use the Hamiltonian Monte Carlo algorithm instead.

Under certain conditions, and as $n \to \infty$, we prove that the maximum empirical likelihood estimator for the linear regression under the BEL framework is consistent. In addition, we provide the asymptotic distribution for the posterior Bayesian empirical likelihood. Moreover, the hierarchical model provides a Bayesian method for selecting the ridge and lasso parameters. As such, we place a diffuse hyperprior on the shrinkage term.

Chapter 2

Bayesian Empirical Likelihood for Linear Regression

Penalized regression and Bayesian inference are gaining an important role in the era of Big Data. Often, due to the volume of the predictor variables, the data suffer from a multicollinearity problem and variables selection is necessary. Ridge regression (Tikhonov and Nikolayevich, 1943) treats the multicollinearity problem, and lasso regression (Tibshirani, 1996) addresses both variable selection and multicollinearity.

Some Bayesian and frequentist approaches for linear regression and penalized regression are based on parametric likelihoods, in which most of the time we assume that data are normally distributed. In the Bayesian approach, the likelihood is paired with the conjugate, non-conjugate, or noninformative parametric priors. Prior parameters are usually assumed to be known and can be estimated by an empirical Bayesian analysis or treated through a hyperprior.

We are interested in deriving a robust approach based on BEL for ridge and lasso models, robust meaning here that the normality assumption on the data is not required. Both models have a close connection to the Bayesian linear model. It suffices to put a prior distribution

on regression parameters to obtain the desired model such that the prior parameters depend on the penalty coefficient; for instance, placing double-exponential and normal priors lead to the lasso and ridge regressions, respectively. Double-exponential distribution is presented in the form of a mixture of normals with an exponential mixing density (Andrews and Mallows, 1974).

In this chapter, we derive BEL for the linear regression, discuss the limitations of Gibbs sampler and Metropolis-Hastings algorithms, introduce Hamiltonian Monte Carlo, and conclude with some examples.

2.1 Profile Empirical Likelihood Ratio for Linear Regression

We begin with notation and definitions. We observe a set of n pairs $(\boldsymbol{z_1}, y_1), \cdots, (\boldsymbol{z_n}, y_n)$. If we believe that the relationship between $\boldsymbol{z_i}$ and y_i is linear, then this association can be explained by the following model:

$$y_i = \theta_0 + \theta_1 z_{i1} + \theta_2 z_{i2} + \cdots + \theta_p z_{ip} + \epsilon_i \tag{2.1}$$

where $\boldsymbol{z_i} = [z_{i1}, \cdots, z_{ip}]^T$ and y_i are the predictor and response variables, respectively, θ_0 is the unknown intercept, θ_j is the unknown slope for explanatory variable z_{ij}, and ϵ_i is the error for data pair $(\boldsymbol{z_i}, y_i)$. We re-write model (2.1) as:

$$y_i = \boldsymbol{x_i}^T \boldsymbol{\theta} + \epsilon_i$$

where $\boldsymbol{x_i} = [1, \boldsymbol{z_i}]^T$ and $\boldsymbol{\theta} = [\theta_0, \theta_1, \cdots, \theta_p]^T$. Let $X = (\mathbf{1}, \boldsymbol{x_1}, \cdots, \boldsymbol{x_p})$ be our design matrix. In the linear model, our objective is to estimate the coefficients by minimizing

$$\sum_{i=1}^{n} \left(y_i - \boldsymbol{x_i}^T\boldsymbol{\theta}\right)^2 \tag{2.2}$$

such that $\sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$ where $\hat{y}_i = \boldsymbol{x_i}^T\hat{\boldsymbol{\theta}}$. Assuming that X^TX is invertible, the value that minimizes (2.2) is

$$\hat{\boldsymbol{\theta}}^{LS} = \left(X^TX\right)^{-1} X^T\boldsymbol{y}.$$

The estimation of regression parameters can also be approached via estimating equations:

$$E\left(X^T\left(\boldsymbol{y} - X\hat{\boldsymbol{\theta}}^{LS}\right)\right) = 0$$

Now, we can define the profile empirical likelihood ratio for $\boldsymbol{\theta}$ as follows:

$$R(\boldsymbol{\theta}) = \max_{w_i}\Big\{\prod_{i=1}^{n} nw_i |\ w_i \geq 0,\ \sum_{i=1}^{n} w_i = 1,\ \sum_{i=1}^{n} w_i\boldsymbol{x_i}(y_i - \boldsymbol{x_i}^T\boldsymbol{\theta}) = 0\Big\} \tag{2.3}$$

where $\boldsymbol{w} = \{w_1, \cdots, w_n\}$ is the vector of weights of $\boldsymbol{y} = \{y_1, \cdots, y_n\}$. Equation (2.3) describes a function on the n-dimensional simplex:

$$\boldsymbol{w} = \{w_1, \cdots, w_n |\ w_i \geq 0,\ \sum_{i=1}^{n} w_i = 1\} \in \Delta_{n-1}.$$

To maximize Equation (2.3), we implement the Lagrange Multiplier method:

$$G = \sum_{i=1}^{n} \log nw_i - n\boldsymbol{\lambda}^T \sum_{i=1}^{n} w_i\boldsymbol{x_i}(y_i - \boldsymbol{x_i}^T\boldsymbol{\theta}) - \gamma(1 - \sum_{i=1}^{n} w_i) \tag{2.4}$$

where $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_{p+1})' \in R^{p+1}$ and $\gamma \in R$ are Lagrange multipliers. Differentiating equation (2.4) with respect to w_i and using the constraint in equation (2.3) :

$$\frac{\partial G}{\partial w_i} = \frac{1}{w_i} - n\boldsymbol{\lambda}^T \boldsymbol{x_i}(y_i - \boldsymbol{x_i}^T\boldsymbol{\theta}) - \gamma \tag{2.5}$$

$$\frac{\partial G}{\partial w_i} = 0 \Leftrightarrow \sum_{i=1}^{n} w_i \frac{\partial G}{\partial w_i} = 0 \tag{2.6}$$

Equations (2.5) and (2.6) imply that $\hat{\gamma} = n$. Then

$$w_i = n^{-1}\Big\{1 + \boldsymbol{\lambda}^T \boldsymbol{x_i}(y_i - \boldsymbol{x_i}^T\boldsymbol{\theta})\Big\}^{-1} \tag{2.7}$$

where $\boldsymbol{\lambda} = \boldsymbol{\lambda(\theta)}$ satisfies $p+1$ equations given by

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\boldsymbol{x_i}(y_i - \boldsymbol{x_i}^T\boldsymbol{\theta})}{1 + \boldsymbol{\lambda}^T \boldsymbol{x_i}(y_i - \boldsymbol{x_i}^T\boldsymbol{\theta})} = \mathbf{0}. \tag{2.8}$$

Substituting the expression for w_i into $\log R(\boldsymbol{\theta})$ yields

$$\log R(\boldsymbol{\theta}) = \log \prod_{i=1}^{n} n w_i = -\sum_{i=1}^{n} \log\Big\{1 + \boldsymbol{\lambda}^T \boldsymbol{x_i}(y_i - \boldsymbol{x_i}^T\boldsymbol{\theta})\Big\} \tag{2.9}$$

This is a particular example of the general approach from Qin and Lawless (1994) with specific estimating equations for multiple regression.

To find the estimate of $\boldsymbol{\theta}$, we follow the Bayesian approach, where we replace the likelihood with the profile empirical likelihood ratio. For a given prior $\pi(\boldsymbol{\theta})$, the empirical posterior

density function is given by

$$\begin{aligned}\pi(\boldsymbol{\theta}|\boldsymbol{y},\ \boldsymbol{X},\ \boldsymbol{\lambda}) &\propto R(\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \\ &\propto \exp\Big\{-\sum_{i=1}^{n}\log(1+\boldsymbol{\lambda}^T\boldsymbol{x_i}(y_i-\boldsymbol{x_i}^T\boldsymbol{\theta}))\Big\}\pi(\boldsymbol{\theta}) \\ &= \exp\left(-\sum_{i=1}^{n}\log(1+\boldsymbol{\lambda}^T\boldsymbol{x_i}(y_i-\boldsymbol{x_i}^T\boldsymbol{\theta}))+\log\pi(\boldsymbol{\theta})\right).\end{aligned} \tag{2.10}$$

2.2 Estimation

In this section, we select a prior distribution for BEL for the linear regression as well as the method of estimating the Lagrange multiplier. The normal distribution has the maximum entropy (Jaynes, 1957). Thus, if mean and variance are given, a normal prior in some sense has minimum information. We place the following priors:

$$\pi(\boldsymbol{\theta}|\sigma^2) \sim N(\mathbf{0},\ \frac{1}{\sigma^2}A)$$
$$\pi(\sigma^2) \sim IG(a_1, b_1)$$

which is a normal distribution with vector mean $\mathbf{0}$ and covariance matrix A, whose pdf is $p(\boldsymbol{x}) \propto (2\pi)^{-p/2}|A|^{-1/2}\exp(-\frac{\boldsymbol{x}^tA^{-1}\boldsymbol{x}}{2\sigma^2}),\ -\infty < \boldsymbol{x} < \infty$, where A is assumed to be known and positive definite. $IG(a_1, b_1)$ is the inverse gamma distribution whose pdf is $p(x) \propto x^{-(a+1)}\exp(-b/x)$. $\boldsymbol{\lambda}$ is the vector of the Lagrange multipliers, which is the root of equation (2.8). To find its solution, we use the modified Newton-Raphson method. There is another approach that finds an analytic solution to $\boldsymbol{\lambda}$ suggested by Chen and Van Keilegom (2009). However, this approach fails to provide the optimum values for $\boldsymbol{\lambda(\theta)}$ when $\boldsymbol{\theta}$ is not around the maximum likelihood estimator.

One can easily see that the minus derivative of equation (2.9) is equal to equation (2.8). Thus, $\boldsymbol{\lambda}$ is the minimizer of equation $L(\boldsymbol{\lambda}) := -\sum_{i=1}^{n}\log\Big\{1+\boldsymbol{\lambda}^T\boldsymbol{x_i}(y_i-\boldsymbol{x_i}^T\boldsymbol{\theta})\Big\}$. As discussed

in Qin and Lawless (1994), the existence and uniqueness of $\boldsymbol{\lambda}$ are guaranteed provided that the following conditions are satisfied:

1. The vector $\mathbf{0} \in R^{p+1}$ is within the convex hull of $\left\{\boldsymbol{x_i}(y_i - \boldsymbol{x_i}^T\boldsymbol{\theta}),\ i = 1, \cdots, n\right\}$.
2. The matrix $\sum_{i=1}^{n} \frac{\boldsymbol{A_i}\boldsymbol{A_i}^T}{[1+\lambda^T\boldsymbol{A_i}]^2}$ is positive definite where $\boldsymbol{A_i} = \boldsymbol{x_i}(y_i - \boldsymbol{x_i}^T\boldsymbol{\theta})$.

The domain L must exclude any $\boldsymbol{\lambda}$ for which some $w_i \leq 0$. Thus, we imposed the following constraints:

$$1 + \boldsymbol{\lambda}^T\boldsymbol{x_i}\,(y_i - \boldsymbol{x_i}\boldsymbol{\theta}) > 0,\ i = 1, \cdots, n. \tag{2.11}$$

The original n-dimensional optimization problem is equivalent to a p+1-dimensional problem of minimizing $L(\cdot)$ subject to the constraint (2.11) (Owen, 2001). It is easy shown that $L(\cdot)$ is a convex function on any connected sets satisfying the above constraint. But, unfortunately, $L(\boldsymbol{\lambda})$ is not defined on the sets:

$$1 + \boldsymbol{\lambda}^T\boldsymbol{x_i}\,(y_i - \boldsymbol{x_i}\boldsymbol{\theta}) \leq 0,\ i = 1, \cdots, n.$$

Owen (2001) used a pseudo-logarithm function, which extends $L(\boldsymbol{\lambda})$ outside the convex set:

$$\log_*(z) = \begin{cases} \log(z) & ,\ \text{if } z \geq \frac{1}{n}, \\ \log(\frac{1}{n}) - 1.5 + 2nz - \frac{(nz)^2}{2} & ,\ \text{if } z < \frac{1}{n}. \end{cases} \tag{2.12}$$

The objective function becomes:

$$L_*(\boldsymbol{\lambda}) = -\sum_{i=1}^{n} \log_*\left(1 + \boldsymbol{\lambda}^T\boldsymbol{x_i}\,(y_i - \boldsymbol{x_i}\boldsymbol{\theta})\right).$$

That is, minimizing $L(\boldsymbol{\lambda})$ is equivalent to minimizing $L_*(\boldsymbol{\lambda})$ over $\boldsymbol{\lambda} \in R^{p+1}$ without constraint. Taking the derivative of $\log_*(z)$ with respect to z, we obtain

$$\log_*^{'}(z) = \begin{cases} \frac{1}{z} & , \text{ if } z \geq \frac{1}{n}, \\ 2n - n^2 z & , \text{ if } z < \frac{1}{n} \end{cases} \tag{2.13}$$

and taking derivative of equation (2.13) with respect to z gives

$$\log_*^{''}(z) = \begin{cases} -\frac{1}{z^2} & , \text{ if } z \geq \frac{1}{n}, \\ -n^2 & , \text{ if } z < \frac{1}{n}. \end{cases} \tag{2.14}$$

In general, one can easily obtain the gradient and Hessian matrix of $L_*(\boldsymbol{\lambda})$ for the linear model:

$$L_*^{\cdot}(\boldsymbol{\lambda}) \equiv \frac{\partial L_*(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} = -\sum_{i=1}^{n} \log_*^{'}(1 + \boldsymbol{\lambda}^T \boldsymbol{x_i}(y_i - \boldsymbol{x_i}^T \boldsymbol{\theta}))(\boldsymbol{x_i}(y_i - \boldsymbol{x_i}^T \boldsymbol{\theta})).$$

$$L_*^{\cdot\cdot}(\boldsymbol{\lambda}) \equiv \frac{\partial^2 L_*(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda} \partial \boldsymbol{\lambda}^T} = -\sum_{i=1}^{n} \log_*^{''}(1 + \boldsymbol{\lambda}^T \boldsymbol{x_i}(y_i - \boldsymbol{x_i}^T \boldsymbol{\theta}))(\boldsymbol{x_i}(y_i - \boldsymbol{x_i}^T \boldsymbol{\theta}))(\boldsymbol{x_i}(y_i - \boldsymbol{x_i}^T \boldsymbol{\theta}))^T > 0.$$

We use the Newton-Raphson method to compute $\boldsymbol{\lambda}$ iteratively:

$$\boldsymbol{\lambda}_{k+1} \ = \ \boldsymbol{\lambda}_k \ - \ [L_*^{\cdot\cdot}(\boldsymbol{\lambda_k})]^{-1} L_*^{\cdot}(\boldsymbol{\lambda_k}).$$

The process can be repeated until it converges to a fixed point. A convenient initial value for $\boldsymbol{\lambda}$ is a zero vector, which corresponds to $w_i = \dfrac{1}{n}$, for $i = 1, \cdots, n$. This method works as follows:

Step 0: Let $\boldsymbol{\lambda}_0 = \mathbf{0}$. Set $k = 0$, $\gamma_0 = 1$ and $\varepsilon = 10^{-8}$

Step 1: Calculate $\Delta_1(\boldsymbol{\lambda_k})$ and $\Delta_2(\boldsymbol{\lambda_k})$ where

$$\Delta_1(\boldsymbol{\lambda_k}) = \sum_{i=1}^{n} \frac{\boldsymbol{x_i}(y_i - \boldsymbol{x_i}^T\boldsymbol{\theta})}{1 + \boldsymbol{\lambda}^T\boldsymbol{x_i}(y_i - \boldsymbol{x_i}^T\boldsymbol{\theta})}$$

and

$$\Delta_2(\boldsymbol{\lambda_k}) = \left(-\frac{\left[\boldsymbol{x_i}(y_i - \boldsymbol{x_i}^T\boldsymbol{\theta})\right]\left[\boldsymbol{x_i}(y_i - \boldsymbol{x_i}^T\boldsymbol{\theta})\right]^T}{\left(1 + \boldsymbol{\lambda}^T\boldsymbol{x_i}(y_i - \boldsymbol{x_i}^T\boldsymbol{\theta})\right)^2} \right)^{-1} \Delta_1(\boldsymbol{\lambda_k})$$

if $||\Delta_2(\boldsymbol{\lambda_k})|| < \varepsilon$, stop the algorithm and report $\boldsymbol{\lambda_k}$; otherwise go to Step 2.

Step 2: Calculate $\delta_k = \gamma_k \Delta_2(\boldsymbol{\lambda_k})$. If $1 + (\boldsymbol{\lambda_k} - \delta_k)^T\boldsymbol{x_i}(y_i - \boldsymbol{x_i}^T\boldsymbol{\theta}) \leq 0$ for some i, let $\gamma_k = \gamma_k/2$ and repeat Step 2.

Step 3: Set $\boldsymbol{\lambda_{k+1}} = \boldsymbol{\lambda_k} - \delta_k$, $k = k + 1$ and $\gamma_k = (k+1)^{-\frac{1}{2}}$.

We have determined the method to estimate the Lagrange multipliers. To estimate our parameter of interest, $\boldsymbol{\theta}$, in the linear regression, we use the Bayesian approach based on EL. That is, the posterior empirical distribution of $\boldsymbol{\theta}$ is proportional to the profile EL ratio multiplied by priors. The EL ratio is given by:

$$R(\boldsymbol{\theta}) = \exp\Big\{ -\sum_{i=1}^{n} \log(1 + \boldsymbol{\lambda}^T\boldsymbol{x_i}(y_i - \boldsymbol{x_i}^T\boldsymbol{\theta})) \Big\}$$

and in combination with the priors defined above yields the following posterior distribution:

$$\begin{aligned} \pi(\boldsymbol{\theta}, \sigma^2 | X, \boldsymbol{y}) \propto \exp\Big\{ -\sum_{i=1}^{n} \log(1 + \boldsymbol{\lambda}^T\boldsymbol{x_i}(y_i - \boldsymbol{x_i}^T\boldsymbol{\theta})) \Big\} \times \\ \exp\left(-\frac{1}{2\sigma^2}\boldsymbol{\theta}^T A^{-1}\boldsymbol{\theta} \right) \times \left(\sigma^2\right)^{(a_1+1)} \exp\left(-b_1/\sigma^2\right) \end{aligned} \tag{2.15}$$

Equation (2.15) gives rise to the following sampling scheme:

- Sample $\boldsymbol{\theta}$ from

$$\pi(\boldsymbol{\theta}|\sigma^2\,\boldsymbol{y},\ X) \propto \exp\Big\{-\sum_{i=1}^{n}\log(1+\boldsymbol{\lambda}^T\boldsymbol{x_i}(y_i-\boldsymbol{x_i}^T\boldsymbol{\theta}))-\frac{1}{2\sigma^2}\boldsymbol{\theta}^T A^{-1}\boldsymbol{\theta}\Big\}. \tag{2.16}$$

- Sample σ^2 from $IG\left(a_1+1, b_1+\dfrac{\boldsymbol{\theta}^T A^{-1}\boldsymbol{\theta}}{2}\right)$

The posterior empirical likelihood of $\boldsymbol{\theta}$ does not have a closed form, which makes the implementation of the Gibbs sampler impossible. One can think of implementing the Metropolis-Hastings algorithm as it is suited to generate samples from a distribution that lacks an analytic form. However, the implementation of MH is challenging and fails to achieve convergence due to the nature of the posterior density support, which complicates the process of finding an efficient proposal density for the MH algorithm. The surface of the posterior empirical likelihood is rigid and not smooth with many local optimums. If we select initial values far from the global optimum, the MH algorithm, often, get trapped in cycles. For instance, we apply BEL for the linear model on Old Faithful Geyser data (Härdle, 1991) in Yellowstone National Park, Wyoming, USA. The objective of that experiment was to study the waiting time between eruptions and the duration of the eruption for the Old Faithful geyser. The MLE of the slope is 0.0756. Figure 2.1 shows the trace plots of the slope, based on Metropolis-Hastings, by using different initial values. We choose the Normal distribution as jumping density. It is clear that MH is sensitive to the starting value under the EL framework. In addition, the chains do not mix well. The problem of convergence is due to the intricacy of its support. Often, the chain gets trapped in a region and never reaches the global optimum. To observe this, we consider 100 independent and identically distributed bivariate observations $x_i = (x_{i1}, x_{i2})$, $i = 1, \cdots, 100$; we assume that $y_i = \theta_1 x_{i1} + \theta_2 x_{i2} + e_i$ where $\theta_1 = 2$, $\theta_2 = 5$, and e_i is the error term. Figure 2.2 depicts the perspective plot of $\log\left(\pi(\boldsymbol{\theta}|\sigma^2, X, \boldsymbol{y})\right)$ for various values of θ_1 and θ_2. We can see that the support is non-convex

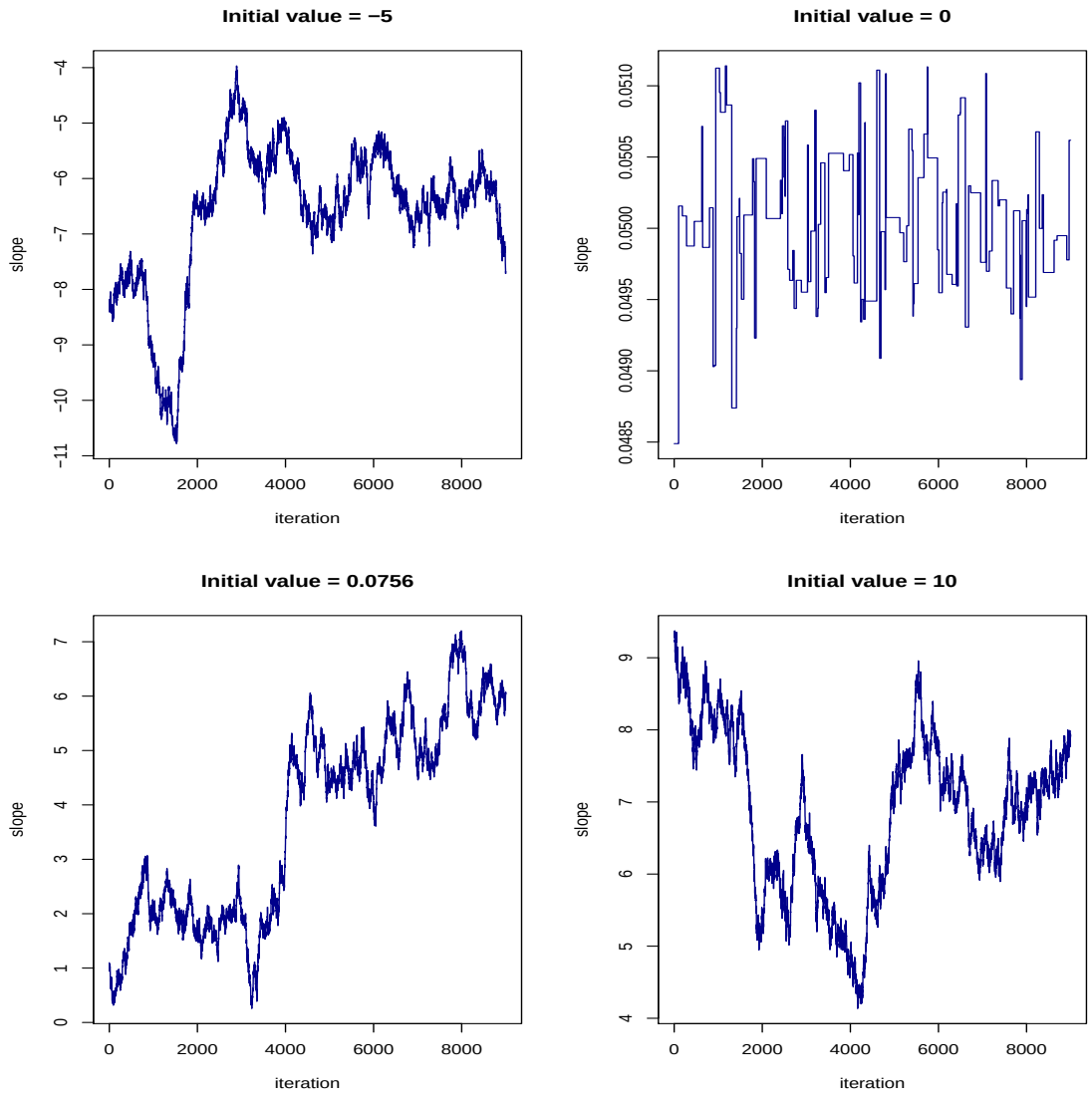

Figure 2.1: Trace plots for the slope using different initial values. Top left: $\theta_1^{(0)}$=-5. Top right: $\theta_1^{(0)} = 0$. Bottom left: $\theta_1^{(0)}$ =0.0756. Bottom right: $\theta_1^{(0)} = 10$.

where its surface is rigid and not smooth. That is, if we start from values far from the global optimum, the chain gets trapped in some local optimum. Therefore, we are required to tune the Metropolis-Hastings to find a good proposal density with the appropriate variance that allows us to reach all states frequently and provides a high acceptance rate. Instead, we use the Hamiltonian Monte Carlo algorithm (HMC), which converges quickly towards target distribution. In HMC, distances between successive generated points are large. Thus, it requires fewer iterations to get the representative sampling.

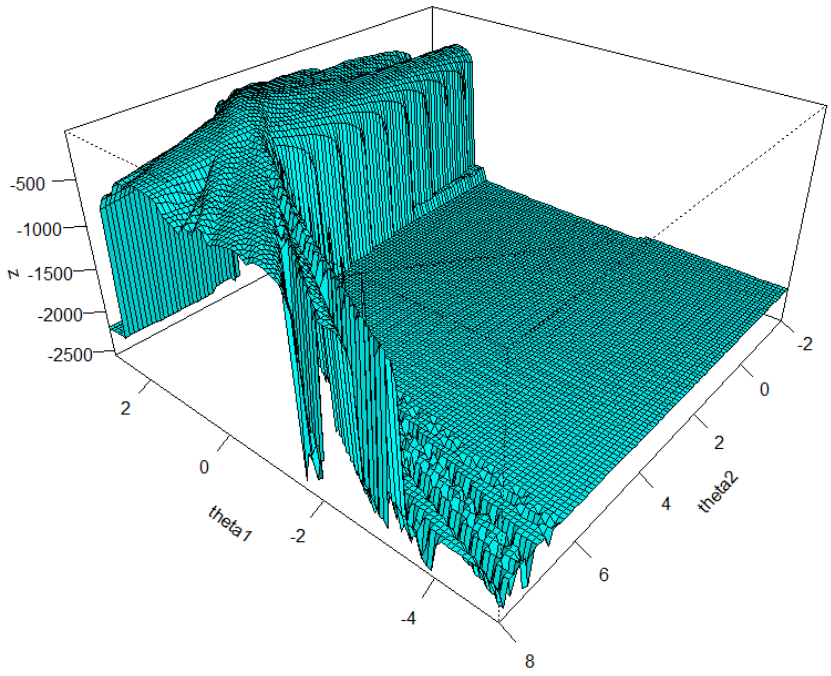

Figure 2.2: Perspective plot of $\log(\pi(\boldsymbol{\theta}, X, \boldsymbol{y}))$ for various values of θ_1 and θ_2.

2.3 Hamiltonian Monte Carlo for Bayesian Empirical Likelihood

In this section, we present a summary of Neal's (2011) Hamiltonian Monte Carlo. HMC, also known as Hybrid Monte Carlo, is an MCMC method to generate posterior samples for which direct sampling is difficult. It borrows an idea from physics to apply to the local random walk behavior in the Metropolis algorithm, thus allowing it to move much more rapidly through the target distribution (Gelman et al., 2013). HMC uses the gradient of the density and the Hamiltonian system to sample successive states for the Metropolis-Hastings algorithm with a high jump and large acceptance probability. This reduces the correlation between successive sampled states, which allows for a quicker convergence. The Hamiltonian system is a dynamic system controlled by Hamilton's equation. To better understand the Hamiltonian system, let us consider a physical interpretation in the two-dimensional case. Imagine that, under a gravitational field, a particle is moving over a continuous surface of varying heights. That is, the state of this evolution consists of the position of the particle, given by a two-dimensional vector v, and the momentum of the particle, given by a two-dimensional vector u. In physics, the momentum of the particle is equal to its mass times its velocity. The total energy of the particle is equal to its potential energy $U(v)$ plus its kinetic energy $K(u)$. Moreover, the potential energy of the particle is proportional to the height of the surface at its current position, and its kinetic energy is equal to $||u||^2/2m$. u is the momentum of the particle at its current position, and m is its mass. One interesting property of the Hamiltonian system is that the total energy of the particle, as it moves up or down, remains constant, and what changes are its potential energy and its kinetic energy. The underlying logic of HMC sampling is as follows. To sample from a posterior distribution $\pi(\theta|x)$, we treat the parameter θ as a particle and denote its value as its current position.

We define the potential energy and the kinetic energy as

$$U(\theta) = -\log\{\pi(\theta|x)\}$$
$$K(u) = \frac{1}{2}u^T M^{-1} u$$

where $u = (u_1,\ u_2, \cdots, u_s)^T$ is the momentum vector and M is the mass matrix, also known as the dispersion matrix. $K(u)$ rises from Gaussian distribution $N(0,\ M)$ where M is a symmetric positive definite matrix. We choose M. In this dissertation we set M equals the identity matrix. The total energy or Hamiltonian system is defined as

$$H(\theta,\ u) = U(\theta) + K(u).$$

The position of θ and the momentum u of the particle change over time and are determined by the partial derivatives of the Hamiltonian system. One should keep in mind, as mentioned above, that the total energy remains constant as the particle moves. These partial derivatives give rise to the so-called Hamiltonian equations of motion

$$\begin{aligned} \frac{d\theta}{dt} &= \frac{\partial H}{\partial u} = M^{-1}u, \\ \frac{du}{dt} &= -\frac{\partial H}{\partial \theta} = -\frac{\partial U(\theta)}{\partial \theta} = \frac{\nabla \pi(\theta|x)}{\pi(\theta|x)}. \end{aligned} \tag{2.17}$$

Neal (2011) showed that these Hamiltonian equations are reversible, invariant, and volume-preserving, which make the Hamiltonian system suitable for MCMC sampling schemes. When $\pi(\theta|x)$ lacks a closed form, the equations (2.17) lack analytic solutions. Thus, the solution is approximated at discrete time steps. Following Neal (2011), we apply the leapfrog integration method to approximate the solution of the Hamiltonian equations. First, a small step size ϑ is selected. Then, given the current value of θ and u at time t, the position and

momentum at time $t + \vartheta$ are updated as follows:

$$u\left(t + \frac{1}{2}\vartheta\right) = u(t) - \frac{1}{2}\vartheta\frac{\partial U(\theta(t))}{\partial\theta},$$
$$\theta\left(t + \vartheta\right) = \theta(t) + \vartheta M^{-1}u\left(t + \frac{1}{2}\vartheta\right),$$
$$u\left(t + \vartheta\right) = u\left(t + \frac{1}{2}\vartheta\right) - \frac{1}{2}\vartheta\frac{\partial U(\theta(t+\vartheta))}{\partial\theta}$$

Sometimes the approximation introduces errors, and an accept-reject algorithm is required to conserve the invariant property of HMC (Neal, 2011). The procedure works as follows: In the first step, new values for the momentum vector u are randomly drawn from a Gaussian distribution $N(0,\ M)$, independently of the current values of θ. In the second step, starting with the current state, (θ, u), a Hamiltonian system is simulated for L steps using the leapfrog method, with a step size of ϑ. At the end of this L-step trajectory, the proposed state (θ^*, u^*) is accepted with probability

$$\min\left[1, \exp\left(-H(\theta^*, u^*) + H(\theta, u)\right)\right] = \min\left[1, \exp\left(-U(\theta^*) + U(\theta) - K(u^*) + K(u)\right)\right] \quad (2.18)$$

where $U(\theta) = -\log(\pi(\theta|x))$ and $K(u) = \frac{1}{2}u^T M^{-1}u$. If the proposed state is rejected, the next state is the same as the current state. Gelman et al. (2013) suggested that the HMC is optimally efficient when its acceptance rate is approximately 65%. For a reader who would like to see in detail the properties of the above method, Neal (2011), Section 2, might be a useful reference. Next, we introduce two illustrative examples. In the first example, we use the HMC to sample from a bivariate normal distribution. The second example is BEL for the mean using Darwin's data (Darwin, 1876).

Example: Bivariate Normal

In this example, we use the Hamiltonian Monte Carlo to sample from a bivariate normal distribution:

$$\pi(\boldsymbol{x}) \sim N\left(\boldsymbol{\mu} = [1,5]^T, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$$

To sample from $\pi(\boldsymbol{x})$, we need to determine the expression for $U(\boldsymbol{x})$ and $\dfrac{\partial U(\boldsymbol{x})}{\partial \boldsymbol{x}}$. The potential energy function $U(\boldsymbol{x})$ can be defined as $U(\boldsymbol{x}) = -\log(\pi(\boldsymbol{x}))$. That is,

$$U(\boldsymbol{x}) \propto \frac{(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}{2}$$
$$\frac{\partial U(\boldsymbol{x})}{\partial \boldsymbol{x}} = (\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1}$$

Using the expressions of $U(\boldsymbol{x})$ as the potential energy and $\dfrac{\partial U(\boldsymbol{x})}{\partial \boldsymbol{x}}$ as the kinetic energy, we implement the HMC method for the bivariate normal in R. Figure 2.3 displays a simulation

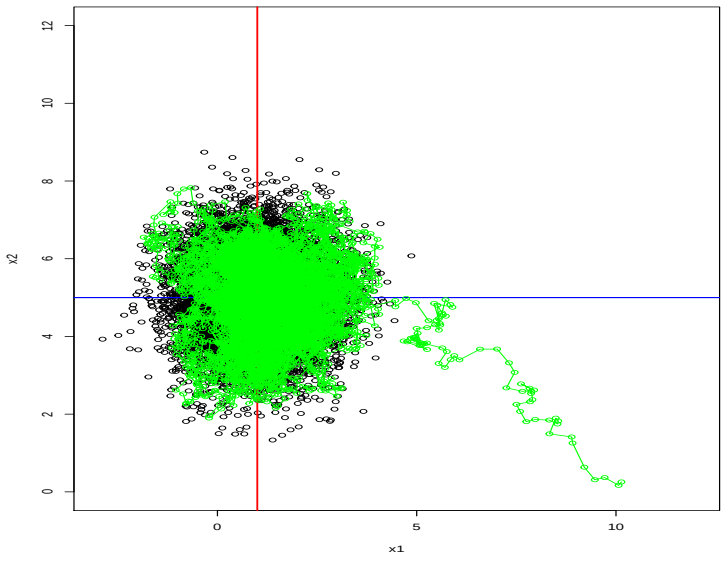

Figure 2.3: Hamiltonian Monte Carlo samples from bivariate normal with mean $[1,5]^T$, variances equal to 1, and correlation zero.

of 10000 samples from bivariate normal distribution. We implement HMC to sample from the same distribution in which the initial values are, with $\boldsymbol{\mu} = [10, 0]^T$, far from the mean. The green points are the first 5000 HMC samples. It is obvious that the HMC estimate is rapidly approaching the area of high density. In addition, HMC samples explore the space well and are scattered all around the data cloud.

Example: Bayesian Empirical Likelihood for the Mean

In this example, we apply BEL for the mean on Darwin's data (Table 2.1), which was meant to determine if cross-fertilized plants grew taller than self-fertilized plants (Darwin, 1876). To estimate the mean, we use the results obtained in Section 1.2. We place $N(a_0, \sigma_0)$ prior on μ where $a_0 = 2.6067$ (equals the sample mean) and $\sigma_0 = 0.1$ (small variance) are assumed to be known such that. The empirical posterior density of μ is

$$\pi(\mu|\boldsymbol{x}) \propto \exp\left(-\sum_{i=1}^{n} \log\left[1 + \lambda(x_i - \mu)\right] - \frac{1}{2\sigma_0^2}(\mu - a_0)^2\right).$$

Table 2.1: The outcome of a classic experiment by Darwin (1876).

Cross	Self	Height	Cross	Self	x
23.5	17.4	6.1	18.3	16.5	1.8
12.0	20.4	-8.4	21.6	18.0	3.6
21.0	20.0	1.0	23.3	16.3	7.0
22.0	20.0	2.0	21.1	18.0	3.0
19.1	18.4	0.7	22.1	12.8	9.3
21.5	18.6	2.9	23.0	15.5	7.5
22.1	18.6	3.5	12.0	18.0	-6.0
20.4	15.3	5.1			

To sample from the above distribution, we implement HMC where the negative logarithm posterior density and the gradient of the negative logarithm posterior density are

$\sum_{i=1}^{n} \log\left[1+\lambda(x_i-\mu)\right] + \frac{1}{2\sigma_0^2}(\mu-a_0)^2$ and $\sum_{i=1}^{n} \frac{1}{1+\lambda(x_i-\mu)} + \frac{\mu-a_0}{\sigma_0^2}$, respectively. The number of iterations used is 5000 with 1000 burn-in. We set $L = 10$ and $\vartheta = 0.1$.

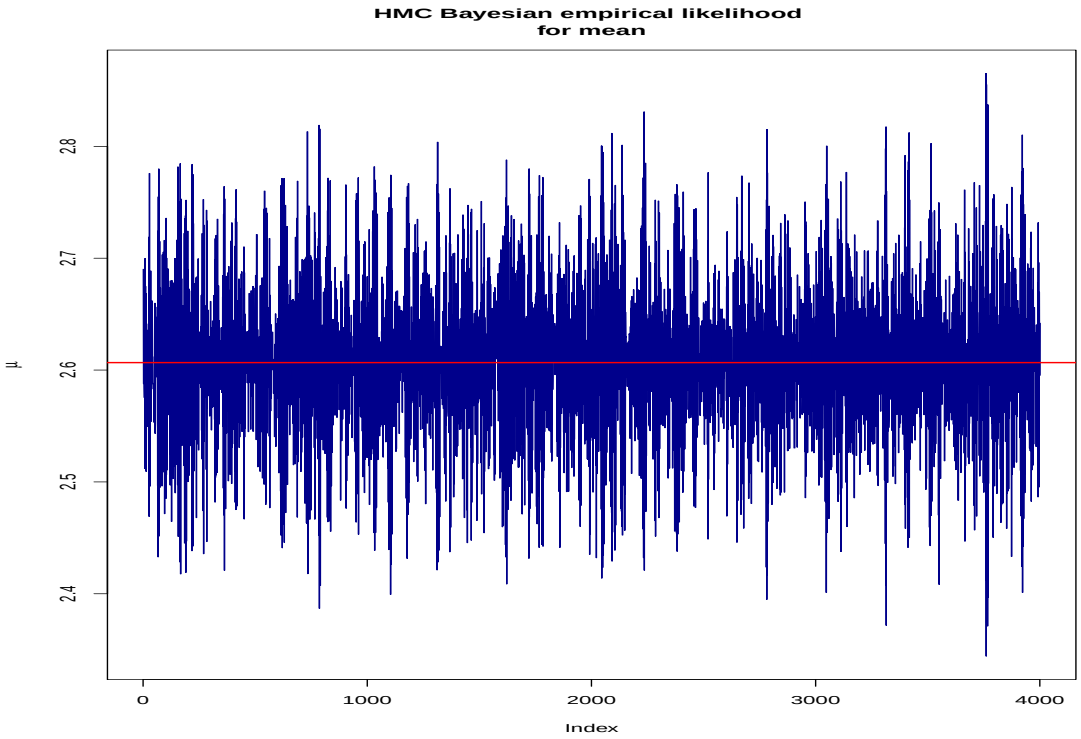

Figure 2.4: Trace plot for the parameter μ after 5000 iterations with 1000 burn-in; red line is the OLS estimate.

Figure 2.4 shows the trace plot for the parameter μ after 5000 iterations with 1000 burn-in where the red line is the OLS estimate ($\hat{\mu} = 2.606667$). This plot displays a well-behaved MCMC output, and the center of the chain appears to be around a value with reasonable fluctuations. This indicates that the chain is mixing well. The acceptance rate is 68.73%, which suggests that the HMC is working efficiently. The empirical posterior mean estimate is 2.607813, which is close to the OLS value. The standard error is 0.02058 and the standard deviation of the data is 4.71.

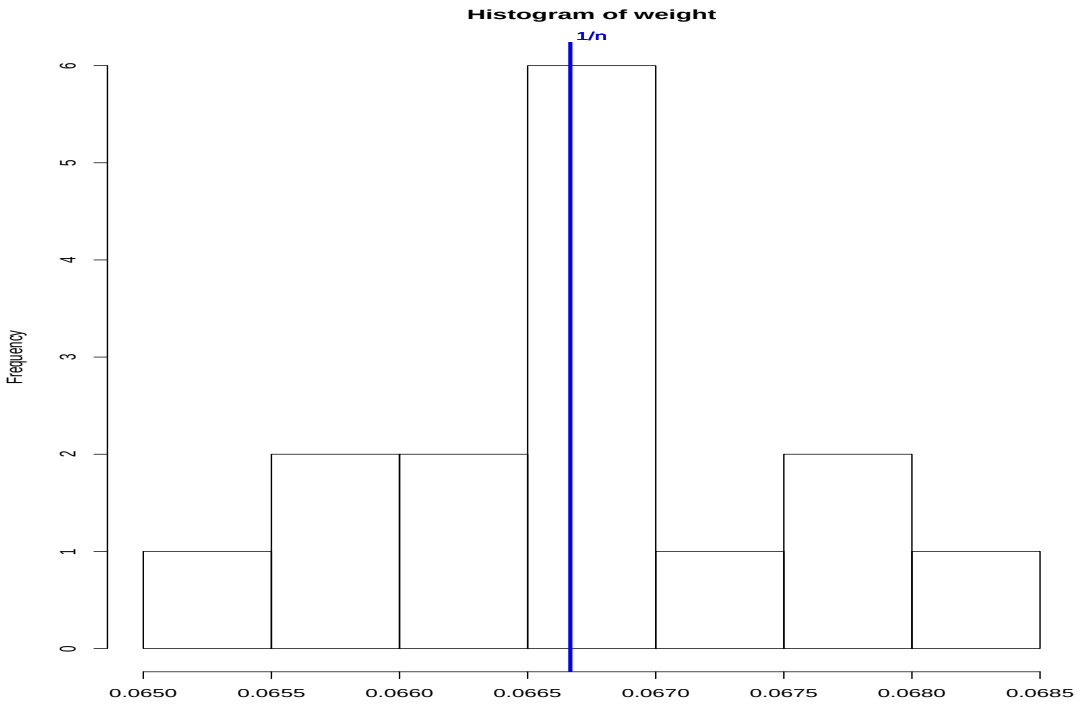

Figure 2.5: Histogram of the weights $\hat{w}_i = \frac{1}{n}\frac{1}{1+\lambda(\hat{\mu})(x_i-\hat{\mu})}$, for $j = 1, \cdots, 15$. The blue vertical line represents the value of $\frac{1}{n} = \frac{1}{15}$.

Figure 2.5 depicts the distribution of the weights when $\hat{\mu} = 2.607813$ and λ evaluated at $\hat{\mu}$. As expected, they are around the value of $\frac{1}{n}$ and they sum to one.

2.4 Illustrative Examples

In this section, we present two examples of linear regression using the Bayesian approach based on EL. We use two real datasets: cancer data (Rice, 1988) and prostate cancer data (Stamey et al., 1989).

We implement a building block of Hamiltonian Monte Carlo and Gibbs sampler to sample from Equation (2.15) in both cases. First, we sample from the distribution of $\pi(\boldsymbol{\theta}|\sigma^2, X, \boldsymbol{y})$ using the HMC approach. This method requires finding the gradients of the negative loga-

rithm empirical posterior density:

$$-log\left(p(\boldsymbol{\theta}|X,\boldsymbol{y}\right) = \sum_{i=1}^{n} \log\left[1+\boldsymbol{\lambda}^T\boldsymbol{x_i}\left(y_i-\boldsymbol{x_i}^T\boldsymbol{\theta}\right)\right] + \frac{1}{2\sigma^2}\boldsymbol{\theta}^T A^{-1}\boldsymbol{\theta}$$

$$-\frac{\partial \log\left(p(\boldsymbol{\theta}|X,\boldsymbol{y})\right)}{\partial\boldsymbol{\theta}} = -\sum_{i=1}^{n}\frac{\boldsymbol{\lambda}^T\boldsymbol{x_i}\boldsymbol{x_i}^T}{1+\boldsymbol{\lambda}^T\boldsymbol{x_i}\left(y_i-\boldsymbol{x_i}^T\boldsymbol{\theta}\right)} + \frac{1}{\sigma^2}\boldsymbol{\theta}^T A^{-1}.$$

Second, we sample σ^2 from Inverse gamma with shape a_1+1 and rate $b_1+\boldsymbol{\theta}^T A^{-1}\boldsymbol{\theta}/2$.

Example 1 : Cancer Data

In this example, we apply the linear regression to the cancer data provided by Rice (1988). We implement the Bayesian approach based on EL. Each data point is from a county in North Carolina, South Carolina, or Georgia. For each county, the number of adult white women living there in 1960 is given, as is the number of deaths due to breast cancer among adult white females from 1950 through 1969 inclusive. There are 301 counties in the dataset; a more detailed description of these data is given in Owen (1991). The response variable is the value for breast cancer mortality, and the predictor variable is the adult white female population. We implement an MCMC sampling scheme using 5000 iterations with 1000 burn-in. We use a building block of HMC and Gibbs sampler. To sample $\boldsymbol{\theta}$, we implement the HMC scheme where the step size and the number of leapfrog steps are set to $\vartheta = 0.025$ and $L = 10$, respectively. On the other hand, we use the Gibbs sampler to generate samples from the distribution of σ^2.

Figure 2.6 suggests that the slope of the adult white female population parameter has a unimodal bell-shaped distribution. Also, it appears from the trace plot in Figure 2.7 that the center of the chain is around the OLS estimate with reasonable fluctuation. This indicates that the chain is mixing well. From the autocorrelation plot in Figure 2.8, it is clear that the HMC sample has a correlation that decreases quickly as the lag increases. Table 2.2

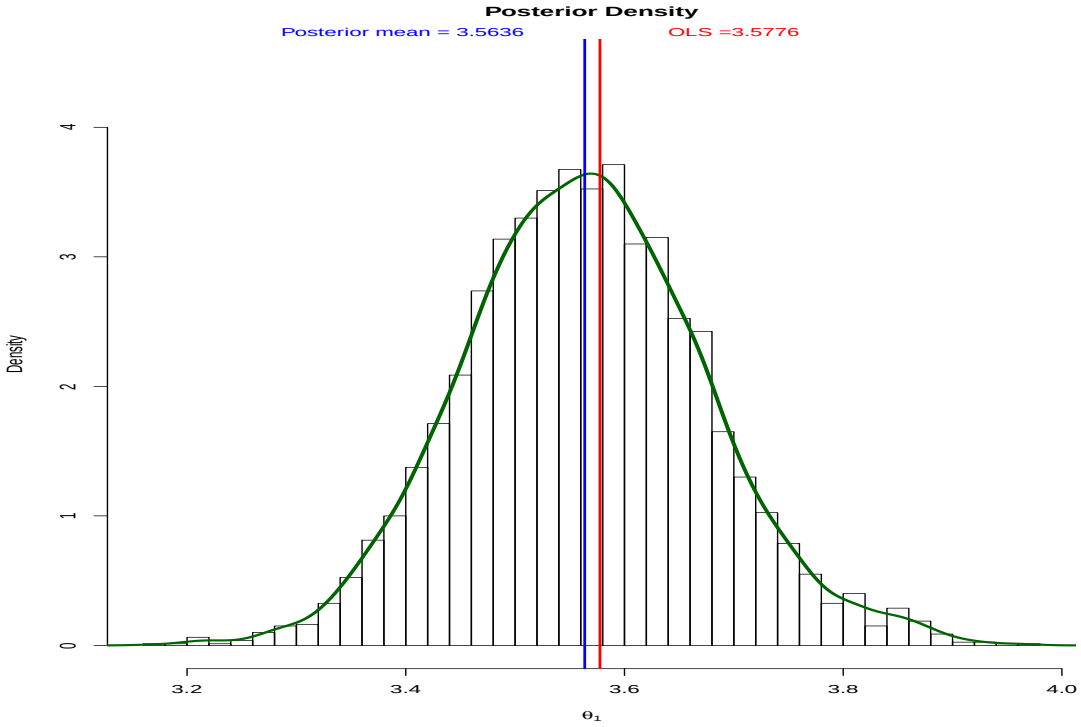

Figure 2.6: Histogram of θ along with kernel density curve.

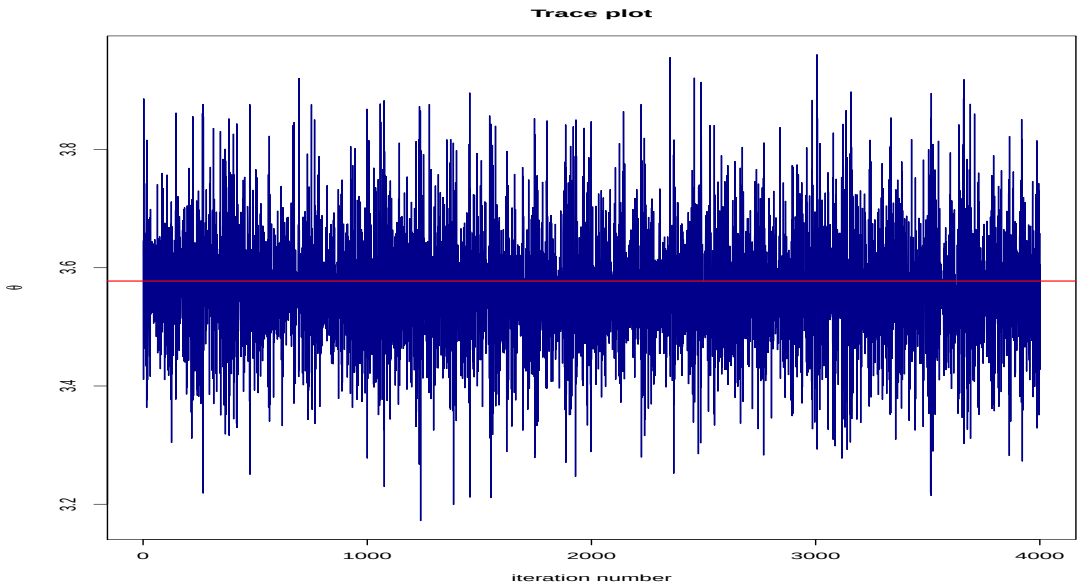

Figure 2.7: Trace plot for the parameter θ using 5000 iterations with 1000 burn-in; red line is the OLS estimate.

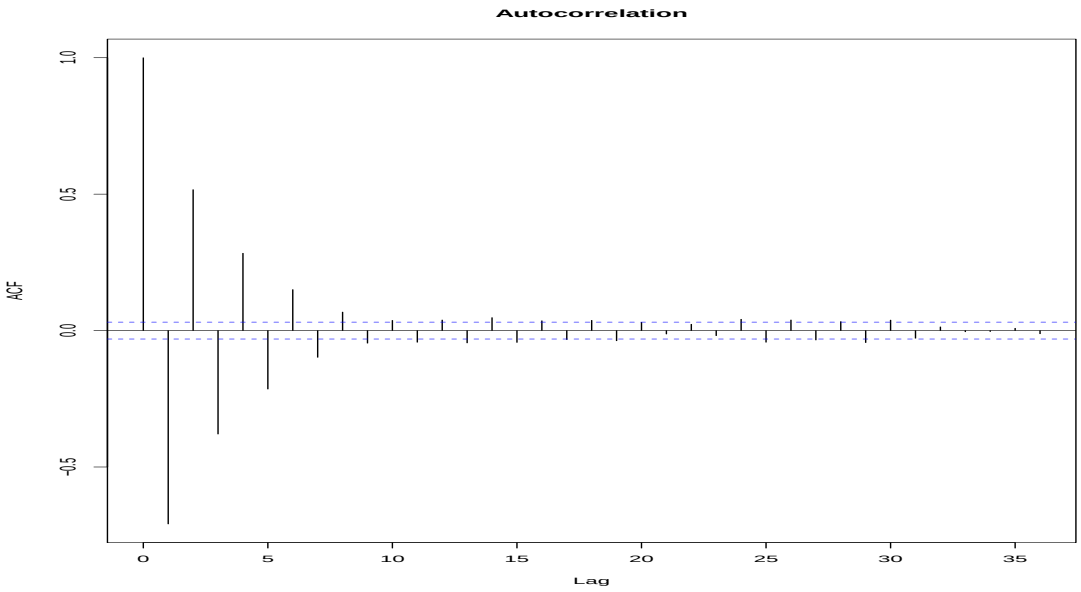

Figure 2.8: Autocorrelation plot of θ.

gives the estimate of the slope, based on the HMC, and the OLS estimate. We provide the estimate based on the posterior mean where the burn-in is set to 1000. It is evident that the posterior mean estimate is relatively close to the OLS estimate. The acceptance rate is approximately 70%, which suggests that the HMC is working efficiently. By the results presented above, we conclude that the rate of cancer deaths per 1000 population is 3.564. And because deaths were counted over 20 years, the annualized rate is $3.564/20 = 0.178$ per thousand. The intercept, $\hat{\theta_0} = -0.0526$, is obtained by subtracting the product of the average of the population and the slope from the average of deaths due to breast cancer. In addition, the first quantile, second, and third quantiles are 3.489, 3.561, and 3.635, respectively.

Our approach is semi-parametric where we did not assume any distribution to data. We obtained similar results compared to the OLS approach, but our procedure provides small standard error, which indicates that the sample of the slope of the adult white female popu-

lation is an accurate reflection of the population. Also, another advantage is that we have the entire posterior distribution of our parameter of interest, which can be summarized through mean, median, standard deviation, quantiles, etc.

Table 2.2: Posterior summary statistics for cancer data provided by Rice (1988).

Parameter	OLS estimate	Posterior mean estimate	S.E. of the posterior mean	Posterior percentiles		
				25%	50%	75%
Slope	3.578	3.564	0.006	3.489	3.561	3.635

Example 2: Prostate Cancer Data

In this example, we use the prostate cancer data from a study by Stamey et al. (1989). The data examined the correlation between the level of prostate specific antigen (PSA) and a number of clinical measures in 97 men who were about to receive a radical prostatectomy. The aim of this study is to predict the log of PSA (lpsa) from a number of measurements including log cancer volume (lcavol), log prostate weight (lweight), age, log of benign prostatic hyperplasia amount (lbph), seminal vesicle invasion (svi), log of capsular penetration (lcp), Gleason score (gleason), and percent of Gleason scores 4 or 5 (pgg45). We implement the MCMC sampling scheme using 5000 iterations and 1000 burn-in. For HMC, the step size and the number of leapfrogs are set to $\vartheta = 0.04$ and $L = 10$, respectively. The data are centered and scaled. Table **??** presents the BEL posterior estimates of the eight variables described above along with their OLS estimates, standard errors, and percentiles. The posterior means are quite close to the OLS estimates with small standard errors. The acceptance rate is approximately 62%. The trace plots of the posterior quantities are displayed in Figure 2.9. It is clear that the center of each chain is around the OLS estimate with reasonable fluctuation, which indicates the chain is mixing well. Figure 2.10 shows that all parameters

have unimodal bell-shaped distributions. The autocorrelation plots depicted in Figure 2.11 suggest that the HMC samples, for each predictor variable, have a correlation that decreases quickly as the lag increases.

Parameter	OLS estimate	Posterior mean estimate	S.E. of the posterior mean	Posterior percentiles		
				25%	50%	75%
lcavol	0.5994	0.5936	0.0080	0.5416	0.5932	0.6466
lweight	0.1955	0.2119	0.0082	0.1548	0.2066	0.2622
age	-0.1267	-0.1224	0.0066	-0.1664	-0.1231	-0.0797
lbph	0.1346	0.1180	0.0075	0.068	0.1190	0.1667
svi	0.2748	0.2732	0.0084	0.2182	0.2764	0.3301
lcp	-0.1278	-0.1178	0.0102	-0.1861	-0.1190	-0.0514
gleason	0.0282	0.0242	0.0091	-0.0365	0.0260	0.0851
pgg45	0.1106	0.1052	0.0109	0.0301	0.1024	0.1758

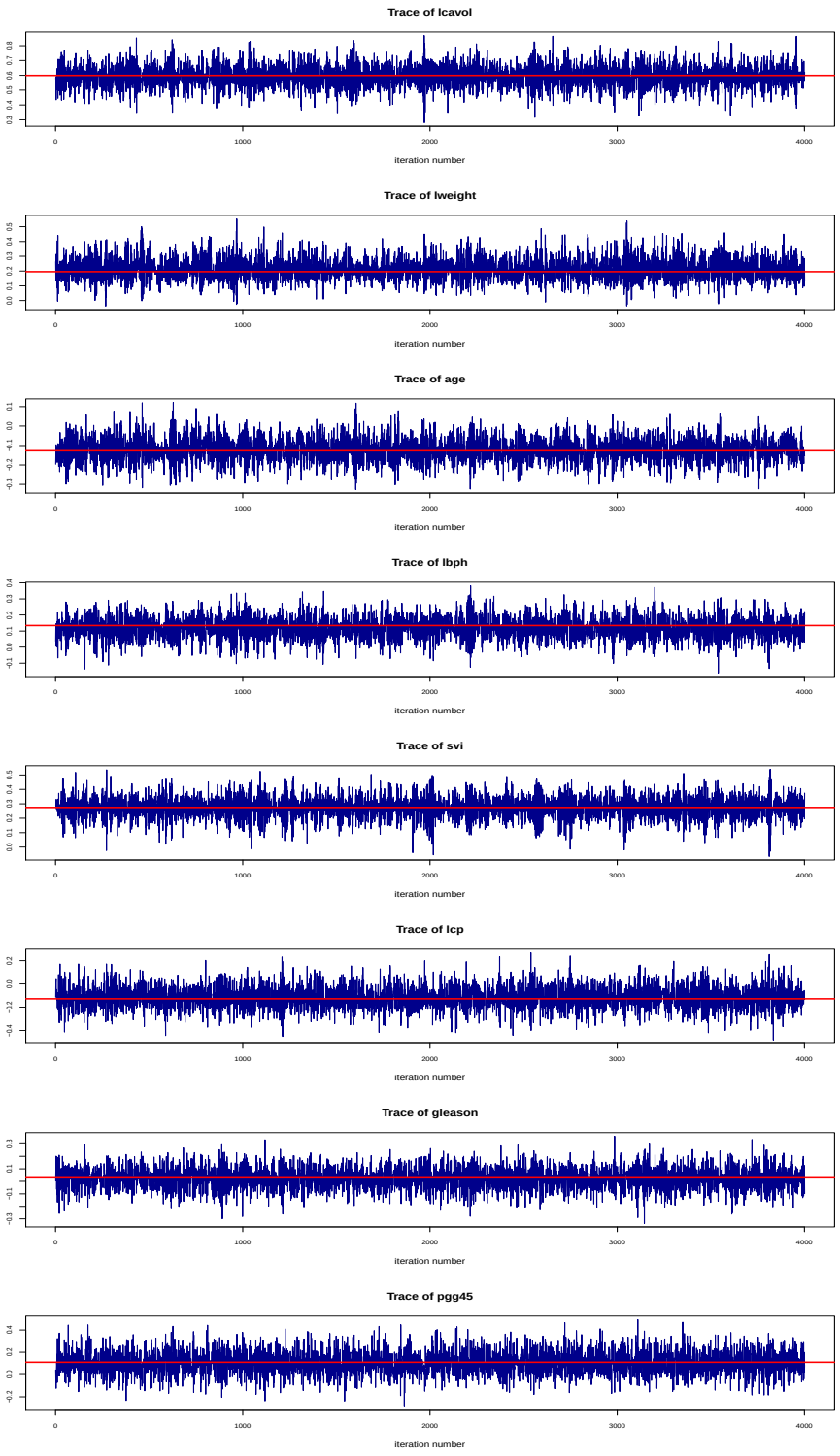

Figure 2.9: Trace plots for variables lcavol, lweight, age, lbph, svi, lcp, gleason, and pgg45 for prostate cancer data (Stamey et al., 1989). The red line in each trace plot represents the OLS estimate.

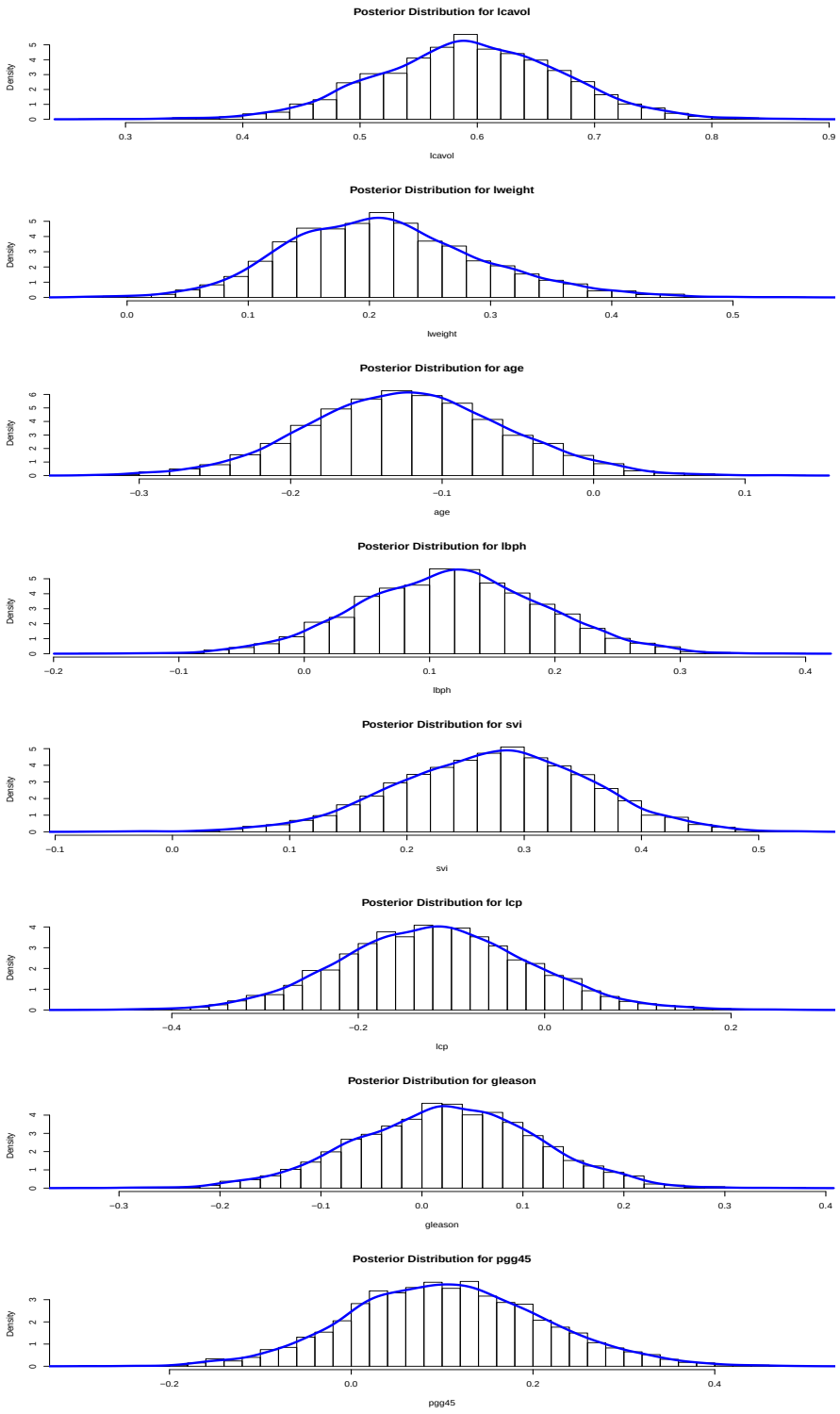

Figure 2.10: Histograms of the posterior distribution for variables lcavol, lweight, age, lbph, svi, lcp, gleason, and pgg45 along with the kernel density curve for prostate cancer data (Stamey et al., 1989).

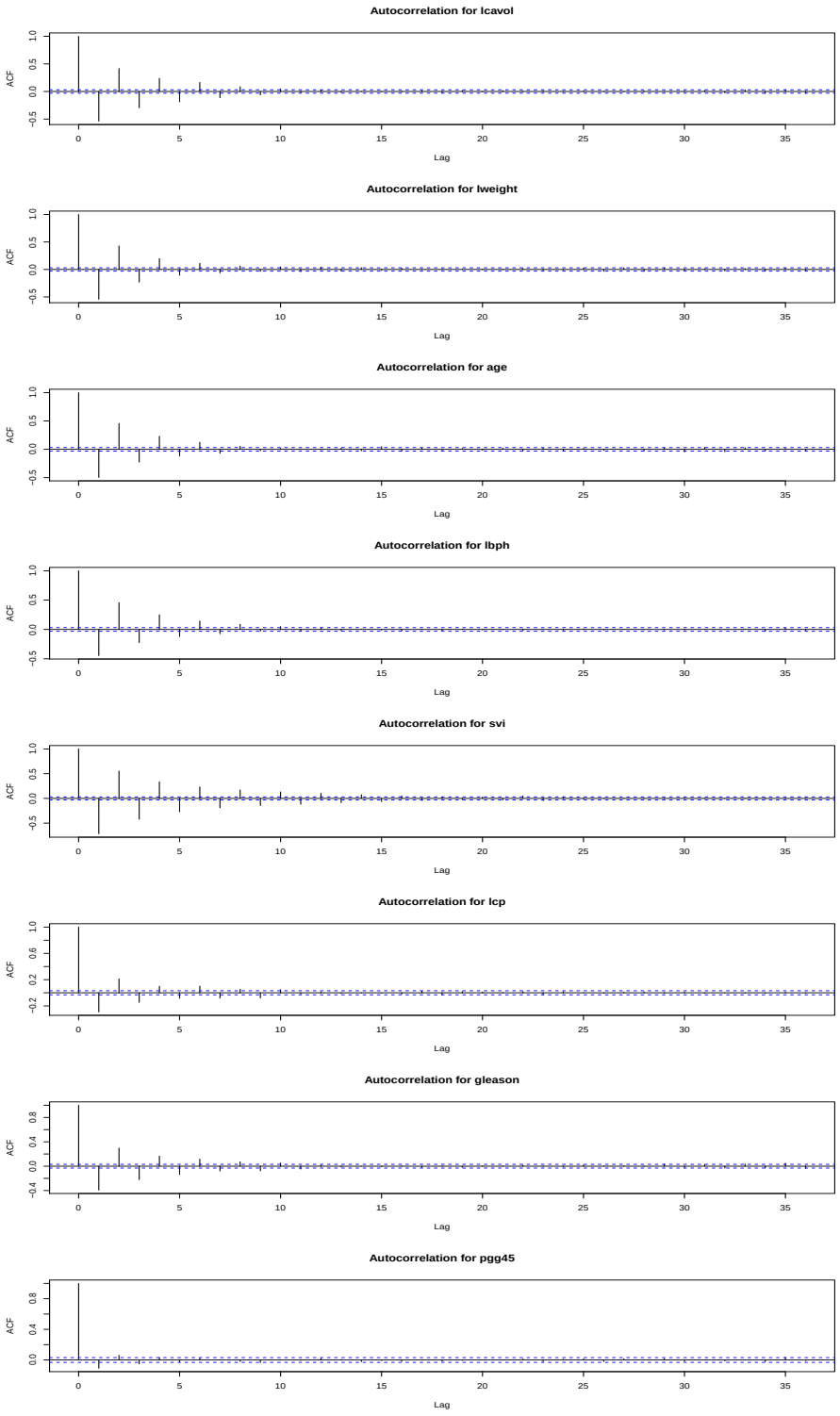

Figure 2.11: Autocorrelation plots of the posterior distribution for variables lcavol, lweight, age, lbph, svi, lcp, gleason, and pgg45 for prostate cancer data.

2.5 Summary

In this Chapter, we proposed an alternative approach to linear regression by using the Bayesian method based on EL. The implementation of MCMC algorithms such as the Gibbs sampler and Metropolis-Hastings was challenging. The resulting posterior distribution lacked an analytic form, and therefore we could not apply the Gibbs sampler. In addition, due to the intricacy of the support of the posterior empirical density, implementation of Metropolis-Hastings is a daunting task. We used instead the Hamiltonian Monte Carlo algorithm that exploits information from the gradients to avoid random walk and move faster toward regions of high density. The implementation of HMC is easy as it only requires the derivation of the gradient. It is not recommended to use the numerical approach to compute the gradient because it makes the algorithm too slow to calculate. Therefore, we should be cautious with its derivation.

In the next Chapter, we prove that the maximum empirical likelihood estimator is consistent. Also, we show that if we place a normal prior on $\boldsymbol{\theta}$, and under certain assumptions, the posterior EL for regression parameters is asymptotically normal. This applies to the linear regression, ridge regression, and lasso regression. As discussed previously, the penalized regression has a close connection to BEL for linear regression. The penalty term is presented in the form of a hyperprior. Note that in the lasso case we use the Laplace distribution as a prior using the representation of Andrews and Mallows (1974), which has the form of a mixture of normals with an exponential mixing density.

Chapter 3

Properties of the Regression Parameters Under Bayesian Empirical Likelihood

In this Chapter, we show that the maximum empirical likelihood estimator, $\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmax}} - \sum_{i=1}^{n} \log\left(1 + \boldsymbol{\lambda}^T g(X_i, y_i, \boldsymbol{\theta})\right)$, is consistent, and the posterior empirical density for $\boldsymbol{\theta}$, under certain conditions, and as $n \to \infty$, is asymptotically normal. Also, we show that the asymptotic distribution of minus the logarithm-posterior EL is chi-square with p degrees of freedom, where p is the number of covariates. The consistency is an asymptotic property, which is important because it guarantees that the estimator becomes more precise and accurate when we collect more data.

The posterior empirical density of $\boldsymbol{\theta}$ does not have a closed form, and a good approximation is required because the asymptotic distribution has a theoretical importance. For instance, one can derive Bayesian credible regions. Now, we introduce the definition of a consistent estimator. Suppose a random sample $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n) = \boldsymbol{\theta}_n$ has a joint density $\pi(\boldsymbol{\theta}_n|X)$. We denote an estimator $T(\boldsymbol{\theta})$ for a sample $\boldsymbol{\theta}_n$ by $T_n = T(\boldsymbol{\theta}_n)$. In studying the behavior of T_n for large sample size, we will consider the sequence of estimator $\{T_n\}$. For

example, if $T(\boldsymbol{\theta})$ is the sample mean, then the sequence of estimators is

$$\left\{\theta_1, \frac{\theta_1+\theta_2}{2}.\frac{\theta_1+\theta_2+\theta_3}{3}, \cdots, \bar{\theta}_n, \cdots\right\}.$$

Consistency is the property of a sequence of estimators rather than a single estimator, although we say "consistent estimator".

3.1 Consistency of the Maximum Empirical Likelihood Estimator

To prove the consistency of the MELE, Yang and He (2012) used the theorem of consistency of M-estimators in Van der Vaart (1989), the quadratic expansion approximating the EL function (Molanes Lopez et al., 2009), and P-measurable class of measurable functions (Kosorok, 2008). Our proof uses the theorem of M-estimators in Van der Vaart (1989), but it is completely different than the approach of Yang and He (2012).

Let Θ denote the parameter space. We assume that Θ is compact. Let $\boldsymbol{\theta_0}$ be the true parameter. Assume that $f(X, \boldsymbol{y}, \boldsymbol{\theta}) = -\log\left(1 + \boldsymbol{\lambda}^T g(\boldsymbol{\theta})\right)$ is a continuous function of $\boldsymbol{y}$ at each $\boldsymbol{\theta}$ and $g(X, \boldsymbol{y}, \boldsymbol{\theta})$ is the estimating equations. Assume that there exists a dominating function $d(X, \boldsymbol{y})$ such that $E\left[d(X, \boldsymbol{y})\right] < \infty$, and $||f(X, \boldsymbol{y}, \boldsymbol{\theta})|| \leq d(X, \boldsymbol{y})$. Then the MELE

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\, R_n(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}} - \sum_{i=1}^{n} \log\left(1 + \boldsymbol{\lambda}^T g(X_i, y_i, \boldsymbol{\theta})\right)$$

is a consistent estimator of $\boldsymbol{\theta_0}$.

Proof. To prove the consistency of $\boldsymbol{\theta}$, we use Theorem (5.7) of Van der Vaart (1989):

Theorem 3.1.1. *Let M_n be random functions and let M be a fixed function of $\boldsymbol{\theta}$ such that*

for every $\varsigma > 0$

$$\sup_{\boldsymbol{\theta}\in\Theta}|M_n(\boldsymbol{\theta}) - M(\boldsymbol{\theta})| \xrightarrow{p} 0$$

$$\sup_{||\boldsymbol{\theta}-\boldsymbol{\theta}_0||>\varsigma} M(\boldsymbol{\theta}) < M(\boldsymbol{\theta}_0).$$

Then any sequence of estimators $\hat{\boldsymbol{\theta}}_n$ *with* $M_n(\hat{\boldsymbol{\theta}_n}) \geq M_n(\boldsymbol{\theta}_0) - o_P(1)$ *converges in probability to* $\boldsymbol{\theta}_0$.

We have

$$R_n(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \log\left(1 + \boldsymbol{\lambda}^T g(X_i, y_i, \boldsymbol{\theta})\right) = -\sum_{i=1}^{n} \log\left(1 + \boldsymbol{\lambda}^T g_i(\boldsymbol{\theta})\right)$$

where $g_i(\boldsymbol{\theta}) = g(X_i, y_i, \boldsymbol{\theta})$ and $\boldsymbol{\lambda}$ satisfies:

$$\sum_{i=1}^{n} \frac{g_i(\boldsymbol{\theta})}{1 + \boldsymbol{\lambda}^T g_i(\boldsymbol{\theta})} = \mathbf{0}.$$

Let $M_n(\boldsymbol{\theta}) = \dfrac{R_n(\boldsymbol{\theta})}{n} = -\dfrac{1}{n}\sum_{i=1}^{n} \log\left(1 + \boldsymbol{\lambda}^T g_i(\boldsymbol{\theta})\right)$, and $M(\boldsymbol{\theta})$ be the expected value of $M_n(\boldsymbol{\theta})$. That is,

$$\begin{aligned} M(\boldsymbol{\theta}) &= E(M_n(\boldsymbol{\theta})) \\ &= -\frac{1}{n}\sum_{i=1}^{n} E\left[\log\left(1 + \boldsymbol{\lambda}^T g_i(\boldsymbol{\theta})\right)\right] \\ &= -E\left[\log\left(1 + \boldsymbol{\lambda}^T g_n(\boldsymbol{\theta})\right)\right], \text{ by i.i.d} \\ &= -E\left[\log\left(1 + \boldsymbol{\lambda}^T g(\boldsymbol{\theta})\right)\right] \end{aligned}$$

where $g(\boldsymbol{\theta}) = g_n(X_n, y_n, \boldsymbol{\theta})$ and $\boldsymbol{\lambda}$ satisfies:

$$E\left(\frac{g(\boldsymbol{\theta})}{1 + \boldsymbol{\lambda}^T g(\boldsymbol{\theta})}\right) = \mathbf{0}. \tag{3.1}$$

Now we show that for $\varsigma > 0$

$$\sup_{||\boldsymbol{\theta}-\boldsymbol{\theta_0}||>\varsigma} M(\boldsymbol{\theta}) < M(\boldsymbol{\theta_0}).$$

We have $M(\boldsymbol{\theta_0}) = -E\left[\log\left(1+\boldsymbol{\lambda}^T g(\boldsymbol{\theta_0})\right)\right]$ and by definition $E(g(\boldsymbol{\theta_0})) = \mathbf{0}$, which implies that $\boldsymbol{\lambda}^T = \mathbf{0}$. Therefore, $M(\boldsymbol{\theta_0}) = -E\left[\log\left(1+0\right)\right] = 0$. Too see this, we use Chen and Van Keilegom (2009)'s approach. The equation $\dfrac{g_i(\boldsymbol{\theta})}{1+\boldsymbol{\lambda}^T g_i(\boldsymbol{\theta})} = \mathbf{0}$ can be simplified to

$$\frac{1}{n}\sum_{i=1}^{n} g_i(\boldsymbol{\theta})\left(1+\boldsymbol{\lambda}^T g_i(\boldsymbol{\theta})\right) + \frac{1}{n}\sum_{i=1}^{n} g_i(\boldsymbol{\theta})\frac{\boldsymbol{\lambda}^T g_i(\boldsymbol{\theta}) g_i(\boldsymbol{\theta})^T \boldsymbol{\lambda}}{1+\boldsymbol{\lambda}^T g_i(\boldsymbol{\theta})} = \mathbf{0}.$$

The last term on the left hand side is $O_p(1/n)$, which is negligible relative to the first term. Therefore,

$$\begin{aligned}\boldsymbol{\lambda} &= \frac{\sum_{i=1}^n g_i(\boldsymbol{\theta})}{\sum_{i=1}^n g_i(\boldsymbol{\theta}) g_i(\boldsymbol{\theta})^T} + o_p(n^{-1/2}) \\ &= \frac{\sum_{i=1}^n X_i\left(y_i - X_i^T\boldsymbol{\theta}\right)}{\sum_{i=1}^n \left[X_i\left(y_i - X_i^T\boldsymbol{\theta}\right)\right]\left[X_i\left(y_i - X_i^T\boldsymbol{\theta}\right)\right]^T}.\end{aligned}$$

Hence when $\boldsymbol{\theta} = \boldsymbol{\theta_0}$, $\boldsymbol{\lambda} = \mathbf{0}$.

Let $\Gamma(\boldsymbol{\theta}, r)$ denote an open sphere centered at $\boldsymbol{\theta}$ with radius r such that for $\boldsymbol{\theta} \neq \boldsymbol{\theta_0}$:

$$\begin{aligned}\lim_{r\to 0}\sup_{\boldsymbol{\theta}^*\in\Gamma(\boldsymbol{\theta},r)} -E\left[\log\left(1+\boldsymbol{\lambda}^T g(\boldsymbol{\theta}^*)\right)\right] &= \sup_{\boldsymbol{\theta}^*\in\Gamma(\boldsymbol{\theta},0)} -E\left[\log\left(1+\boldsymbol{\lambda}^T g(\boldsymbol{\theta}^*)\right)\right] \\ &= -E\left[\log\left(1+\boldsymbol{\lambda}^T g(\boldsymbol{\theta})\right)\right] \\ &= E\left[\log\left(\frac{1}{1+\boldsymbol{\lambda}^T g(\boldsymbol{\theta})}\right)\right].\end{aligned}$$

We know that $log(x) \leq x - 1$ for all $x > 0$. Thus for $\dfrac{1}{1+\boldsymbol{\lambda}^T g(\boldsymbol{\theta})} > 0$, we have:

$$\begin{aligned} \log\left(\frac{1}{1+\boldsymbol{\lambda}^T g(\boldsymbol{\theta})}\right) &\leq \frac{1}{1+\boldsymbol{\lambda}^T g(\boldsymbol{\theta})} - 1 \\ &= \frac{-\boldsymbol{\lambda}^T g(\boldsymbol{\theta})}{1+\boldsymbol{\lambda}^T g(\boldsymbol{\theta})}. \end{aligned} \tag{3.2}$$

By applying the expectation, we obtain

$$\begin{aligned} \sup_{||\boldsymbol{\theta}-\boldsymbol{\theta_0}||>\varsigma} M(\boldsymbol{\theta}) &= -E\left[\log\left(1+\boldsymbol{\lambda}^T g(\boldsymbol{\theta})\right)\right] \\ &= E\left[\log\left(\frac{1}{1+\boldsymbol{\lambda}^T g(\boldsymbol{\theta})}\right)\right] \\ &\leq -\boldsymbol{\lambda}^T E\left[\log\left(\frac{g(\boldsymbol{\theta})}{1+\boldsymbol{\lambda}^T g(\boldsymbol{\theta})}\right)\right] \\ &= 0 \text{ (by Equation (3.1))}. \end{aligned}$$

We showed that $M(\boldsymbol{\theta_0}) = 0$. Therefore, $\sup_{||\boldsymbol{\theta}-\boldsymbol{\theta_0}||>\varsigma} M(\boldsymbol{\theta}) < 0 = M(\boldsymbol{\theta_0})$ for $\boldsymbol{\theta} \neq \boldsymbol{\theta_0}$. Thus, the second condition of Theorem (5.7) of Van der Vaart (1989) is satisfied.

Now, we need to show that the first condition in Van der Vaart's (1998) theorem is fulfilled. Under the following assumptions:

1 Θ is compact.

2 $f(X, \boldsymbol{y}, \boldsymbol{\theta}) = -\log\left(1+\boldsymbol{\lambda}^T g(\boldsymbol{\theta})\right)$ is continuous at each $\boldsymbol{\theta} \in \Theta$ for almost all $\boldsymbol{y}$'s and measurable function of $\boldsymbol{y}$ at each $\boldsymbol{\theta}$. Actually, the function $\log\left(1+\boldsymbol{\lambda}^T g(\boldsymbol{\theta})\right)$ is continuous and defined when $1+\boldsymbol{\lambda}^T g(\boldsymbol{\theta}) > 0$.

3 There exists a dominating function $d(X, \boldsymbol{y})$ such that $E(d(X, \boldsymbol{y})) < \infty$ and $||f(X, \boldsymbol{y}, \boldsymbol{\theta})|| < d(X, \boldsymbol{y})$. From (3.2), one can see that $f(X, \boldsymbol{y}, \boldsymbol{\theta})$ is dominated by

$$d(X, \boldsymbol{y}) = \frac{-\boldsymbol{\lambda}^T g(\boldsymbol{\theta})}{1+\boldsymbol{\lambda}^T g(\boldsymbol{\theta})},$$

and by uniform law of large numbers (Jennrich, 1969), we have that $E(f(X, \boldsymbol{y}, \boldsymbol{\theta}))$ is continuous in $\boldsymbol{\theta}$, and

$$\begin{aligned}\sup_{\boldsymbol{\theta}\in\Theta}||M_n(\boldsymbol{\theta}) - M(\boldsymbol{\theta})|| &= \sup_{\boldsymbol{\theta}\in\Theta}|| - \frac{1}{n}\sum_{i=1}^{n}\log\left(1+\boldsymbol{\lambda}^T g_i(\boldsymbol{\theta})\right) - \left(-E\left[\log\left(1+\boldsymbol{\lambda}^T g(\boldsymbol{\theta})\right)\right]\right)|| \\ &= \sup_{\boldsymbol{\theta}\in\Theta}||\frac{1}{n}\sum_{i=1}^{n} f(X_i, y_i, \boldsymbol{\theta}) - E\left[f\left(X, \boldsymbol{y}, \boldsymbol{\theta}\right)\right]|| \xrightarrow{a.s} 0\end{aligned}$$

We know that convergence almost surely implies convergence in probability. Therefore, $\sup_{\boldsymbol{\theta}\in\Theta}||M_n(\boldsymbol{\theta}) - M(\boldsymbol{\theta})|| \xrightarrow{p} 0$. Then, $\hat{\boldsymbol{\theta}}$ is a consistent estimator. □

To prove the first condition of Van der Vart's theorem, we used the uniform law of large number that implies the convergence in probability. In contrast, Yang and He (2012) relied on the empirical process theory and the concept of the P-measurable class of measurable function (Kosorok, 2008). Yang and He (2012) assumed that the estimating equation is twice continuously differentiable and applied the Taylor expansion to prove the second condition of Van der Vart's theorem. In contrast, we used the concept of a bounded function. In addition, Yang and He (2012)'s proof requires assumptions about the smoothness of the estimating equation because it involves an indicator function.

Next, we demonstrate that, under certain regularity conditions and $n \to \infty$, the posterior empirical likelihood is asymptotically normal.

3.2 Asymptotic Distribution of the Posterior Empirical Likelihood

First, assume that we place a normal prior on $\boldsymbol{\theta}$ with mean $\boldsymbol{\theta_0}$ and covariance matrix A. We assume that A is known and is positive definite. Under certain regularity conditions, and as $n \to \infty$, the posterior distribution of $\boldsymbol{\theta}$ converges to normal, with mean m_n and covariance

J_n, where

$$J_n = J(\hat{\boldsymbol{\theta}_n}) + A^{-1}$$
$$m_n = J_n^{-1}\left[A^{-1}\boldsymbol{\theta_0} + J(\hat{\boldsymbol{\theta}_n})\hat{\boldsymbol{\theta}_n}\right]$$

and $\hat{\boldsymbol{\theta}_n}$ is the profile maximum empirical likelihood estimate of $\boldsymbol{\theta}$, $\boldsymbol{\theta_0}$ is the prior mean, and $J(\hat{\boldsymbol{\theta}_n})$ is minus the second derivative of the log empirical likelihood evaluated at $\hat{\boldsymbol{\theta}_n}$.

Proof. The posterior empirical distribution of $\boldsymbol{\theta}$ is

$$\begin{aligned}\pi(\boldsymbol{\theta}|X,\boldsymbol{y}) &\propto R(\boldsymbol{\theta})\pi(\boldsymbol{\theta}).\\ &\propto \exp\left(\log\prod_{i=1}^{n}\left[\frac{1}{1+\boldsymbol{\lambda^T x_i}(y_i-\boldsymbol{x_i}^T\boldsymbol{\theta})}\right]-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta_0})^T A^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta_0})\right).\\ &= \exp\left(\log\pi(X,\boldsymbol{y}|\boldsymbol{\theta})-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta_0})^T A^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta_0})\right)\end{aligned}$$

where,

$$\pi(X,\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n}\frac{1}{1+\boldsymbol{\lambda^T x_i}(y_i-\boldsymbol{x_i}^T\boldsymbol{\theta})}.$$

Similar to Bernardo and Smith (1994), we expand the logarithm term about its maxima $\hat{\boldsymbol{\theta}_n}$, assumed to be determined by setting $\triangledown\log\pi(X,\boldsymbol{y}|\boldsymbol{\theta}) = \mathbf{0}$, we obtain:

$$\log\pi(\boldsymbol{\theta}|X,\boldsymbol{y}) = \log\pi(X,\boldsymbol{y}|\hat{\boldsymbol{\theta}_n}) - \frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}_n})^T A^{-1}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}_n}) + R_n$$

where R_n is the reminder and is small for large n.
In addition, we have: $\log\left(\pi(X,\boldsymbol{y}|\hat{\boldsymbol{\theta}_n})\right) = -\frac{1}{2}\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}_n}\right)^T J(\hat{\boldsymbol{\theta}_n})\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}_n}\right)$ where :

$$J(\hat{\boldsymbol{\theta}_n}) = \left(-\frac{\partial^2\log\pi(X,\boldsymbol{y}|\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right)_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}_n}}.$$

If we assume n is large and ignore constants of proportionality, we have:

$$\begin{aligned}
\pi(\boldsymbol{\theta}|X,\boldsymbol{y}) &\propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}_n}\right)^T J(\hat{\boldsymbol{\theta}_n})\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}_n}\right)-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta_0})^T A^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta_0})\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}_n}\right)^T J(\hat{\boldsymbol{\theta}_n})\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}_n}\right)+(\boldsymbol{\theta}-\boldsymbol{\theta_0})^T A^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta_0})\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\boldsymbol{\theta^T} J(\hat{\boldsymbol{\theta}_n})\boldsymbol{\theta}-\boldsymbol{\theta^T} J(\hat{\boldsymbol{\theta}_n})\hat{\boldsymbol{\theta}_n}-\hat{\boldsymbol{\theta}_n} J(\hat{\boldsymbol{\theta}_n})\boldsymbol{\theta}+\hat{\boldsymbol{\theta}_n} J(\hat{\boldsymbol{\theta}_n})\hat{\boldsymbol{\theta}_n}+\boldsymbol{\theta^T} A^{-1}\boldsymbol{\theta}-\boldsymbol{\theta^T} A^{-1}\boldsymbol{\theta_0}\right)\right\} \\
&\quad + \exp\left\{-\frac{1}{2}\left(-\boldsymbol{\theta_0^T} A^{-1}\boldsymbol{\theta}+\boldsymbol{\theta_0^T} A^{-1}\boldsymbol{\theta_0}\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\theta^T}\left[J(\hat{\boldsymbol{\theta}_n})+A^{-1}\right]\boldsymbol{\theta}-2\boldsymbol{\theta^T} J(\hat{\boldsymbol{\theta}_n})\hat{\boldsymbol{\theta}_n}+2\boldsymbol{\theta^T} A^{-1}\boldsymbol{\theta_0}\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\boldsymbol{\theta^T}\left[J(\hat{\boldsymbol{\theta}_n})+A^{-1}\right]\boldsymbol{\theta}-2\boldsymbol{\theta^T}\left[J(\hat{\boldsymbol{\theta}_n})\hat{\boldsymbol{\theta}_n}+A^{-1}\boldsymbol{\theta_0}\right]\right)\right\}.
\end{aligned}$$

Setting $J_n = J(\hat{\boldsymbol{\theta}_n}) + A^{-1}$ and $m_n = J_n^{-1}\left[A^{-1}\boldsymbol{\theta_0} + J(\hat{\boldsymbol{\theta}_n})\hat{\boldsymbol{\theta}_n}\right]$, we have:

$$\pi(\boldsymbol{\theta}|X,\boldsymbol{y}) \propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\theta^T} J_n\boldsymbol{\theta}-2\boldsymbol{\theta^T} J_n m_n\right)\right\}.$$

We complete the square above by adding and subtracting $m_n^T J_n m_n$. Therefore,

$$\pi(\boldsymbol{\theta}|X,\boldsymbol{y}) \propto \exp\left\{-\frac{1}{2}\left(\left[\boldsymbol{\theta}-m_n\right]^T J_n\left[\boldsymbol{\theta}-m_n\right]\right)\right\},$$

is the kernel of $N_p\left(m_n,\ J_n\right)$, with m_n and J_n defined above. □

Next, we derive the Bayesian credible intervals. First, we need to find the asymptotic distribution of minus the logarithm of the posterior empirical likelihood.

3.3 Bayesian Credible Regions

Bayesian credible regions are intervals in the domain of the posterior probability distribution. Recall that the frequentist confidence intervals do not have straightforward probabilistic

interpretations; however, the Bayesian credible regions can be interpreted as having a high probability of containing the unknown quantity. In this section, we derive the Bayesian credible intervals for the posterior empirical distribution of $\boldsymbol{\theta}$. We prove, under certain regularity conditions, and as $n \to \infty$, that $-2\log(\pi(\boldsymbol{\theta}|X,\boldsymbol{y}))$ converges in distribution to χ^2_p as $n \to \infty$.

Proof. We need the following theorem:

Theorem 3.3.1. *If $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a vector of order p and $\boldsymbol{\Sigma}$ is positive definite, then*

$$(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}) \xrightarrow[n\to\infty]{} \chi^2_p.$$

We showed, under certain regularity conditions and as $n \to \infty$, that $\boldsymbol{\theta} \sim N(m_n, J_n)$. Therefore, $-2\log(\pi(\boldsymbol{\theta}|X,\boldsymbol{y})) \propto (\boldsymbol{\theta}-m_n)^T J_n (\boldsymbol{\theta}-m_n)$. Now, it suffices to show that $J_n = A^{-1} + J(\hat{\boldsymbol{\theta}_n})$ is positive definite matrix. A^{-1} is positive definite because , by assumption, A is positive definite. Now, we compute the second derivative of negative $\log \pi(X,\boldsymbol{y}|\boldsymbol{\theta})$:

$$\frac{\partial}{\partial \boldsymbol{\theta}^T}\left(\sum_{i=1}^{n} \log\left[1+\lambda^T \boldsymbol{x_i}(y_i - \boldsymbol{x_i}^T\boldsymbol{\theta})\right]\right) = -\sum_{i=1}^{n} \frac{\lambda^T \boldsymbol{x_i}\boldsymbol{x_i}^T}{1+\lambda^T \boldsymbol{x_i}(y_i-\boldsymbol{x_i}^T\boldsymbol{\theta})}$$
$$\frac{\partial}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\left(-\sum_{i=1}^{n} \log\left[1+\lambda^T \boldsymbol{x_i}(y_i - \boldsymbol{x_i}^T\boldsymbol{\theta})\right]\right) = \sum_{i=1}^{n} \frac{(\boldsymbol{x_i}\boldsymbol{x_i}^T)^T \lambda\lambda^T \boldsymbol{x_i}\boldsymbol{x_i}^T}{(1+\lambda^T \boldsymbol{x_i}(y_i-\boldsymbol{x_i}^T\boldsymbol{\theta}))^2} > 0$$

because the denominator is positive and the numerator has a quadratic form. Therefore, it implies that J_n is positive definite because the sum of two positive definite matrices is positive definite. Thus, $-2\log(\pi(\boldsymbol{\theta}|X,\boldsymbol{y})) \xrightarrow[n\to\infty]{D} \chi^2_p$. □

For $0 < \alpha < 1$, the property presented above provides an asymptotic justification for tests that reject the value of $\boldsymbol{\theta}$ at level α, when $-2log(\boldsymbol{\theta}|X,\boldsymbol{y}) > \chi^{2,1-\alpha}_p$. The unrejected values of $\boldsymbol{\theta}$ form a 100(1-α)% Bayesian empirical credible regions. For numbers $0 < \alpha_1 < \alpha_2 < 1$ where $\alpha = \alpha_1 + \alpha_2$, we find quantiles $0 < c_1 < c_2 < \infty$ of the χ^2_p distribution that satisfy

$p\left[\chi_p^{2,1-\alpha} \leq c_j\right] = \alpha_j$, for $j = 1, 2$; then

$$\begin{aligned}
\alpha_2 - \alpha_1 &= p\left[c_1 < -2\log(\boldsymbol{\theta}) < c_2 | X, \boldsymbol{y}\right] \\
&= p\left[-\frac{c_2}{2} < \log(\boldsymbol{\theta}) < -\frac{c_1}{2} | X, \boldsymbol{y}\right] \\
&= p\left[e^{-c_2/2} < \boldsymbol{\theta} < e^{-c_1/2} | X, \boldsymbol{y}\right].
\end{aligned}$$

The shortest possible interval enclosing $(1-\alpha)\%$ of the posterior mass is known as the Highest Posterior Density (HPD) confidence interval. Usually, HPDs are found by a numerical search. To find the HPD, we use function *hdi* in package *HDInterval* (Meredith and Kruschke, 2016).

3.4 Example

We apply BEL for linear regression to the prostate cancer data introduced in Section 2.4. We run an MCMC sampling scheme with 5000 iterations. For the HMC algorithm, the step size and the number of leapfrogs are 0.04 and 10, respectively. The posterior inferences about $\boldsymbol{\theta}$ are exhibited in Table 3.1. The second, third, fourth, and fifth columns of the Table represent the posterior mean, the highest 95% probability density intervals, the 95% equal-tailed credible regions, and the 95% confidence intervals of each clinical variable, respectively. For instance, the posterior mean of the slope of age is -0.1224 and the 95% highest posterior density interval is [-0.2492, 0.0099]. That is, we are 95% sure that the value of the slope of age is between -0.2492 and 0.0099. Note that the predictor age has the shortest range. In contrast, the predictor pgg45 has the highest range. Figure 3.1 depicts the posterior distributions of 5000 draws for each clinical variable along with kernel density. The vertical lines in blue are the lower and upper values of the highest density interval. One can conclude that these intervals are approximately symmetric. Consequently, the 95% equal-tailed credible regions are quite similar to the 95% HPD intervals.

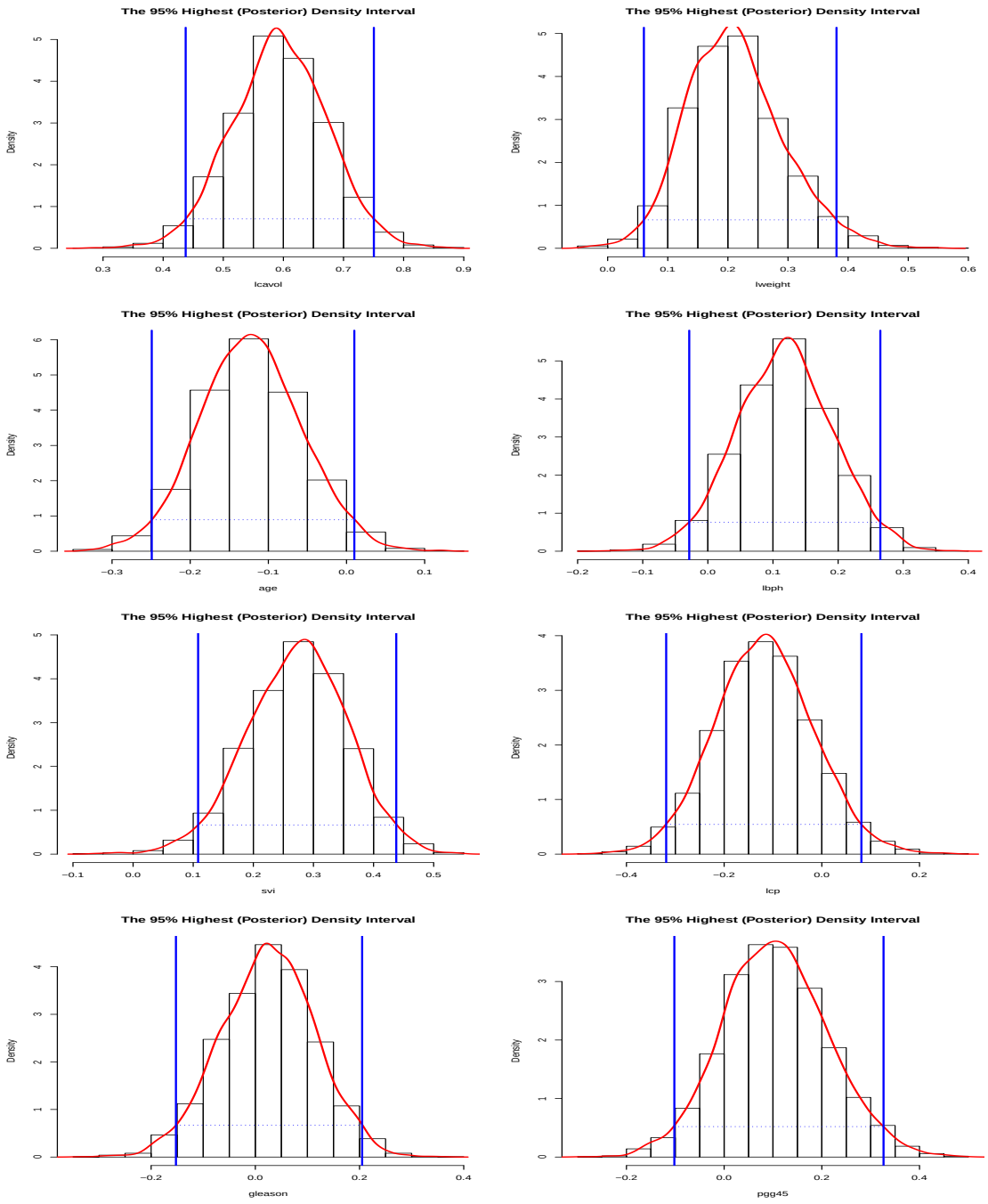

Figure 3.1: The 95% highest (posterior) density region for each clinical predictor in the prostate data (Stamey et al., 1989).

Table 3.1: Summaries of the posterior distribution of coefficients in linear regression using the prostate cancer data (Stamey et al., 1989), along with the 95% highest (posterior) density intervals, the 95% equal-tailed credible regions, and the 95% confidence intervals.

Predictor variable	Posterior mean for θ_j	95% HPD for θ_j	95% equal-tailed credible region for θ_j	95% confidence interval for θ_j
lcavol	0.5936	[0.4375, 0.7505]	[0.4419, 0.7488]	[0.4220, 0.7767]
lweight	0.2119	[0.0603, 0.3807]	[0.0599, 0.3736]	[0.0510, 03401]
age	-0.1224	[-0.2492, 0.0099]	[-0.2408, 0.0144]	[-0.2690, 0.0157]
lbph	0.1180	[-0.0286, 0.2648]	[-0.0200, 0.2683]	[-0.0106, 0.2797]
svi	0.2732	[0.1082, 0.4378]	[0.1137, 0.4370]	[0.1017, 0.4479]
lcp	-0.1178	[-0.3187, 0.0810]	[-0.3113, 0.0812]	[-0.3456, 0.0901]
gleason	0.0242	[-0.1524, 0.2049]	[-0.1451, 0.2049]	[-0.1664, 0.2229]
pgg45	0.1052	[-0.1019, 0.3263]	[-0.0964, 0.3229]	[-0.1029, 0.3240]

Chapter 4

Bayesian Empirical Likelihood for Lasso and Ridge Regression

4.1 Introduction

The objective of statistical inference is to find estimates that improve prediction accuracy and model interpretability. Constantly, when we have many predictors, certain variables are highly correlated and using the ordinary least square yields estimates with high variances. When the predictor variables are highly correlated, it is normally impossible to interpret estimates of individual coefficients. The multicollinearity problem and variable selection have been handled in a variety of different ways. For instance, one can use the variance inflation factors (VIF) (Kutner et al., 2004) and remove predictors with VIF higher than 10. VIF is a measure that determines how much the variance of an estimated regression parameter is increased because of collinearity. Another approach is to use partial least squares (PLS) regression (Wold, 1966) or principal components analysis (PCA) (Pearson, 1901). PLS regression is a method that reduces the predictor variables to a smaller set of an uncorrelated component by projecting the predicted variables and the predictor variables to

a new space. On the other hand, PCA is a technique that reduces the number of predictor variables by using an orthogonal transformation. It transforms a set of predictor variables to a set of values of linearly uncorrelated variables known as principal components. Also, one can consider using stepwise regression (Efroymson, 1960) or best subsets regression (Kutner et al., 2004). Stepwise regression is an automatic method of selecting predictor variables that proposes a single regression model. In each step, a predictor is considered for addition or deletion from the set of explanatory variables based on some criterion like R-squared, Mallows Cp, PRESS, Akaike information criterion, Bayesian information criterion, or false discovery rate. The best subsets regression is a technique that works similarly to stepwise regression, and the main difference is that it provides multiple regression models. Another approach to tackle this type of problem is to use ridge regression or lasso regression. Both methods impose constraints on the regression parameters. The constraint is presented in the form of a vector norm, where lasso regression uses the l_1 norm and ridge regression uses the l_2 norm. The key difference between those two norms is the shape of the constraint. The constraint has a diamond shape under the l_1 norm. However, it has a circle shape under the l_2 norm. The ridge and lasso regressions have a close connection to the Bayesian linear model when the regression parameters have independent Normal and Laplace priors, respectively. In this Chapter, we propose an alternative semi-parametric Bayesian approach based on empirical likelihood, which does not require the assumption of a parametric likelihood for the errors. It is semi-parametric because it combines the profile empirical likelihood ratio and priors, which are non-parametric and parametric, respectively.

Introduction to Ridge

Ridge regression (Tikhonov and Nikolayevich, 1943), also known as the method of linear regularization, penalizes the size of the regression coefficients by imposing an l_2 penalty.

That is, it minimizes a penalized residual sum of squares,

$$\min_{\boldsymbol{\theta}} \left(\frac{1}{2} ||\boldsymbol{y} - X\boldsymbol{\theta}||_2^2 + \alpha ||\boldsymbol{\theta}||_2^2 \right) \tag{4.1}$$

where

$$\alpha \geq 0,$$

$\boldsymbol{\theta}$ is a $p \times 1$ vector,

$\boldsymbol{y}$ is a $n \times 1$ vector,

X is a $p \times p$ matrix.

α is a complexity parameter that controls the amount of shrinkage. It is used to overcome the multicollinearity problem in data by adding a small positive value ($\alpha \geq 0$) to the diagonal element of the $X^T X$ matrix from multiple regression. The larger the value of α, the greater the amount of shrinkage (Hastie et al., 2009). The l_2 norm of a coefficient vector $\boldsymbol{\theta}$ is given by $||\boldsymbol{\theta}||_2^2 = \sum_{j=1}^{p} \theta_j^2$. The term $\alpha ||\boldsymbol{\theta}||_2^2$ is referred to as the ridge penalty. The solution to the ridge regression problem is given by:

$$\hat{\boldsymbol{\theta}}^{\text{ridge}} = \left(X^T X + \alpha \boldsymbol{I} \right)^{-1} X^T \boldsymbol{y}. \tag{4.2}$$

The ridge solution is quite similar to the ordinary least squares solution (OLS) but with a value, α, added to the diagonal of $X^T X$. It is easy to note that the solution presented in equation (4.2) equals to OLS when α=0 and equals to zero when $\alpha \rightarrow \infty$. As shown in Figure 4.1, the constraint has a circle form with no sharp points. That is, the intersection between the constraint area and the contour of ellipses will not occur on an axis, and so some ridge regression coefficients do not shrink to zero. Moreover, $\hat{\boldsymbol{\theta}}^{\text{ridge}}$ is a biased estimator of $\boldsymbol{\theta}$ but with a small variance compared to the variance of the OLS estimate. In the case of an

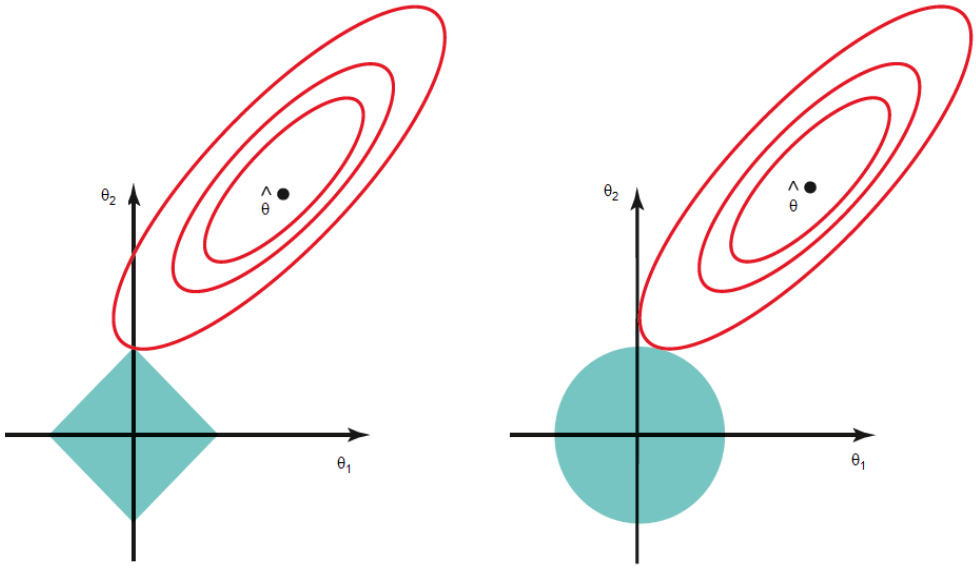

Figure 4.1: The geometry underlying the estimation of the lasso (left) and ridge regression (right). The solid blue area is the constraint region $|\theta_1| + |\theta_2| \leq t$ and $\theta_1^2 + \theta_2^2 \leq t$, respectively, while the red ellipses are the level sets of the loss function $||y - x\theta||_2^2$ (Source: James et al. (2013))

orthonormal design matrix, the ridge estimator scales the OLS estimator by $\frac{1}{1+\alpha}$. Next, we introduce the lasso regression.

Introduction to Lasso

The least absolute shrinkage and selection operator is a regression method that involves penalizing the absolute size of the regression coefficient and was introduced by Tibshirani (1996). It performs both variable selection and regularization. Given the vector of predictors $X = \boldsymbol{x_1}, \cdots, \boldsymbol{x_p}$, we would like to predict n observed response $\boldsymbol{y}$ via a linear model. The lasso solves the following regularized optimization problem:

$$\min_{\boldsymbol{\theta}} \left(\frac{1}{2} ||\boldsymbol{y} - X\boldsymbol{\theta}||_2^2 + \alpha ||\boldsymbol{\theta}||_1 \right), \tag{4.3}$$

where

$$\alpha \geq 0,$$

$\boldsymbol{\theta}$ is a $p \times 1$ vector,

$\boldsymbol{y}$ is a $n \times 1$ vector,

X is a $p \times p$ matrix.

by using l_1 penalty. α is a complexity parameter that controls the amount of shrinkage. The l_1 norm of a coefficient vector $\boldsymbol{\theta}$ is given by $||\boldsymbol{\theta}||_1 = \sum_{j=1}^{p} |\theta_j|$. It has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter α is sufficiently large (James et al., 2013). That is, this penalty term leads to feature/model selection. Fan and Li (2001) showed that the lasso is the only model that produces a sparse solution among l_q penalized estimators ($q \geq 1$). As depicted in Figure 4.1, the lasso solution occurs where the boundary of the feasible set first coincides with the level sets of the loss function.

Note that the lasso penalty contains the absolute value; thus, the objective function in equation (4.3) is not differentiable. Therefore, in general, the lasso solution lacks a closed form. This requires implementation of an optimization algorithm to find the minimizing solution. In the special case of an orthonormal design matrix, a closed form solution for lasso can be derived

$$\hat{\theta}_j^{\,lasso} = S\left(\hat{\theta}_j^{\,OLS}, \alpha\right)$$

where S, the soft-thresholding operator, is defined as

$$S(x, \alpha) = \begin{cases} x - \alpha & \text{if } x > \alpha \\ 0 & \text{if } |x| \leq \alpha \\ x + \alpha & \text{if } x < -\alpha \end{cases}$$

When $\alpha = 0$, then the lasso simply gives the ordinary least squares fit. On the other hand, when α is sufficiently large, the lasso method provides a model in which all coefficient estimates equal zero. Similar to ridge regression, lasso produces a biased estimator with a small variance.

To sum up, the lasso produces interpretable models that retain a subset of predictors and generate more accurate predictions compared to ridge regression. Moreover, the shrinkage term makes the lasso and ridge regression estimates biased, but it reduces the variance, which results in a bias/variance trade-off. It is worth noting that the OLS estimates are scale equivariant; however, the penalized regression coefficients can change when multiplying a given predictor by a constant because of the penalty term in the objective function. This change is why it is necessary to apply lasso and ridge after standardizing the predictors. In the next section, we present the Bayesian empirical likelihood for ridge regression.

4.2 Bayesian Empirical Likelihood for Ridge Regression

Ridge regression has a close connection to Bayesian linear regression. Noting the form of the penalty term in (4.1), one can conclude that the ridge regression parameters have independent and identical Normal priors. The shrinkage parameter, α, is introduced in the model in the form of a hyperparameter. Encouraged by this connection, we consider a semi-parametric Bayesian model using a Normal prior of the form

$$\begin{aligned} \pi(\boldsymbol{\theta}|\sigma^2, \alpha) &= \prod_{j=1}^{p} \sqrt{\frac{\alpha}{2\pi\sigma^2}} e^{-\frac{\alpha}{2\sigma^2}\theta_j^2} \\ \sigma^2 &\sim IG(a,\ b) \end{aligned} \tag{4.4}$$

where IG denotes the inverse gamma distribution with shape parameter a and scale param-

eter b. Note that α plays the role of prior precision. For example, a small (large) value of α leads to a wider (more concentrated) prior. By replacing the likelihood function with the profile EL ratio in the Bayesian setting, we have the following hierarchical representation of the full model:

$$\begin{aligned} R(\boldsymbol{\theta}) &\sim \exp\left(-\sum_{i=1}^{n}\log\left(1+\boldsymbol{\lambda}^T\boldsymbol{x_i}\left(y_i-\boldsymbol{x_i}^T\boldsymbol{\theta}\right)\right)\right), \\ \boldsymbol{\theta}|\sigma^2,\ \alpha &\sim N(\mathbf{0_{p\times 1}},\ \frac{\sigma^2}{\alpha}\boldsymbol{I_{p\times p}}), \\ \sigma^2 &\sim IG(a,\ b), \\ \sigma,\ \alpha &>\ 0. \end{aligned} \tag{4.5}$$

To estimate the intercept, θ_0, we could place a flat prior. But because we standardized both the predictor and response variable, the intercept is zero. The full conditional distribution of $\boldsymbol{\theta}$ and σ^2 is given by:

$$\pi\left(\boldsymbol{\theta},\sigma^2|X,\boldsymbol{y},\alpha\right)\propto\exp\left(-\sum_{i=1}^{n}\log\left(1+\boldsymbol{\lambda}^T\boldsymbol{x_i}\left(y_i-\boldsymbol{x_i}^T\boldsymbol{\theta}\right)\right)\right)\left(\frac{1}{\sigma^2}\right)^{p/2+a+1}\exp\left(-\frac{1}{\sigma^2}\left[b+\frac{\alpha}{2}\boldsymbol{\theta^T\theta}\right]\right)$$

The full conditional for σ^2 is inverse-gamma with shape parameter $p/2+a$ and scale parameter $b+\frac{\alpha}{2}\boldsymbol{\theta^T\theta}$. The full conditional distribution for $\boldsymbol{\theta}$ does not have a closed form:

$$\pi\left(\boldsymbol{\theta}|\sigma^2,\alpha,X,\boldsymbol{y}\right)\propto\exp\left(-\sum_{i=1}^{n}\log\left(1+\boldsymbol{\lambda}^T\boldsymbol{x_i}\left(y_i-\boldsymbol{x_i}^T\boldsymbol{\theta}\right)\right)-\frac{\alpha}{2\sigma^2}\boldsymbol{\theta^T\theta}\right) \tag{4.6}$$

We use a building block of HMC and Gibbs sampler to sample θ and σ^2, respectively. The implementation of the HMC requires the gradient of minus the logarithm empirical posterior

density of $\boldsymbol{\theta}$:

$$\text{negative log-likelihood: } -\log\left(\pi\left(\boldsymbol{\theta}|\sigma^2,\alpha\right)\right) = \sum_{i=1}^{n}\log\left(1+\boldsymbol{\lambda}^T\boldsymbol{x_i}\left(y_i-\boldsymbol{x_i}^T\boldsymbol{\theta}\right)\right)+\frac{\alpha}{2\sigma^2}\boldsymbol{\theta^T\theta}$$

$$\text{gradient: } -\frac{\partial\log\left(\pi\left(\boldsymbol{\theta}|\sigma^2,\alpha\right)\right)}{\partial\boldsymbol{\theta}} = \sum_{i=1}^{n}\frac{-\boldsymbol{\lambda^T x_i x_i}^T}{1+\boldsymbol{\lambda}^T\boldsymbol{x_i}\left(y_i-\boldsymbol{x_i}^T\boldsymbol{\theta}\right)}+\frac{\alpha}{\sigma^2}\boldsymbol{\theta^T}$$

We implemented a function in R for the Bayesian ridge based on empirical likelihood. Table 4.1 describes arguments and outputs of this function.

Table 4.1: Summary of the function implemented in R for the Bayesian ridge based on empirical likelihood.

Arguments	Outputs
x: Design matrix. y: Predictor variable. *nsim*: Number of iterations. *nwarm*: Number of iteration for burn-in. e: Stepsize for the leapfrog steps. L: Number of leapfrog steps. *shrinkage*: Penalty coefficient.	*posteriorbeta*: Posterior mean estimate of $\boldsymbol{\theta}$. *beta*: HMC samples of $\boldsymbol{\theta}$. *Sigma2*: Sample of σ^2. *weights*: Empirical likelihood weights: $w_i = n^{-1}\left(\frac{1}{1+\hat{\boldsymbol{\lambda}}^T\boldsymbol{x_i}(y_i-\boldsymbol{x_i}^T\hat{\boldsymbol{\theta}})}\right)$ *-2LLR*: -2 log likelihood ratio. *p.value*: The observed p-value by χ^2 approximation. *lambda*: Lagrange multiplier value evaluated at the posterior mean estimate. *grad*: Gradient value of lambda. *hess*: Hessian matrix of lambda.

The asymptotic distribution of $\boldsymbol{\theta}$ under the Bayesian ridge model based on EL is easily obtained. Let $\hat{\boldsymbol{\theta}_n}$ be the profile maximum likelihood estimate of $\boldsymbol{\theta}$ and let

$$J(\hat{\boldsymbol{\theta}_n}) = \left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\sum_{i=1}^{n}\log\left(1+\boldsymbol{\lambda}^T\boldsymbol{x_i}(y_i-\boldsymbol{x_i}^T\boldsymbol{\theta})\right)\right)_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}_n}}. \tag{4.7}$$

By the result presented in Section 3.2, the posterior distribution of $\boldsymbol{\theta}$ converges to normal,

with mean m_{n1} and covariance J_{n1} where

$$J_{n1} = J(\hat{\boldsymbol{\theta}_n}) + \frac{\sigma^2}{\alpha} I_{p\times p}$$
$$m_{n1} = J_{n1}^{-1} J(\hat{\boldsymbol{\theta}_n})\hat{\boldsymbol{\theta}_n}.$$

Next, we derive the Bayesian empirical likelihood for lasso.

4.3 Bayesian Empirical Likelihood for Lasso Regression

Similar to the ridge, lasso has a close connection to the Bayesian linear model. Tibshirani (1996) suggested that the lasso estimates can be interpreted as posterior mode estimates. That is, using a hierarchical model, one can place an independent identical double-exponential prior, also known as Laplace distribution, on the parameters of the model. Several authors suggested using Laplace distribution as a prior (Figueiredo, 2003; Bae and Mallick, 2004; Yuan and Lin, 2005). Motivated by this, Park and Casella (2008) considered a fully Bayesian analysis using a conditional double-exponential prior. We consider a conditional prior specification of the form

$$\pi(\boldsymbol{\theta}|\sigma^2,\ \alpha)\ =\ \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} \exp\left(-\alpha|\theta_j|/\sqrt{\sigma^2}\right). \tag{4.8}$$

Andrews and Mallows (1974) showed that the Laplace distribution can be represented as a scale mixture of normals with an exponential mixing density (See Appendix A):

$$\frac{a}{2}e^{-a|z|} = \int_0^\infty \frac{1}{2\pi s} e^{-z^2/(2s)} \frac{a^2}{2} e^{-a^2 s/2} ds, \quad a\ >\ 0. \tag{4.9}$$

Figure 4.2 shows that the Laplace distributions are sharply peaked at their mean where a high scale value yields a probability density near to zero. Another notable feature is that

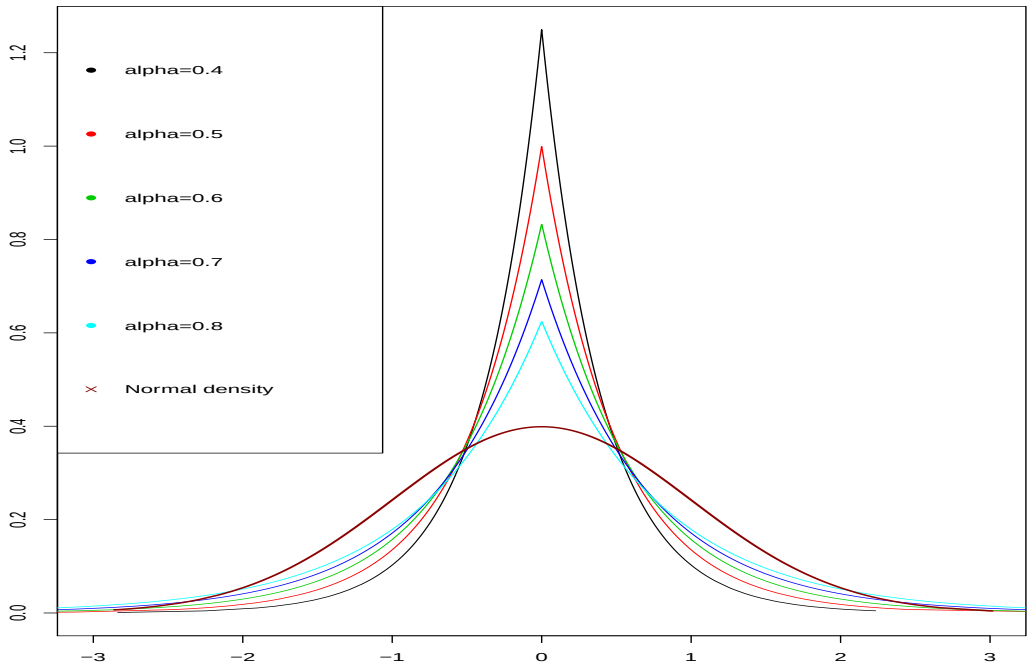

Figure 4.2: Laplace distribution with mean zero and different values of the scale parameter along with normal density with mean zero and standard deviation 1.

the Laplace distribution assigns a higher density around its mean compared to the Normal density. Using the hierarchical representation of Park and Casella (2008) where we replace the likelihood function by the profile empirical likelihood ratio for linear model, our hierarchical representation of the full model becomes

$$
\begin{aligned}
R(\boldsymbol{\theta}) &\sim \exp\left(-\sum_{i=1}^{n}\log\left(1+\boldsymbol{\lambda}^T\boldsymbol{x_i}\left(y_i-\boldsymbol{x_i}^T\boldsymbol{\theta}\right)\right)\right),\\
\boldsymbol{\theta}|\sigma^2,\tau_1^2,\cdots,\tau_p^2 &\sim N_p(\mathbf{0},\sigma^2 D_\tau),\\
D_\tau &= \operatorname{diag}\left(\tau_1^2,\cdots,\tau_p^2\right),\\
\sigma^2,\tau_1^2,\cdots,\tau_p^2|\alpha &\sim \pi\sigma^2 d\sigma^2\prod_{j=1}^{p}\frac{\alpha^2}{2}e^{-\alpha^2\tau_j^2}d\tau_j^2,\\
\sigma^2,\tau_1^2,\cdots,\tau_p^2 &> 0.
\end{aligned}
\tag{4.10}
$$

We avoid placing a prior distribution on the intercept by standardizing both the predictor variables and the response variable. After integrating out $\tau_1^2,\cdots,\tau_p^2$, the conditional prior on $\boldsymbol{\theta}$ has the desired form (4.8). We choose $\pi(\sigma^2)=\mathrm{IG}(a,b)$. One can also impose a noninformative prior $\pi(\sigma^2)=1/\sigma^2$ on σ^2. Conditioning on σ^2 guarantees the unimodality of the full posterior distribution (See Appendix B). The parameter $\boldsymbol{\tau^2}$ can be viewed as a latent parameter that assigns different weights to the p covariates. The full empirical posterior distribution is:

$$
\begin{aligned}
\pi(\boldsymbol{\theta},\sigma^2,\tau_1^2,\cdots,\tau_p^2|X,\boldsymbol{y},\alpha)\propto&\exp\left(-\sum_{i=1}^{n}\log\left(1+\boldsymbol{\lambda}^T\boldsymbol{x_i}\left(y_i-\boldsymbol{x_i}^T\boldsymbol{\theta}\right)\right)\right)\left(\frac{1}{\sigma^2|D_\tau|}\right)^{1/2}\exp\left(-\frac{1}{2\sigma^2}\boldsymbol{\theta}^T D_\tau^{-1}\boldsymbol{\theta}\right)\\
&(\sigma^2)^{-a-1}\exp\left(-\frac{b}{\sigma^2}\right)\prod_{i=1}^{p}\frac{\alpha^2}{2}\exp\left(-\alpha^2\tau_j^2/2\right).
\end{aligned}
\tag{4.11}
$$

Equation (4.11) gives rise to the following sampling scheme:

1. Sample $\boldsymbol{\theta}$ from

$$
\pi(\boldsymbol{\theta}|\sigma^2,\tau_1^2,\cdots,\tau_p^2)\propto\exp\left(-\sum_{i=1}^{n}\log\left(1+\boldsymbol{\lambda}^T\boldsymbol{x_i}\left(y_i-\boldsymbol{x_i}^T\boldsymbol{\theta}\right)\right)\right)\exp\left(-\frac{1}{2\sigma^2}\boldsymbol{\theta}^T D_\tau^{-1}\boldsymbol{\theta}\right)
$$

This is a nonstandard distribution. We use the Hamiltonian Monte Carlo algorithm

to sample from it.

2. Sample σ^2 from inverse-gamma with shape parameter $p/2 + a$ and scale parameter $b + \frac{1}{2}\boldsymbol{\theta}^T D_\tau^{-1}\boldsymbol{\theta}$. One has to be cautious in selecting a and b. The smaller the values, the better the estimates are, because a large precision allows sampling from a probability density that is near to zero. That is, when choosing a large penalty, it forces the estimates to shrink toward zero.

3. Sample $1/\tau_j^2$ from inverse-Gaussian with mean and shape equals to $\sqrt{\frac{\alpha^2\sigma^2}{\theta_j^2}}$ and α^2, respectively (derivation is presented in Appendix C).

We implement a building block of the HMC and Gibbs sampler to sample $\boldsymbol{\theta}$, σ^2, and $\tau_1^2, \cdots, \tau_p^2$. The gradient of minus the logarithm empirical likelihood density of $\boldsymbol{\theta}$:

$$-\log\text{-likelihood:}\quad -\log\left(\,\pi\left(\boldsymbol{\theta}|\sigma^2,\tau_1^2,\cdots,\tau_p^2,\alpha\right)\right) = \sum_{i=1}^{n}\log\left(1+\boldsymbol{\lambda}^T\boldsymbol{x_i}\left(y_i-\boldsymbol{x_i}^T\boldsymbol{\theta}\right)\right) + \frac{\boldsymbol{\theta}^T D_{\tau^{-1}}\boldsymbol{\theta}}{2\sigma^2}$$

$$\text{gradient:}\quad -\frac{\partial\log\left(\,\pi\left(\boldsymbol{\theta}|\sigma^2,\alpha\right)\right)}{\partial\boldsymbol{\theta}} = \sum_{i=1}^{n}\frac{-\boldsymbol{\lambda^T x_i x_i}^T}{1+\boldsymbol{\lambda}^T\boldsymbol{x_i}\left(y_i-\boldsymbol{x_i}^T\boldsymbol{\theta}\right)} + \frac{1}{\sigma^2}\boldsymbol{\theta}^T D_\tau^{-1}$$

We have also implemented a function in R that performs the Bayesian lasso based on empirical likelihood. Table 4.2 describes arguments and outputs of this function.

The asymptotic distribution of $\boldsymbol{\theta}$ under the Bayesian lasso model based on EL is easily obtained. Similarly to the ridge, let $\hat{\boldsymbol{\theta}_n}$ be the profile maximum likelihood estimate of $\boldsymbol{\theta}$ and $J(\hat{\boldsymbol{\theta}_n})$ be as defined in Equation (4.7). By the result presented in Section 3.2, the posterior distribution of $\boldsymbol{\theta}$ converges to normal, with mean m_{n2} and covariance J_{n2} where

$$J_{n2} = J(\hat{\boldsymbol{\theta}_n}) + \sigma^2 D_\tau$$

$$m_{n2} = J_{n2}^{-1} J(\hat{\boldsymbol{\theta}_n})\hat{\boldsymbol{\theta}_n}.$$

Table 4.2: Summary of the function implemented in R for the Bayesian lasso based on empirical likelihood.

Arguments	Outputs
x: Design matrix y: Predictor variable. *nsim*: Number of iterations. *nwarm*: Number of burn-in e: Stepsize for the leapfrog steps. L: Number of leapfrog steps. *shrinkage*: Penalty coefficient. PriorS1: Shape parameter of the prior on σ^2. PriorS2: Scale parameter of the prior on σ^2.	*posteriorbeta*: Posterior mean estimate of $\boldsymbol{\theta}$. *beta*: HMC samples of $\boldsymbol{\theta}$. *Sigma2*: Sample of σ^2. *weights*: Empirical likelihood weights: $w_i = n^{-1}\left(\frac{1}{1+\hat{\boldsymbol{\lambda}}^T \boldsymbol{x_i}(y_i - \boldsymbol{x_i}^T\hat{\boldsymbol{\theta}})}\right)$ *-2LLR*: -2 log likelihood ratio. *p.value*: The observed p-value by χ^2 approximation. estimate. *lambda*: Lagrange multiplier evaluated at the posterior mean *grad*: Gradient value of lambda. *hess*: Hessian matrix of lambda. invTau2 : Samples of $1/\tau_j^2$.

4.4 Illustrative Examples

In this Section, we provide three illustrative examples of the methods derived in Sections 4.2 and 4.3. The first example uses simulated data, and the remaining examples are based on real datasets. In Section 4.5, we provide credible interval regions, HPD, and posterior distribution for the Bayesian lasso and ridge methods based on empirical likelihood.

Simulation

In this example, we use simulated data to investigate the performance of the Bayesian ridge and the Bayesian lasso based on empirical likelihood. We simulate a data set that consists

of 150 observations and 40 covariates from the following model:

$$\boldsymbol{y} = \boldsymbol{\theta}^{\boldsymbol{T}}\boldsymbol{X} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\theta} = \left(-3, -3, 4, 3, -1, \underbrace{0, 0, 0, 0, 0, 0, 0, 0, \cdots, 0}_{\text{35 coefficients are equal to 0}}\right)$ and ϵ is normal with mean 0 and standard deviation 4. Figure 4.3 shows the Bayesian lasso (top) and the Bayesian ridge (bottom) posterior mean estimates based on empirical likelihood over a grid of α values, using 5000 iterations with 1000 burn-in for each value of α. For lasso, the range of α is [0, 20], whereas, for the ridge, the range of α is [0, 100]. For the HMC method, the step size and the number of leapfrog steps are set to $\vartheta = 0.01$ and $L = 10$, respectively. At each iteration, we use the modified Newton-Raphson to estimate the Lagrange vector $\boldsymbol{\lambda}$. We scale both predictors and the response variable so that the intercept is 0. The HMC Bayesian empirical likelihood lasso and ridge estimates were posterior means computed over a grid of α. Each curve corresponds to a predictor variable. The Figure shows the path of each variable against the range of values of α. In the lasso case, it is evident that as we increase the value of α, the coefficients shrink to zero. More specifically, the predictors x_2 and x_5 shrink toward zero faster compared to predictors x_1, x_3 and x_4; x_5 has the quickest decrease rate. Similarly, in the ridge case, the coefficients of predictors shrink to the neighborhood of zero as we increase the value of α but never attain zero.

Prostate Cancer Data

We apply the Bayesian lasso and ridge methods, based on empirical likelihood, on the prostate cancer data set presented in Example 2.4. We compare our results with the fully parametric Bayesian approach. To sample $\boldsymbol{\theta}$, we use the HMC scheme with 5000 iterations and 1000 burn-in. For the lasso case, we choose $L = 10$ and $\vartheta = 0.025$. In contrast, for the ridge case we set $\vartheta = 0.01$ and $L = 10$. At each iteration, we use the modified Newton-

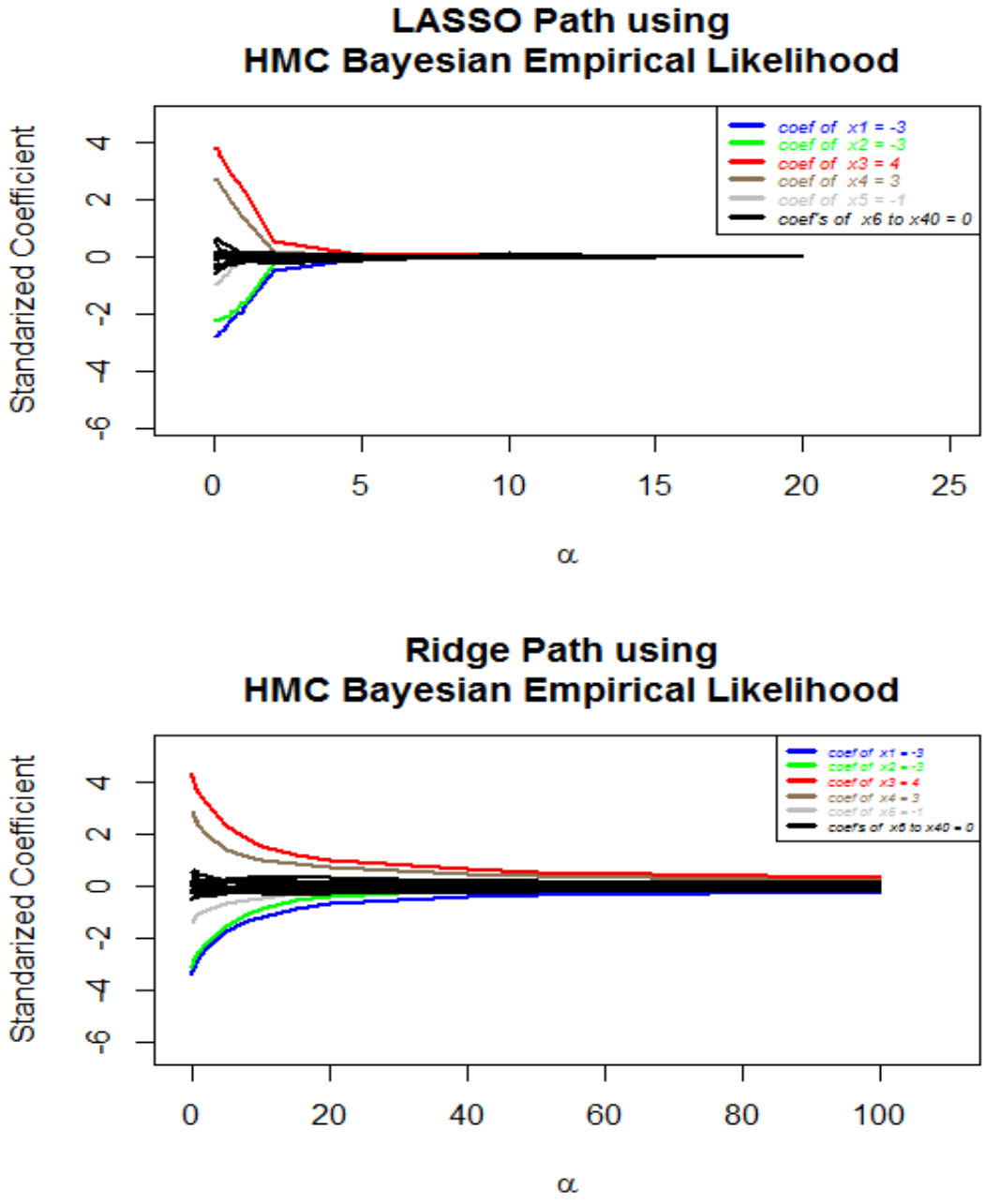

Figure 4.3: Lasso and ridge path for the simulated data using HMC Bayesian empirical likelihood.

Raphson to estimate the Lagrange vector $\boldsymbol{\lambda}$. We scale both predictors and the response variable so that the intercept is 0. Figure 4.4 compares the Bayesian lasso based on empirical likelihood and the Bayesian lasso on the prostate cancer data. In both approaches, the estimates were posterior means computed over a grid of α. For α, we used a range of [0, 1500]. The Figure shows paths of these estimates as their respective shrinkage parameter changes. It is evident that both methods provide almost identical results. In addition, one can conclude that the clinical predictors log cancer volume, log prostate weight, and seminal vesicle invasion have more influence on the log of prostate specific antigen. Moreover, all coefficients shrink to zero for a shrinkage parameter larger than 100. Similarly, Figure 4.5 compares posterior means estimate for the Bayesian ridge based on empirical likelihood and Bayesian ridge on the prostate cancer data. The Figure depicts the paths of these estimates over a range of values of the shrinkage parameter. Both methods provide similar results. As expected, coefficients do not shrink exactly to zero. Moreover, it is obvious that the predictor variables log cancer volume, seminal vesicle invasion, and log prostate weight have the lowest decrease rate compared to other predictor variables and are more influential on the level of prostate-specific antigen.

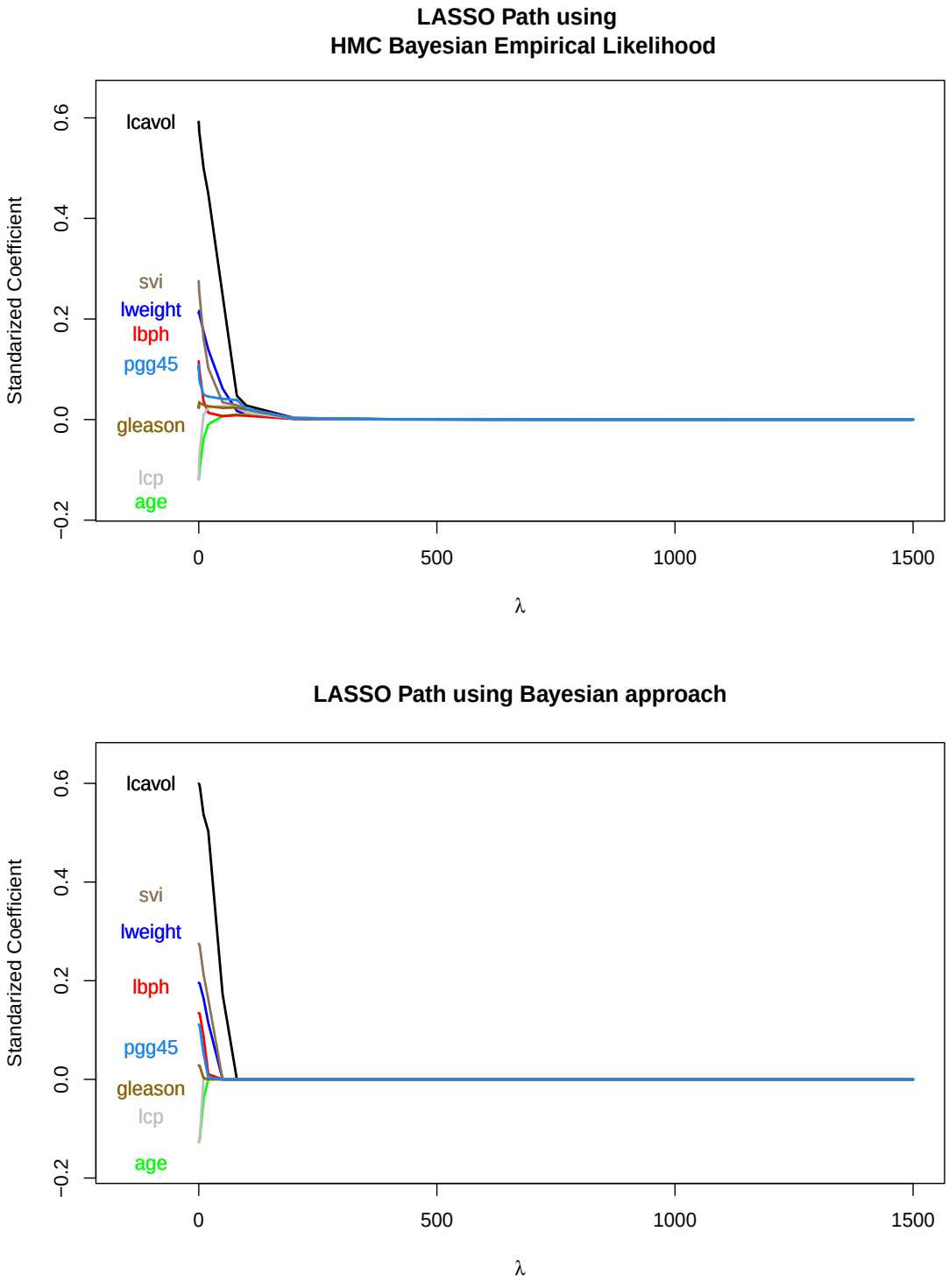

Figure 4.4: Lasso path for the prostate cancer data using HMC Bayesian empirical likelihood (top) and Bayesian method (bottom)

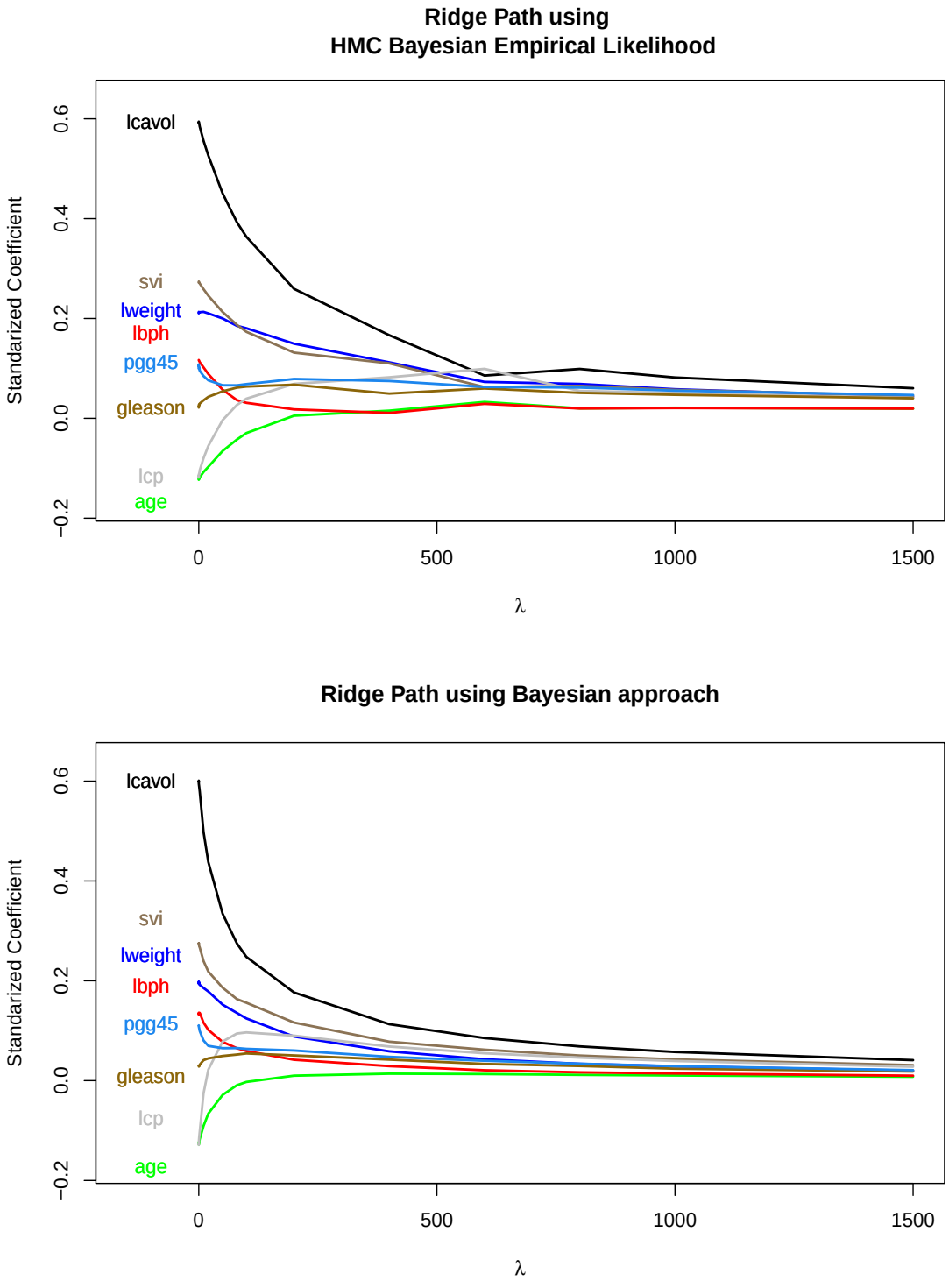

Figure 4.5: Ridge path for the prostate cancer data using HMC Bayesian empirical likelihood (top) and Bayesian method (bottom)

Diabetes Data

We apply the Bayesian lasso and ridge methods, based on empirical likelihood, on the diabetes data provided by Efron et al. (2004). Data are scaled, consist of 442 diabetes patients, and examined the relationship between 10 baseline variables and a quantitative measure of disease progression one year after baseline. These variables are age, sex, body mass index, average blood pressure, and six blood measurements.

To sample $\boldsymbol{\theta}$, we use the HMC scheme with 5000 iterations and 1000 burn-in. For the lasso method, we choose $\vartheta = 0.025$ and $L = 10$, whereas, for the ridge method, we choose $\vartheta = 0.01$ and $L = 10$. Figure 4.6 compares posterior mean estimates for the Bayesian lasso based on empirical likelihood and Bayesian lasso on the diabetes data. The Figure shows the paths of these estimates as their respective shrinkage parameter changes. For α, we used a range of [0, 1500]. It is evident that both methods provide almost identical results. Also, one can conclude that the predictor variables sex, age, s1, s2, s4, and s6 have the faster decrease rate and are less influential on the disease progress compared to other predictors. Similarly, Figure 4.7 compares posterior mean estimates for the Bayesian ridge based on empirical likelihood and the Bayesian ridge. The Figure depicts the path of these estimates over a range of values for the shrinkage parameter. Similarly to lasso, the range of α is [0, 1500]. Both methods provide similar results. As expected, these coefficients do not shrink exactly to zero.

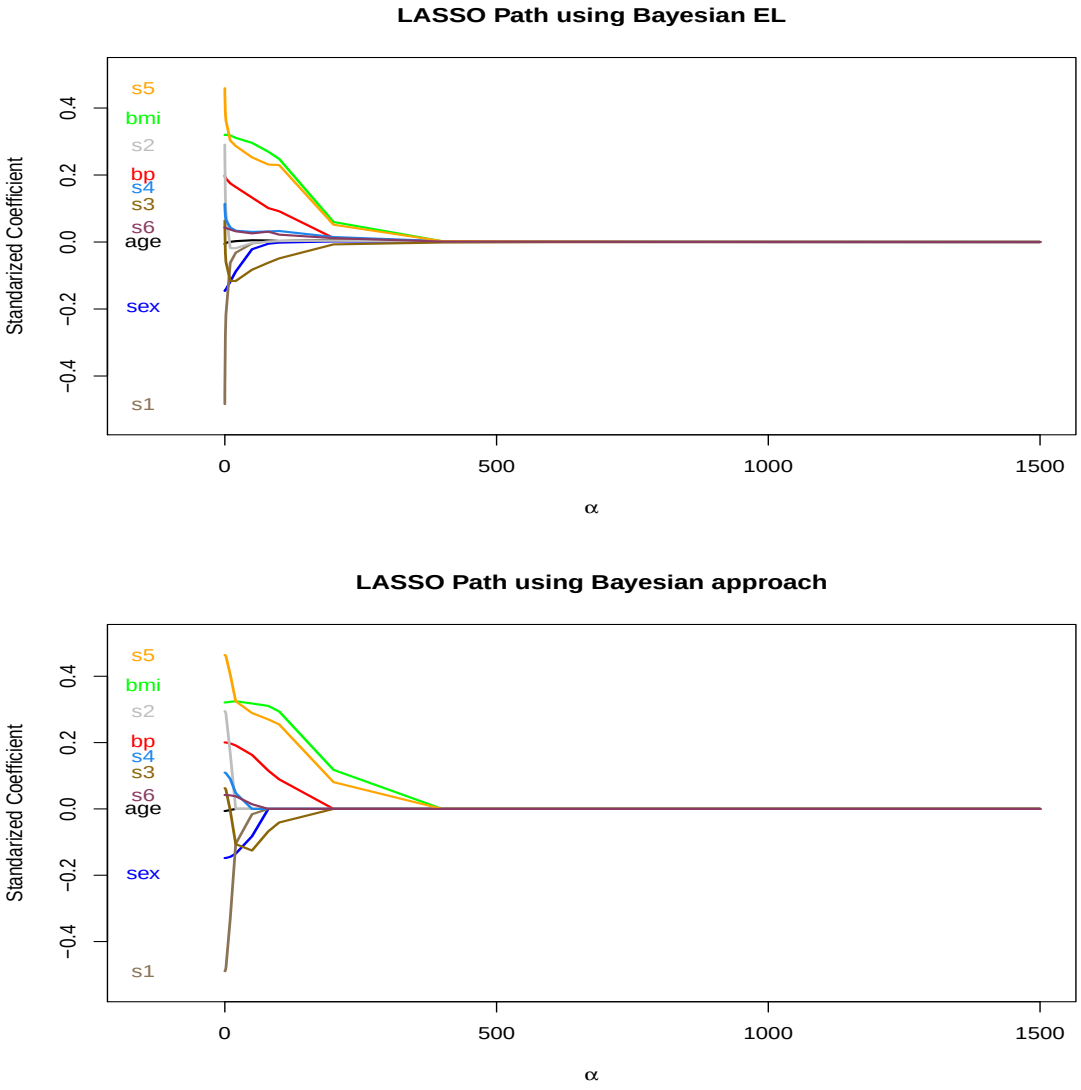

Figure 4.6: Lasso path for the diabetes data using HMC Bayesian empirical likelihood (top) and Bayesian method (bottom)

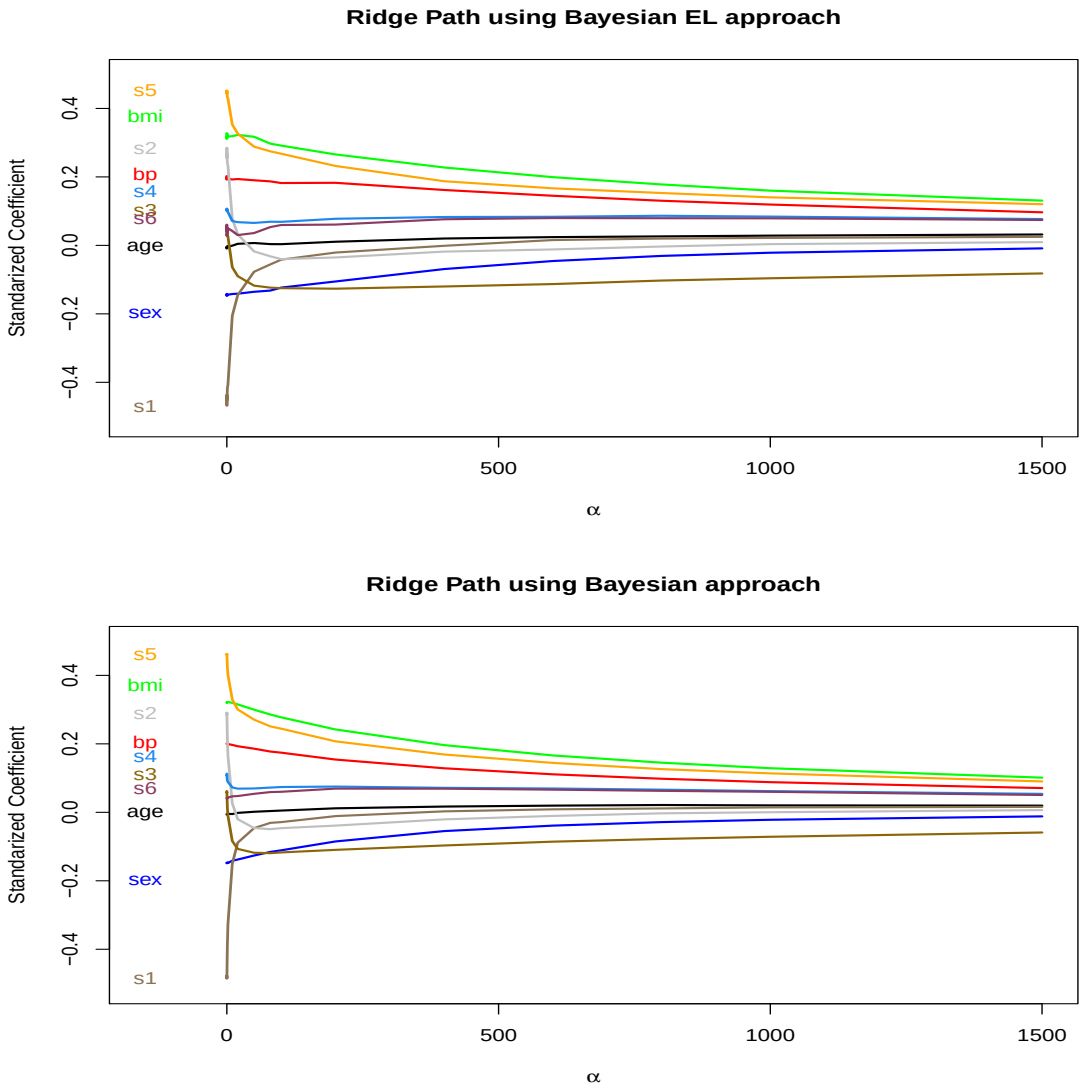

Figure 4.7: Ridge path for the diabetes data using HMC Bayesian empirical likelihood (top) and Bayesian method (bottom)

4.5 Estimation of the Shrinkage Parameter

The selection of the penalty, α, is crucial because each value of α corresponds to a fitted model. To select an optimal value of α for lasso and ridge regression, one can use certain empirical approaches, such as Akaike information criterion AIC (Akaike, 1974), Bayes infor-

mation criterion BIC (Schwarz, 1978), cross-validation CV (Geisser, 1993), or Generalized cross-validation GCV (Craven and Wahba, 1978). The most frequent method used is K-fold cross-validation, which works as follows. Given data $(X, \boldsymbol{y})$, we partition them into K parts. This gives us K pairs: $(X_1, \boldsymbol{y_1}), \cdots, (X_K, \boldsymbol{y_K})$. Let n_i be the number of points in the i^{th} pair $(X_i, \boldsymbol{y_i})$ and let $\boldsymbol{\theta}^{(-i)}_{penalized}$ be the lasso or ridge solution obtained using data pair $(X^{(-i)}, \boldsymbol{y}^{(-i)})$. Given a range of plausible values for α, we define the average cross-validation mean squared error as:

$$\overline{CV}_{MSE}(\alpha) \;=\; \frac{1}{K}\sum_{i=1}^{K}\frac{1}{n_i}\left|\left|\left(\boldsymbol{y_i}^{(-i)} - X_i^{(-i)}\boldsymbol{\theta}^{\boldsymbol{(-i)}}_{\boldsymbol{penalized}}\right)\right|\right|_2^2$$

where $(-i)$ means that the i^{th} cross-validation mean squared error is calculated without the i^{th} partition. Every data point appears in the testing set exactly once, and $k-1$ times in the training set. The way how we partition the data does not matter; we use $k = 5$. The disadvantage of this method is that it is computationally costly.

The optimal value of α, denoted by α^*, is the value that minimizes the average cross-validated mean squares error:

$$\alpha^* \;=\; \text{argmin}_\alpha \overline{CV}_{MSE}(\alpha)$$

Note that α is a continuous parameter and considering all its possible values is not practically feasible. Thus, one has to be cautious in the discretization of its range. In the Bayesian setting, Park and Casella (2008) discussed the implementation of an empirical Bayes approach with an EM algorithm. In this dissertation, we use an approach that treats the shrinkage coefficient as a random parameter by placing a hyperprior on it. The disadvantage of this method is that the conditional posterior distribution of the shrinkage parameter does not involve data at all.

For lasso, we use three different hyperpriors: gamma distribution, uniform distribution, and beta distribution. On the other hand, for the ridge model, we only use a gamma distribution as a hyperprior. Placing a uniform hyperprior on α results in unknown truncated function,

which makes its implementation very challenging. Also, one has to be careful in choosing the hyper-hyperparameters because, in the Bayesian lasso approach, α controls the distribution of $1/\tau_j^2$ that regulates the weight of the covariates.

As a remedy, we divide data into two sets: training and validation. Apply K-fold cross-validation on training data and retrieve the value, $\alpha^{training}$, that results in a small prediction error. After that, we place a prior on the shrinkage parameter and choose its hyper-hyperparameters such that the posterior estimate is around the $\alpha^{training}$.

Lasso Case

Park and Casella (2008) placed a gamma distribution with shape r and rate d on α^2. In this case, factoring the equation (4.11) by this prior leads to the following posterior conditional distribution:

$$\pi(\alpha|\tau_1^2, \cdots, \tau_p^2) \propto (\alpha^2)^{p+r-1} \exp\left(-\alpha^2 \left[d + \frac{1}{2}\sum_{j=1}^{p} \tau_j^2\right]\right),$$

which is a gamma distribution with shape $p+r$ and rate $d + \frac{1}{2}\sum_{i=1}^{p} \tau_j^2$. An alternative way is to consider a class of uniform priors on α^2 of the form:

$$\pi(\alpha^2) = \frac{1}{\eta_2 - \eta_1};\ \alpha^2 > 0, 0 \leq \eta_1 < \eta_2.$$

When this prior is used in equation (4.11), the full conditional distribution of α^2 is a truncated gamma

$$\pi(\alpha^2|\tau_1^2, \cdots, \tau_p^2) \propto G(p,\ \frac{1}{2}\sum_{j=1}^{p} \tau_j^2) I_{(\eta_1 \leq \alpha^2 \leq \eta_2)}.$$

To sample from the above function, we use the fact that the distribution function for the truncated gamma is just a linear function of the gamma distribution. That is, if $S(x)$ is the cumulative distribution function of the gamma distribution and $s(x)$ is the density, and we

are truncating such that $\eta_1 \leq x \leq \eta_2$, then the density function is :

$$s_{[\eta_1,\eta_2]}(x) = \frac{s(x)}{s(\eta_2) - s(\eta_1)}$$

and

$$S_{[\eta_1,\eta_2]}(x) = \frac{S(x) - S(a)}{S(\eta_2) - S(\eta_1)} \sim \text{uniform}[0,1].$$

To sample the truncated gamma random variable, we generate a random uniform $u \sim$ uniform$[0,1]$ and compute

$$S^{-1}\left[S(\eta_1 + u\left(S(\eta_2) - S(\eta 1)\right))\right].$$

A different selection is to use a more flexible distribution. Los Campos et al. (2009) placed a beta distribution with shapes ν_1 and ν_2 on $\tilde{\alpha} = \frac{\alpha}{u}$, where $u > 0$ is an upper bound on α. That is,

$$\pi(\alpha) = \text{Beta}\left(\tilde{\alpha}(\alpha)|\nu_1,\nu_2\right)\left|\frac{\partial\tilde{\alpha}(\alpha)}{\partial}\right| \propto \text{Beta}\left(\frac{\alpha}{u}|\nu_1,\nu_2\right).$$

If we know that the shrinkage parameter $\alpha \in (0,1)$, we can use a beta distribution without the constraint. When the constrained beta distribution is used in (4.11), the full conditional distribution of α (not on α^2) is:

$$\pi(\alpha|\tau_1^2,\cdots,\tau_p^2) \propto (\alpha)^{2p+\nu_1-1}(u-\alpha)^{\nu_2-1}\exp\left(-\frac{\alpha^2}{2}\sum_{j=1}^{p}\tau_j^2\right),$$

which lacks a closed form. For its implementation, we use the Hamiltonian Monte Carlo where the negative logarithm empirical posterior distribution is:

$$-\log\left(\pi(\alpha|\tau_1^2,\cdots,\tau_p^2)\right) \propto (1-2p-\nu_1)\log\alpha + (1-\nu_2)\log(u-\alpha) + \frac{\alpha^2}{2}\sum_{j=1}^{p}\tau_j^2$$

and the gradient is

$$-\frac{\partial}{\partial\alpha}\log\left(\pi(\alpha|\tau_1^2,\cdots,\tau_p^2)\right) = \frac{1-2p-\nu_1}{\alpha}+(\nu_2-1)\frac{1}{u-\alpha}+\alpha\sum_{j=1}^{p}\tau_j^2.$$

Note that the posterior distributions of α and α^2 presented above do not depend on data.

Example

We use the diabetes data from Efron et al. (2004) presented in Section 4.4. As discussed above, choosing the hyper-hyperparameters is challenging because they control the value of α and, hence, the amount of shrinkage. We split the data into 50% training and 50% testing, and applied 5-fold cross-validation over a grid of α values on the training set. The lasso parameter that minimized the average cross-validation mean square was 0.25. Thus, we choose hyper-hyperparameters such that the mean of the hyperprior is around 0.25. We run an MCMC with 5000 iterations and 1000 burn-in.

Table 4.3 presents the posterior mean estimates of the shrinkage parameter under three different hyperpriors. All values are near to 0.25. Figure 4.8 displays the trace plot and the kernel density for the shrinkage parameter when gamma is chosen as the hyperprior. The estimates are around the center with reasonable fluctuation, which indicates that the chain is mixing well. From the kernel density plot, one can see that the distribution is unimodal. Figure 4.9 presents the trace plot and the kernel density for the shrinkage parameter when the hyperprior is distributed as beta. The estimates are around the center with reasonable fluctuation, which indicates that the chain is mixing well. From the kernel density plot, one can see that the distribution is unimodal and right-skewed. Figure 4.10 exhibits the trace plot and the kernel density when the shrinkage hyperprior follows the uniform distribution. The estimates are around the center with values a bit shifted to the left of the center and with reasonable fluctuation, which indicates that the chain is mixing well. From the kernel

density plot, one can see that the distribution is unimodal and right-skewed with a heavy tail.

Table 4.4 represents the posterior mean estimates of the Bayesian lasso based on empirical likelihood using different hyperpriors. The values of the estimates, in each case, look quite similar. From Figures 4.11, 4.12, and 4.13, it appears that the trace plots are around their centers with reasonable fluctuations, which indicate that the chains have good mixing. Table 4.5 provides the 95% highest posterior density intervals and the 95% equal-tailed credible regions. It seems that, under different hyperpriors, they are quite similar.

Table 4.3: Posterior mean of the shrinkage parameter for the Bayesian lasso based on empirical likelihood under gamma, beta, and uniform hyperpriors.

Hyperprior	Gamma	Beta	Uniform
Estimate of the shrinkage parameter	0.2265	0.2324	0.2177

Table 4.4: Posterior mean estimates for the Bayesian lasso based on empirical likelihood method, using gamma, beta, uniform distributions as hyperprior on the penalty term.

Hyperprior	AGE	SEX	BMI	BP	S1	S2	S3	S4	S5	S6
Gamma	-0.012	-0.083	0.287	0.079	-0.032	-0.005	-0.104	0.026	0.370	0.054
Beta	-0.011	-0.082	0.287	0.079	-0.027	-0.007	-0.106	0.024	0.367	0.054
Uniform	-0.012	-0.085	0.288	0.080	-0.030	-0.007	-0.108	0.025	0.370	0.054

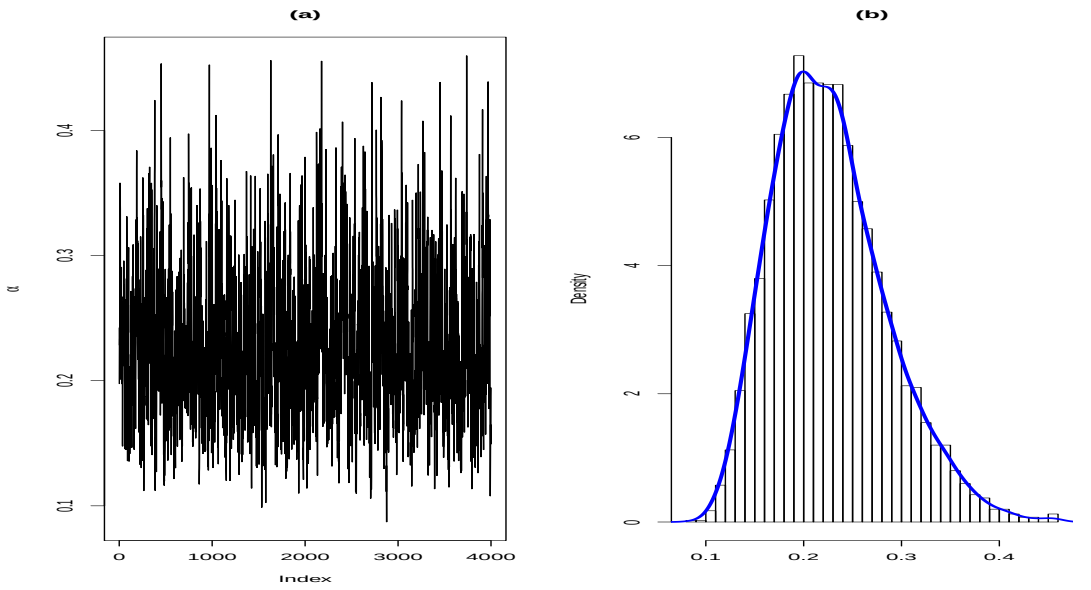

Figure 4.8: Trace plot (a) and histogram along with the kernel density (b) of the posterior mean estimates for the shrinkage parameter under gamma hyperprior in the BEL lasso; using 5000 iterations with 1000 burn-in

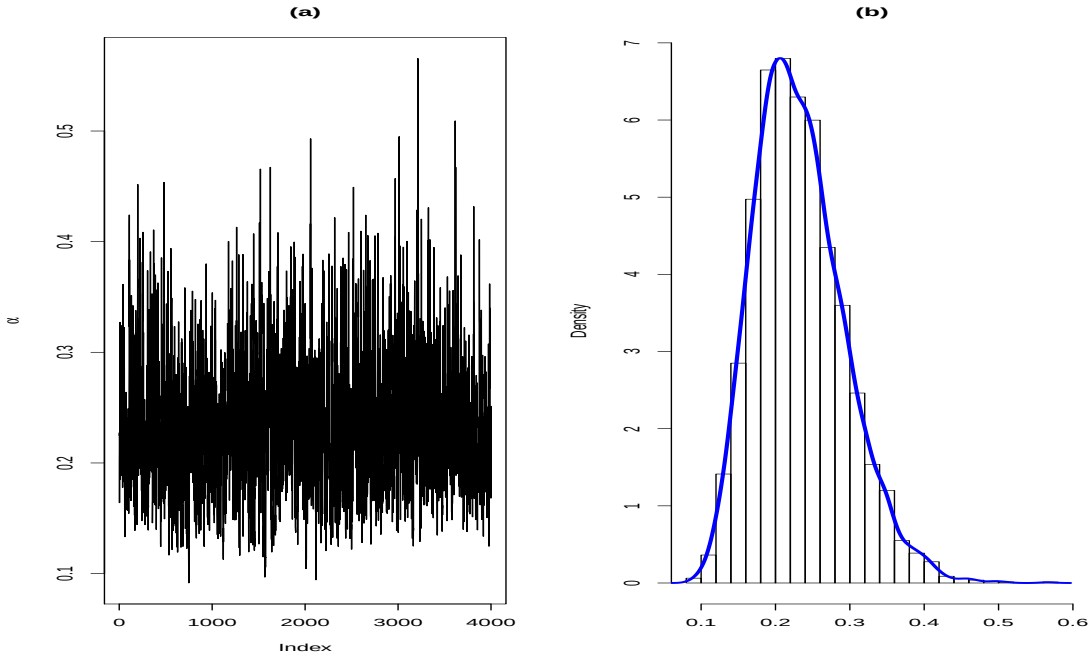

Figure 4.9: Trace plot (a) and histogram along with the kernel density (b) of the posterior mean estimates for the shrinkage parameter under beta hyperprior in the BEL lasso; using 5000 iterations with 1000 burn-in

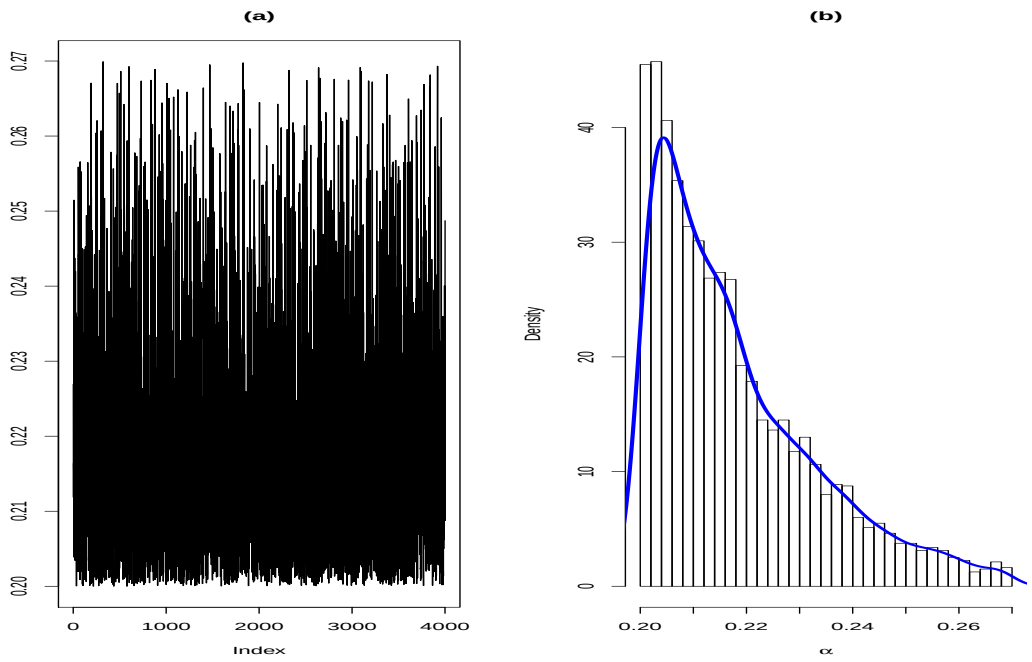

Figure 4.10: Trace plot (a) and histogram along with the kernel density (b) of the posterior mean estimates for the shrinkage parameter under uniform hyperprior in the BEL lasso, using 5000 iterations with 1000 burn-in

Table 4.5: The 95% highest posterior density intervals and the 95% credible regions of the posterior distribution of the coefficients in the lasso model using the diabetes data, under gamma, beta, uniform distributions as hyperprior on the penalty term.

	Gamma hyperprior	Beta hyperprior	Uniform hyperprior
Variables	95% HPD **(95% credible regions)**	95% HPD **(95% credible regions)**	95% HPD **(95% credible regions)**
AGE	[-0.092, 0.066] **([-0.089, 0.066])**	[-0.092, 0.067] **([-0.089, 0.067])**	[-0.097, 0.069] **([-0.102, 0.062])**
SEX	[-0.183, 0.013] **([-0.180, 0.010])**	[-0.178, 0.015] **([-0.176, 0.012])**	[-0.177, 0.009] **([-0.172, 0.009])**
BMI	[0.167, 0.406] **([0.173, 0.407])**	[0.170, 0.410] **([0.175, 0.409])**	[0.169, 0.406] **([0.170, 0.400])**
BP	[-0.024, 0.196] **([-0.017, 0.197])**	[-0.02, 0.194] **([-0.023, 0.192])**	[-0.026, 0.194] **([-0.027, 0.188])**
S1	[-0.166, 0.088] **([-0.159, 0.091])**	[-0.161, 0.090] **([-0.162, 0.085])**	[-0.160, 0.086] **([-0.158, 0.083])**
S2	[-0.123, 0.106] **([-0.122, 0.105])**	[-0.118, 0.104] **([-0.118, 0.101])**	[-0.129, 0.097] **([-0.134, 0.089])**
S3	[-0.223, 0.015] **([-0.222, 0.008])**	[-0.231, 0.012] **([-0.225, 0.011])**	[-0.231, 0.016] **([-0.228, 0.014])**
S4	[-0.087, 0.160] **([-0.081, 0.163])**	[-0.088, 0.149] **([-0.086, 0.147])**	[-0.085, 0.151] **([-0.085, 0.148])**
S5	[0.235, 0.504] **([0.233, 0.496])**	[0.233, 0.493] **([0.237, 0.492])**	[0.235, 0.504] **([0.240, 0.502])**
S6	[-0.029, 0.151] **([-0.029, 0.147])**	[-0.029, 0.149] **([-0.027, 0.146])**	[-0.034, 0.151] **([-0.029, 0.152])**

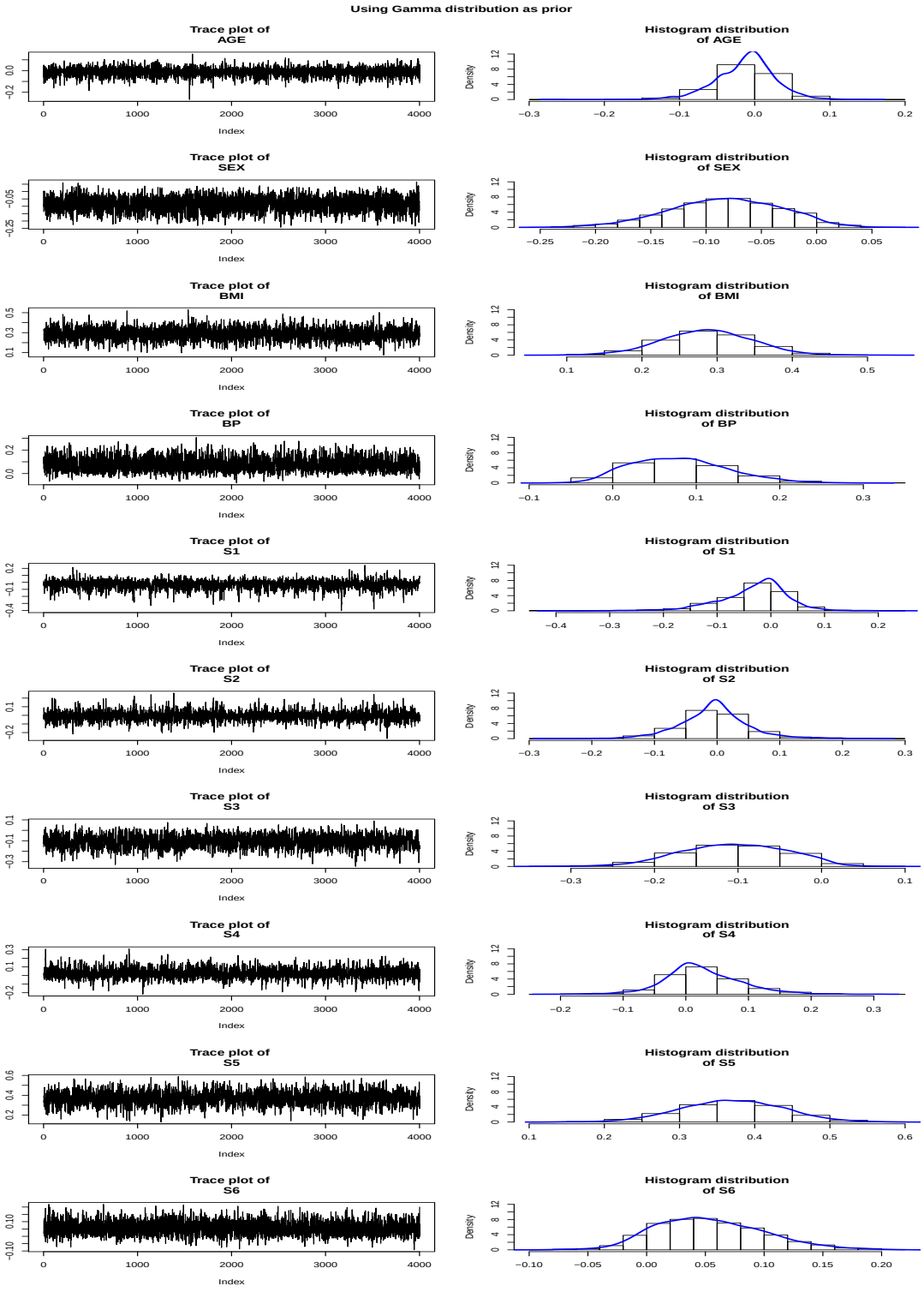

Figure 4.11: Trace plot and histogram along with the kernel density of the posterior mean estimates for the BEL lasso coefficients under gamma hyperprior; using 5000 iterations with 1000 burn-in

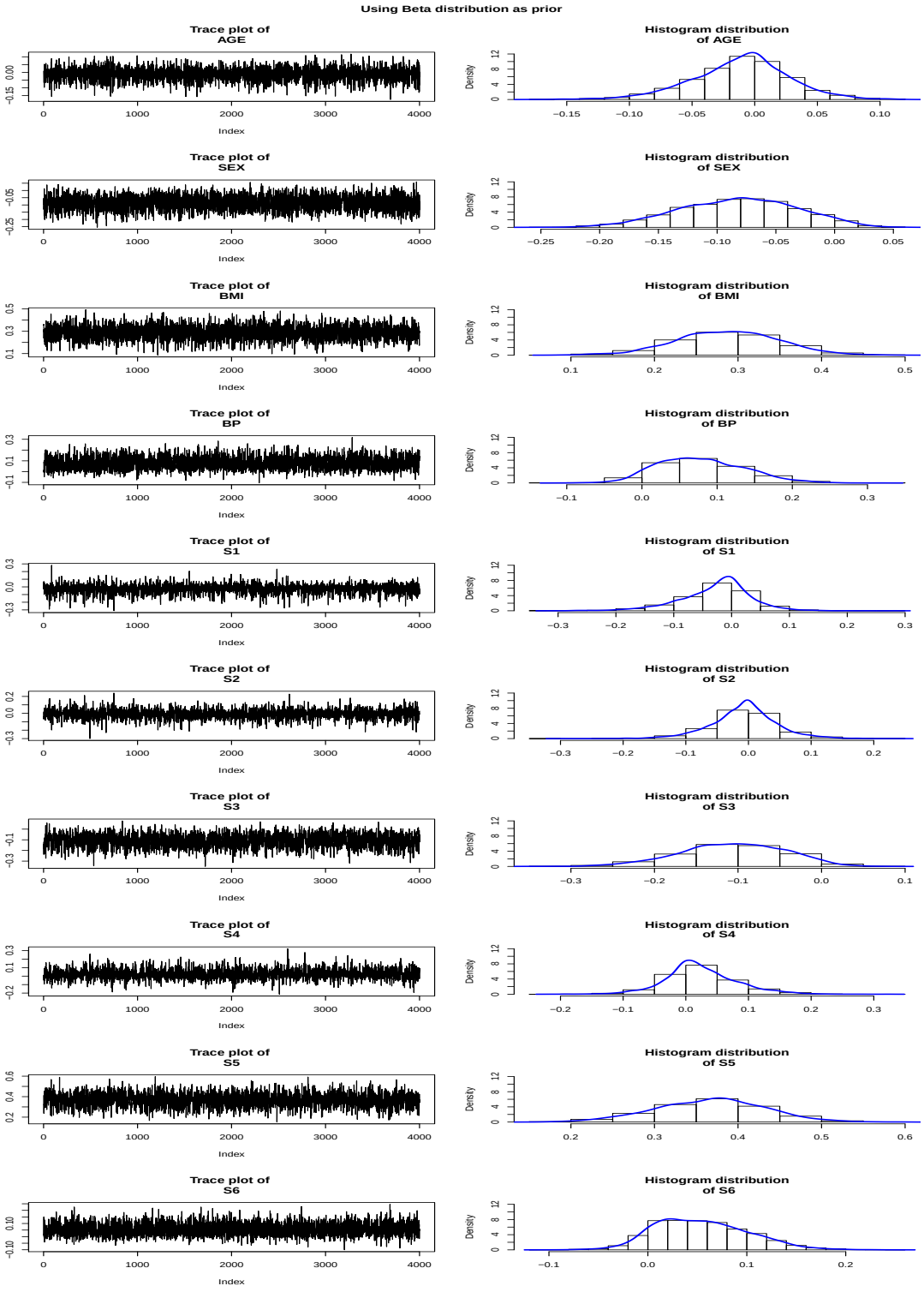

Figure 4.12: Trace plot and histogram along with the kernel density of the posterior mean estimates for the BEL lasso coefficients under beta hyperprior; using 5000 iterations with 1000 burn-in

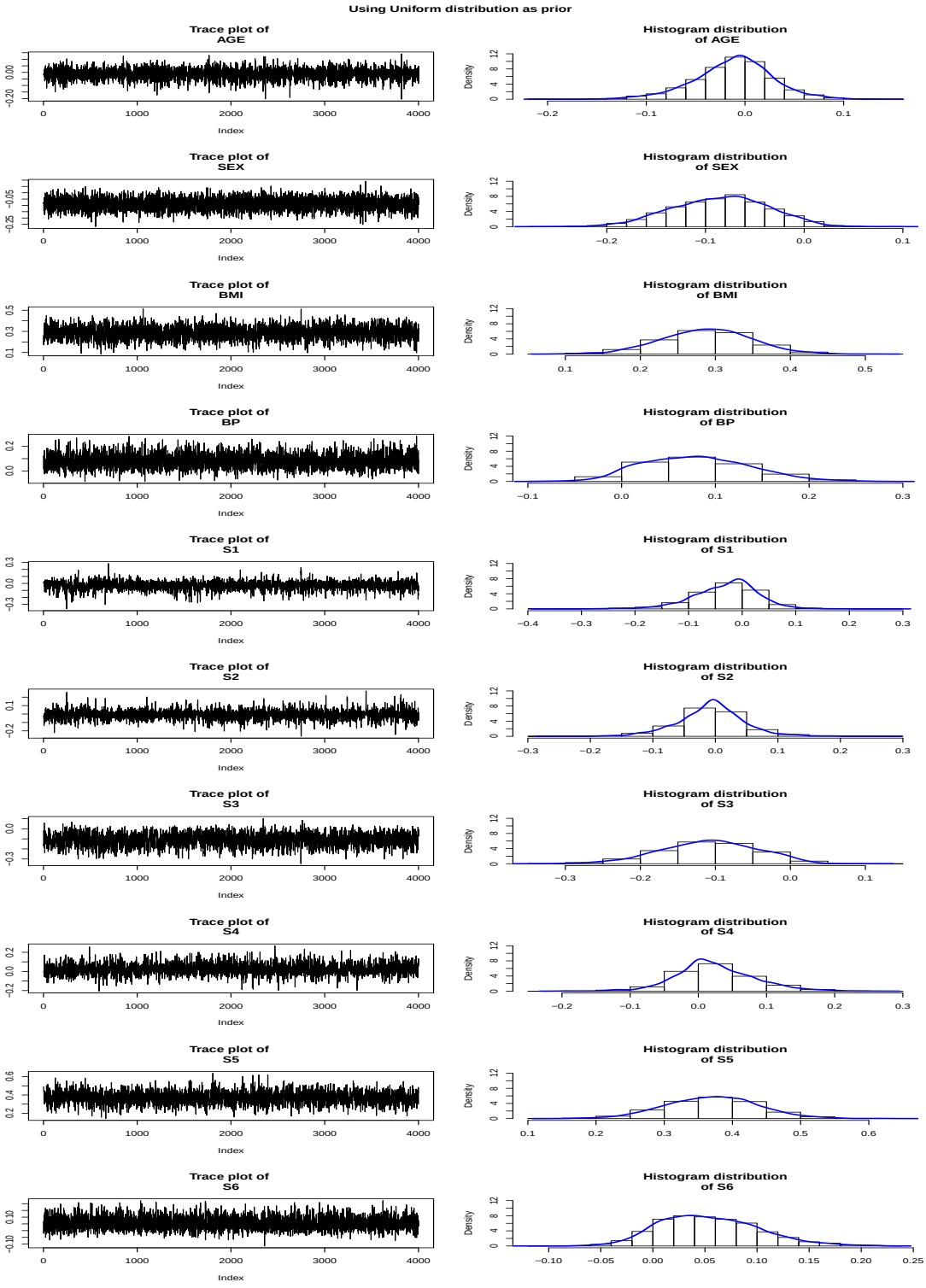

Figure 4.13: Trace plot and histogram along with the kernel density of the posterior mean estimates for the BEL lasso coefficients under uniform hyperprior, using 5000 iterations with 1000 burn-in

Ridge Case

Contrary to the lasso case, a hyperprior is placed on α (not α^2). That is, if we assume that α follows a gamma distribution with shape r and rate d, then in combination with (4.5) the resulting conditional posterior distribution is:

$$\pi(\alpha|\boldsymbol{\theta}) \propto \alpha^{r-1} \exp\left(-\alpha\left[d + \frac{\boldsymbol{\theta}^{\boldsymbol{T}}\boldsymbol{\theta}}{2\sigma^2}\right]\right),$$

which is a gamma distribution with shape r and rate $d + \dfrac{\boldsymbol{\theta}^{\boldsymbol{T}}\boldsymbol{\theta}}{2\sigma^2}$.

Example

We use the same example presented in Example 4.5. We split the data into 50% training and 50% testing, and performed 5-fold cross-validation over a range of values of α on the training set. The ridge parameter that minimizes the average cross-validation mean square was 0.32. Thus, the hyper-hyperparameters are chosen such that the posterior mean is around 0.32. We run an MCMC with 5000 iterations and 1000 burn-in. The posterior mean of the shrinkage parameter is 0.2912, which is close to 0.32. Figure 4.14 presents the trace plot and the kernel density for the shrinkage parameter when the hyperprior distribution is gamma. The estimates are around the center with reasonable fluctuation, which indicates that the chain is mixing well. From the kernel density, one can see that the distribution is right-skewed and has the shape of gamma distribution.

Table 4.6 represents the posterior mean estimates of the Bayesian ridge based on empirical likelihood under the gamma hyperprior. From Figure 4.15 it appears that the trace plots are around their centers with reasonable fluctuations, which indicate that the chains have good mixing. Table 4.7 provides the 95% highest posterior density intervals and the 95% equal-tailed credible regions. It seems that both the HPD and credible regions are quite similar.

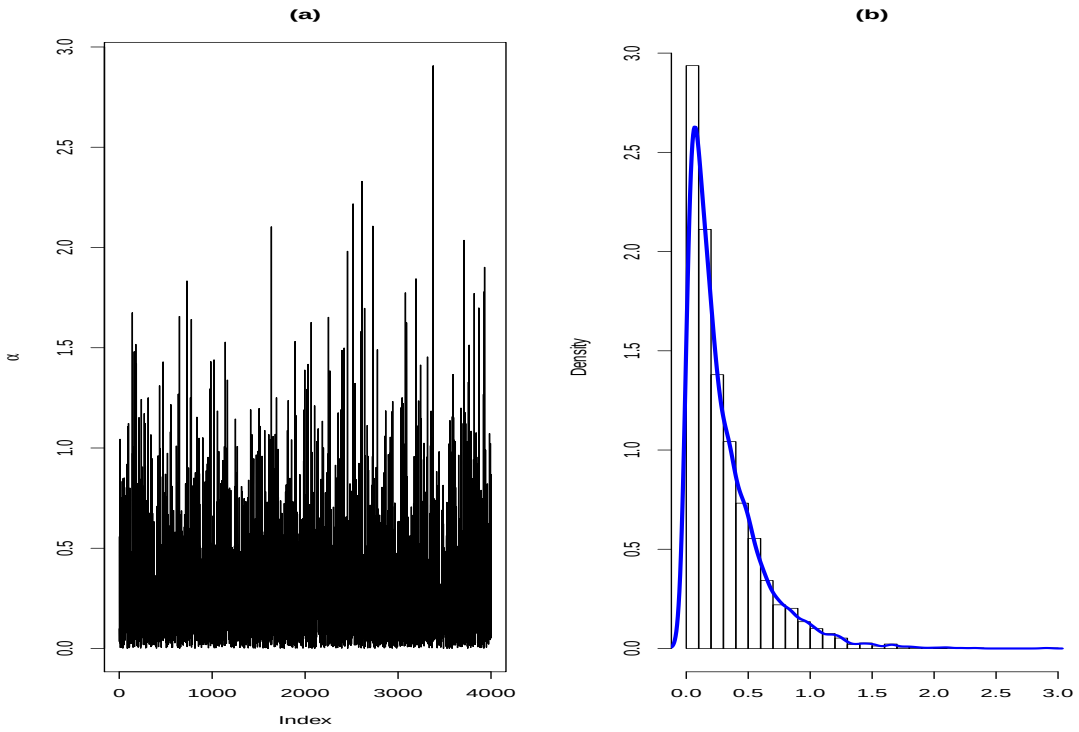

Figure 4.14: Trace plot (a) and histogram along with the kernel density (b) of the posterior mean estimates for the shrinkage parameter under gamma hyperprior in the BEL ridge; using 5000 iterations with 1000 burn-in

Table 4.6: The posterior mean estimates for the Bayesian ridge based on the empirical likelihood method using gamma distribution as hyperprior on the penalty term.

Hyperprior	AGE	SEX	BMI	BP	S1	S2	S3	S4	S5	S6
Gamma	-0.037	-0.131	0.300	0.117	-0.113	0.043	-0.117	0.012	0.397	0.073

Table 4.7: The 95% highest posterior density intervals and the 95% credible regions of the posterior distribution of the coefficients in the ridge model using gamma distribution as hyperprior on the penalty term.

Variables	95% HPD	95% credible regions
AGE	[-0.142, 0.065]	[-0.142, 0.060]
SEX	[-0.228, -0.031]	[-0.220, -0.027]
BMI	[0.186, 0.409]	[0.187, 0.406]
BP	[-0.012, 0.237]	[-0.010, 0.232]
S1	[-0.500, 0.245]	[-0.501, 0.234]
S2	[-0.258, 0.356]	[-0.257, 0.346]
S3	[-0.323, 0.097]	[-0.321, 0.090]
S4	[-0.205, 0.222]	[-0.201, 0.216]
S5	[0.230, 0.574]	[0.235, 0.573]
S6	[-0.026, 0.175]	[-0.025, 0.171]

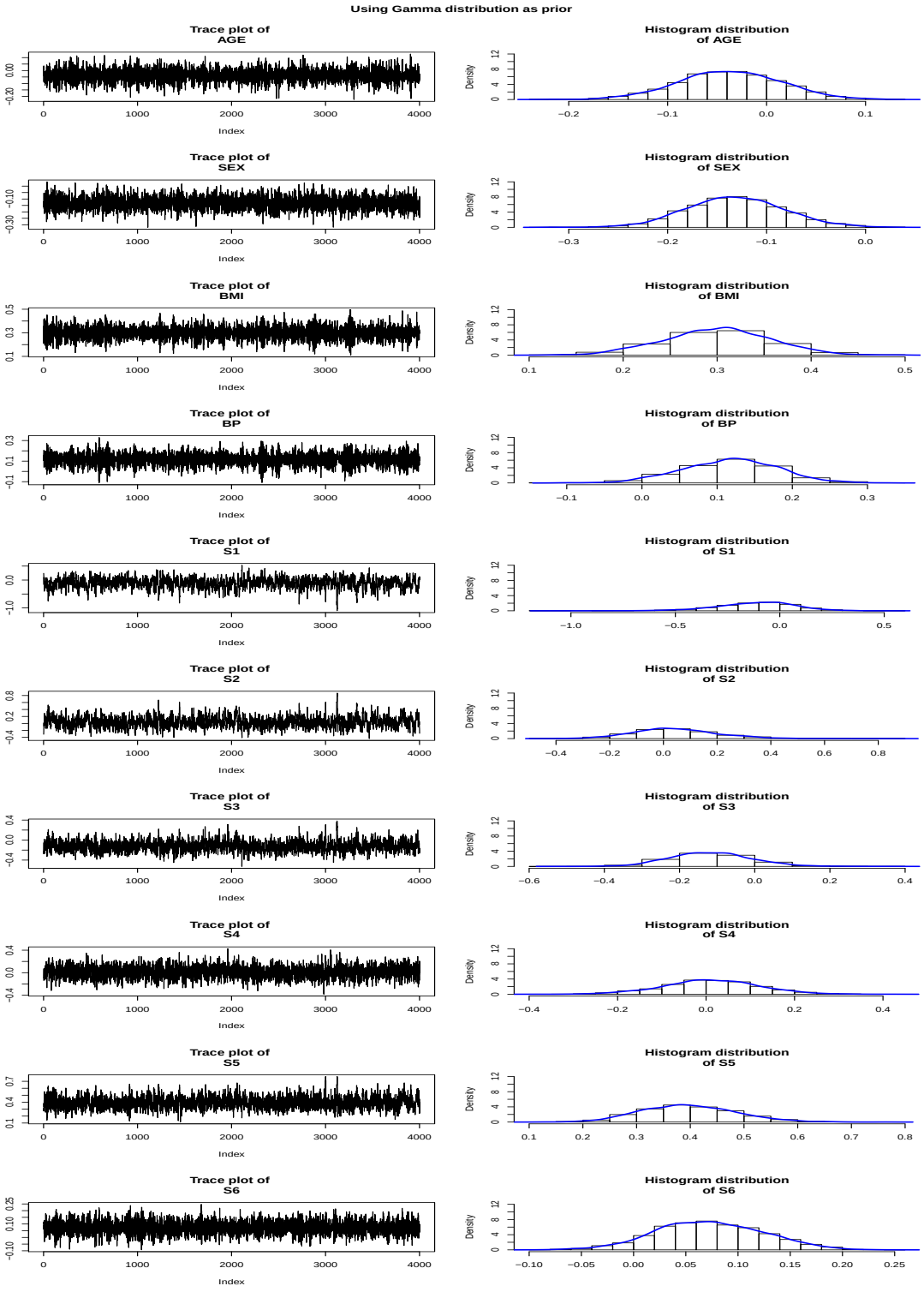

Figure 4.15: Trace plot and histogram along with the kernel density of the posterior mean estimates for the BEL ridge coefficients under gamma hyperprior; using 5000 iterations with 1000 burn-in

4.6 Summary

In this Chapter, we derived the Bayesian lasso and ridge based on empirical likelihood. We compared them to the full parametric Bayesian approach, and the results were similar. We discussed the estimation of the penalty term. Using the Bayesian approach, placing a hyperprior on the penalty term yields a conditional posterior distribution for α that does not depend on data. This makes it too sensitive to the hyper-hyperparameters and a bit challenging to determine the class of distributions to place on α. Another approach is to use the empirical Bayes by Marginal Maximum Likelihood (Casella, 2001).

Following Casella's (2001) approach, one can use a Monte Carlo expectation-maximization algorithm that complements the HMC and Gibbs sampler and provides marginal maximum likelihood. Each iteration involves running the block of HMC and Gibbs sampler using α estimated from the sample of the previous iteration. That is, iteration k uses α estimated in iteration $k-1$. For instance, in the lasso case, Equation (4.11) yields the complete-data log-likelihood:

$$
\begin{aligned}
&-\sum_{i=1}^{n}\log\left(1+\boldsymbol{\lambda}^T\boldsymbol{x_i}(\boldsymbol{y_i}-\boldsymbol{x_i}^T\boldsymbol{\theta})\right)-\frac{1}{2}\log\sigma^2--\frac{1}{2}\log|\boldsymbol{D_\tau}|-\frac{1}{2\sigma^2}\boldsymbol{\theta}^T\boldsymbol{D_\tau^{-1}}\boldsymbol{\theta}\\
&-(a+1)\log\sigma^2-\frac{b}{\sigma^2}+p\log\alpha^2-\frac{\alpha^2}{2}\sum_{j=1}^{p}\tau_j^2.
\end{aligned}
$$

The expectation step involves taking the expected value of the above distribution, conditional on $\boldsymbol{y}$ and under $\alpha^{(k-1)}$, to obtain:

$$
Q(\alpha|\alpha^{(k-1)})=p\log\alpha^2-\frac{\alpha^2}{2}\sum_{j=1}^{p}E_{\alpha^{(k-1)}}\left[\tau_j^2|\boldsymbol{y}\right]+\text{ term not involving }\alpha.
$$

The maximization step maximized $Q(\alpha|\alpha^{(k-1)})$ over α to obtain the next estimate. In this case we have

$$\alpha^{(k)} = \sqrt{\frac{2p}{\sum_{j=1}^{p} E_{\alpha^{(k-1)}}\left[\tau_j^2|\boldsymbol{y}\right]}}.$$

Park and Casella (2008) suggested to use the following initial value:

$$\alpha^{(0)} = \frac{p\sqrt{\hat{\sigma^2_{LS}}}}{\sum_{j=1}^{p}|\hat{\theta_j^{LS}}|}$$

where $\hat{\sigma^2_{LS}}$ and $\hat{\theta_j^{LS}}$ are estimated from the least squares approach.

Chapter 5

Summary and Directions for Future Research

5.1 Summary

We proposed a new approach for linear regression, ridge regression, and lasso regression based on Bayesian empirical likelihood. These are considered semi-parametric models because they combined a non-parametric and a parametric part. By using the profile EL ratio instead of likelihood, we avoided the potential problem of model misspecification.

The Bayesian EL ratio approach was straightforward. However, the resulting posterior distribution was intractable with complex and non-convex support. The nature of the support made the implementation of the traditional Markov Chain Monte Carlo algorithms difficult. First, the application of the Gibbs sampler was impossible because the kernel of the posterior distributions was unknown. Second, the implementation of the Metropolis-Hastings sampler was very challenging because it required the estimation of the proposal density and its parameters. In fact, even if we proposed a correct jumping density, the algorithm converged only when the parameters were equal to the maximum likelihood esti-

mator. If we started with a value far from the global optimum, the chain got trapped and never converged. To overcome this, we used the Hamiltonian Monte Carlo (Neal, 2011). The HMC used the gradient information to reduce random walk behavior, which led to faster convergence. It only required the derivation of the gradient and the values that control the HMC process. These values can be found by trial and error. Chaudhuri et al. (2017) showed that, under certain assumptions, once the parameters are inside the support, they never go outside and always jump back to the interior of the support if they reach its boundary.

We discussed that the penalized regression has a close connection to the Bayesian linear model. It sufficed to place a prior distribution on the parameters of the model where the penalty was introduced in the form of a hyperprior. For this reason, we started first with the derivation of BEL for the linear regression by deriving its empirical likelihood ratio. That is, to obtained the ridge regression and lasso regression, we multiplied the empirical likelihood ratio by the appropriate priors. For instance, we obtained the lasso model and ridge model if we placed the double exponential, using Andrews and Mallows' (1974) representation, and the normal distribution on the regression parameters, respectively. We compared our approach to a pure Bayesian approach, and we obtained similar results but the BEL was more robust because it did not rely on making, e.g., normality assumption on the data.

The penalty term α plays a major role in the shrinkage of the parameters. That is, they shrink to zero as we increase the value of α. To estimate its value, one can use cross-validation. In this dissertation, we treated it as a random parameter, and we placed a hyperprior distribution on it. We used a family of gamma, uniform and beta distributions. The disadvantage of this approach was that the resulting posterior distribution for α did not depend on the data, which made it very sensitive to the hyper-hyperparameters. For instance, let us assume that we placed a gamma prior with shape a and rate b on the shrinkage coefficient. Then, the posterior conditional distribution of α is an updated gamma distribution with shape $a + constant1$ and rate $b + constant2$. Then, the mean posterior

will be around $\frac{a + constant1}{b + constant2}$, which will be feed into the rest of the model. To investigate the value of α we divided data into two datasets: training and validation. We ran a 5-fold cross-validation on the training set and determined the value α^* that minimized the mean squared error. Then we chose the hyper-hyperparameters a and b based on the value of α^*.

Moreover, the estimating equations forced the introduction of the Lagrange multiplier $\boldsymbol{\lambda}$ in the convex hull, which depends on the values of the regression parameters $\boldsymbol{\theta}$. In our approach, at each iteration, we sampled $\boldsymbol{\theta}$ and found $\boldsymbol{\lambda}$ that solved equation (2.8). Thus, it required a careful design of the algorithm to find $\boldsymbol{\lambda}$ because of the constraints imposed by the weights in equation (2.7). We followed the same approach presented by Owen (2001), which uses the concept of the convex duality.

5.2 Directions for Future Research

There are many various directions for future research that extend our work. We now list some future research directions based on the results we obtained.

Pure Hamiltonian Monte Carlo Approach

In the course of this research, we used a building block of MCMC methods to estimate our parameters of interest. Precisely, we implemented a block of HMC and Gibbs sampler. An alternative approach is to use only the HMC approach. We have seen that the Hamiltonian Monte Carlo method leads to a quicker convergence by rapidly reaching the space of high density. Subsequently, it will increase the speed and improve the efficiency of the algorithm. For instance, we have seen that the full joint empirical posterior distribution for the ridge

regression is

$$\pi(\boldsymbol{\theta}, \sigma^2 | X, \boldsymbol{y}, \alpha) \propto \left(\frac{1}{\sigma^2}\right)^{p/2+a+1} \exp\left(-\sum_{i=1}^{n} \log\left[1 + \boldsymbol{\lambda}^T \boldsymbol{x_i} (y_i - \boldsymbol{x_i}^T \boldsymbol{\theta})\right] - \frac{1}{\sigma^2}\left[b + \frac{\alpha \boldsymbol{\theta}^T \boldsymbol{\theta}}{2}\right]\right). \tag{5.1}$$

The negative log of equation (5.1) is:

$$\sum_{i=1}^{n} \log\left[1 + \boldsymbol{\lambda}^T \boldsymbol{x_i} (y_i - \boldsymbol{x_i}^T \boldsymbol{\theta})\right] + \frac{1}{\sigma^2}\left[b + \frac{\alpha \boldsymbol{\theta}^T \boldsymbol{\theta}}{2}\right] + \frac{1}{\sigma^2}\left(b + \frac{\alpha \boldsymbol{\theta}^T \boldsymbol{\theta}}{2}\right) \tag{5.2}$$

Let $\boldsymbol{\theta_1} = (\boldsymbol{\theta},\ \sigma^2)^T$ be our parameters of interest. The gradient of equation (5.2) with respect to each parameter is

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\theta_1}}\left(-\log(\pi(\boldsymbol{\theta}, \sigma^2 | X, \boldsymbol{y}, \alpha))\right) &\propto \left(\frac{\partial}{\partial \boldsymbol{\theta}}\left(-\log(\pi(\boldsymbol{\theta}, \sigma^2 | X, \boldsymbol{y}, \alpha))\right),\ \frac{\partial}{\partial \sigma^2}\left(-\log(\pi(\boldsymbol{\theta}, \sigma^2 | X, \boldsymbol{y}, \alpha))\right)\right)^T \\ &= \left(\sum_{i=1}^{n} \frac{-\boldsymbol{\lambda} \boldsymbol{x_i} \boldsymbol{x_i}^T}{1 + \boldsymbol{\lambda}^T \boldsymbol{x_i} (y_i - \boldsymbol{x_i}^T \boldsymbol{\theta})} + \frac{\alpha}{\sigma^2} \boldsymbol{\theta}^T,\ \frac{p/2 + a + 1}{\sigma^2} + \frac{b + \alpha \boldsymbol{\theta}^T \boldsymbol{\theta}/2}{\sigma^4}\right)^T.\end{aligned}$$

In this case, our model has $p+2$ parameters, which are represented by a single vector where the parameter σ^2 is restricted under the model to be positive.

Similarly, the full joint empirical posterior distribution for the lasso regression is

$$\begin{aligned}\pi(\boldsymbol{\theta}, \sigma^2, \tau_1^2, \cdots, \tau_p^2 | X, \boldsymbol{y}) \propto &\exp\left(-\sum_{i=1}^{n} \log\left(1 + \boldsymbol{\lambda}^T \boldsymbol{x_i}\left(y_i - \boldsymbol{x_i}^T \boldsymbol{\theta}\right)\right)\right)\left(\frac{1}{\sigma^2 |D_\tau|}\right)^{1/2} \exp\left(-\frac{1}{2\sigma^2} \boldsymbol{\theta}^T D_\tau^{-1} \boldsymbol{\theta}\right) \\ &(\sigma^2)^{-a-1} \exp\left(-\frac{b}{\sigma^2}\right) \prod_{i=1}^{p} \frac{\alpha^2}{2} \exp\left(-\alpha^2 \tau_j^2 / 2\right).\end{aligned} \tag{5.3}$$

The negative log of equation (5.3) is

$$\sum_{i=1}^{n} \log\left(1 + \boldsymbol{\lambda}^T \boldsymbol{x_i}\left(y_i - \boldsymbol{x_i}^T \boldsymbol{\theta}\right)\right) + \frac{\log(\sigma^2 |D_\tau|)}{2} + \frac{\boldsymbol{\theta}^T D_\tau^{-1} \boldsymbol{\theta}}{2\sigma^2} + (a+1)\log(\sigma^2) + \frac{b}{\sigma^2} + \sum_{j=1}^{p} \frac{\alpha^2 \tau_j^2}{2}. \tag{5.4}$$

Let $\boldsymbol{\theta_2} = (\boldsymbol{\theta},\ \boldsymbol{\tau},\ \sigma^2)^T$ be our parameters of interest where $\boldsymbol{\tau} = (\tau_1^2, \cdots, \tau_p^2)$. The gradient of equation (5.4) with respect to each parameter is

$$\frac{\partial}{\partial \boldsymbol{\theta_2}} \left(-\log \pi(\boldsymbol{\theta}, \sigma^2, \boldsymbol{\tau} | X, \boldsymbol{y}) \right) = \begin{pmatrix} -\dfrac{\partial}{\partial \boldsymbol{\theta}} \log \pi(\boldsymbol{\theta}, \sigma^2, \boldsymbol{\tau} | X, \boldsymbol{y}) \\ -\dfrac{\partial}{\partial \boldsymbol{\tau}} \log \pi(\boldsymbol{\theta}, \sigma^2, \boldsymbol{\tau} | X, \boldsymbol{y}) \\ -\dfrac{\partial}{\partial \sigma^2} \log \pi(\boldsymbol{\theta}, \sigma^2, \boldsymbol{\tau} | X, \boldsymbol{y}) \end{pmatrix}$$

where

$$\begin{aligned} -\frac{\partial}{\partial \boldsymbol{\theta}} \log \pi(\boldsymbol{\theta}, \sigma^2, \boldsymbol{\tau} | X, \boldsymbol{y}) &= \sum_{i=1}^{n} \frac{-\boldsymbol{\lambda} \boldsymbol{x_i} \boldsymbol{x_i}^T}{1 + \boldsymbol{\lambda}^T \boldsymbol{x_i} (y_i - \boldsymbol{x_i}^T \boldsymbol{\theta})} + \frac{\boldsymbol{\theta}^T D_\tau}{\sigma^2} \\ -\frac{\partial}{\partial \sigma^2} \log \pi(\boldsymbol{\theta}, \sigma^2, \boldsymbol{\tau} | X, \boldsymbol{y}) &= \frac{1}{2\sigma^2} - \frac{\boldsymbol{\theta}^T D_\tau \boldsymbol{\theta}}{2\sigma^4} - \frac{b}{\sigma^4} \\ -\frac{\partial}{\partial \tau_j} \log \pi(\boldsymbol{\theta}, \sigma^2, \tau_j | X, \boldsymbol{y}) &= \frac{1}{2\tau_j^2} + \frac{\tau_j^2}{2} \left[\frac{\theta_j^2}{\sigma^2} + \alpha^2 \right]. \end{aligned}$$

Our model has $2p + 2$ parameters, which are represented by a single vector where the parameters $\tau_1^2, \cdots, \tau_p^2, \sigma^2$ are restricted under the model to be positive.

Ridge Regression

The l_2 penalty in ridge regression has a nice form such that the solution of $\boldsymbol{\theta}$ has a closed form:

$$\hat{\boldsymbol{\theta}} = (X^T X + \alpha \boldsymbol{I_{p \times p}})^{-1} X^T \boldsymbol{y}. \tag{5.5}$$

Therefore, instead of including the penalty term in the form of a hyperprior, one can the use

solution in (5.5) as the estimating equation to maximize our profile likelihood:

$$R(\boldsymbol{\theta}) = \max_{w_i}\Big\{\prod_{i=1}^{n} nw_i |\ w_i \geq 0,\ \sum_{i=1}^{n} w_i = 1,\ \boldsymbol{w}^T\left(X^TX + \alpha \boldsymbol{I_{p\times p}}\right)\boldsymbol{\theta} = \boldsymbol{w}^T X^T \boldsymbol{y}\Big\}$$

and place a normal prior on $\boldsymbol{\theta}$ with known parameters. Note that:

$$\begin{aligned}
\boldsymbol{w}^T\left(X^TX + \alpha \boldsymbol{I_{p\times p}}\right)\boldsymbol{\theta} - \boldsymbol{w}^T X^T \boldsymbol{y} &= \boldsymbol{w}\left(X^TX + \alpha \boldsymbol{I_{p\times p}}\boldsymbol{I_{p\times p}}^T\right)\boldsymbol{\theta} - \boldsymbol{w}^T X^T \boldsymbol{y} \\
&= \sum_{i=1}^{n} w_i\left(\left[\boldsymbol{x_i}\boldsymbol{x_i}^T + \alpha \boldsymbol{i_i}\boldsymbol{i_i}^T\right]\boldsymbol{\theta} - \boldsymbol{x_i} y_i\right) \\
&= \sum_{i=1}^{n} w_i\left(\left[\boldsymbol{x_i}\boldsymbol{x_i}^T + \alpha\right]\boldsymbol{\theta} - \boldsymbol{x_i} y_i\right)
\end{aligned}$$

where $\boldsymbol{i_i}$ is the i^{th} row of $I_{p\times p}$ and $\boldsymbol{i_i}\boldsymbol{i_i}^T = 1$. Using steps similar to those presented in Section 2.2, the profile likelihood ratio is given by $-\sum_{i=1}^{n}\log\left(1 + \boldsymbol{\lambda}^T\left[\boldsymbol{x_i} y_i - \left(\boldsymbol{x_i}\boldsymbol{x_i^t} + \alpha\right)\boldsymbol{\theta}\right]\right)$ where $\boldsymbol{\lambda}$ satisfies p equations given by:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{x_i} y_i - \left(\boldsymbol{x_i}\boldsymbol{x_i^t} + \alpha\right)\boldsymbol{\theta}}{1 + \boldsymbol{\lambda}^T\left[\boldsymbol{x_i} y_i - \left(\boldsymbol{x_i}\boldsymbol{x_i^t} + \alpha\right)\boldsymbol{\theta}\right]} = \boldsymbol{0}.$$

Bayesian Empirical Likelihood by Placing a Prior Distribution on Weights $w_1, \cdots, w_n$

The method presented in this dissertation is based on the empirical distribution of $\boldsymbol{\theta}$ under an informative prior on $\boldsymbol{\theta}$. The profile log EL ratio for the linear regression is

$$l_{EL}(\boldsymbol{\theta}) = -\sum_{i=1}^{n}\log\left\{1 + \boldsymbol{\lambda}^T\boldsymbol{x_i}\left[y_i - \boldsymbol{x_i}^T\boldsymbol{\theta}\right]\right\} = n\log n + \sum_{i=1}^{n}\log w_i \propto \log\prod_{i=1}^{n} w_i$$

and depends on weights. An alternative approach is to treat $(w_1, \cdots, w_p)$ as unknown parameters by placing a prior distribution on $w_1, \cdots, w_n$ instead of $\boldsymbol{\theta}$. The maximum of the profile empirical likelihood ratio, $R(\boldsymbol{w}) \propto \prod_{i=1}^{n} w_i$, is computed by maximizing w_i subject

to $\sum_{i=1}^{n} w_i = 1$ and $w_i \in (0,1)$. One can consider placing a Dirichlet distribution prior $D(\gamma_1, \cdots, \gamma_n)$ on $(w_1, \cdots, w_n)$. The Dirichlet distribution of order $n \geq 2$ with parameters $\gamma_1, \cdots, \gamma_n > 0$ has a probability density function:

$$\pi(w_1, \cdots, w_n | \gamma_1, \cdots, \gamma_n) = \frac{1}{B(\boldsymbol{\gamma})} \prod_{i=1}^{n} w_i^{\gamma_i - 1}, \ \forall w_i \in (0,1),$$

where $\sum_{i=1}^{n} w_i = 1$ and $B(\boldsymbol{\gamma}) = \dfrac{\Gamma(\gamma_1, \cdots, \gamma_n)}{\Gamma(\gamma_1)\Gamma(\gamma_2), \cdots, \Gamma(\gamma_n)}$. The posterior distribution of $(w_1, \cdots, w_n)$ given X is a Dirichlet distribution $D(\gamma_1 + 1, \gamma_2 + 1, \cdots, \gamma_n + 1)$ and is given by:

$$\pi(w_1, \cdots, w_n | X) \propto \prod_{i=1}^{n} w_i^{\gamma_i}. \tag{5.6}$$

This approach is similar to the Bayesian bootstrap (Rubin, 1981) that places a Dirichlet prior on the parameter of interest. An appropriate choice of the Dirichlet prior is the improper Dirichlet-Haldane prior (Aitkin, 2008) corresponding to $\gamma_i = 0 \ , \ \forall \ i = 1, \cdots, n$. The distribution in equation (5.6) can be approximated by Markov Chain Monte Carlo; however, it is difficult to translate knowledge about $\boldsymbol{w}$ into knowledge of $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$. Recall that:

$$\hat{w}_i = n^{-1} \left\{ 1 + \hat{\boldsymbol{\lambda}}^T \boldsymbol{x_i} (y_i - \boldsymbol{x_i}^T \hat{\boldsymbol{\theta}}) \right\}^{-1}.$$

The relationship between the weights and $\boldsymbol{\theta}$ is not a straightforward transformation. It is not easy to estimate the value of $\boldsymbol{\theta}$ from the weights, which are $p+1$ and n dimensional vectors, respectively.

When $p \gg n$

When the sample size is smaller than the number of predictor variables, we encounter two problems. First, the curse of dimensionality is acute. Second, there is insufficient informa-

tion, degrees of freedom, to estimate the full model. Penalized regressions were introduced to overcome the sparsity problem and to deal with data where $p >> n$. For example, the lasso regression imposes a l_2 penalty on regression parameters and tends to find an estimate of $\boldsymbol{\theta}$ that is equal to zero.

Our approach is Bayesian based on the EL method, and we believe it will fail in the case when $p >> n$. We incorporated the estimating equations in the convex hull to maximize our profile empirical likelihood ratio. However, $X^T X$ is not invertible when $p > n$ because it is not of full rank. Hence, $w_i = 0$, for $i = 1, \cdots, n$. The estimating equations for linear regression are $X^T X\boldsymbol{\theta} - X^T \boldsymbol{y}$. One can set $\tilde{\Sigma} = X^T X + cI$ by adding a small value c to the diagonal of $X^T X$. Then, use $\tilde{\Sigma}$ in the estimating equations instead of $X^T X$. But the main question is how we choose a value for c.

Bibliography

Aitkin, M. "Applications of the Bayesian bootstrap in finite populations inference." *Journal of Official Statistics*, 24(1):21–51 (2008).

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

Andrews, D. F. and Mallows, C. I. "Scale mixtures of normal distributions." *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(1):99–102 (1974).

Bae, K. and Mallick, B. K. "Gene selection using a two-level hierarchical Bayesian model." *Bioinformatics*, 20(18):3424–3430 (2004).

Baggerly, K. A. "Empirical likelihood as a goodness-of-fit measure." *Biometrika*, 85(3):535–547 (1998).

Bernardo, J. M. and Smith, A. F. M. *Bayesian Theory*. John Wiley & Sons, Chichester, New York, USA (1994).

Casella, G. "Empirical Bayes Gibbs sampling." *Biostatistics*, 2(4):485–500 (2001).

Chaudhuri, S., Drton, M., and Richardson, T. S. "Estimation of a covariance matrix with zeros." *Biometrika*, 94(1):199–216 (2007).

Chaudhuri, S. and Ghosh, M. "Empirical likelihood for small area estimation." *Biometrika*, 98(2):473–480 (2011).

Chaudhuri, S., Handcock, M. S., and Rendall, M. S. "Generalized linear models incorporating population level information: An empirical-Likelihood-based approachs." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 70(2):311–328 (2008).

Chaudhuri, S., Mondal, D., and Yin, T. "Hamiltonian Monte Carlo sampling in Bayesian empirical likelihood computation." *Royal Statistical Society Statistical Methodology Series B*, 79(1):293–320 (2017).

Chen, J. and Qin, J. "Empirical likelihood estimation for finite populations and the effective usage of auxiliary information." *Biometrika*, 80(1):107–116 (1993).

Chen, S. X. and Van Keilegom, I. "A review on empirical likelihood methods for regression." *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, 18(3):415–447 (2009).

Chhikara, R. S. and Folks, J. L. *The Inverse Gaussian Distribution: Theory, Methodology, and Applications*. New York, USA: Marcel Dekker, Inc. (1989).

Cooper, N. G., Eckhardt, R., and Shera, N. *From Cardinals to Chaos: Reflection on the Life and Legacy of Stanislaw Ulam*. New York, USA: Cambridge University Press (1989).

Craven, P. and Wahba, G. "Smoothing noisy data with spline functions." *Numerische Mathematik*, 31(4):377–403 (1978).

Darwin, C. *The Effects of Cross and Sefl Fertilisation in the Vegetebal Kingdom*. London: John Murray (1876).

Deville, J. C. and Sarndal, C. E. "Calibration estimators in survey sampling." *Journal of the American Statistical Association*, 87(418):376–382 (1992).

DiCiccio, T., Hall, P., and Romano, J. "Empirical likelihood is Bartlett-correctable." *The Annals of Statistics*, 19(2):1053–1061 (1991).

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. "Least angle regression." *The Annals of Statistics*, 32(2):407–451 (2004).

Efroymson, M. A. *Multiple Regression Analysis*. New York: In A. Ralston and H. S. Wilf (Eds.), Mathematical Methods for Digital Computers (1960).

Fan, J. and Li, R. "Variable selection via nonconcave penalized likelihood and its oracle properties." *Journal of the American Statistical Association*, 96(456):1348–1360 (2001).

Figueiredo, M. A. T. "Adaptive sparseness for supervised learning." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1150–1159 (2003).

Geisser, S. *Predictive Inference: An Introduction*. Chapman and Hill, New York (1993). Monographs on Statistics and Apllied probability 55.

Gelman, A. "Prior distributions for variance parameters in hierarchical models (comment on article by Brown and Draper)." *Bayesian Analysis*, 1(3):515–534 (2006).

Gelman, A., Carlin, J. B., Stern, H. C., Dunson, D. B., Vehtari, A., and Donald, R. B. *Bayesian Data Analysis*. Chapman and Hall/CRC (2013).

Geman, S. and Geman, D. "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741 (1984).

Godambe, V. P. "An optimum property of regular maximum likelihood estimation." *The Annals of Mathematical Statistics*, 31(4):1208–1211 (1960).

Grendár, M. and Judge, G. "Asymptotic equivalence of empirical likelihood and Bayesian MAP." *The Annals of Statistics*, 37(5a):2445–2457 (2009).

Hartley, H. O. and Rao, J. N. K. "A new estimation theory for sample surveys." *Biometrika*, 55(3):547–557 (1968).

Hastie, T., Tibshirani, J., Robert, and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. New York: Springer, second edition (2009).

Hastings, W. "Monte Carlo sampling methods using Markov chains and their applications." *Biometrika*, 57(1):97–109 (1970).

Härdle, W. *Smoothing Techniques with Implementation in S*. Spring series in statistics. Springer-Verlag, Berlin and Heidelberg GmbH & Co. K (1991).

James, G., Witten, D., Hastie, T., and Tibshirani, R. *An Introduction to Statistical Learning: with Applications in R*. Springer texts in statistics. Springer New York, first edition (2013).

Jaynes, E. T. "Information theory and statistical mechanics." *Physical Review. Series II*, 106(4):620–630 (1957).

Jennrich, R. I. "Asymptotic properties of non-linear least squares estimators." *The Annals of Mathematical Statistics*, 40(2):633–643 (1969).

Kolaczyk, E. D. "Empirical likelihood for generalized linear models." *Statistica Sinica*, 4(1):199–218 (1994).

Kosorok, M. R. *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York (2008).

Kuk, A. Y. C. and Mak, T. K. "Median estimation in the presence of auxiliary information." *Journal of the Royal Statistical Society. Series B (Methodological))*, 51(2):261–269 (1989).

Kutner, M., Nachtsheim, C., and Neter, J. *Applied Linear Regression Models*. Irwin series: Operations and decision sciences. McGraw-Hill/Irwin, fourth edition (2004).

Lazar, N. A. "Bayesian empirical likelihood." *Biometrika*, 90(2):319–326 (2003).

Los Campos, G., Naya, H., Gianola, D., Grossa, J., Legarra, A., Manfredi, E., Weigel, K., and Cotes, M. J. "Predicting quantitative traits with regression models for dense molecular markers and pedigree." *Genetics Society of America*, 182(1):375–385 (2009).

Lynch, S. M. *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. Statistics for Social and Behavioral Scince. New York: Springer (2007).

Mengersen, K. L., Pudlo, P., and Robert, C. P. "Bayesian computation via empirical likelihood." *Proceedings of the National Academy of Sciences of the United States of America*, 110(4):1321–1326 (2013).

Meredith, M. and Kruschke, J. *HDInterval: Highest (Posterior) Density Intervals* (2016). R package version 0.1.3.
URL `https://CRAN.R-project.org/package=HDInterval`

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. "Equation of state calculations by fast computing machines." *The Journal of Chemical Physics*, 21(6):1087–1092 (1953).

Molanes Lopez, E. M., Van Keilegom, I., and Veraverbeke, N. "Empirical likelihood for non-smooth criterion functions." *Scandinavian Journal of Statistics Theory and Applications*, 36(3):413–432 (2009).

Monahan, J. F. and Boos, D. D. "Proper likelihoods for Bayesian analysis." *Biometrika*, 79(2):271–278 (1992).

Neal, R. *Handbook of Markov Chain Monte Carlo, chapter 5: MCMC using Hamiltonian Dynamics*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Chapman & Hall/CRC (2011).

Owen, A. B. "Empirical likelihood ratio confidence intervals for a single functional." *Biometrika*, 75(2):237–249 (1988).

—. "Empirical likelihood ratio confidence regions." *The Annals of Statistics*, 18(1):90–120 (1990).

—. "Empirical likelihood for linear models." *The Annals of Statistics*, 19(4):1725–1747 (1991).

—. *Empirical Likelihood*. Chapman & Hall/CRC, Boca Raton (2001).

Park, T. . and Casella, G. "The Bayesian lasso." *Journal of the American Statistical Association*, 103(482):681–686 (2008).

Pearson, K. "On lines and planes of closest fit to systems of points in space." *Philosophical Magazine*, 2(6):559–572 (1901).

Qin, J. and Lawless, J. "Empirical likelihood and general estimating equations." *The Annals of Statistics*, 22(1):300–325 (1994).

Rao, J. N. K. and Wu, C. "Bayesian pseudo-empirical-likelihood intervals for complex surveys." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):533–544 (2010).

Rice, J. A. *Mathematical Statistics and Data Analysis*. Pacific Grove, California: Wadsworth and Brooks/Cole (1988).

Rockafellar, R. T. "Lagrange multipliers and optimality." *Society for Industrial and Applied Mathematics*, 35(2):183–238 (1993).

Rubin, D. B. "The Bayesian bootstrap." *The Annals of Statistics*, 9(1):130–134 (1981).

Schennach, S. M. "Point estimation with exponentially tilted empirical likelihood." *The Annals of Statistics*, 35(2):634–672 (2007).

Schwarz, G. "Estimating the dimension of a model." *The Annals of Statistics*, 6(2):461–464 (1978).

Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E., and Yang, N. "Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate II radical prostatectomy treated patients." *Journal of Urology*, 141(5):1076–1083 (1989).

Tibshirani, R. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 58(1):267–288 (1996).

Tikhonov, A. N. and Nikolayevich, N. "On the stability of inverse problems." *Doklady Akademii Nauk SSSR*, 39(5):195–198 (1943).

Van der Vaart, A. W. *Asymptotic Statistics*. Cambridge series in Statistical and Probablistic Mathematics 3. Cambridge: Cambridge University Press (1989).

Wilks, S. S. "The large-sample distribution of the likelihood ratio for testing composite hypotheses." *The Annals of Mathematical Statistics*, 9(1):60–62 (1938).

Wold, H. *Estimation of Principal Components and Related Models by Iterative Least squares*, 391–420. New York: Academic Press (1966).

Wu, C. "Weighted empirical likelihood inference." *Statistics & Probability Letters*, 66(1):67–79 (2004).

Yang, Y. and He, X. "Bayesian empirical likelihood for quantile regression." *The Annals of Statistics*, 40(2):1102–1131 (2012).

Yuan, M. and Lin, Y. "Efficient empirical Bayes variable selection and estimation in linear models." *Journal of the American Statistical Association*, 100(472):1215–1225 (2005).

Appendix A

MIXTURE OF NORMALS WITH AN EXPONENTIAL MIXING DENSITY

Let $z|s \sim N(0, s)$ and $s \sim \exp(\lambda^2/2)$. Then we have:

$$f_Z(z) = \int_0^\infty f_{Z|s}(z) f_s(s) d(s).$$

Now, find the moment generating function for Z:

$$\begin{aligned} M_Z(t) &= E_Z(e^{zt}) \\ &= \int_{-\infty}^{\infty} e^{zt} \int_0^\infty \frac{1}{\sqrt{2\pi s}} e^{-z^2/2s} \frac{\lambda^2}{2} e^{-s\lambda^2/2} ds dz \\ &= \int_0^\infty \frac{\lambda^2}{2} e^{-s\lambda^2/2} \int_{-\infty}^{\infty} \frac{e^{zt - z^2/2s}}{\sqrt{2\pi s}} dz ds. \end{aligned}$$

Note that:

$$\begin{aligned}\int_{-\infty}^{\infty} \frac{e^{zt-z^2/2s}}{\sqrt{2\pi s}} dz &= \frac{1}{\sqrt{2\pi s}} \int_{-\infty}^{\infty} e^{-\frac{1}{2s}\left[z^2-2szt\right]} dz \\ &= \frac{1}{\sqrt{2\pi s}} \int_{-\infty}^{\infty} e^{-\frac{1}{2s}\left[z^2-2tsz+(ts)^2-(ts)^2\right]} dz \\ &= \left(e^{\frac{(ts)^2}{2s}}\right)\left(\frac{1}{\sqrt{2\pi s}} \int_{-\infty}^{\infty} e^{-\frac{1}{2s}(z-ts)^2} dz\right) = e^{(ts)^2/2s}.\end{aligned}$$

The integral is equal to one because the second term is $N(ts, s)$. We have:

$$\begin{aligned}M_Z(t) &= \int_0^{\infty} \frac{\lambda^2}{2} e^{-\frac{-s\lambda^2}{2}+\frac{st^2}{2}} ds \\ &= \frac{\lambda^2}{2} \int_0^{\infty} e^{-s/2\left[\lambda^2-t^2\right]} ds \\ &= \frac{\lambda^2}{\frac{2(\lambda^2-t^2)}{2}} \left(e^{-s/2\left(\lambda^2-t^2\right)}\right)_0^{\infty} \\ &= \frac{\lambda^2}{\lambda^2-t^2} \\ &= \frac{1}{1-\frac{t^2}{\lambda^2}},\ t^2 < \frac{1}{\lambda^2},\end{aligned}$$

which is the moment generating function of the Laplace distribution with mean 0 and scale $\frac{1}{\lambda}$.

Appendix B

UNIMODALITY UNDER PRIOR

The heuristic proof in Park and Casella (2008) applies equally to the Bayesian empirical likelihood. We show that the joint posterior distribution $\pi(\boldsymbol{\theta}, \sigma^2|\boldsymbol{y})$ of $\boldsymbol{\theta}$ and $\sigma^2 > 0$ under the prior

$$\pi(\boldsymbol{\theta}, \sigma^2) = \pi(\sigma^2) \prod_{j=1}^{p} \frac{\alpha}{2\sqrt{\sigma^2}} \exp\left(-\frac{\alpha|\theta_j|}{\sqrt{\sigma^2}}\right)$$

is unimodal for typical choices of $\pi(\sigma^2)$ and any choice of $\alpha \geq 0$, such that for every $c > 0$ the upper level set

$$\left\{(\boldsymbol{\theta}, \sigma^2) : \pi(\boldsymbol{\theta}, \sigma^2|\boldsymbol{y}) > c,\ \sigma^2 > 0\right\}$$

is connected. The log-posterior distribution for $\boldsymbol{\theta}$ and σ^2 is proportional to:

$$\log(\pi(\sigma^2)) + \left(-\sum_{i=1}^{n} \log\left[1 + \boldsymbol{\lambda}^T X_i(y_i - X_i^T\boldsymbol{\theta})\right]\right) - \frac{\alpha}{\sqrt{\sigma^2}}||\boldsymbol{\theta}||.$$

The second term is clearly concave in $\boldsymbol{\theta}$ because its second derivative with respect to $\boldsymbol{\theta}$ is negative:

$$\frac{\partial}{\partial \boldsymbol{\theta}^T}\left(-\sum_{i=1}^{n} \log\left[1 + \boldsymbol{\lambda}^T X_i(y_i - X_i^T\boldsymbol{\theta})\right]\right) = \sum_{i=1}^{n} \frac{\boldsymbol{\lambda}^T X_i X_i^T}{1 + \boldsymbol{\lambda}^T X_i(y_i - X_i^T\boldsymbol{\theta})}$$

$$\frac{\partial}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\left(-\sum_{i=1}^{n} \log\left[1 + \boldsymbol{\lambda}^T X_i(y_i - X_i^T\boldsymbol{\theta})\right]\right) = -\sum_{i=1}^{n} \frac{\boldsymbol{\lambda}(X_i X_i^T)^T X_i X_i^T \boldsymbol{\lambda}^T}{\left(1 + \boldsymbol{\lambda}^T X_i(y_i - X_i^T\boldsymbol{\theta})\right)^2} < 0$$

because the denominator is positive and the numerator has a quadratic form and thus pos-

itive. The third term is clearly concave in σ^2 (the second derivative of l_1-norm is 0). If we chose σ^2 to be an inverse gamma or invariant prior $\frac{1}{\sigma^2}$, then $\log(\sigma^2)$ is clearly a concave function. The sum of concave functions is also concave. Therefor, the log posterior function for $(\boldsymbol{\theta}, \sigma^2)$ is unimodal.

Appendix C

DISTRIBUTION of τ_j^2

For $j = 1, \cdots, p$, the conditional posterior density of τ_j^2 is

$$\begin{aligned}
\left(1/\tau_j^2\right)^{1/2} \exp\left(-\frac{1}{2}\left(\alpha^2\tau_j^2 + \frac{\theta_j^2\sigma^2}{\tau_j^2}\right)\right) &= \left(1/\tau_j^2\right)^{1/2} \exp\left(-\frac{1}{2\tau_j^2}\left[\alpha^2(\tau_j^2)^2 + \frac{\theta_j^2}{\sigma^2}\right]\right) \\
&= \left(1/\tau_j^2\right)^{1/2} \exp\left(-\frac{\alpha^2}{2\tau_j^2}\left[(\tau_j^2)^2 + \frac{\theta_j^2}{\alpha^2\sigma^2}\right]\right).
\end{aligned}$$

For simplicity, let $\mu = \sqrt{\dfrac{\alpha^2\sigma^2}{\theta_j^2}}$ and $\tau_j^2 = w$, then

$$\begin{aligned}
\left(1/\tau_j^2\right)^{1/2} \exp\left(-\frac{1}{2}\left(\alpha^2\tau_j^2 + \frac{\theta_j^2\sigma^2}{\tau_j^2}\right)\right) &= (1/w)^{1/2} \exp\left(-\frac{\alpha^2}{2w}\left[w^2 + \frac{1}{\mu^2}\right]\right) \\
&= (1/w)^{1/2} \exp\left(-\frac{\alpha^2}{2\mu^2 w}\left[(\mu w)^2 + 1\right]\right) \\
&= (1/w)^{1/2} \exp\left(-\frac{\alpha^2}{2\mu^2 w}\left[1 - \mu w\right]^2 - \frac{\alpha^2}{2\mu^2 w} 2\mu w\right) \\
&\propto \left(\frac{\alpha^2}{2\pi w}\right)^{1/2} \exp\left(-\frac{\alpha^2}{2\mu^2 w}\left[1 - \mu w\right]^2\right).
\end{aligned}$$

Chhikara and Folks (1989) showed that if X followed an Inverse Gaussian (IG) distribution with mean μ and shape α^2 then its inverse, $\dfrac{1}{w}$, has the density above. Hence,

$$f_W(w; \mu, \alpha^2) = w\mu^2 f_X(w; \frac{1}{\mu}, \frac{\alpha^2}{\mu^2}).$$

Therefore, $\dfrac{1}{\tau_j^2} \sim IG(\sqrt{\dfrac{\alpha^2\sigma^2}{\theta_j^2}}, \alpha^2)$.