Durham E-Theses

High Dimensional Statistical Modelling with Limited Information

BASU, TATHAGATA

How to cite:

BASU, TATHAGATA (2021) *High Dimensional Statistical Modelling with Limited Information*, Durham theses, Durham University. Available at Durham E-Theses Online: http://etheses.dur.ac.uk/13920/

Use policy

The full-text may be used and/or reproduced, and given to third parties in any format or medium, without prior permission or charge, for personal research or study, educational, or not-for-profit purposes provided that:

- a full bibliographic reference is made to the original source
- a link is made to the metadata record in Durham E-Theses
- the full-text is not changed in any way

The full-text must not be sold in any format or medium without the formal permission of the copyright holders.

Please consult the full Durham E-Theses policy for further details.

Academic Support Office, The Palatine Centre, Durham University, Stockton Road, Durham, DH1 3LE
e-mail: e-theses.admin@durham.ac.uk Tel: +44 0191 334 6107
http://etheses.dur.ac.uk

High Dimensional Statistical Modelling with Limited Information

Tathagata Basu

A Thesis presented for the degree of
Doctor of Philosophy

Probability & Statistics
Department of Mathematical Sciences
Durham University
Durham, United Kingdom

March, 2021

Dedicated to

My family, friends and teachers.

High Dimensional Statistical Modelling with Limited Information

Tathagata Basu

Submitted for the degree of Doctor of Philosophy
March, 2021

Abstract

Modern scientific experiments often rely on different statistical tools, regularisation being one of them. Regularisation methods are usually used to avoid overfitting but we may also want use regularisation methods for variable selection, especially when the number of modelling parameters are higher than the total number of observations. However, performing variable selection can often be difficult under limited information and we may get a misspecified model. To overcome this issue, we propose a robust variable selection routine using a Bayesian hierarchical model.

We adapt the framework of Narisetty and He to propose a novel spike and slab prior specification for the regression coefficients. We take inspiration from the imprecise beta model and use a set of beta distributions to specify the prior expectation of the selection probability. We perform a robust Bayesian analysis over this set of distributions in order to incorporate expert opinion in an efficient manner.

We also discuss novel results on likelihood-based approaches for variable selection. We exploit the framework of the adaptive LASSO to propose sensitivity analyses of LASSO-type problems. The sensitivity analysis also gives us a novel non-deterministic classifier for high dimensional problems, which we illustrate using real datasets.

Finally, we illustrate our novel robust Bayesian variable selection using synthetic and real-world data. We show the importance of prior elicitation in variable selection as well as model fitting and compare our method with other Bayesian approaches for variable selection.

Declaration

The work in this thesis is based on the research carried out under the supervision of Dr. Jochen Einbeck and Professor Matthias C.M. Troffaes within the Department of Mathematical Sciences, Durham University, England. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all my own work unless referenced to the contrary in the text.

Copyright © 2021 by Tathagata Basu.
"The copyright of this thesis rests with the author. No quotations from it should be published without the author's prior written consent and information derived from it should be acknowledged".

Acknowledgements

First and foremost, I would like to thank my supervisors Dr. Jochen Einbeck and Prof. Matthias C. M. Troffaes for their constant support, guidance and encouragement throughout my doctoral studies. I would also like to thank Dr. Kumaresh Dhara for his valuable inputs and Daniel Krpelik for many interesting discussions which have helped shape this work.

I would also like to acknowledge the financial support provided by the European Commissions H2020 programme, through the UTOPIAE Marie Curie Innovative Training Network.

Finally, I would like to thank my friends and family for their constant support and motivation. Besides this, I wish to thank all the essential workers, who have been a part of my life and helped me to survive through this global pandemic.

Contents

List of Symbols

$\mathbb{R}^n$	Space of n-tuples of real numbers
$\boldsymbol{x}$	Matrix of predictor (input) variables
y	Vector of response (output) variable
β	Vector of regression parameters
ϵ	Vector of random noises
$\mathcal{N}(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2
$E(\cdot)$	Expectation operator
$P(\cdot)$	Probability density (mass for discrete random variables) function
$L(\cdot)$	Likelihood function
$\hat{\beta}$	Vector of regression coefficient estimates
$\lVert \cdot \rVert_2$	Standard Euclidean norm, that is $\lVert a \rVert_2 = \sqrt{\sum_j a_j^2}$
$\hat{\beta}_{\mathrm{OLS}}$	Vector of ordinary least square estimates
∂	Differential operator
$\hat{\beta}_{\mathrm{R}}$	Vector of ridge estimates
$\mathbf{I}_p$	$p \times p$ dimensional identity matrix
λ	Regularisation parameter
$\mathrm{diag}(m)$	Diagonal matrix with m as the diagonal element(s)
ℓ_1 or $\lVert \cdot \rVert_1$	Absolute norm, that is $\lVert a \rVert_1 = \sum_j \lvert a_j \rvert$
$\hat{\beta}_{\mathrm{NG}}$	Vector of non-negative Garrote estimates
$\mathbb{I}$	Indicator function
$\hat{\beta}_{\mathrm{L}}$	Vector of LASSO estimates
c	Vector of class (output) variable
$\hat{\beta}_{\mathrm{lr}}$	Vector of logistic regression estimates

$\hat{\beta}_{\text{plr}}$	Vector of penalised logistic regression estimates
π	Decision probability
$\hat{\pi}$	Estimated decision probability
$\ell(\cdot, \lambda)$	Lagrangian function
$\hat{\beta}_{\text{AL}}$	Vector of adaptive LASSO estimates
$\hat{\beta}_{\text{aplr}}$	Vector of adaptive penalised logistic regression estimates
$\mathcal{S}$	Index set of the true active covariates
p^*	Total number of true active covariates
β^*	Vector of true regression coefficients
$\beta^*_{\mathcal{S}}$	Vector of β corresponding to true active covariates
$\xrightarrow{d}$	Convergence in distribution
$\xrightarrow{p}$	Convergence in probability
β_{MAP}	Maximum a posteriori estimate of β
z	Vector of indicators, used for covariate selection
q	Vector of selection probabilities
$\delta_\eta(\cdot)$	Discrete mass concentrated at η
$\underline{P}(\cdot)\ \big(\overline{P}(\cdot)\big)$	Lower (upper) probability
$\underline{E}(\cdot)\ \big(\overline{E}(\cdot)\big)$	Lower (upper) expectation
$\mathcal{P}$	Subset of p-dimensional unit hypercube of probability measures
$\overset{t}{\propto}$	Porportional to a function of t

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

In this thesis, we focus on high-dimensional statistical modelling with limited data. That is, we try to find a mathematical relation between a response (or, output) variable and predictor (or, input) variable(s) in a regressional context and the number of observations is less than the number of predictors. High-dimensional statistical modelling is an integral part of several scientific and socio-economic problems such as space exploration, clinical trials, climate modelling, stock analysis, *etc.* However, in many of these fields, we have to model with limited information as the experiments are often expensive and time consuming. We, therefore, are interested in high-dimensional statistical models that are not sensitive with respect to perturbation in data and performs well in prediction as well.

The concept of statistical modelling dates back to the early nineteenth century. Legendre used the method of least squares and proposed a formulation, which is vaguely related to linear models [50]. The notion of high-dimensional statistical modelling is relatively new within the scientific community and became popular in late twentieth century. However, an elicitation-based method has not been proposed to tackle the lack of information. In this thesis, we will draw inspiration from Bayesian variable selection approaches to develop a novel variable selection approach. We first investigate different regularisation methods to understand the sensitivity of variable selection with respect to the regularisation term. We then address this issue of variable selection in a Bayesian paradigm. Parts of the sensitivity analysis have been published [5, 6].

Variable selection is a popular topic among both frequentist and Bayesian statisticians. Large datasets, such as gene micro-arrays often contain more predictors than the total number of observations. These datasets are often highly correlated and require variable selection methods to avoid overfitting. One of the foremost works in Bayesian variable selection was presented by Mitchell and Beauchamp [56]. The authors used a two-component prior to specify the regression coefficients. They proposed a point mass at 0 and a uniform distribution elsewhere. Later, George and McCulloch [41] proposed a Gibbs sampling method for variable selection, where they used latent variables to identify the active variables. Later Ishwaran and Rao [48] provided a more generalised framework for two-component priors and coined the term spike and slab prior. They used a continuous bi-modal prior for the regression coefficient.

The frequentist approach for variable selection became popular after Tibshirani [69] introduced the LASSO (or, least absolute selection and shrinkage operator). In the LASSO, an ℓ_1 penalty-term is added to the log-likelihood of the linear model. This type of penalty keeps the penalised likelihood convex unlike subset selection, where ℓ_0 is used as a penalty. Introduction of the LASSO led to several other works on the theoretical properties of variable selection. Fan and Li [34] worked on the oracle properties of the LASSO and showed that it can be inconsistent in variable selection and introduced the SCAD (or, smoothly clipped absolute deviation). Later in 2006, Zou [86] introduced the adaptive LASSO and showed that simple use of data driven weights can result in consistent variable selection and asymptotic unbiased estimates.

Introduction of the LASSO led to several Bayesian variable selection methods as well. Tibshirani [69] noted that a Laplace (double exponential) prior can be used to specify regression coefficient as a Bayesian alternative for the LASSO. Park and Casella [60] exploited the use of Laplace prior and proposed a hierarchical setup for variable selection. Lykou and Ntzoufras [54] proposed a modification of the model by Park and Casella [60] and introduced selection indicators. Other notable works are (i)the Dirichlet-LASSO by Bhattacharya et al. [11], using global-local mixtures of Gaussians for prior specification and (ii)the spike and slab LASSO by Ročková and George [64], using a Laplace prior to construct the spike and slab distributions.

A key issue for high-dimensional problems is the data sparsity. Most of the methods are often based on the assumption that the observed data is the true representation of the problem. However, in real life this may not be the case. Further observations may change the variability of the predictors and may suggest that more (or, less) predictors should be in the model. To overcome these issues, we will adapt a spike and slab model with robust Bayesian analysis.

A robust Bayesian analysis [10] considers a set of priors instead of fixing a single prior. It emphasises the fact that it is almost impossible to capture prior evidence by using a single prior and it is better to consider all the priors which are reasonable. A set of priors can be chosen based on several criteria. In this setting, we will focus on the range of prior hyper-parameters to obtain the set. We consider an imprecise beta model, which is a special case of Walley's imprecise Dirichlet model [76]. We use a set of imprecise beta distributions to specify our hyperprior for the selection probability of a co-variate.

We exploit the use of conjugate priors in our robust Bayesian analysis. Our use of a set of priors gives us a set of posteriors for efficient computation. We investigate the posterior estimates of the regression coefficients and selection indicators to obtain a robust variable selection. This is the first instance of an imprecise variable selection scheme for high-dimensional statistical modelling. Our evaluation of imprecision works in two levels. Our method allows a variable to be indeterminate which shows the imprecision in the variability of the co-variates. The second instance of imprecision is addressed through the posteriors of selected variables. The sets of posteriors provide a range for the posterior estimates instead of single values. This helps us to understand the indeterminacy in the model fitting. However, this type of imprecision is often dependent on the scale of the dataset and we use a relative measure to characterise this type of imprecision.

1.2 Contribution and outline

This thesis revolves around two novel methods for high-dimensional problems. One of these methods is sensitivity analysis for LASSO-type problemsChapter 5 and the other one is robust Bayesian variable selectionChapter 8. The span of the contributions motivates us to present this thesis from a unified frequentist-Bayesian perspective and we discuss the relevant developments accordingly. In this chapter, we have discussed some of the works in the high-dimensional statistics and the methods that we use to present our contributions. The rest of thesis is organised as follows. In Chapter 2, we explain statistical modelling in a formal manner and discuss different aspects of statistical modelling with a special focus on linear models.

After Chapter 2, this work is split into two broad categories. Chapter 3 deals with likelihood-based parameter estimation, followed by our novel contribution on LASSO-type problems in Chapter 5. On the other hand, Chapter 6 and 7 are focused on the Bayesian methodologies, followed by Chapter 8 where we present a novel variable selection scheme using robust Bayesian analysis. We also discuss the mathematics behind numerical optimisation along with different optimisation methods for likelihood-based approaches in Chapter 4.

In Chapter 3, we discuss several likelihood-based parameter estimation techniques. We introduce the notion of likelihood and maximum likelihood estimates as a frequentist point estimate and show the use of likelihood-based parameter estimation for linear models and regularisation techniques. We then discuss different variable selection routines using likelihood based estimation and model selection techniques for the best fit. Finally, we conclude this chapter by introducing inferential methods for regularisation techniques.

Chapter 4 is focused on the several optimisation methods, which are required for likelihood-based estimation. The chapter starts with basic notions of convexity and duality for constrained optimisation followed by a discussion on different optimisation techniques for non-smooth objective functions, which occur frequently in variable selection methods.

In Chapter 5, we discuss a sensitivity analysis for LASSO-type problems. We exploit the notion of adaptive LASSO to show how we can assess the variability of predictors and their effect on variable selection. We use this idea to introduce a novel credal classification routine for logistic regression. Parts of this work were published in [5, 6].

Chapter 6 is focused on the statistical inference in Bayesian paradigm. We discuss the role of subjective belief in statistical analysis and its effect in choosing a prior. We explore different Bayesian regression models and discuss their analogy with frequentist counterparts through maximum a posteriori estimates. Chapter 6 also gives a formal definition of spike and slab models, which are the basis of our robust Bayesian variable selection approach. We discuss different types of spike and slab models and their mathematical formulations.

Chapter 7 starts with the notion of robust Bayesian analysis. We discuss the scope of robust Bayesian analysis in high-dimensional models and introduce the imprecise beta model. Later, we formulate our model for variable selection using robust Bayesian analysis. We use this formulation in Chapter 8 to demonstrate our contribution in robust Bayesian variable selection. We analyse our model for various sets of selection probabilities. We discuss the effects of these selection probabilities through posterior means of the regression coefficients and selection indicators. We illustrate our method using several synthetic datasets and also perform a dedicated analysis of real datasets in Chapter 9

Finally in Chapter 10, we conclude this thesis. We discuss our findings and issues while investigating the problem.

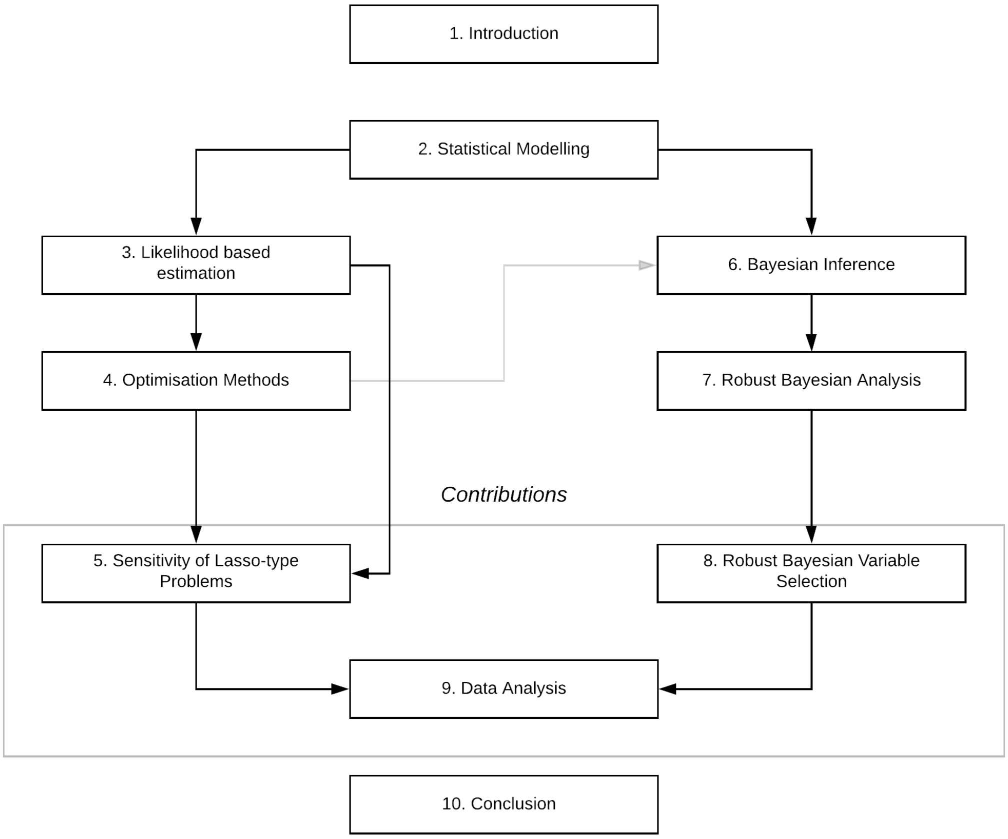
1. Introduction
2. Statistical Modelling
3. Likelihood based estimation
6. Bayesian Inference
4. Optimisation Methods
7. Robust Bayesian Analysis
Contributions
5. Sensitivity of Lasso-type Problems
8. Robust Bayesian Variable Selection
9. Data Analysis
10. Conclusion

Figure 1.1: Flow diagram of the thesis

Chapter 2

Statistical Modelling

For an efficient statistical inference from a population, we need a suitable model that describes the characteristics of the population. The population may contain several random or deterministic variables. In statistical modelling, we establish a mathematical relationship between these variables by using statistical assumptions. In a regression context, these variables can be categorised as response variables and predictor variables. We can also characterise predictor variables as independent variables and response variables as dependent variables. We can describe predictors and responses in the following way:

1. Predictor (or, independent) variables are characteristics of the system which directly control the properties of the system.

2. Response (or, dependent) variables are characteristics of the system which depend on the predictor variables. In other words, they respond to a change of values of the predictors in some systematic fashion.

Assume, we have a dataset containing n independent and identically distributed (i.i.d.) observations of responses $y_1, \ldots, y_n \in \mathbb{R}$, along with corresponding vector-valued predictors $x_1, \ldots, x_n \in \mathbb{R}^p$. We consider each x_i to be a column vector. We also use another type of variables in a regression setting. These are p unknown parameters $\beta := (\beta_1, \cdots, \beta_p)^T$. One of the objectives of statistical modelling is to identify a functional relationship ('model') between the responses and the predictor variables:

$$E(y_i|x_i) = \phi(x_i, \beta) \tag{2.1}$$

where ϕ is a function that depends β.

2.1 Linear Regression

Linear regression is one of the most popular forms of statistical modelling. Here, the functional relationship between the response and predictor is linear i.e. $\phi(x_i, \beta) := x_i^T \beta$, and usually the assumption $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ is made for the random errors. The linear model can be written in a matrix form for all cases $i \in \{1, \ldots, n\}$ simultaneously as follows:

$$y = \boldsymbol{x}\beta + \epsilon \tag{2.2}$$

where

$$y := \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \qquad \boldsymbol{x} := \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} \qquad \beta := \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \qquad \epsilon := \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}. \tag{2.3}$$

The matrix $\boldsymbol{x}$ is called the *design matrix.* Remember that each $x_i \in \mathbb{R}^p$ is considered as a column-vector, so $\boldsymbol{x}$ is an $n \times p$ matrix.

This can be extended to non-parametric approaches which do not assume an explicit parametric shape, but most of such approaches achieve this by simply introducing a large number of basis vectors, so that they still can be expressed as in Eq. (2.1).

Example 2.1 (Gaia Dataset). *Gaia*[1] *is a mission by the European Space Agency (ESA) to formulate a three dimensional map of our galaxy [32]. The data depicted in Fig. 2.1 are part of a dataset that was generated prior to the launch of the mission by computer simulations [31, 4]. The data contain essentially spectral information divided into* $p = 16$ *wavelength bands (intervals), along with certain stellar parameters, which are to be inferred from the spectral data. That is, each observation in*

[1] This dataset is openly available and has been loaded from the `R` package `LPCM` [30] for illustration.

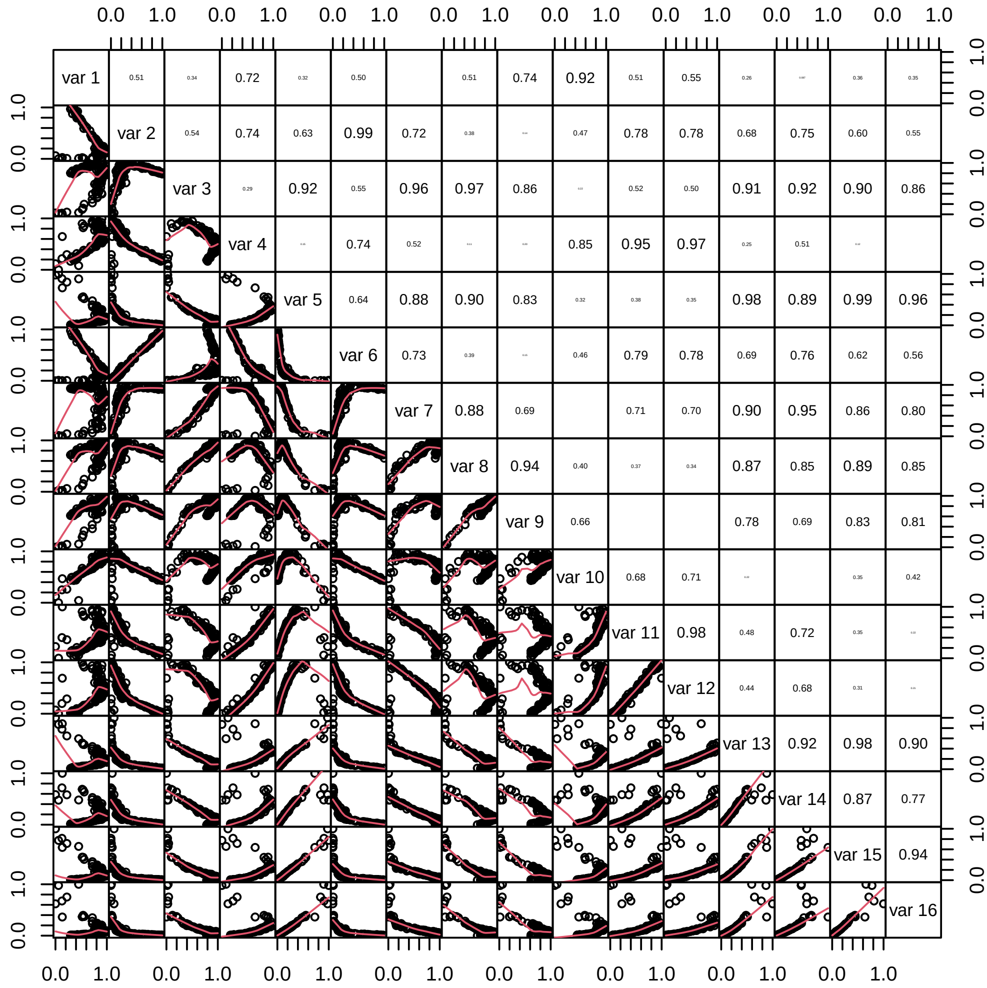

Figure 2.1: Correlation between the predictors in Gaia dataset.

the data set represents a stellar object, and the measurement for each 'band' is the energy flux (photon counts) emitted from that object within that wavelength interval. In the dataset that we have available, a total of $n = 8286$ *observations (stellar objects) are recorded. In our example, we consider steller temperature as the response variable. We will discuss this in Section 9.2.*

We scale the predictors such that the range is 0 to 1 and show the scatterplot matrix of the predicors in Fig. 2.1. We observe strong correlations between the predictors suggesting that they carry redundant information.

An important aspect of a statistical model is the presence of randomness within the model. In Fig. 2.1, we observe the presence of random noise along with the trend. Therefore, besides model fitting, our goal is also to quantify the randomness present within the model. For that, we rely on statistical inference techniques.

2.2 Statistical Inference

Statistical inference is the process by which we use the available data to gain knowledge about the model parameters, as well as their uncertainties. In a wider sense it will also include methods by which we quantify and validate our assumptions on the model. Statistical inference deals with the estimation of parameters that are used to specify the family of probability distributions which underlie the statistical model for $y_i|x_i$. There are several methods available to do statistical inference. However, we will discuss statistical inference using two approaches: the likelihood-based approach and the Bayesian approach.

The likelihood-based approach (Casella and Berger [15], Cox [22]) is a widely used estimation method. The estimation can be a point estimate, where we simply try to find the best numerical value for the parameter of the model. Alternatively, we may seek an interval which covers the unknown parameter value with high probability (generally 0.95). We call this a 95% confidence interval.

While several point estimators are available, the maximum likelihood estimator (or, MLE) is the most popular method because of its simple and wide implementability and its consistency properties. Maximum likelihood estimator finds the parameter value which maximises the probability density of the sample given the parameter, i.e. the likelihood. For linear regression models with Gaussian errors, MLE is equivalent to ordinary least squares.

The Bayesian approach (Berger [9], Gelman et al. [39]) starts from Bayes's rule for conditional probability. Let y denote the data. For example, in our setting, y is simply the vector of observed response values $(y_1, \ldots, y_n)^T$. The statistical model is specified through a likelihood function $P(y \mid \beta)$. In the context of the regression model, this likelihood would be considered conditional on the observed values of the predictors, that is, the observed values of the predictors are considered as fixed. Finally, we need a prior distribution $P(\beta)$ for the model parameters β

to incorporate our prior knowledge. Bayes's rule then tells us that the posterior distribution $P(\beta \mid y)$ is given by

$$P(\beta \mid y) \propto P(\beta)P(y \mid \beta). \tag{2.4}$$

The normalisation constant can be calculated from the law of total probability if necessary. However, this calculation may not be trivial and simulation methods like MCMC need to be employed. The posterior distribution is then used for further inference. For instance, we can look at its mean, mode, or other characteristics. In many cases the posterior mode corresponds to the maximum likelihood estimate.

Chapter 3

Likelihood-based estimation

In Chapter 2, we have introduced the notion of statistical modelling and statistical inference. This chapter focuses on the likelihood-based approaches in the context of linear models. We define maximum likelihood estimation (or, MLE) in Section 3.1 and use the notion of MLE to show its relation with ordinary least squares in Section 3.2. After Section 3.2, we focus on the regularisation methods for high-dimensional models. We discuss Ridge regression and it's properties in Section 3.3. Section 3.4 is focused on variable selection techniques, where the non-negative garrote and regularisation under ℓ_q penalties have been investigated. These type of methods are closely related to 'Least Absolute Shrinkage and Selection Operator' or LASSO, which we discuss in Section 3.5 along with the LASSO for logistic regression in Section 3.6. Regularisation methods often require model selection, which is discussed in Section 3.7. Finally, we discuss inference for these regularisation techniques in Section 3.8.

3.1 Likelihood Function

In Chapter 2, we discussed the notion of random noise within the observations and the underlying distributional assumption. We use this distributional assumption and treat these observations as random variables. This treatment allows us to form a joint probability density function with respect to the unknown parameters. We call this joint probability density function as the likelihood function. Therefore, for a sequence of observations $y := (y_1, \cdots, y_n)^T$ and parameters $\beta := (\beta_1, \cdots, \beta_p)^T$, we define the likelihood function in the following way:

$$L(\beta; y) = \prod_{i=1}^{n} P(y_i \mid \beta) \tag{3.1}$$

where P is a probability density function that comes from the distributional assumption.

Definition 3.1 (Law of Likelihood)**.** *Let $\beta' := (\beta'_1, \ldots, \beta'_p)^T$ be a vector of parameters. Then the observations $y := (y_1, \cdots, y_n)^T$ support β over β' if $L(\beta; y) > L(\beta'; y)$. Alternatively if,*

$$r := \frac{\prod_{i=1}^{n} P(y_i \mid \beta)}{\prod_{i=1}^{n} P(y_i \mid \beta')} > 1. \tag{3.2}$$

The evidence is indifferent to the parameters β and β' if the ratio is equal to 1.

Note that, law of likelihood allows us to interpret likelihoods but it's not sufficients as there can be other set of observations which are more informative. A detailed discussion on different aspects of likelihood can be in *Likelihood* by Edwards [28]. The book also provides a formal discussion on the notion of 'support', which we do not intend to cover in this thesis. For the sake of interpretation, we use the term 'support' as a measure of evidence produced by the data as suggested in the *Cambridge Dictionary of Statistics* [33].

3.1.1 Maximum Likelihood Estimation

If the parameters β are unknown, we can exploit law of likelihood to estimate these parameters. We maximise the likelihood function to get estimates of the unknown parameters β. These maximum likelihood estimates are given by:

$$\hat{\beta} := \arg\max_{\beta} \prod_{i=1}^{n} P(y_i \mid \beta) \tag{3.3}$$

In some cases, we may take logarithm on the likelihood function for the sake of calculation as it is a monotone operator.

For linear models, we assume that the random noises follow normal distribution. This gives us the following likelihood function.

$$L(\beta; y, \boldsymbol{x}) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} \prod_{i=1}^{n} \exp\left(-\frac{(y_i - x_i^T\beta)^2}{2\sigma^2}\right) \tag{3.4}$$

$$= \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\left(-\frac{\|y - \boldsymbol{x}\beta\|_2^2}{2\sigma^2}\right). \tag{3.5}$$

Taking logarithm on both sides gives us,

$$\ln L(\beta; y, \boldsymbol{x}) = -n \ln \sqrt{2\pi\sigma^2} - \frac{\|y - \boldsymbol{x}\beta\|_2^2}{2\sigma^2}. \tag{3.6}$$

Since the first term in the right hand side of Eq. (3.6) is independent of β, therefore maximising $\ln L(\beta; y, \boldsymbol{x})$ is equivalent to minimising the sum of the squared error given by:

$$R(\beta) := \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - x_i^T\beta)^2 = \|y - \boldsymbol{x}\beta\|_2^2. \tag{3.7}$$

We use $\|\cdot\|_2$ to denote the standard Euclidean norm, that is $\|z\|_2 := \sqrt{\sum_{i=1}^n z_i^2}$.

3.2 Ordinary Least Squares

Minimising the sum of the squared errors in Eq. (3.7) gives us ordinary least squares estimates [26]. We can express this in the following way:

$$\hat{\beta}_{\text{OLS}} := \arg\min_{\beta} R(\beta). \tag{3.8}$$

A necessary condition to have a minimum for Eq. (3.7) is

$$\frac{\partial}{\partial\beta} R(\beta) = -2\boldsymbol{x}^T y + 2(\boldsymbol{x}^T\boldsymbol{x})\beta = 0. \tag{3.9}$$

To ensure that the solution to Eq. (3.9) gives us the minimum, we need to investigate the second derivative, which is given by:

$$\frac{\partial^2}{\partial\beta^2} R(\beta) = 2(\boldsymbol{x}^T\boldsymbol{x}) \tag{3.10}$$

Now, the solution to Eq. (3.9) exists when $\boldsymbol{x}^T\boldsymbol{x}$ is invertible. This also assures that $\boldsymbol{x}^T\boldsymbol{x}$ is positive definite (see Lemma A.1) and satisfies the sufficient condition for minimum, that is the second derivative is positive definite. Then the ordinary least squares estimates are given by

$$\hat{\beta}_{\text{OLS}} = (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T y, \tag{3.11}$$

where $(\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T$ is the Moore-Penrose inverse of $\boldsymbol{x}$. The *Gauss-Markov theorem* states that when the errors are uncorrelated with expectation zero and constant variance, then the ordinary least squares estimator is the best linear unbiased estimator as it has lowest sampling variance.

Ordinary least squares estimates also give us closed-form expressions for the variances of the regression coefficients, which allow us to perform inference. However, two issues that often arise are:

1. If $p > n$ then $\boldsymbol{x}^T\boldsymbol{x}$ is singular, hence Eq. (3.9) has no unique solution.
2. Even if $p \leq n$, p may still be much larger than needed, and we may wish to identify sparse solutions where unnecessary parameters are set to zero. In other words, we may wish to perform variable selection as part of our statistical inference.

Illustration

We illustrate ordinary least squares estimation using a synthetic dataset. This allows us to investigate the efficiency of the method and compare ordinary least squares estimates with true regression coefficients.

Example 3.2. *We construct a synthetic dataset to illustrate least squares method with* 10 *predictors and 100 observations. We generate this dataset from a standard normal distribution, so that is* $x_{i,j} \sim \mathcal{N}(0,1)$ *for* $i = 1,\ 2,\ldots,\ 100$ *and* $j = 1,\ 2,\ \ldots,\ 10$*. We generate the response vector* y *such that,* $y_i = x_i^T\beta + \epsilon_i$*, where* $\beta = (-18, -79, -23, 59, 54, -1, 64, 41, 98, -20)^T$ *and* $\epsilon_i \sim \mathcal{N}(0, 0.01)$*.*

We fit an ordinary least squares model using the function `lm` from the `R` package called `stats` [63]. We provide the summary in Table 3.1. The first row in the table represents the intercept term in the linear model and rest represent the regression coefficients. We present the least square estimates in the left most column followed by the standard error of these regression coefficients. In the third column we present the t- value or simply the ratio of the least square estimates and their standard errors. In the right most column we provide the p-value or the probability of the observed data when the null hypothesis is true. A detailed discussion on these terms can be found in the books authored by (Casella and Berger [15], Cox [22]).

We observe that the least squares estimates for the regression coefficients are in good agreement with the true value that we used to generate the synthetic dataset in Example 3.2. We notice that we have a non-zero intercept term. However, the p-value is significantly high for the intercept term. Therefore we may consider the null hypothesis to be true and take intercept term as zero.

	Estimate	Std. Error	t-value	p-value
Int	-0.01	0.01	-1.2e+00	2.33e-01
β_1	-18	0.01	-1.8e+03	<2e-16
β_2	-79	0.01	-8.4e+03	<2e-16
β_3	-23	0.01	-2.3e+03	<2e-16
β_4	59	0.01	6.0e+03	<2e-16
β_5	54	0.01	5.8e+03	<2e-16
β_6	-1	0.01	-1.1e+02	<2e-16
β_7	64	0.01	6.9e+03	<2e-16
β_8	41	0.01	4.7e+03	<2e-16
β_9	98	0.01	9.4e+03	<2e-16
β_{10}	-20	0.01	-1.8e+03	<2e-16

Table 3.1: Summary of ordinary least squares estimates using Example 3.2.

3.3 Ridge Regression

In Section 3.2, we discuss ordinary least squares estimates and its properties. However, ordinary least squares are not applicable for correlated datasets or high-dimensional problems. This issue can be resolved by adding a suitable regularisation term in the negative log-likelihood of the linear model. Tikhonov [70] introduced ridge regression by adding an ℓ_2 penalty term to the squared error. Therefore, we formulate ridge estimates in the following way:

$$\hat{\beta}_{\mathrm{R}}(\lambda) \coloneqq \arg\min_{\beta} \left(R(\beta) + \lambda \|\beta\|_2^2 \right). \tag{3.12}$$

A necessary condition to have a minimum for Eq. (3.12) is

$$\frac{\partial}{\partial \beta} \left(R(\beta) + \lambda \|\beta\|_2^2 \right) = -2\boldsymbol{x}^T y + 2(\boldsymbol{x}^T \boldsymbol{x})\beta + 2\lambda \mathbf{I}_p \beta = 0. \tag{3.13}$$

Now, second derivative is given by:

$$\frac{\partial^2}{\partial \beta^2} \left(R(\beta) + \lambda \|\beta\|_2^2 \right) = 2 \left(\boldsymbol{x}^T \boldsymbol{x} + \lambda \mathbf{I}_p \right). \tag{3.14}$$

The introduction of regularisation term ensures that $\left(\boldsymbol{x}^T \boldsymbol{x} + \lambda \mathbf{I}_p \right)$ is invertible and positive definite. Therefore from Eq. (3.13), we have the Ridge estimates as

$$\hat{\beta}_{\mathrm{R}}(\lambda) = \left(\boldsymbol{x}^T \boldsymbol{x} + \lambda \mathbf{I}_p \right)^{-1} \boldsymbol{x}^T y. \tag{3.15}$$

Unlike, ordinary least squares, these estimates are dependent on an additional parameter, λ. Therefore, we need to select an optimal value of λ through a suitable model selection technique, which we discuss in Section 3.7. Ridge regression also gives us closed-form expressions for variences of the regression coefficients and asymptotically unbiased (check Lehmann and Casella [51, p. 438]) under suitable regularity conditions (see Lemma A.3).

Illustration

We illustrate ridge regression using a similar dataset to the one we see in Example 3.2. However, for ridge regression, we are interested in the case when $\boldsymbol{x}^T \boldsymbol{x}$ is not invertible. To achieve that, we introduce collinearity in the design matrix.

Example 3.3. *We generate this synthetic dataset by simulating first 9 predictors from standard normal distribution so that $x_{i,j} \sim \mathcal{N}(0,1)$, where $i = 1, \cdots, 100$ and $j = 1, \cdots, 9$. We construct another predictor $x_{i,10} = \sum_{j=1}^{9} x_{i,j}$. This ensures that the matrix $\boldsymbol{x}^T \boldsymbol{x}$ is singular. We obtain the response vector in the similar fashion to*

Example 3.2. Therefore, $y_i = x_i^T\beta + \epsilon_i$, where $\beta = (-18, -79, -23, 59, 54, -1, 64,$ 41, 98, $-20)^T$ and $\epsilon_i \sim \mathcal{N}(0, 0.01)$.

Since we introduced collinearity by constructing the 10-th predictor as $x_{i,10} = \sum_{j=1}^{9} x_{i,j}$, we can show that for the first 9 predictors, true regression coefficients are $\beta = (-38, -99, -43, 39, 34, -21, 44, 21, 78)^T$ and the 10-th regression coefficient is equivalent to zero.

	Estimate	Std. Error	t-value	p-value
Int	-3			
β_1	-35.4	1.5	2.41e+01	<2e-16
β_2	-90.1	1.3	6.91e+01	<2e-16
β_3	-38.0	1.4	2.69e+01	<2e-16
β_4	34.2	1.4	2.45e+01	<2e-16
β_5	26.8	1.3	2.08e+01	<2e-16
β_6	-17.0	1.3	1.34e+01	<2e-16
β_7	40.2	1.3	3.15e+01	<2e-16
β_8	18.1	1.2	1.48e+01	<2e-16
β_9	69.9	1.5	4.67e+01	<2e-16
β_{10}	0.7	0.3	2.75e+00	6e-03

Table 3.2: Summary of the ridge estimates obtained from Example 3.3.

We fit the ridge regression model using the package `ridge` [24], which performs an automatic tuning of λ using the method proposed by Cule and Iorio [23]. The optimal λ by this method is given by 0.09. We provide the summary of ridge regression estimates in Table 3.2. We follow the same convention as of Table 3.1. We provide the estimates in left most columns followed by the standard errors, t-value and p-value. In ridge regression, we usually consider the intercept term to be constant. Therefore, the first row in Table 3.2 remains empty except for the estimate of this intercept.

We can also perform variable selection based on the p-values on the right most column. We assume that the regression coefficients are equal to zero under the null hypothesis [25]. In this example we see that the p-value of the 10th co-variate is significantly higher than the p-values of other regression coefficients. Therefore, we may argue that this estimate can be considered as zero.

3.4 Sparse Regression

In our previous example, we discussed how we can perform variable selection using p-value. However, to decide on the threshold for p-value is subjective and therefore we may seek for a method which performs automatic variable selection. Several variable selection routines are available to obtain sparse estimates or simply zero as estimated value for some of the regression coefficients. In this section, we discuss these variable selection methods using likelihood-based approaches.

3.4.1 Non-Negative Garrote

The non-negative garrote was introduced by Breiman [14]. It is a two stage procedure that gives a sparse solution. It has a close relationship to the LASSO [44, 69]. However, as a starting point of the problem ordinary least squares estimates are required. Given the initial estimate $\hat{\beta}_{\text{OLS}} \in \mathbb{R}^p$, we solve the following optimisation problem over $m = (m_1, m_2, \cdots, m_p)^T$:

$$\hat{m} = \arg \min_{\substack{m \geq 0 \\ \|m\|_1 \leq t}} \|y - \boldsymbol{x}\boldsymbol{m}\hat{\beta}_{\text{OLS}}\|_2^2 \tag{3.16}$$

where $\boldsymbol{m} := \text{diag}(m) \in \mathbb{R}^{p \times p}$, and $\| \cdot \|_1$ denotes the ℓ_1-norm; that is $\|m\|_1 := \sum_{i=1}^{p} |m_i|$. We get the final non-negative garrote parameter estimate $\hat{\beta}_{\text{NG}}$ by setting $\hat{\beta}_{\text{NG},\, j} = \hat{m}_j \hat{\beta}_{\text{OLS},\, j}$ for each $j \in \{1, 2, \ldots, p\}$.

Solution for the non-negative garrote

The non-negative garrote can be formulated as a constrained optimisation problem. Therefore, we can get the non-negative garrote estimates by using the notion of duality, which we will explain in Section 4.1. We introduce a Lagrangian multiplier λ for the constraint $\|\boldsymbol{m}\|_1 - t \leq 0$ [44]. This gives us the following objective function:

$$\max_{\lambda \geq 0} \min_{\boldsymbol{m} \geq 0} \Big(\|y - \boldsymbol{x}\boldsymbol{m}\hat{\beta}_{\text{OLS}}\|_2^2 + \lambda(\|\boldsymbol{m}\|_1 - t) \Big) \tag{3.17}$$

Effectively, we thus need to solve

$$\hat{\boldsymbol{m}}_\lambda = \arg\min_{\boldsymbol{m}\geq 0} \Big(\|y - \boldsymbol{x}\boldsymbol{m}\hat{\beta}_{\mathrm{OLS}}\|_2^2 + \lambda\|\boldsymbol{m}\|_1 \Big) \tag{3.18}$$

where the Lagrange multiplier $\lambda \geq 0$ can be interpreted as a regularisation parameter. If $\|\hat{\boldsymbol{m}}_\lambda\|_1 \leq t$ for $\lambda = 0$, then we are done. Otherwise, λ is calibrated until $\|\hat{\boldsymbol{m}}_\lambda\|_1 = t$, as we will show in Section 4.1. This value for λ is also the value that achieves the maximum in Eq. (3.17). When $\boldsymbol{x}^T\boldsymbol{x} = \mathbf{I}_p$, we have an explicit solution of Eq. (3.18) as given by Yuan and Lin [81]:

$$\hat{m}_{\lambda,j} = \max\left\{0, 1 - \frac{\lambda}{(\hat{\beta}_{\mathrm{OLS},\, j})^2}\right\}. \tag{3.19}$$

Consequently, in this case, if the coefficient $\hat{\beta}_{\mathrm{OLS},\, j}$ of a predictor is less than $\sqrt{\lambda}$, then $\hat{m}_{\lambda,j} = 0$, and therefore also $\hat{\beta}_{\mathrm{NG},\, j} = \hat{m}_{\lambda,j}\hat{\beta}_{\mathrm{OLS},\, j} = 0$. In this way, larger λ will produce sparser solutions.

The starting point of this method depends on the least squares estimates $\hat{\beta}_{\mathrm{OLS}}$. Therefore, if $p > n$, then no unique solution is available. However, alternative initial estimators such as the LASSO can be used in this case [81].

Illustration

We illustrate the non-negative garrote using Example 3.2. For the computation of regression coefficient estimates `nngarrote` [18] has been used. The coefficient estimates are provided in Table 3.3. To obtain these estimates, we perform model selection over different values of λ. We us cross-validation method to find this optimal λ which is equal to 0.009. For this value of λ, we see that β_6 is considered as non-important.

Int	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
-0.9	-16.4	-77.0	-20.6	56.8	51.5	0	62.5	39.3	96.4	-18.1

Table 3.3: Non-negative garrote estimates obtained from Example 3.2.

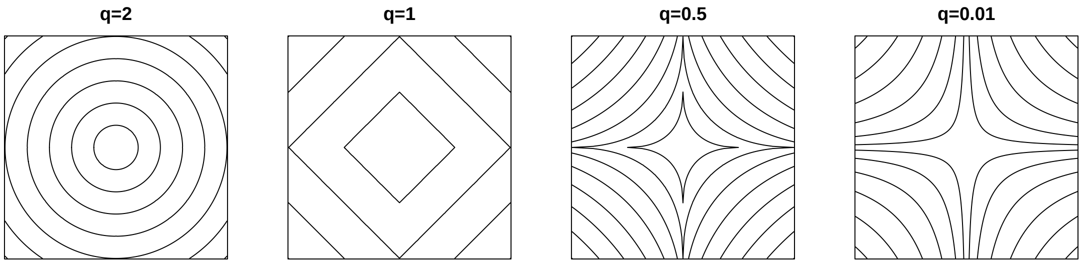

Figure 3.1: Contour plots of different ℓ_q penalty functions.

3.4.2 Regularisation under ℓ_q penalty

Unfortunately, the non-negative garrote in Eq. (3.16) still fails to deliver when we have no least squares estimate to start from, which happens for instance when we have more predictors than observations. To solve this, we can use a different method, where no initial estimate is needed. The basic idea is to add a penalty term to the least squares problem, in order to penalise non-zero parameter values. This can be done in the following way:

$$\hat{\beta}_\lambda = \arg\min_\beta \left(\frac{1}{2} \|y - \boldsymbol{x}\beta\|_2^2 + \lambda \|\beta\|_q^q \right) \tag{3.20}$$

where $q \geq 0$ determines the shape of the penalty, and $\lambda \geq 0$ determines the strength of the penalty. Here,

$$\|z\|_q^q := \begin{cases} \sum_{i=1}^n |z_i|^q & \text{if } q > 0 \\ \sum_{i=1}^n \mathbb{I}_{z_i \neq 0} & \text{if } q = 0 \end{cases} \tag{3.21}$$

where $\mathbb{I}_{z_i \neq 0} = 1$ if $z_i \neq 0$, and 0 otherwise. So, $\|z\|_0^0$ simply counts the number of non-zero components of z.

For different values of q we have different types of regularisation. This leads to ridge regression for $q = 2$, LASSO for $q = 1$, and subset selection method for $q = 0$ [44].

In Fig. 3.1, we illustrate some contour plots of the ℓ_q penalty function, for different values of q. As will be illustrated in Section 3.5, it is the 'spiked' shape of the contours on the co-ordinate axis that leads to sparse estimates; in other words all penalties with $q \leq 1$ will lead to sparse estimators. However, for $q < 1$, the ℓ_q

penalty function is no longer convex, as can be seen from the contour plots. Therefore, $q = 1$ is the only value for which the problem is convex and allows sparse solutions. Further discussion on the non-convex penalties of the form can be found in the book *Statistical Learning with Sparsity* by Hastie et al. [46, p. 84].

3.5 LASSO

The LASSO estimator was first proposed by Tibshirani [69]. The objective is to solve the ordinary least squares problem, but subject to an additional constraint on the ℓ_1 norm of the parameters, as follows:

$$\min_{\beta :\, \|\beta\|_1 \le t} \left(\frac{1}{2} \|y - \boldsymbol{x}\beta\|_2^2 \right). \tag{3.22}$$

It is usually assumed that $\boldsymbol{x}$ and y are standardised to mean 0. Otherwise, they can always be standardised without any loss of generality.

Solution for the LASSO

By strong duality (see Theorem 4.1 in Section 4.1), equivalently, we can solve the dual problem, by introducing a Lagrangian multiplier λ for the constraint $\|\beta\|_1 - t \le 0$:

$$\max_{\lambda \ge 0} \min_{\beta} \left(\frac{1}{2} \|y - \boldsymbol{x}\beta\|_2^2 + \lambda(\|\beta\|_1 - t) \right). \tag{3.23}$$

For the inner minimisation problem, we need to find

$$\hat{\beta}_{\mathrm{L}}(\lambda) := \arg\min_{\beta} \left(\frac{1}{2} \|y - \boldsymbol{x}\beta\|_2^2 + \lambda \|\beta\|_1 \right). \tag{3.24}$$

Eq. (3.24) is solved using numerical optimisation methods. However, when the columns of $\boldsymbol{x}$ are standardised such that $\boldsymbol{x}^T \boldsymbol{x} = \mathbf{I}_p$, the solution to this system can be expressed as a thresholded version of the ordinary least squares [44]:

$$\hat{\beta}_{\mathrm{L},\, j}(\lambda) = S_\lambda(\hat{\beta}_{\mathrm{OLS},\, j}) \tag{3.25}$$

with *soft-thresholding operator* (see Fig. 3.2)

$$S_\lambda(\beta_j) := \operatorname{sign}(\beta_j) \max\{0, |\beta_j| - \lambda\} \tag{3.26}$$

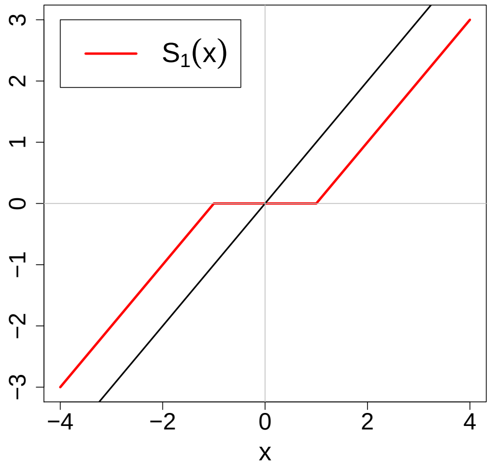

Figure 3.2: Soft-thresholding function $S_\lambda(x)$ for $\lambda = 1$.

where

$$\operatorname{sign}(\beta_j) := \begin{cases} -1 & \text{if } \beta_j < 0 \\ 0 & \text{if } \beta_j = 0 \\ 1 & \text{if } \beta_j > 0. \end{cases} \tag{3.27}$$

The contour lines in Fig. 3.3 illustrate the way LASSO works. The contours refer to the ordinary least squares problem, and the diamond corresponds to the constraint $\|\beta\|_1 = t$. We search for the point on the diamond closest to the ordinary least squares. This is likely to lie on the axes, hence setting smaller parameters to 0.

Illustration

We illustrate LASSO using the dataset in Example 3.3, where we have 10 predictors and 100 observations. In this case, singularity comes from the collinearity introduced in the predictors. We use the package `glmnet` [36] to perform cross validation for model selection which gives us the optimal $\lambda = 1.682$. We provide these LASSO estimates in Table 3.4. We observe that the LASSO considers β_{10} to be non-important for the optimal value of λ.

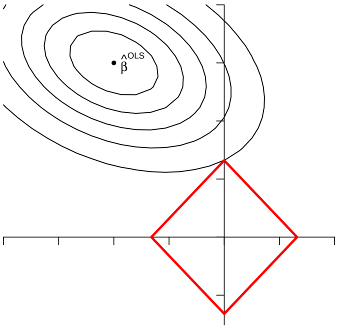

Figure 3.3: Relationship between the OLS estimate and the ℓ_1 constraint imposed by the LASSO (red); adapted from [46].

Int	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
-0.8	-36.2	-97.1	-40.1	37.3	32.6	-20.6	42.0	19.1	76.8	0

Table 3.4: The LASSO estimates obtained from Example 3.3.

3.6 LASSO for Classification

Classification is a method for assigning a new object to a class or a group based on the observed features or attributes of the object. Classification is used in many applications such as pattern recognition for hand writing, disease treatment, facial recognition, chemical analysis, and so on. In general, a classifier can be seen as a function that maps a set of continuous or discrete variables into a categorical class variable. Constructing a classifier from random samples is an important problem in statistical inference. In our work, we will restrict ourselves to the case where there are only two classes to choose from, i.e. ‘binary classification’.

Let c be a random variable that takes values in $\{0, 1\}$. Let x be a p-dimensional vector that denotes the attributes of an object and let $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$ denote the vector of regression coefficients. In a regression setting, we construct a classifier through a generalised linear model (GLM) as follows:

$$E(c \mid x) = h\left(x^T \beta\right) \tag{3.28}$$

where h acts as a 'link' function and E stands for expectation. We define

$$\pi(x) := E(c \mid x) = P(c = 1 \mid x). \tag{3.29}$$

3.6.1 Logistic Regression

Logistic regression is a well-used special case of the GLM, which is suitable for classification with continuous attributes. Now, consider the generalised model in Eq. (3.28). For logistic regression, we use the following link function:

$$h(a) := \frac{\exp(a)}{1 + \exp(a)}. \tag{3.30}$$

We define a vector $c := (c_1,\ c_2,\ \ldots,\ c_n)^T$ denoting n observed classes such that, $c_i \in \{0, 1\}$. The c_i's are thus Bernoulli random variables. Let $\boldsymbol{x} := [x_1, x_2, \ldots, x_n]^T$, with $x_i \in \mathbb{R}^p$, denote the observed attributes for n objects, so that $\boldsymbol{x}$ corresponds to the design matrix in the terminology of classical statistical modelling. It is easy to see that the log likelihood of the data is:

$$\log(L(c, \boldsymbol{x}; \beta)) = \sum_{i=1}^{n} \Bigg(c_i \left(x_i^T \beta \right) - \log\left(1 + \exp(x_i^T \beta)\right) \Bigg). \tag{3.31}$$

Therefore, the maximum likelihood estimate of the unknown parameter β is equivalent to:

$$\hat{\beta}_{\text{lr}} := \arg\min_{\beta} \{ -\log(L(c, \boldsymbol{x}; \beta)) \}. \tag{3.32}$$

3.6.2 Penalised Logistic Regression (PLR)

In the high-dimensional case, that is when the number of attributes is more than the number of observations ($p > n$), the performance of logistic regression is often not satisfactory. Apart from over-fitting, numerical optimisation methods often converge to local solutions because of multi-collinearity. Several techniques have been proposed to deal with this. Generally, a penalty term is introduced in the negative log-likelihood, leading to penalised logistic regression. A LASSO-type penalty [69] is very popular because of its variable selection property [67, 85]. The penalised logistic regression (PLR) as a regularisation method is defined by:

$$\hat{\beta}_{\text{plr}}(\lambda) := \arg\min_{\beta} \left\{ -\log(L(c, \boldsymbol{x}; \beta)) + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\}, \tag{3.33}$$

We get sparse estimate for β when $0 \leq q \leq 1$.

Once we have the estimate, we can then define, for any new object with known attributes $x_* \in \mathbb{R}^p$ and unknown class c_*,

$$\hat{\pi}(x_*, \lambda) := P\left(c_* = 1 \mid x_*; \hat{\beta}_{\text{aplr}}(\lambda)\right) = h\left(x_*^T \hat{\beta}_{\text{aplr}}(\lambda)\right). \tag{3.34}$$

We can then for instance classify the object as 0 if $\hat{\pi}(x_*, \lambda) < 1/2$, as 1 if $\hat{\pi}(x_*, \lambda) > 1/2$, and as either if $\hat{\pi}(x_*, \lambda) = 1/2$. The value of λ is chosen through cross-validation (explained in Section 3.7.1), where λ acts as a tuning parameter.

3.7 Model Selection

In the previous sections, we show how we require model selection to determine the optimal value for the regularisation parameter λ. For ridge regression, we can find this optimal value using an automatic method as described in [23]. For the non-negative garrote or the LASSO, we rely on a different model selection method called cross-validation. This is also used for model validation in some cases.

3.7.1 Cross-Validation

Cross-validation is a commonly used method to identify the optimal value of a tuning parameter, which is in our case the penalty parameter λ. It is based on minimising an estimate of the prediction error. In cross-validation, we use one part of the data to fit the LASSO model, and the other part of the data to validate it [46].

We fix initially a dense grid of values of λ, that is λ is discretised with small step-sizes over a suitable range which reflects the scope of the regularisation trade-off that we are willing to consider. The dataset is then divided into K equally sized partitions. We assume for simplicity that K is a divisor of n so that each partition contains n/K elements. For each fixed value of λ of the grid, and the k'th partition, $k = 1, \ldots, K$, we fit the regression model using the remaining $K - 1$ parts and calculate the prediction error of the fitted model. Specifically, denote $\hat{\beta}_\lambda^{-k}$ the parameter vector obtained under a penalty of λ when omitting the k'th partition, so that $x_i^T \hat{\beta}_\lambda^{-k}$ is the corresponding fitted model under predictor x_i. Then the averaged loss for the k'th partition is

$$R_k(\lambda) = \frac{K}{n} \sum_{i=1}^{\frac{n}{K}} d(y_i, x_i^T \hat{\beta}_\lambda^{-k}) \tag{3.35}$$

where for the linear model (Eq. (2.2)), the loss function 'd' is just the squared error. We repeat this step for every $k = 1, 2, \cdots, K$ and combine the values of $R_k(\lambda)$ to find the average loss, $R(\lambda) = K^{-1} \sum_{k=1}^{K} R_k(\lambda)$. This is then repeated for every value of λ in the grid, and we choose the value of λ which minimises $R(\lambda)$ [44].

Typically for the LASSO, smaller values of λ result to more predictors in the model, which may lead to an over-fitted model. However, for larger values of λ, the model has fewer predictors leading to sparsity and producing a more easily interpretable model.

To avoid misunderstandings, it is noted that the problem of finding the optimal λ (in the sense of minimal prediction error), as discussed in this subsection, is very different from, and entirely unrelated to, the problem of maximising over λ as, for instance, in Eq. (3.23). The latter is a purely formal operation which ensures mathematical equivalence of the two dual versions of the LASSO optimisation problem, and does not imply any statement on the best choice of λ.

3.8 Inference for Regularisation Techniques

Regularisation techniques such as the non-negative garrote or the LASSO don't have any closed-form expression for variance. Therefore we need other methods to perform inference on the parameter estimation. In the context of linear regression, we perform inference in two major ways, using the refit-based method and using the bootstrap method.

3.8.1 Refit-based Methods

Refit-based methods are usually used for sparse regression. Once we attain sparsity within a model we select the non-zero co-variates to form a new predictor matrix. We then apply ordinary least squares on these co-variates which allows us to obtain different statistical quantities of these parameter estimates such as standard error, p-value, *etc.*

Refit for non-negative garrote In Section 3.4.1, we illustrated the non-negative garrote using Example 3.2. We observed that 6-th predictor was non-important based on the optimal λ which we obtained through cross-validation. Now, we discard the 6-th predictor and construct a new design matrix using the other predictors from Example 3.2. By construction of this design matrix, this is non-singular and therefore, we can fit least squares to obtain standard errors of the estimates. We show our results in Table 3.5. From left to right, we provide estimated value, standard error, t-value and p-value, similar to our analysis with ordinary least squares. We notice that the refit estimates are in good agreement with the true regression coefficients. We also observe an additional non-zero intercept term in the model. However, high p- value suggests that we can consider this intercept term to be zero.

	Estimate	Std. Error	t-value	p-value
Int	-0.2	0.1	-1.6e+00	1.17e-01
β_1	-18	0.1	-1.6e+02	<2e-16
β_2	-79	0.1	-7.6e+02	<2e-16
β_3	-23	0.1	-2.1e+02	<2e-16
β_4	59	0.1	5.3e+02	<2e-16
β_5	54	0.1	5.2e+02	<2e-16
β_7	64	0.1	6.2e+02	<2e-16
β_8	41	0.1	4.2e+02	<2e-16
β_9	98	0.1	8.6e+02	<2e-16
β_{10}	-20	0.1	-1.6e+02	<2e-16

Table 3.5: Refit estimates after performing the non-negative garrote on Example 3.2.

Refit for the LASSO To illustrate the LASSO, we used a dataset with collinearity in it. Now, from the illustration in Section 3.5, we observe that the LASSO estimator considers the 10-th predictor as non-important therefore we discard this predictor to construct a new design matrix. Clearly, this modified dataset is non-singular and hence we can apply ordinary least squares. We show the result obtained from the refit model in Table 3.6. The columns from left to right represents the estimated value, standard error, t-value and p-value as discussed earlier in our illustration for the non-negative garrote. From the table, we see that refit estimates are in good agreement with the true regression coefficients which we provided in Example 3.3.

	Estimate	Std. Error	t-value	p-value
Int	-0.01	0.01	-1.2e+00	2.33e-01
β_1	-38	0.01	-3.6e+03	<2e-16
β_2	-99	0.01	-1.0e+04	<2e-16
β_3	-43	0.01	-4.2e+03	<2e-16
β_4	39	0.01	3.8e+03	<2e-16
β_5	34	0.01	3.5e+03	<2e-16
β_6	-21	0.01	-2.2e+03	<2e-16
β_7	44	0.01	4.6e+03	<2e-16
β_8	21	0.01	2.3e+03	<2e-16
β_9	78	0.01	7.3e+03	<2e-16

Table 3.6: Refit estimates after performing the LASSO on Example 3.3.

3.8.2 Bootstrap

Bootstrap is a general frequentist method to quantify statistical accuracy, where one randomly draws samples from a given training dataset with replacement, the sample size being equal to that of the original training dataset. This is done for B times. Then one fits the model to each of these B datasets and examines the empirical distributions of the estimated parameters. We illustrate this method using the package called `bootstrap` [59].

Bootstrap for non-negative garrote We perform bootstrapping using non-negative garrote estimator. In Table 3.7 we provide the summary. In the left most column, we provide the averaged estimate from the bootstrap samples; followed by 1st quartile, Median, 3rd quartile and standard deviation. We see that the non-negative garrote estimate for the 6-th predictor is equal to zero for every bootstrap sample. We also notice that both means and medians of the regression coefficients are in good agreement with true regression coefficients.

Bootstrap for LASSO Similar to the non-negative garrote, we perform bootstrapping for the LASSO and provide our results in Table 3.8. We notice that the 10-th predictor remains non-important in every bootstrap samples and therefore is in good agreement with our illustration of the LASSO using Example 3.3.

	Mean	1st Qu	Median	3rd Qu	Sd
Int	-0.8	-1.2	-0.8	-0.5	0.6
β_1	-16.3	-16.7	-16.3	-15.9	0.6
β_2	-77.0	-77.3	-76.9	-76.7	0.5
β_3	-20.6	-21.0	-20.6	-20.2	0.6
β_4	56.8	56.5	56.8	57.2	0.5
β_5	51.4	51.0	51.4	51.7	0.4
β_6	0	0	0	0	0
β_7	62.4	62.1	62.4	62.7	0.5
β_8	39.2	38.9	39.2	39.6	0.5
β_9	96.4	96.0	96.4	96.8	0.5
β_{10}	-18.1	-18.5	-18.1	-17.6	0.6

Table 3.7: Bootstrap summary for the non-negative garrote.

	Mean	1st Qu	Median	3rd Qu	Sd
Int	-0.9	-1.3	-1.0	-0.6	0.6
β_1	-36.4	-36.8	-36.5	-36.1	0.5
β_2	-97.1	-97.3	-97.1	-96.8	0.4
β_3	-40.7	-40.9	-40.7	-40.4	0.5
β_4	37.1	36.7	37.2	37.4	0.5
β_5	31.4	31.1	31.4	31.7	0.4
β_6	-18.9	-19.1	-18.9	-18.6	0.5
β_7	42.7	42.4	42.6	43.0	0.5
β_8	19.5	19.2	19.6	19.9	0.5
β_9	76.1	75.8	76.1	76.4	0.5
β_{10}	0	0	0	0	0

Table 3.8: Bootstrap summary for the LASSO.

Chapter 4

Optimisation Methods

In Chapter 3, we noticed how optimisation is an important part of likelihood based approaches. Methods such as ordinary least squares or ridge regression are easy to optimise due to their closed form expressions. However, methods like LASSO need efficient numerical optimisation techniques and can not be solved using classical methods such us gradient descent method. This motivates us to inspect the theory behind the optimisation algorithms for LASSO-type problems. This is also beneficial for an in-house software implementation for optimisation with piece-wise differentiable functions.

In this chapter, we first discuss the mathematical foundations of non-linear optimisation in Section 4.1. We present the notion of subgradient of a function followed by the necessary conditions for optimality in constrained optimisation problems. Later, in Section 4.2, we derive the necessary conditions mentioned in Section 4.1 for LASSO and discuss different numerical schemes to obtain optimal solution for LASSO-type problems.

4.1 Strong Duality Conditions

In this section, we briefly give the main duality result for non-linear optimisation that we mentioned in Chapter 3. An extensive detail on the following topic can be found in the book authored by Boyd and Vandenberghe [13]. Here, we will only present the fundamental concepts that are required to present optimisation for piece-wise differentiable functions.

Assume we aim to minimise a function $f(\beta)$, where $\beta \in B \subseteq \mathbb{R}^p$ subject to a constraint $h(\beta) \leq 0$. In the following sections, we will have either $B = \mathbb{R}^p$ or $B = \mathbb{R}^p_+$ (i.e. the set of non-negative vectors in $\mathbb{R}^p$), although in principle B can be an arbitrary convex set. So, we try to find

$$f^* := \min_{\substack{\beta \in B \\ h(\beta) \leq 0}} f(\beta). \tag{4.1}$$

One may think of the function $f(\cdot)$ as a least squares criterion or a negative (log-)likelihood. Define now the *Lagrangian*:

$$\ell(\beta, \lambda) := f(\beta) + \lambda h(\beta) \tag{4.2}$$

and the *Lagrange dual function*:

$$g(\lambda) := \min_{\beta \in B} \ell(\beta, \lambda). \tag{4.3}$$

Note that

$$\begin{aligned} \max_{\lambda \geq 0} g(\lambda) = \max_{\lambda \geq 0} \min_{\beta \in B} \ell(\beta, \lambda) &\leq \max_{\lambda \geq 0} \min_{\substack{\beta \in B \\ h(\beta) \leq 0}} \ell(\beta, \lambda) && (4.4) \\ &\leq \min_{\substack{\beta \in B \\ h(\beta) \leq 0}} f(\beta) = f^*. && (4.5) \end{aligned}$$

This inequality holds in general. Strong duality tells us that, under certain conditions, the inequality becomes an equality [13, §5.2.3].

Theorem 4.1 (Strong Duality). *If f and h are convex functions, and $h(\beta) < 0$ for at least one $\beta \in B$, then*

$$\max_{\lambda \geq 0} g(\lambda) = \min_{\substack{\beta \in B \\ h(\beta) \leq 0}} f(\beta) = f^* \tag{4.6}$$

So, under strong duality, to minimise $f(\beta)$ over β subject to $h(\beta) \leq 0$, we can also instead maximise the Lagrange dual function over $\lambda \geq 0$. In that case, the *Karush-Kuhn-Tucker conditions* provide necessary and sufficient conditions for optimality.

Definition 4.2 (Subgradient)**.** *For any function F on B, we say that $v \in \mathbb{R}^p$ is a* subgradient *of F at β whenever*

$$F(\beta') - F(\beta) \geq v^T(\beta' - \beta) \tag{4.7}$$

for all $\beta' \in B$. The set of all subgradients of F at β is denoted by $\partial F(\beta)$.

Theorem 4.3 (Karush-Kuhn-Tucker)**.** *If f and h are convex functions, and $h(\beta) < 0$ for at least one $\beta \in B$, then $f(\beta) = f^*$ if and only if*

$$\mathbf{0} \in \partial f(\beta) + \lambda \partial h(\beta) \tag{4.8}$$

$$\lambda h(\beta) = 0 \tag{4.9}$$

$$h(\beta) \leq 0 \tag{4.10}$$

$$\lambda \geq 0 \tag{4.11}$$

Eq. (4.8) is just a fancy way of writing that β is a global minimum of $f + \lambda h$, for a fixed value of λ. Equation (4.8) is called the *stationarity condition*. Equation (4.9) is called the *complementary slackness condition*, and implies that either $\lambda = 0$ or $h(\beta) = 0$. The inequality $h(\beta) \leq 0$ is called *primal feasibility*, and the inequality $\lambda \geq 0$ is called *dual feasibility*.

To solve the Karush-Kuhn-Tucker conditions, we split the problem into two cases as per Eq. (4.9), $\lambda = 0$ and $h(\beta) = 0$. We then solve Eq. (4.8) under each equality constraint. We throw away any solution that does not satisfy primal or dual feasibility, and then choose the solution that achieves the lowest value.

For the case $\lambda = 0$, we need to find the global unconstrained minimum of f. If the primal feasibility constraint $h(\beta) \leq 0$ is satisfied at the global minimum of f, then we have found a solution. Obviously, this solution must be the optimal solution of the original constrained problem as well.

If $h(\beta) > 0$ at the global minimum of f, then we need to find the minimum of f under the constraint that $h(\beta) = 0$. We could do so by finding a joint solution to the system of equations formed by Eq. (4.8) and $h(\beta) = 0$. Alternatively, we could gradually increase λ until the global unconstrained minimum $g(\lambda)$ of $f + \lambda h$ satisfies $h(\beta) = 0$. Indeed, due to the form of the objective function, increasing λ will favour β that have lower values for $h(\beta)$, so eventually, $h(\beta) = 0$. By strong duality, we also know that finding this λ is equivalent to maximising the Lagrange dual function $g(\lambda)$ over $\lambda \geq 0$.

4.2 Optimisation for LASSO

For LASSO, the Lagrangian is given by

$$\frac{1}{2}\|y - \mathbf{x}\beta\|_2^2 + \lambda(\|\beta\|_1 - t). \tag{4.12}$$

From the discussion in Section 4.1, we know that if $\|\hat{\beta}_0\|_1 \leq t$, then the solution is immediately given by $\hat{\beta}_0$ (note that $\hat{\beta}_0 = \hat{\beta}_{\text{OLS}}$). If $\|\hat{\beta}_0\|_1 > t$, then we need find that value for $\lambda \geq 0$ for which $\|\hat{\beta}_\lambda\|_1 = t$, and the solution is then given by the corresponding $\hat{\beta}_\lambda$. In either case, this λ is also the λ which achieves the maximum in Eq. (3.23), and which solves the Karush-Kuhn-Tucker conditions in Theorem 4.3.

As we can see, along with complementary slackness (either $\lambda = 0$ or $\|\beta\|_1 = t$) and feasibility ($\lambda \geq 0$ and $\|\beta\|_1 \leq t$), this condition fully characterises the optimality of our solution. The stationarity condition (Eq. (4.8) in Section 4.1) says that the subgradient with respect to β of this Lagrangian must contain the origin. Therefore, we derive the stationarity condition of the Karush-Kuhn-Tucker equations for LASSO in the following way:

$$\mathbf{0} \in -\mathbf{x}^T(y - \mathbf{x}\beta) + \lambda\partial\|\beta\|_1. \tag{4.13}$$

It can be shown that [58, §3.1.5]

$$\partial|\beta_j| \coloneqq \begin{cases} \{-1\} & \text{if } \beta_j < 0 \\ [-1, 1] & \text{if } \beta_j = 0 \\ \{1\} & \text{if } \beta_j > 0, \end{cases} \tag{4.14}$$

for $j = 1, 2, \cdots, p$. Therefore, we can write Eq. (4.13) in the following way

$$\mathbf{x}^T(y - \mathbf{x}\beta) = \lambda s \tag{4.15}$$

where $s = (s_1, s_2, \ldots, s_p)$ are auxiliary variables subject to the constraint $s_j \in \partial|\beta_j|$.

Note that, for LASSO, it is sufficient to minimise the following objective function:

$$J(\beta) = \frac{1}{2}\|y - \mathbf{x}\beta\|_2^2 + \lambda(\|\beta\|_1). \tag{4.16}$$

This formulation of the objective function also allows us to express it as a decomposible function.

Definition 4.4 (Decomposible)**.** *A convex function $J(\beta)$ is decomposible if it can be written as sum of two convex function*

$$J(\beta) = f(\beta) + h(\beta) \tag{4.17}$$

where $f(\beta)$ is differentiable but $h(\beta)$ is not.

Therefore, for LASSO, the term obtained from the likelihood is differentiable, however the penalty term is not.

4.2.1 Sub-gradient Method

Subgradient method is an alternative to gradient based optimisation methods for non-differentiable functions (Shor et al. [68]). If a function is convex but not necessarily differentiable then we can apply sub-gradient method for minimisation. Then for any subgradient $g(\beta)$ and sequence of stepsize $t^{(k)}$, the algorithm for subgradient method can be shown in the following way:

Algorithm

- Initial guess: β^0
- Increment step: $\beta^{(k+1)} = \beta^{(k)} - t^{(k+1)}g(\beta^{(k)})$
- Updating: $\beta_{\text{best}}^{(k+1)} = \arg\min\{J(\beta^{(1)}), \cdots, J(\beta^{(k+1)})\}$

In the updating step we have to make sure that we are keeping track of the best solutions as subgradient method is not necessarily descent.

Convergence The convergence rate of subgradient method is based on the Lipschitz continuity of the objective function $J(\beta)$. That is, let $L_1 > 0$ be a constant such that,

$$|J(\beta) - J(\gamma)| \leq L_1 \|\beta - \gamma\|_2 \tag{4.18}$$

then for a sequence of step sizes $t^{(k)}$ and optimal solution β^*, we can derive the following inequality:

$$J(\beta^{(k+1)}_{\text{best}}) - J(\beta^*) \leq \frac{d^2 + L_1^2 \sum_{i=1}^{k} (t^{(i)})^2}{2 \sum_{i=1}^{k} t^{(i)}} \tag{4.19}$$

where $d^2 = \|\beta^{(0)} - \beta^*\|_2^2$. Therefore the convergence rate is dependent on the the choice of the sequence of step sizes.

4.2.2 Proximal Gradient Method

Proximal gradient method exploits the notion of decomposible function for the minimisation process. It uses a quadratic approximation of the differentiable term and keeps the non-differentiable term as it is.

Therefore, we can use proximal mapping of $J(\beta)$ (Eq. (4.16)) given by:

$$\text{prox}_t(\beta) = \arg\min_{\gamma} \frac{1}{2t} \|\beta - \gamma\|_2^2 + \|\gamma\|_1 \tag{4.20}$$

This operator makes sure that the solution remains close to β as well as minimises the ℓ_1 penalty term. Beck and Teboulle [8] used this proximal operator to propose the following proximal gradient algorithm:

Algorithm

- Initial guess: β^0
- Increment step: $\beta^{\text{temp}} = \beta^{(k-1)} - t\nabla f(\beta^{(k-1)})$
- Updating: $\beta^{(k)} = \text{prox}_t\left(\beta^{\text{temp}}\right)$

where t denotes a fixed step size. We can also do a lines search technique to accelerate the optimisation problem. We add an intermediate step in the following way:

$$v = \beta^{(k-1)} + \frac{k-2}{k-1}(\beta^{(k-1)} - \beta^{(k-2)}) \tag{4.21}$$

and increment

$$\beta^{\text{temp}} = v - t\nabla f(v) \tag{4.22}$$

Convergence Unlike the subgradient method, the convergence of the proximal gradient method is based on Lipschitz continuity of the gradient of the differentiable term $f(\beta)$. That is, let $L_2 > 0$ be constant such that

$$|\nabla f(\beta) - \nabla f(\gamma)| \leq L_2 \|\beta - \gamma\|_2 \tag{4.23}$$

then for a fixed step-size $t \leq 1/L_2$

$$J(\beta^{(k)}) - J(\beta^*) \leq \frac{\|\beta^0 - \beta^*\|_2^2}{2kt} \tag{4.24}$$

where β^* is the optimal solution. This convergence rate can be improved for accelerated variant such that

$$J(\beta^{(k)}) - J(\beta^*) \leq \frac{2\|\beta^0 - \beta^*\|_2^2}{k(t+1)^2}. \tag{4.25}$$

4.2.3 Co-ordinate Descent Method

The coordinate descent method successively minimises a multivariate function along each coordinate [66] and achieves global minimum. Tseng [73] showed that the solution obtained through coordinate descent method converges to the optimal solution. The algorithm is straight forward and simple.

Algorithm

- Initial guess: β^0
- Updating:

$$\begin{aligned}
\beta_1^{(k+1)} &= \arg\min_{\beta_1} J(\beta_1, \beta_2^{(k)}, \beta_3^{(k)}, \cdots, \beta_p^{(k)}) \\
\beta_2^{(k+1)} &= \arg\min_{\beta_2} J(\beta_1^{(k+1)}, \beta_2, \beta_3^{(k)}, \cdots, \beta_p^{(k)}) \\
&\vdots \\
\beta_p^{(k+1)} &= \arg\min_{\beta_p} J(\beta_1^{(k+1)}, \beta_2^{(k+1)}, \beta_3^{(k+1)}, \cdots, \beta_p)
\end{aligned}$$

Convergence The convergence rate for coordinate descent methods have not been explored much in the literature. However, for LASSO, the convergence is given by Saha and Tewari [65]. The convergence rate for coordinate descent method is also dependent on the Lipschitz continuity of the gradient of the dfferentiable component of the objective funtion. That is, if

$$|\nabla f(\beta) - \nabla f(\gamma)| \leq L_2 \|\beta - \gamma\|_2 \tag{4.26}$$

for some constant $L > 0$, then under some suitable regularity condition

$$J(\beta^{(k)}) - J(\beta^*) \leq \frac{L_2 \|\beta^0 - \beta^*\|_2^2}{2k}. \tag{4.27}$$

Saha and Tewari [65] showed that for LASSO, coordinate descent performs much faster than the other two methods. However, for a general optimisation problem, the performance depends on several parameters and we do not have a single best method. Besides these three methods, there is also a dedicated optimisation method for LASSO-type regression problems called 'LAR' or least angle regression developed by Efron et al. [29]. However, we omit 'LAR' in our discussion as the other three methods are easily interpretable and applicable to any optimisation problem without much modification. These three methods are also useful in the context of Bayesian inference where we may want to compute maximum a posteriori (MAP) estimates.

Chapter 5

Sensitivity Analysis of LASSO-type Problems

In Chapter 3, we have discussed different likelihood-based approaches for linear regression. We learnt how the LASSO can be used for high-dimensional models because of its efficient variable selection. The introduction of LASSO led to several works on the asymptotic properties of variable selection methods. Fan and Li [34] provided the conditions for consistent variable selection and described these properties as oracle properties for variable selection methods. They showed that LASSO can be inconsistent in variable selection at times. Later, Zou [86] introduced an adaptive version of LASSO that satisfies oracle properties for variable selection. That is, adaptive LASSO is consistent in variable selection and the adaptive LASSO estimates are asymptotically unbiased.

In this chapter, we exploit the framework of adaptive LASSO and present a novel sensitivity analysis of LASSO-type problems. In Section 5.1, we introduce adaptive LASSO for linear and logistic models followed by the consistency properties of these LASSO-type problems in Section 5.2. In Section 5.3. we show our sensitivity analysis on adaptive LASSO along with novel error bounds for adaptive LASSO. Finally, in Section 5.4 we introduce a novel robust classification routine for logistic regression problems. Part of these sensitivity analyses have been published [5, 6].

5.1 Adaptive LASSO

Zou [86] introduced the notion of adaptive LASSO. They proposed the idea of data-driven weights in the penalty which satisfies the oracle properties introduced by Fan and Li [34]. Van De Geer and Bühlmann [75] gave restricted eigen value conditions for the LASSO and Van de Geer et al. [74] provided an error bound for the adaptive lasso for misspecified models.

5.1.1 Adaptive LASSO

Let $\hat{\beta} := (\hat{\beta}_1, \cdots, \hat{\beta}_p)$ be any root-n-consistent estimator (see Lehmann and Casella [51, p. 454]) of β, then the adaptive LASSO estimates [86] are given by

$$\hat{\beta}_{\mathrm{AL}}(\lambda, \gamma) := \arg\min_{\beta} \left(\frac{1}{2} \|y - \boldsymbol{x}\beta\|_2^2 + \lambda \sum_{j=1}^{p} w_j(\gamma) |\beta_j| \right) \tag{5.1}$$

where

$$w_j(\gamma) = \frac{1}{|\hat{\beta}_j|^\gamma}, \quad \text{for} \quad \gamma > 0. \tag{5.2}$$

Note, that $\gamma = 0$ gives us the usual LASSO estimates for β. Zou [86] showed that positive values of this additional parameter γ allows the adaptive LASSO to be a consistent estimator which we will discuss in Section 5.2.

The adaptive LASSO can be computed as regular LASSO by using transformation of variables. We rewrite Eq. (5.1) as

$$\hat{\beta}_{\mathrm{AL}}(\lambda, \gamma) = \boldsymbol{k}(\gamma) \arg\min_{\beta^*(\gamma)} \left(\frac{1}{2} \|y - \boldsymbol{x}\boldsymbol{k}(\gamma)\beta^*(\gamma)\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j^*(\gamma)| \right) \tag{5.3}$$

where

$$\boldsymbol{k}(\gamma) := \operatorname{diag}\left(\frac{1}{w_1(\gamma)}, \cdots, \frac{1}{w_p(\gamma)} \right) = \operatorname{diag}\left(|\hat{\beta}_1|^\gamma, \cdots, |\hat{\beta}_p|^\gamma \right) \tag{5.4}$$

and

$$\beta^*(\gamma) := (w_1(\gamma)\beta_1, \cdots, w_p(\gamma)\beta_p) = [\boldsymbol{k}(\gamma)]^{-1}\beta. \tag{5.5}$$

Therefore, with $\boldsymbol{x}^*(\gamma) := \boldsymbol{x}\boldsymbol{k}(\gamma)$,

$$\hat{\beta}^*(\lambda, \gamma) := \arg\min_{\beta^*(\gamma)} \left(\frac{1}{2} \|y - \boldsymbol{x}^*(\gamma)\beta^*(\gamma)\|_2^2 + \lambda \|\beta^*(\gamma)\|_1 \right), \tag{5.6}$$

from which we can compute adaptive LASSO estimate by $\hat{\beta}_{\text{AL}}(\lambda, \gamma) = \boldsymbol{k}(\gamma)\hat{\beta}^*(\lambda, \gamma)$.

In general, we cannot find an analytical solution to Eq. (5.1) or Eq. (5.6) and we need to use an iterative soft-thresholding operator to get a solution similar to LASSO. For the orthogonal design case, the weights and the parameter γ (Eq. (5.2)) in the adaptive LASSO gives us a modified soft-thresholding operator (see Section 3.5) in the following way:

$$\hat{\beta}_{\text{AL}}(\lambda, \gamma, \hat{\beta}) = \text{Soft}(\hat{\beta}_{\text{OLS}}; \lambda/|\hat{\beta}|^{\gamma}) = \text{sign}(\hat{\beta}_{\text{OLS}}) \cdot \max\left\{0, \left(|\hat{\beta}_{\text{OLS}}| - \frac{\lambda}{|\hat{\beta}|^{\gamma}}\right)\right\}, \quad (5.7)$$

where $\hat{\beta}$ is any root-n-consistent estimate of β in Eq. (2.2). Since ordinary least squares estimates are root-n-consistent, therefore using $\hat{\beta} = \hat{\beta}_{\text{OLS}}$ in Eq. (5.7), we get

$$\hat{\beta}_{\text{AL}}(\lambda, \gamma) = \text{Soft}(\hat{\beta}_{\text{OLS}}; \lambda/\hat{\beta}_{\text{OLS}}^{\gamma}). \quad (5.8)$$

In Fig. 5.1, we illustrate these soft-thresholding operators using Eq. (5.8) for different values of γ. Here, the dotted line represents the true values of β and bold line represents adaptive LASSO estimates. We see that setting $\gamma = 0$ (top left) gives us a constant shift from the true value even for large values of β. However, as we increase γ, we see that the adaptive LASSO estimates becomes close to true value.

5.1.2 Adaptive Penalised Logistic Regression (APLR)

Similar to LASSO, the LASSO-type penalty in PLR can be inconsistent in variable selection and it is also not asymptotically unbiased. We can overcome this issue through the idea of adaptive LASSO. This approach is known to be adaptive penalised logistic regression (APLR) [86, 2].

Let $\hat{\beta} := (\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_p)$ be any root-n-consistent estimate for our logistic regression problem. Then, for any fixed $\gamma > 0$, the APLR [86] estimates are given by:

$$\hat{\beta}_{\text{aplr}}(\lambda, \gamma) := \arg\min_{\beta} \left(-\log(L(c, \boldsymbol{x}; \beta)) + \lambda \sum_{j=1}^{p} w_j(\gamma)|\beta_j| \right) \quad (5.9)$$

where

$$w_j(\gamma) := \frac{1}{|\hat{\beta}_j|^{\gamma}}. \quad (5.10)$$

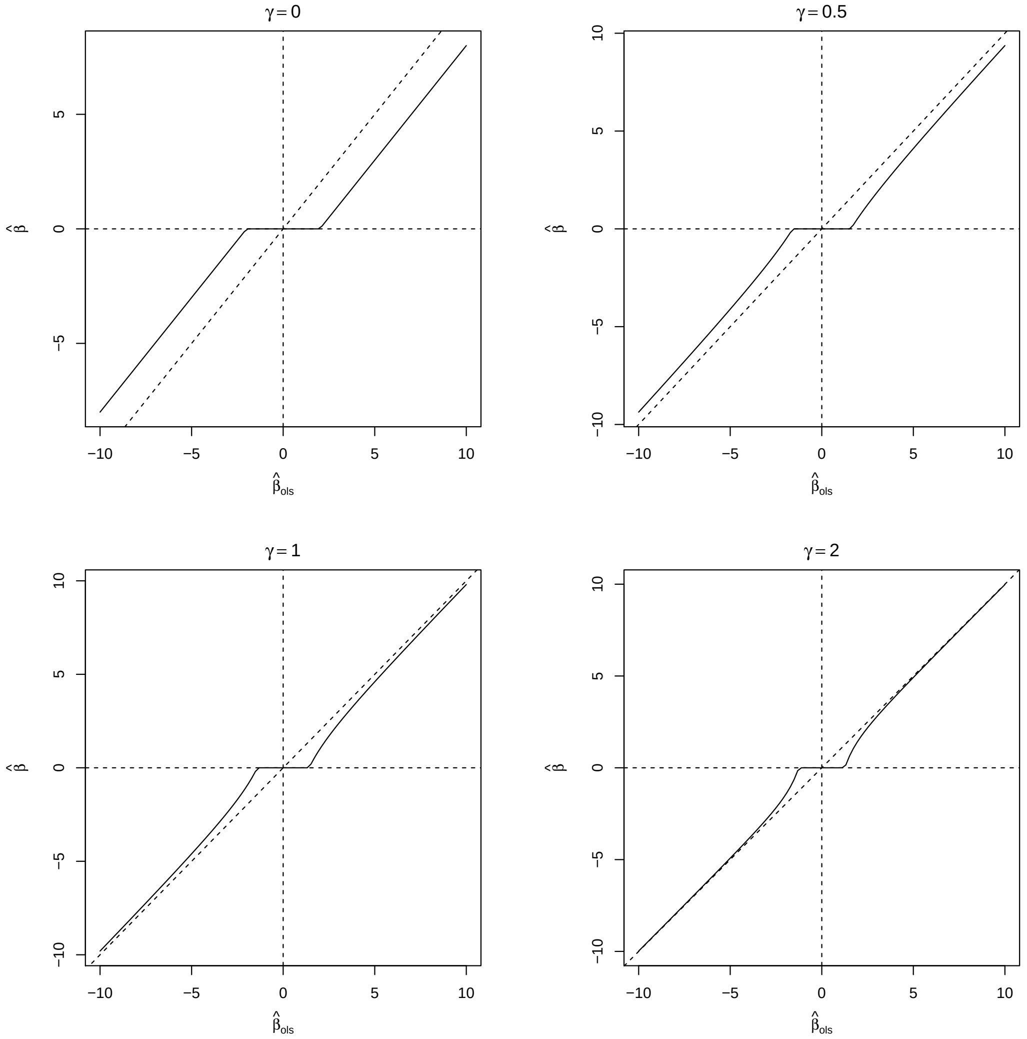

Figure 5.1: Soft thresholding operator for different values of γ and fixed λ (= 2).

Zou [86] showed that with these weights along with some suitable regularity conditions, APLR follows desirable asymptotic properties for high-dimensional problems [34].

Computation: For $\gamma > 0$, the objective function of APLR is given by:

$$J(\beta) := \left(\sum_{i=1}^{m} \left[-c_i \left(x_i^T \beta \right) + \log \left(1 + \exp(x_i^T \beta) \right) \right] + \lambda \sum_{j=1}^{p} w_j(\gamma) |\beta_j| \right), \tag{5.11}$$

where $w_j(\gamma)$ is given by Eq. (5.10). Now, for optimality Eq. (5.11) must satisfy the Karush-Kuhn-Tucker condition. Therefore, we have,

$$0 \in \sum_{i=1}^{m} \left[-x_{ij} c_i + x_{ij} \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)} \right] + \lambda w_j(\gamma) \partial(|\beta_j|), \tag{5.12}$$

where $\partial|\beta_j|$ is as defined in Eq. (4.14).

Let $s := (s_1, s_2, \cdots, s_p)$ be subject to the constraint $s \in \partial|\beta_j|$. Then, $\hat{\beta}_{\text{aplr}}$ satisfies the following:

$$\sum_{i=1}^{m} \left[-x_{ij} c_i + x_{ij} \frac{\exp\left(x_i^T \hat{\beta}_{\text{aplr}}(\lambda, \gamma) \right)}{1 + \exp\left(x_i^T \hat{\beta}_{\text{aplr}}(\lambda, \gamma) \right)} \right] = -\lambda w_j(\gamma) s_j \tag{5.13}$$

$$\sum_{i=1}^{m} x_{ji} \left[c_i - \frac{\exp\left(x_i^T \hat{\beta}_{\text{aplr}}(\lambda, \gamma) \right)}{1 + \exp\left(x_i^T \hat{\beta}_{\text{aplr}}(\lambda, \gamma) \right)} \right] = \lambda w_j(\gamma) s_j. \tag{5.14}$$

Now, let $h(\boldsymbol{x}\hat{\beta}) := \left(h\left(x_1^T \hat{\beta} \right), h\left(x_2^T \hat{\beta} \right), \cdots, h\left(x_n^T \hat{\beta} \right) \right)^T$, where h is the link function defined in Eq. (3.30). Then, we can write Eq. (5.14) as,

$$\boldsymbol{x}^T \left[c - h\left(\boldsymbol{x} \hat{\beta}_{\text{aplr}}(\lambda, \gamma) \right) \right] = \lambda w(\gamma) \cdot s \tag{5.15}$$

where '$\cdot$' denotes component wise multiplication. Note that Eq. (5.15) is not analytically solvable for $\hat{\beta}_{\text{aplr}}$. However, any sub-gradient based numerical optimisation method can be applied to solve it. Then similar to our discussion in Section 3.6, we compute $\hat{\pi}(x_*, \lambda, \gamma)$ for new observation x_* to predict the corresponding class.

5.2 Consistency and Oracle Properties

Let the LASSO estimator be defined by Eq. (3.22). We define the subset $\mathcal{S}$ such that,

$$\mathcal{S} := \{ j : \beta_j \neq 0 \} \quad \text{and} \quad |\mathcal{S}| = p^* < p. \tag{5.16}$$

That is, the true model can be specified by p^* predictors. Then we can rearrange the input matrix $\boldsymbol{x}$ such that first p^* predictors correctly identify the model. Since, $|S| = p^* < p$, then without loss of generality we can write the following

$$\lim_{n\to\infty} \frac{1}{n}\boldsymbol{x}^T\boldsymbol{x} = \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \tag{5.17}$$

such that Σ_{11} is a $p^* \times p^*$ matrix. Now, let for n number of samples $\mathcal{S}_n := \{j : \hat{\beta}_j \neq 0\}$. Then Zhao and Yu [84] and Zou [86] independently showed that the following condition is necessary for the consistency of the LASSO estimator.

Theorem 5.1. *Let,* $\lim_{n\to\infty} P(\mathcal{S}_n = \mathcal{S}) = 1$. *Then there exists some sign vector* $s := (s_1, s_2, \cdots, s_{p^*})$ *such that,*

$$|\Sigma_{21}\Sigma_{11}^{-1}s| \leq 1 \tag{5.18}$$

for each component of the left hand side.

This may not seem convincing at first. However, this holds as Σ_{11} Σ_{21} are limiting values of block diagonal components as described in Eq. (5.17). We suggest to check the articles by Yuan and Lin [80] and Zou [86] for a detailed discussion and proof.

Consistency for Adaptive LASSO

Let $\mathcal{S}_{\text{AL}}^{(n)}$ be the selected subset by the adaptive LASSO when the sample size is n. That is,

$$\mathcal{S}_{\text{AL}}^{(n)} := \{j : \hat{\beta}_{\text{AL};\, j}^{(n)} \neq 0\}. \tag{5.19}$$

Let $\beta^* := (\beta_1^*, \cdots, \beta_p^*)$ be the vector of true regression coefficients. Zou [86] showed that the Adaptive LASSO estimates satisfy the following asymptotic properties:

Definition 5.2 (Oracle Properties)**.** *Let* $\frac{\lambda^{(n)}}{\sqrt{n}} \to 0$ *and* $\lambda^{(n)} n^{(\gamma-1)/2} \to \infty$.

P.1 Consistent variable selection: $\lim_{n\to\infty} P\left(\mathcal{S}_{AL}^{(n)} = \mathcal{S}\right) = 1$

P.2 Asymptotic normality: $\sqrt{n}\left(\hat{\beta}_{AL,\,\mathcal{S}}^{(n)} - \beta_{\mathcal{S}}^*\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2\Sigma_{11}^{-1})$

Here, $\xrightarrow{d}$ denotes the convergence in distribution (see Lehmann and Casella [51] for further readings). Zou [86] also noted that adaptive LASSO estimates can follow oracle properties under even weaker conditions on the convergence of λ.

Consistency for APLR

For a sequence of n observations, where x_i is the attribute vector for the i-th observation, we now denote:

$$\boldsymbol{x}_n := x = [x_1, \cdots, x_n]^T \tag{5.20}$$

in order to make the dependence of this $p \times n$ matrix on n explicit.

Let $\mathcal{S}$ be the true subset as defined in Eq. (5.16) and let $\phi(x) := \log(1+\exp(x))$, then for any observation $x_i \in \mathbb{R}^p$ $(1 \leq i \leq n)$, we define the Fisher information matrix by:

$$I(\beta) := \phi''(x_i^T b)x_i x_i^T = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix} \tag{5.21}$$

where I_{11} is a $p^* \times p^*$ matrix.

Regularity Conditions: We define the following regularity conditions for asymptotic properties of APLR.

LC.1 Let $\lambda_n(\gamma)$ be a sequence such that, for $\gamma > 0$

$$\lim_{n\to\infty} \frac{\lambda_n(\gamma)}{\sqrt{n}} = 0 \quad \text{and} \quad \lim_{n\to\infty} \lambda_n(\gamma) n^{(\gamma-1)/2} = \infty. \tag{5.22}$$

For example, the above holds for $\lambda_n(\gamma) = n^{1/2-\gamma/4}$.

LC.2 The Fisher information matrix is finite and positive definite.

LC.3 Let there exist an open set $\mathcal{B} \subseteq \mathbb{R}^p$, such that $\beta^* \in \mathcal{B}$. Then for every $\beta \in \mathcal{B}$ and observation $x_i \in \mathbb{R}^p$ $(1 \leq i \leq n)$, there exists a function M so that

$$\left|\phi'''(x_i^T \beta)\right| \leq M(x_i) < \infty. \tag{5.23}$$

Let $\mathcal{S}_{\text{aplr}}^{(n)} := \{j : \hat{\beta}_{\text{aplr; } j}^{(n)} \neq 0\}$..

Theorem 5.3. *Under LC.1-LC.3, APLR estimates satisfy the following properties:*

LP.1 Consistency in variable selection, i.e.

$$\lim_{n\to\infty} P\left(\mathcal{S}_{aplr}^{(n)} = \mathcal{S}\right) = 1 \tag{5.24}$$

LP.2 Asymptotic normality, i.e.

$$\sqrt{n}\left(\hat{\beta}_{aplr,\,\mathcal{S}} - \beta^*_{\mathcal{S}}\right) \xrightarrow{d} \mathcal{N}(0, I_{11}^{-1}) \tag{5.25}$$

Note, that here $\hat{\beta}_{\text{aplr},\,\mathcal{S}}$ is dependent on both $\lambda_n(\gamma)$ and γ but we omit these for the sake of notation. The proof is already provided by Zou [86] and therefore we omit.

5.3 Sensitivity Analysis of Adaptive LASSO

The framework of the adaptive LASSO allows us to investigate and understand the sensitivity of the adaptive lasso estimates with respect to the weight parameter γ. For this we apply a two-step approach. We use a root-n-consistent estimate to initialise the adaptive LASSO and consider the weights as function of γ. This allows us to obtain the adaptive lasso estimates as functions of γ and we use these estimates to obtain novel error bounds for special type of problems. Let,

$$y = \boldsymbol{x}\beta^* + \epsilon \tag{5.26}$$

be the model with true regression coefficients β^*, such that $|\beta^*| \gg 1$ and $\boldsymbol{x}$ has full column rank, ie. $\boldsymbol{x}^T\boldsymbol{x}$ is invertible.

Let, $\boldsymbol{k}(\gamma)$ be defined by Eq. (5.4) and for the sake of notation, we write it as $\boldsymbol{k}$. Let $\hat{\beta}$ be any root-n-consistent estimate such that

$$\hat{\beta} = (\hat{\beta}_1, \cdots, \hat{\beta}_p). \tag{5.27}$$

Theorem 5.4. *Then, for large effects models (ie. $|\beta_j| \gg 1$ and $0 < \lambda < \min\{\boldsymbol{k}\boldsymbol{x}^T y\}$, we have,*

$$\left\|\hat{\beta}_{AL}(\lambda,\gamma) - \beta^*\right\|_2^2 \leq \frac{\sigma^2}{n}\left\|\Sigma_n^{-1}\right\| + \frac{\lambda^2 p}{n^2}\left\|\Sigma_n^{-1}\right\|^2 \min_{1\leq j\leq p} |\hat{\beta}_j|^{-2\gamma}, \tag{5.28}$$

$$\left\|y - \boldsymbol{x}\hat{\beta}_{AL}(\lambda,\gamma)\right\|_2^2 \leq \frac{\lambda^2 p}{n}\left\|\Sigma_n^{-1}\right\| \min_{1\leq j\leq p} |\hat{\beta}_j|^{-2\gamma}. \tag{5.29}$$

This shows that we reduce the mean square error by increasing the value of γ. It also indicates that for higher values of γ, λ does not control any shrinkage over large effects and produce unbiased estimates. This happens as the $|\hat{\beta}_j|^\gamma$ becomes close to zero for higher values of γ and therefore the effect of λ gets reduced.

Proof. Let the adaptive LASSO model be defined by Eq. (5.1). We use Ridge estimates as the weights of adaptive lasso. Then the weights are given by:

$$w(\gamma) = \left(\frac{1}{|\hat{\beta}_1|^\gamma}, \cdots \frac{1}{|\hat{\beta}_p|^\gamma} \right). \tag{5.30}$$

Then, applying $w(\gamma)$ as weights in adaptive LASSO estimates we get,

$$\hat{\beta}_{\text{AL}}(\lambda, \gamma) = \arg\min_{\beta} \left(\frac{1}{2} \|y - \boldsymbol{x}\beta\|_2^2 + \lambda \sum_{j=1}^{p} w_j(\gamma) |\beta_j| \right). \tag{5.31}$$

Now, applying Karush-Kahn-Tucker condition in Eq. (5.31), we have

$$0 \in -\boldsymbol{x}^T(y - \boldsymbol{x}\beta) + \lambda \boldsymbol{k}^{-1} \partial \|\beta\|_1, \tag{5.32}$$

with $\partial\|\beta\|_1$ as defined in Eq. (4.14). Note that, for any fixed $\lambda < min\{\boldsymbol{k}\boldsymbol{x}^T y\}$, $\beta_j \neq 0$ for $1 \leq j \leq p$. Then, from Eq. (4.14) we have:

$$\text{sign}(\beta_j) := \begin{cases} \{-1\} & \text{if } \beta_j < 0 \\ \{1\} & \text{if } \beta_j > 0. \end{cases} \tag{5.33}$$

Therefore, we write Adaptive LASSO estimates as:

$$\boldsymbol{x}^T \left(y - \boldsymbol{x}\hat{\beta}_{\text{AL}}(\lambda, \gamma) \right) = \lambda \boldsymbol{k}^{-1} s \tag{5.34}$$

$$\boldsymbol{x}^T \left(\boldsymbol{x}\beta^* + \epsilon - \boldsymbol{x}\hat{\beta}_{\text{AL}}(\lambda, \gamma) \right) = \lambda \boldsymbol{k}^{-1} s \tag{5.35}$$

where $s = (s_1, s_2, \ldots, s_{p^*})$ are auxiliary variables subject to the constraint $s_j \in \text{sign}(\beta_j)$.

Now, from Eq. (5.35), we get

$$\frac{1}{n} \boldsymbol{x}^T \left(\boldsymbol{x}\beta^* + \epsilon - \boldsymbol{x}\hat{\beta}_{\text{AL}}(\lambda, \gamma) \right) = \frac{\lambda}{n} \boldsymbol{k}^{-1} s \tag{5.36}$$

$$\frac{1}{n} \boldsymbol{x}^T \boldsymbol{x} \left(\beta^* - \hat{\beta}_{\text{AL}}(\lambda, \gamma) \right) = \frac{\lambda}{n} \boldsymbol{k}^{-1} s - \frac{1}{n} \boldsymbol{x}^T \epsilon \tag{5.37}$$

Since, inverse of $\boldsymbol{x}^T\boldsymbol{x}$ exists. Then,

$$\left(\beta^* - \hat{\beta}_{\text{AL}}(\lambda, \gamma) \right) = \Sigma_n^{-1} \left(\frac{\lambda}{n} \boldsymbol{k}^{-1} s - \frac{1}{n} \boldsymbol{x}^T \epsilon \right) \tag{5.38}$$

taking norm in both sides,

$$\left\|\hat{\beta}_{\mathrm{AL}}(\lambda,\gamma)-\beta^*\right\|_2^2 = \left\|\Sigma_n^{-1}\left(\frac{1}{n}\boldsymbol{x}^T\epsilon-\frac{\lambda}{n}\boldsymbol{k}^{-1}s\right)\right\|_2^2 \tag{5.39}$$

$$\leq \left\|\Sigma_n^{-1}\frac{1}{n}\boldsymbol{x}^T\epsilon\right\|_2^2+\left\|\frac{\lambda}{n}\Sigma_n^{-1}\boldsymbol{k}^{-1}s\right\|_2^2 \tag{5.40}$$

$$\leq \frac{1}{n}\left\|\Sigma_n^{-1}\right\|^2\left\|\frac{1}{\sqrt{n}}\boldsymbol{x}^T\epsilon\right\|_2^2+\frac{\lambda^2}{n^2}\left\|\Sigma_n^{-1}\boldsymbol{k}^{-1}\right\|^2\|s\|_2^2. \tag{5.41}$$

Here, $\|.\|$ is the induced matrix norm in $\mathbb{R}^p$. Now, since, $\|s\|_2^2=p$

$$\leq \frac{\sigma^2}{n}\left\|\Sigma_n^{-1}\right\|^2\|\Sigma_n\|+\frac{p\cdot\lambda^2}{n^2}\left\|\Sigma_n^{-1}\right\|^2\left\|\boldsymbol{k}^{-1}\right\|^2 \tag{5.42}$$

$$\leq \frac{\sigma^2}{n}\left\|\Sigma_n^{-1}\right\|+\frac{\lambda^2 p}{n^2}\left\|\Sigma_n^{-1}\right\|^2\left\|\boldsymbol{k}^{-1}\right\|^2 \tag{5.43}$$

$$\leq \frac{\sigma^2}{n}\left\|\Sigma_n^{-1}\right\|+\frac{\lambda^2 p}{n^2}\left\|\Sigma_n^{-1}\right\|^2\min_{1\leq i\leq p}|\hat{\beta}_i|^{-2\gamma}. \tag{5.44}$$

Similarly, from Eq. (5.34), we have,

$$\left(\boldsymbol{x}\boldsymbol{x}^T\right)^{-1}\boldsymbol{x}\boldsymbol{x}^T\left(y-\boldsymbol{x}\hat{\beta}_{\mathrm{AL}}(\lambda,\gamma)\right)=\lambda\left(\boldsymbol{x}\boldsymbol{x}^T\right)^{-1}\boldsymbol{x}\boldsymbol{k}^{-1}s \tag{5.45}$$

$$\left(y-\boldsymbol{x}\hat{\beta}_{\mathrm{AL}}(\lambda,\gamma)\right)=\lambda\left(\boldsymbol{x}\boldsymbol{x}^T\right)^{-1}\boldsymbol{x}\boldsymbol{k}^{-1}s \tag{5.46}$$

Taking norm on both sides of Eq. (5.46), we get

$$\left\|y-\boldsymbol{x}\hat{\beta}_{\mathrm{AL}}(\lambda,\gamma)\right\|_2^2=\lambda^2\left\|\left(\boldsymbol{x}\boldsymbol{x}^T\right)^{-1}\boldsymbol{x}\boldsymbol{k}^{-1}s\right\|_2^2 \tag{5.47}$$

applying the Cauchy-Schwartz inequality

$$\leq \frac{\lambda^2}{n}\left\|\Sigma_n^{-1}\right\|\left\|\boldsymbol{k}^{-1}\right\|^2\|s\|_2^2. \tag{5.48}$$

Therefore, we get the following:

$$\left\|y-\boldsymbol{x}\hat{\beta}_{\mathrm{AL}}(\lambda,\gamma)\right\|_2^2\leq\frac{\lambda^2 p}{n}\left\|\Sigma_n^{-1}\right\|\min_{1\leq i\leq p}|\hat{\beta}_i|^{-2\gamma}. \tag{5.49}$$

□

5.3.1 Simulation Study

Example 5.5. *We simulate the predictors from a standard normal distribution such that,* $x_{i,j}\sim\mathcal{N}(0,1)$ *for* $j=1,\cdots,20$ *and* $i=1,\cdots,n$*. We assign the regression coefficients so that* $\beta_j\sim$ *Uniform*$(-15,-1)$ *for* $1\leq j\leq 10$ *and* $\beta_j\sim$ *Uniform*$(1,15)$

for $11 \leq j \leq 20$. *This construction assures that the true regression coefficient values are greater than 1. We consider standard normal noise to construct the response vector* $y_i = \sum_{j=1}^{20} x_{i,j}\beta_j + \epsilon_i$, *where* $\epsilon_i \sim \mathcal{N}(0, 0.01)$ *for* $i = 1, \cdots, n$. *We repeat this experiment for* $n = 50, 100, 1000$.

We analyse the sensitivity of the model for $0 \leq \gamma \leq 10$ ($\gamma = 0$ allows us to obtain regular LASSO estimates). On the right hand side in Fig. 5.2, we show the total number of selected predictors in the model for three different choices of n. The different lines corresponds to different values of λ. We see an interesting feature of the model. The λ forcefully shrinks some true non-zero effects to zero for the smaller values of γ. However, as γ increases, the effect of λ becomes less significant and the predictors are included in the model.

To inspect the effect of γ in model fitting, we use mean squared error as a measure of accuracy. We compute the mean squared error as:

$$\mathrm{MSE} = \frac{1}{n}\|y - \boldsymbol{x}\hat{\beta}\|_2^2. \tag{5.50}$$

In the left hand side of Fig. 5.2, we show these mean squared errors. We notice that MSE becomes smaller as the γ increases. We can also see that as we increase the amount observations, the MSE becomes smaller.

5.4 High-dimensional Credal Classification

In Section 5.1, we discussed how an adaptive version of the LASSO for penalised logistic regression can be used for variable selection, which satisfies suitable asymptotic properties [34]. Several other works can be found in the field of penalised logistic regression. However, there isn't as much work on the cases where we deal with limited information, which requires a robust classification regime. We therefore propose an imprecise probabilistic approach in the context of high-dimensional logistic regression.

Several works related to classification can be found in the imprecise probability literature. Zaffalon [82] introduced the idea of the naive credal classifier related to the imprecise Dirichlet model [76]. Bickis [12] introduced an imprecise logit-normal model for logistic regression. Corani and de Campos [20] proposed the tree

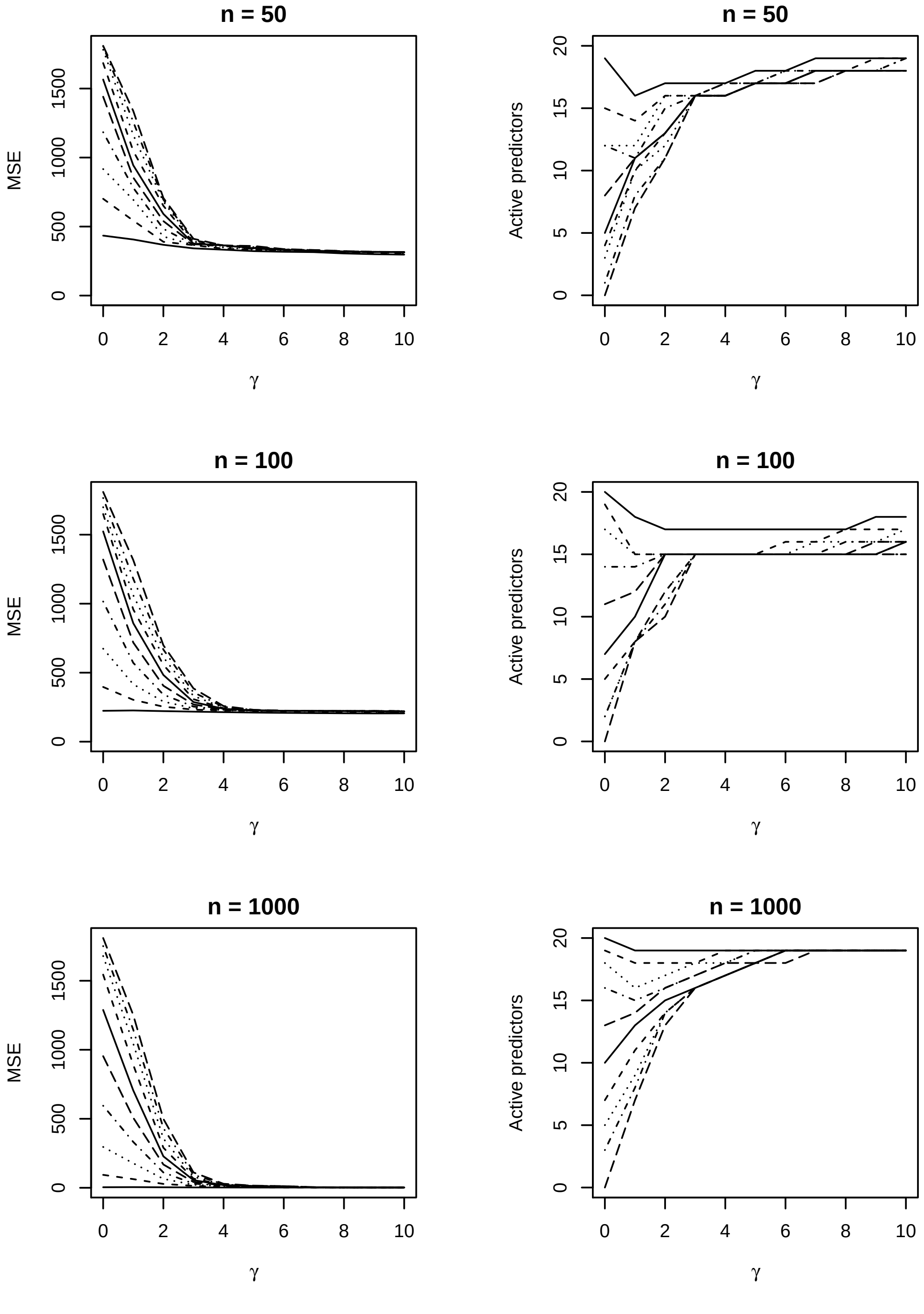

Figure 5.2: Mean squared error (left) and number of active predictors (right) for three different sample size $n = 50$ (top), $n = 100$ (middle) and $n = 1000$ (bottom).

augmented naive classifier based on the imprecise Dirichlet model. Paton *et al.* [61, 62] used a near vacuous set of priors for multinomial logistic regression. José del Coz et al. [49] and Corani and Antonucci [19] investigated rejection-based classifiers for attribute selection. However, high-dimensional problems with automatic attribute selection are yet to be tackled in the context of imprecise probability.

In this section, we introduce a novel imprecise likelihood-based approach for high-dimensional logistic regression problems. We use a set of sparsity constraints through weights in the penalty term. Working with a set of weights relaxes the assumption of preassigned weights and also helps to identify the behaviour of the attributes, whereas sparsity constraints help in variable selection, which is essential for working with high-dimensional problems. We use cross-validation for model validation using different performance measures as suggested by Corani and Zaffalon [21].

5.4.1 Imprecise Adaptive Penalised Logistic Regression

The use of data-driven weights in APLR makes APLR consistent in attribute selection, where the parameter γ is pre-assigned (usually equal to 1) or is estimated through cross-validation. However, high-dimensional problems are sparse in nature, i.e. we have to deal with very limited information and the preliminary estimates (ridge estimates) used for the weights in the adaptive LASSO can be sensitive and unstable. Therefore a single may leads to misclassification, especially when the variability of the attributes is negligible with respect to each other or the observations contain outliers. Sometimes, APLR may also perform poorly during model validation as, a single value of γ can provide two very different vectors of weights for two different parts of a single dataset. For instance, fixing $\gamma = 1$, essentially gives us the inverse of the absolute values of our estimates, which are generally sensitive to the data in sparse regime. So, we propose a sensitivity analysis of APLR over an interval of γ and obtain a non-determinate classifier. We call this method as imprecise adaptive penalised logistic regression or simply IAPLR. This allows the weights to vary in the order of γ providing us a set of sparsity constraints of the form $\sum_{j=1}^{p} |\beta_j|/|\hat{\beta}_j|^{\gamma}$ (see Chapter 4 for constrained optimisation). This set of weight vectors allows the model to be flexible but consistent as we only rely on the data-driven weights.

The sensitivity analysis gives us a set of APLR estimates as a function of γ. We use this set of APLR estimates to obtain a set of estimated probabilities which are used for the decision making.

Decision rule

Consider the APLR estimates defined by Eq. (5.9) and Eq. (5.10). As we described earlier, we perform a sensitivity analysis on the parameter γ. This gives us a set of estimated probabilities dependent on γ, such that $\gamma \in [\underline{\gamma}, \overline{\gamma}]$. We use the notion of credal dominance [82] for the decision criteria.

We can then for instance classify a new object with attributes $x_* \in \mathbb{R}^p$ as $\{0\}$ if $\hat{\pi}(x_*, \lambda, \gamma) < 1/2$ for all $\gamma \in [\underline{\gamma}, \overline{\gamma}]$, as $\{1\}$ if $\hat{\pi}(x_*, \lambda, \gamma) \geq 1/2$ for all $\gamma \in [\underline{\gamma}, \overline{\gamma}]$, and as $\{0, 1\}$ (i.e. indeterminate) otherwise. Note that our classifier now returns non-empty subsets of $\{0, 1\}$ rather than elements of $\{0, 1\}$, to allow indeterminate classifications to be expressed.

5.4.2 Prediction Consistency

We already discussed that IAPLR may give us non-deterministic class as output. The non-deterministic output suggests that we need more samples to obtain a deterministic output. However, for that, we want to be sure that a method is consistent in prediction. That is, if we have infinite amount of data during the decision making process, then we our estimated decision probabilities will be converge to the actual decision probabilities in distribution (see Lehmann and Casella [51] for further discussion on asymptotic concepts).

We define the following:

$$x_{*,\mathcal{S}} := [x_{*,j}]_{j \in \mathcal{S}}, \tag{5.51}$$

i.e., $x_{*,\mathcal{S}}$ is a p^*-dimensional vector.

Theorem 5.6. *Let $x_* \in \mathbb{R}^p$ such that $x_{*,\mathcal{S}}^T x_{*,\mathcal{S}} > 0$. Then for $\gamma > 0$ and under LC.1 -LC.3, we have the following:*

$$\sqrt{n}\left(\hat{\pi}(x_*, \lambda_n(\gamma), \gamma) - \pi(x_*)\right) \xrightarrow{d} \mathcal{N}\left(0, \left[\pi(x_*)\left(1 - \pi(x_*)\right)\right]^2 x_{*,\mathcal{S}}^T I_{11}^{-1} x_{*,\mathcal{S}}\right) \tag{5.52}$$

where I_{11} is the leading block matrix of the Fisher information matrix defined in Eq. (5.21).

Proof. Let $\mathcal{S}_{\text{aplr}}^{(n)^C}$ be the set so that

$$\mathcal{S}_{\text{aplr}}^{(n)^C} \coloneqq \{j : \hat{\beta}_{\text{aplr; } j}^{(n)} = 0\}. \tag{5.53}$$

Then we have,

$$\begin{aligned} x_*^T \hat{\beta}_{\text{aplr}} &= x_{*,\mathcal{S}_{\text{aplr}}^{(n)}} \hat{\beta}_{\text{aplr, } \mathcal{S}_{\text{aplr}}^{(n)}} + x_{*,\mathcal{S}_{\text{aplr}}^{(n)^C}} \hat{\beta}_{\text{aplr, } \mathcal{S}_{\text{aplr}}^{(n)^C}} && (5.54) \\ &= x_{*,\mathcal{S}_{\text{aplr}}^{(n)}} \hat{\beta}_{\text{aplr, } \mathcal{S}_{\text{aplr}}^{(n)}} && (5.55) \\ &= \sum_{j \in \mathcal{S}_{\text{aplr}}^{(n)}} x_* \hat{\beta}_{\text{aplr, } j}. && (5.56) \end{aligned}$$

We know that, under LC.1-LC.3 APLR estimates satifies LP.1. Therefore, as $n \to \infty$,

$$\sum_{j \in \mathcal{S}_{\text{aplr}}^{(n)}} x_* \hat{\beta}_{\text{aplr, } j} \xrightarrow{p} \sum_{j \in \mathcal{S}} x_* \hat{\beta}_{\text{aplr, } j} = x_{*,\mathcal{S}}^T \hat{\beta}_{\text{aplr, } \mathcal{S}} \tag{5.57}$$

where $\xrightarrow{p}$ denotes convergence in probability. Since convergence in probability implies convergence in distribution and h is continuous bounded mapping, we can write the following:

$$\hat{\pi}(x_*, \lambda_n(\gamma), \gamma) \xrightarrow{d} h\left(x_{*,\mathcal{S}}^T \hat{\beta}_{\text{aplr, } \mathcal{S}}\right). \tag{5.58}$$

For a detailed discussion on the above convergence properties we refer to the book authored by Lehmann and Casella [51].

Now, by LP.2, we know that $\hat{\beta}_{\text{aplr, } \mathcal{S}}$ is root-n-consistent. Therefore,

$$\left(\hat{\beta}_{\text{aplr, } \mathcal{S}} - \beta_{\mathcal{S}}^*\right) = O_p(n^{-1/2}). \tag{5.59}$$

Here O_p is used to denote that $\sqrt{n}\left(\hat{\beta}_{\text{aplr, } \mathcal{S}} - \beta_{\mathcal{S}}^*\right)$ is bounded in probability (adapted from *Theory of Point Estimation* [51]). Now, following the approach of Agresti [1] for logistic regression problems, we apply Taylor's series expansion in Eq. (5.58) with respect to the true parameter $\beta_{\mathcal{S}}^*$. Then we have,

$$\begin{aligned} \hat{\pi}(x_*, \lambda_n(\gamma), \gamma) &\xrightarrow{d} h\left(x_{*,\mathcal{S}}^T \beta_{\mathcal{S}}^*\right) + \left(\hat{\beta}_{\text{aplr, } \mathcal{S}} - \beta_{\mathcal{S}}^*\right)^T \frac{\partial h\left(x_{*,\mathcal{S}}^T \beta_{\mathcal{S}}^*\right)}{\partial \beta_{\mathcal{S}}^*} + o_p(n^{-1/2}) && (5.60) \\ &\xrightarrow{d} \pi(x_*) + \left(\hat{\beta}_{\text{aplr, } \mathcal{S}} - \beta_{\mathcal{S}}^*\right)^T \frac{\partial h\left(x_{*,\mathcal{S}}^T \beta_{\mathcal{S}}^*\right)}{\partial \beta_{\mathcal{S}}^*} + o_p(n^{-1/2}). && (5.61) \end{aligned}$$

Here, $o_p(n^{-1/2})$ is used to denote convergence in probability in the order of $n^{-1/2}$. We get this convergence from Eq. (5.59). Now, re-arranging the terms we get,

$$\hat{\pi}(x_*, \lambda_n(\gamma), \gamma) - \pi(x_*) \xrightarrow{d} \left(\hat{\beta}_{\text{aplr},\,\mathcal{S}} - \beta^*_{\mathcal{S}}\right)^T \frac{\partial h\left(x^T_{*,\mathcal{S}}\beta^*_{\mathcal{S}}\right)}{\partial \beta^*_{\mathcal{S}}} + o_p(n^{-1/2}). \tag{5.62}$$

Now, from LP.2 we have,

$$\sqrt{n}\left(\hat{\beta}_{\text{aplr},\,\mathcal{S}} - \beta^*_{\mathcal{S}}\right) \xrightarrow{d} \mathcal{N}\left(0, I_{11}^{-1}\right). \tag{5.63}$$

Then, applying Eq. (5.63) in Eq. (5.61), we get

$$\sqrt{n}\left(\hat{\pi}(x_*, \lambda_n(\gamma), \gamma) - \pi(x_*)\right) \xrightarrow{d} \mathcal{N}\left(0, \left[\frac{\partial h\left(x^T_{*,\mathcal{S}}\beta^*_{\mathcal{S}}\right)}{\partial \beta^*_{\mathcal{S}}}\right]^T I_{11}^{-1} \frac{\partial h\left(x^T_{*,\mathcal{S}}\beta^*_{\mathcal{S}}\right)}{\partial \beta^*_{\mathcal{S}}}\right). \tag{5.64}$$

Now,

$$\begin{aligned}
\frac{\partial h\left(x^T_{*,\mathcal{S}}\beta^*_{\mathcal{S}}\right)}{\partial \beta^*_{\mathcal{S}}} &= \left[\frac{\exp\left(x^T_{*,\mathcal{S}}\beta^*_{\mathcal{S}}\right)\left(1+\exp\left(x^T_{*,\mathcal{S}}\beta^*_{\mathcal{S}}\right)\right) - \exp\left(x^T_{*,\mathcal{S}}\beta^*_{\mathcal{S}}\right)^2}{\left(1+\exp\left(x^T_{*,\mathcal{S}}\beta^*_{\mathcal{S}}\right)\right)^2}\right] x_{*,\mathcal{S}} && (5.65)\\
&= \left[\frac{\exp\left(x^T_{*,\mathcal{S}}\beta^*_{\mathcal{S}}\right)}{\left(1+\exp\left(x^T_{*,\mathcal{S}}\beta^*_{\mathcal{S}}\right)\right)^2}\right] x_{*,\mathcal{S}} && (5.66)\\
&= h\left(x^T_{*,\mathcal{S}}\beta^*_{\mathcal{S}}\right)\left[1 - \frac{\exp\left(x^T_{*,\mathcal{S}}\beta^*_{\mathcal{S}}\right)}{1+\exp\left(x^T_{*,\mathcal{S}}\beta^*_{\mathcal{S}}\right)}\right] x_{*,\mathcal{S}} && (5.67)\\
&= h\left(x^T_{*,\mathcal{S}}\beta^*_{\mathcal{S}}\right)\left(1 - h\left(x^T_{*,\mathcal{S}}\beta^*_{\mathcal{S}}\right)\right) x_{*,\mathcal{S}} && (5.68)\\
&= h\left(x^T_{*}\beta^*\right)\left(1 - h\left(x^T_{*}\beta^*\right)\right) x_{*,\mathcal{S}} && (5.69)\\
&= \pi(x_*)\left(1 - \pi(x_*)\right) x_{*,\mathcal{S}}. && (5.70)
\end{aligned}$$

Therefore, using Eq. (5.70) in Eq. (5.64), we have

$$\sqrt{n}\left(\hat{\pi}(x_*, \lambda_n(\gamma), \gamma) - \pi(x_*)\right) \xrightarrow{d} \mathcal{N}\left(0, \left[\pi(x_*)\left(1 - \pi(x_*)\right)\right]^2 x^T_{*,\mathcal{S}} I_{11}^{-1} x_{*,\mathcal{S}}\right) \tag{5.71}$$

□

5.4.3 Model Validation

In our method, we perform a sensitivity analysis over γ. This gives us a set of estimated probabilities for each fixed value of λ. Depending on these values in this set, the predicted class will be either unique or both '0' and '1'. Therefore, the classical measures of accuracy will not be applicable in this context. So we use the following performance measures, proposed by Corani and Zaffalon [21] for Naive Credal Classifier (NCC).

Measures of Accuracy

We use cross-validation for model selection and validation where λ is used as a tuning parameter. We consider the following performance measures [21, 62] for credal classification.

Definition 5.7 (Determinacy)**.** *Determincay is the performance measure that counts the percentage of classifications with unique output.*

Definition 5.8 (Single accuracy)**.** *Single accuracy is accuracy of the classifications when the output is determinate.*

There are two other performance measures called *indeterminate output size* and *set accuracy.* However, in the context of binary credal classification, indeterminate output size is always equal to 2 and set accuracy is always equal to 1.

The above mentioned performance measures will be used for both model selection and model validation, we first need to choose an optimal λ, i.e. a value of λ that maximises the performance of our model. For this purpose, we need to use a trade-off between determinacy and single accuracy. We use u_{65} utility on the discounted accuracy, as proposed by Zaffalon et al. [83]. Zaffalon et al. [83] suggest several other discounted utility. However, we notice that u_{65} gives us a good balance between single accuracy and determinacy unlike u_{80}, which puts less weight on single accuracy. Therefore u_{80} tends to give higher score to non-deterministic classifiers and comparison with classical methods can be misinterpreted. To avoid these issues, we use u_{65}, which we show in Table 5.1, where each row stands for predicted class and each column stands for the actual class.

	$\{0\}$	$\{1\}$
$\{0\}$	1	0
$\{1\}$	0	1
$\{0, 1\}$	0.65	0.65

Table 5.1: Discounted utility (u_{65}) table for binary credal classification

Note that, for binary credal classification, we can formulate this unified u_{65}

accuracy measure in the following way:

$$\text{Accuracy} = \text{Determinacy} \times \text{Single accuracy} + 0.65 \times (1 - \text{Determinacy}) \quad (5.72)$$

Model Selection and validation

We use nested loop cross-validation for model selection and validation. We first split the dataset $\mathcal{D}$ in 2 equal parts $\mathcal{D}_1$ and $\mathcal{D}_2$. We take $\mathcal{D}_1$ and split it in 5 equal parts. We use 4 of them to train our IAPLR model and use the remaining part for the selection of λ. We do this for each of the 5 parts to get an optimal λ based on the averaged performance measure. After obtaining the optimal λ though cross-validation, we validate our model with $\mathcal{D}_2$.

We repeat the same for $\mathcal{D}_2$, we use $\mathcal{D}_2$ to obtain an optimal λ for model selection and then validate it using $\mathcal{D}_1$. This way, we use each observation exactly twice for testing. This also gives a comparison between these two models obtained from $\mathcal{D}_1$ and $\mathcal{D}_2$ and gives us an idea of variability of the attributes.

5.4.4 Illustration

We use two different datasets for illustration, the Sonar dataset [42] and the LSVT dataset [71]. In both cases, we normalise the attributes to avoid scaling issues and split the datasets in two equal parts $\mathcal{D}_{\mathrm{S},1}$ $\mathcal{D}_{\mathrm{S},2}$ (Sonar) and $\mathcal{D}_{\mathrm{L},1}$, $\mathcal{D}_{\mathrm{L},2}$ (LSVT). We first select our model using $\mathcal{D}_{\mathrm{S},1}$ ($\mathcal{D}_{\mathrm{L},1}$). We vary our set of weights through 20 different γ's ranging from 0.01 to 1. We take a grid of 50 λ values where the bounds are taken following the suggestion by Friedman et al. [35]. We find optimal λ by 5-fold cross validation. We use this optimal λ for model selection.

We compare our results with the naive credal classifier (NCC) [82]. For this, we first categorise the attributes in 5 factors. We train our model in a grid of the concentration parameter s with 50 entries ranging from 0.04 to 2. We run a 5-fold cross-validation for the choice of optimal s and use this value of s for model selection. We also compare our result with the naive Bayes classifier (NBC) [55] and APLR [86, 2]. For APLR select the value of optimal λ through a 5-fold cross-validation. We use `glmnet` [36] for training APLR and IAPLR model. We validate our model using $\mathcal{D}_{\mathrm{S},2}$ ($\mathcal{D}_{\mathrm{L},2}$). We then select our model using $\mathcal{D}_{\mathrm{S},2}$ ($\mathcal{D}_{\mathrm{L},2}$) and validate using $\mathcal{D}_{\mathrm{S},1}$ ($\mathcal{D}_{\mathrm{L},1}$) to capture interaction between the observations.

We show a summary of our results in Table 5.2. The left-most column denotes the training set. We show determinacy in the second column. In third and fourth column, we display the single accuracy and utility based (u_{65}) accuracy, respectively and in the right-most column we display the range of active attributes.

Sonar Dataset

We use the Sonar dataset[1] for the illustration of our method. The dataset consists of 208 observations on 60 attributes in the range of 0 to 1. Sonar signals are reflected by either a metallic cylinder (M) or a roughly cylindrical rock (R), and the attributes represent the energy of the reflected signal within a particular frequency band integrated over time. We use these attributes to classify the types of the reflectors. Q-Q plot suggests that these attributes can be treated as Gaussian. Therefore, we can easily apply IAPLR for variable selection. To do so, we first scale the data so that mean of each attribute is equal to zero and standard deviation of each attribute is equal to 1. For NBC and NCC, we simply cut these attributes in five different levels to treat these as categorical variables. We perform a weighted random sampling to split the dataset in two equal parts. This ensures that ratios of M and R remain close in both parts.

In the top row of Fig. 5.3, we provide the cross validation plots with respect to λ. The shaded grey area denotes the one standard deviation from the averaged accuracy. The vertical dotted line in each plot denotes the optimal λ. For $\mathcal{D}_{\mathrm{S},1}$, the optimal λ is found to be 0.039 and for $\mathcal{D}_{\mathrm{S},2}$ the value is equal to 0.087. We also show the number of active attributes in Fig. 5.4 for these fixed optimal values of λ. We observe that for both partitions the method tends to select more attributes as the value of γ increases.

We show the summary of our illustration in Table 5.2. The left-most column denote the method followed by training dataset, determinacy, single accuracy, u_{65}

[1] This dataset is publicly available for use and can be found in UCI machine learning repository [27].

utility measure and range of active attributes. For Sonar dataset, IAPLR outperforms the rests in terms of determinacy and the u_{65} utility measure. It also has a good agreement in model validation with respect to the datasets unlike NCC or NBC, which are sensitive with respect to the training dataset. It performs an automatic variable selection like APLR. For IAPLR, we have a range of active attributes unlike APLR, which is computed using $\gamma = 1$. We observe that for $\mathcal{D}_{\mathrm{S},1}$, the sparsity of the model is more sensitive than the sparsity of the model trained by $\mathcal{D}_{\mathrm{S},2}$.

LSVT Dataset

We use the LSVT (Lee Silverman Voice Treatment) dataset[2] for the illustration with high-dimensional data. The dataset consists of 126 observations on 310 attributes. The attributes are 310 different biomedical signal processing measures which are obtained through 126 voice recording signals of 14 different persons diagnosed with Parkinson's disease. The responses denote acceptable (1) vs unacceptable (2) phonation during LSVT rehabilitation. We follow a similar data preparation method as of Sonar dataset. We perform a weighted random sampling and split the dataset in two equal halves. We also factorise the data in 5 levels for using NBC and NCC.

We show the cross validation plots in the bottom row of Fig. 5.3. For $\mathcal{D}_{\mathrm{L},1}$, the optimal λ is found to be 0.018 and for $\mathcal{D}_{\mathrm{L},2}$ the value is equal to 0.014. We show the number of active attributes in the bottom row of Fig. 5.4. We observe that for both partitions the method tends to select fewer attributes as the value of γ increases unlike our experiment with Sonar dataset.

We provide the summary of our analyses in Table 5.2 We observe that IAPLR performs much better than the other methods for LSVT dataset. It also has a good agreement in model validation with respect to the datasets unlike NCC, NBC and APLR. However, we notice that the sparsity levels are significantly different for different partitions of the dataset unlike APLR, which selects only 11 attributes for both the partitions.

[2] The dataset is openly available in the UCI machine learning repository [72].

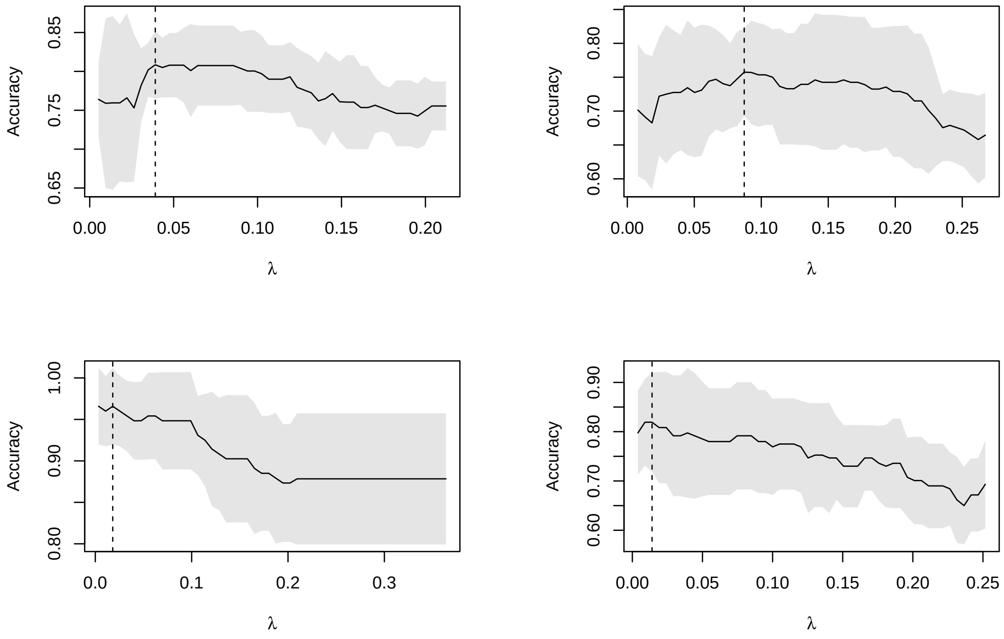

Figure 5.3: Cross-validation curve with respect to the tuning parameter λ. The top row represents the results obtained for $\mathcal{D}_{\mathrm{S},1}$ (left), $\mathcal{D}_{\mathrm{S},2}$ (right) and the bottom row represents that of $\mathcal{D}_{\mathrm{L},1}$ (left), $\mathcal{D}_{\mathrm{L},2}$ (right).

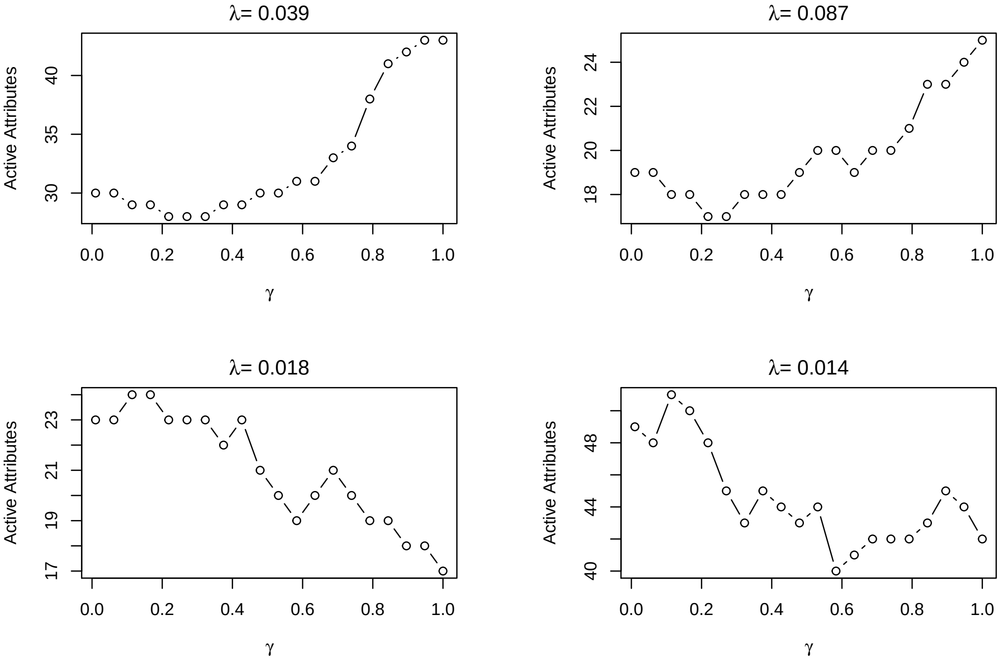

Figure 5.4: Sensitivity of sparsity with respect to γ. The top row represents the results obtained for $\mathcal{D}_{\mathrm{S},1}$ (left), $\mathcal{D}_{\mathrm{S},2}$ (right) and the bottom row represents that of $\mathcal{D}_{\mathrm{L},1}$ (left), $\mathcal{D}_{\mathrm{L},2}$ (right).

Method	Training	Deter.(%)	Single Acc.(%)	u_{65}(%)	Active
Sonar dataset; 60 active predictors					
IAPLR ($\lambda = 0.039$)	$\mathcal{D}_{\mathrm{S},1}$	87	73	72	28–43
IAPLR ($\lambda = 0.087$)	$\mathcal{D}_{\mathrm{S},2}$	87	77	75	17–25
NCC ($s = 0.02$)	$\mathcal{D}_{\mathrm{S},1}$	77	68	67	–
NCC ($s = 0.56$)	$\mathcal{D}_{\mathrm{S},2}$	49	78	72	–
NBC	$\mathcal{D}_{\mathrm{S},1}$	–	–	59	–
NBC	$\mathcal{D}_{\mathrm{S},2}$	–	–	74	–
APLR ($\lambda = 0.104$)	$\mathcal{D}_{\mathrm{S},1}$	–	–	71	12
APLR ($\lambda = 0.189$)	$\mathcal{D}_{\mathrm{S},2}$	–	–	72	9
LSVT dataset; 310 active predictors					
IAPLR ($\lambda = 0.018$)	$\mathcal{D}_{\mathrm{L},1}$	98	82	82	17–24
IAPLR ($\lambda = 0.014$)	$\mathcal{D}_{\mathrm{L},2}$	83	85	81	40–51
NCC ($s = 0.08$)	$\mathcal{D}_{\mathrm{L},1}$	14	78	67	–
NCC ($s = 0.04$)	$\mathcal{D}_{\mathrm{L},2}$	25	88	71	–
NBC	$\mathcal{D}_{\mathrm{L},1}$	–	–	51	–
NBC	$\mathcal{D}_{\mathrm{L},2}$	–	–	40	–
APLR ($\lambda = 0.052$)	$\mathcal{D}_{\mathrm{L},1}$	–	–	81	11
APLR ($\lambda = 0.285$)	$\mathcal{D}_{\mathrm{L},2}$	–	–	76	11

Table 5.2: Summary of model selection and validation

Chapter 6

Bayesian Inference

In Chapter 3, we learnt how likelihood-based approaches can be used in the inference for high dimensional statistical modelling. In likelihood based approaches, we rely on the data and underlying distributional assumptions to perform statistical analysis. However, in high dimensional modelling, the problems are inherently sparse and come with limited information. Therefore, we need to be cautious while performing inference and should consider prior information. The Bayesian paradigm allows us to incorporate this prior belief in the model by exploiting Bayes's rule.

In this chapter, we discuss this Bayesian approach in the regressional setting. We introduce the basic notions of Bayesian inference in Section 6.1. In Section 6.2, we discuss basic Bayesian regression models using different choices of priors. After that, we investigate Bayesian variable selection in Section 6.3, where we discuss the method proposed by George and McCulloch [41]. In Section 6.4, we discuss the Bayesian alternative of LASSO as proposed by Park and Casella [60] and finally in Section 6.5 we explore different Spike and Slab priors for variable selection.

6.1 Foundation

Here, in this section, we provide the fundamental concepts of Bayesian inference. We briefly discuss different notions of Bayesian inference. Further discussion on Bayesian inference can be found in the books authored by Berger [9], Gelman et al. [39], Casella and Berger [15] etc. As we discussed earlier in Section 2.2, the Bayesian approach is based on Bayes's rule [7] given by

$$P(\beta \mid y) \propto P(y \mid \beta)P(\beta) \tag{6.1}$$

where $P(\beta \mid y)$ is our posterior or the probability density of the unknown parameter β conditional on the observed data y. This is directly proportional (up to a constant that may depend on y but does not depend on β) to the product of our prior belief $P(\beta)$ on β and our likelihood function $P(y \mid \beta)$.

6.1.1 Prior

In Bayesian statistics, the choice of prior plays an important role in inference. A prior can be considered as the statistician's belief or knowledge on the modelling parameter. Therefore, the choice of prior can often be subjective and we can not find a best choice. However, we would like to consider a prior which agrees well with the parameter support as well as helps us to incorporate our prior information about the problem. In general, we may categorise these priors in two major ways: subjective priors and objective priors. However, the classification of priors is controversial and many researchers prefer different ways of categorising them.

Subjective Priors

Subjective priors are usually used to incorporate one's subjective belief about the modelling parameter. Subjective priors are often elicitation-based and allow us to gather information from previous analysis. There are several ways of choosing a subjective prior. Garthwaite et al. [37] provided a detailed discussion on elicitation-based approach for choosing a prior. Berger [9] suggested that use of histogram based approaches or empirical cumulative distribution function to construct a subjective prior for a continuous random variable. We can also use point estimates to construct a subjective prior from a conjugate class of priors.

Prior Predictive

Before the data y is observed, we can look into the distribution of this unknown but observable data y, which is given by:

$$P(y) := \int_{\beta} P(y \mid \beta)P(\beta)d\beta \tag{6.2}$$

where $P(y \mid \beta)$ refers to our sampling distribution of some observable quantity y and $P(\beta)$ refers to our prior on the parameter β. We call this distribution $P(y)$ the prior predictive distribution. This is useful to understand, if our choice of prior is consistent to the observable data.

Objective Priors

Objective prior is an alternative method for describing a prior where we usually use objective source of information about the modelling parameter such as parameter support or sign of the modelling parameter. We often consider these priors as non-informative priors as they do not posses any other descriptive information. However, we may argue that our knowledge of parameter support is also relevant information and therefore some researchers coin these type of priors as weakly-informative priors. We usually consider flat priors for this kind of analysis. One of such priors is the uniform distribution on the parameter support which assigns equal density to each point within the parameter support.

Improper Priors

Improper priors can also be classified as objective priors. However, improper priors may not integrate to 1. To give some intuition, we can consider an unbounded parameter, then a uniform distribution will result to an improper prior. In general, improper priors are chosen so that the posterior density function remains proper. However, improper priors are particularly useful for conditional analysis, which we will show while discussing Bayesian regression.

Conjugate Priors

In Bayesian inference, if the posterior and prior belong to the same family of probability distributions then the prior and posterior are called conjugate distributions with respect to the likelihood and the prior belongs to a class of conjugate priors. For regression analysis we usually work with Gaussian assumption on the noise and so for the likelihood. Therefore, we consider exponential family distributions for choice of priors.

Definition 6.1 (Exponential Family). *Let $\beta \coloneqq (\beta_1, \cdots, \beta_r)$ be a vector of parameters. Then the exponential family of distributions is defined by:*

$$f(y \mid \beta) = h(y) \exp\left(\sum_{i=1}^{r} a_i(\beta) T_i(y) - b(\beta)\right) \tag{6.3}$$

where h, a, T and b are fixed functions for each probability distribution.

For instance, in case of a normal distribution, the probability density function is given by:

$$\begin{aligned} f(y \mid \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) && (6.4)\\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2 - 2y\mu + \mu^2}{2\sigma^2} - \ln\sigma\right) && (6.5)\\ &= \frac{1}{\sqrt{2\pi}} \exp\left(\left(-\frac{1}{2\sigma^2}, \frac{\mu}{\sigma^2}\right) \cdot (y^2, y)^T - \frac{\mu^2}{2\sigma^2} - \ln\sigma\right). && (6.6) \end{aligned}$$

Therefore for normal distribution, $h(y) \coloneqq \frac{1}{\sqrt{2\pi}}$, $a(\mu, \sigma^2) \coloneqq \left(-\frac{1}{2\sigma^2}, \frac{\mu}{\sigma^2}\right)$, $T(y) \coloneqq (y^2, y)$ and $b(\mu, \sigma^2) \coloneqq (\frac{\mu^2}{2\sigma^2} + \ln\sigma)$.

6.1.2 Estimation

In the Bayesian paradigm, we rely on the posterior distribution for the parameter estimation. From the posterior distribution, we can learn about the parameter in three different ways. The most common and convenient way to learn from the posterior distribution is to check the posterior mean given by:

$$E(\beta \mid y) = \int \beta P(\beta \mid y) d\beta. \tag{6.7}$$

However, this only works when there exist a finite mean. Besides posterior mean, we sometimes look for the maximum a posteriori (MAP) estimates. That is we look for the value that achieves greatest posterior density. We look for MAP in the following way:

$$\beta_{\text{MAP}} := \arg\max_{\beta} P(\beta \mid y). \tag{6.8}$$

In some cases, we also check for the posterior median as a robust estimate especially if we suspect that the data contain some outliers.

Posterior Predictive

The posterior predictive is the distribution of a future data point, conditional on the data already observed. That is, let y^* be new observed variables then we can define posterior predictive in the following way:

$$P(y^* \mid y) := \int_{\beta} P(y^* \mid \beta) P(\beta \mid y) d\beta. \tag{6.9}$$

6.2 Bayesian Regression

As we discussed earlier, the choice of prior plays an important role in Bayesian inference. This is applicable for Bayesian regression as well. We usually perform Bayesian regression in two different ways, based on the choice of priors.

6.2.1 Notation of the Model

We construct the likelihood from the normality assumption of the noise. We write this likelihood in the following way:

$$y \mid \mu, \beta, \sigma^2, \boldsymbol{x} \sim \mathcal{N}(\mu + \boldsymbol{x}\beta, \sigma^2 \mathbf{I}_n). \tag{6.10}$$

Here μ denotes the intercept term of the regression model and $\boldsymbol{x}$ denotes the matrix of predictors. We assume μ to be known, and scale the data set accordingly, so without loss of generality, μ can be assumed zero. We also consider $\boldsymbol{x}$ to be non-random and therefore we can drop the conditional on $\boldsymbol{x}$. That is, we can simply write this as

$$y \mid \beta, \sigma^2 \sim \mathcal{N}(\boldsymbol{x}\beta, \sigma^2 \mathbf{I}_n). \tag{6.11}$$

6.2.2 Improper Prior

For the priors on the β and σ^2, we can consider non-informative priors. Gelman et al. [39] discussed the use of improper priors to specify model parameters. They use a uniform prior on the joint density of $(\beta, \log \sigma)$, which gives us the following improper prior:

$$P(\beta, \sigma^2) \propto \frac{1}{\sigma^2}. \tag{6.12}$$

This setting allows us to learn completely from the data points with the joint posterior given by:

$$\begin{align} P(\beta, \sigma^2 \mid y) &\propto P(y \mid \beta, \sigma^2) P(\beta, \sigma^2) \tag{6.13} \\ &\propto \frac{1}{\sigma^{n+2}} \exp\left(-\frac{1}{2\sigma^2} \|y - \boldsymbol{x}\beta\|_2^2\right). \tag{6.14} \end{align}$$

We can exploit the conjugacy property of this prior to get the following posterior distribution of β conditional σ^2 (see Gelman et al. [39])

$$\beta \mid \sigma^2, y \sim \mathcal{N}\left(\hat{\beta}_{\mathrm{OLS}}, (\boldsymbol{x}^T\boldsymbol{x})^{-1}\sigma^2\right) \tag{6.15}$$

where $\hat{\beta}_{\mathrm{OLS}} := (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T y$ are the ordinary least squares estimates. Therefore the posterior expectations of the regression coefficients are equal to the ordinary least squares estimates.

Now, for ordinary least squares, we know that,

$$\|y - \boldsymbol{x}\beta\|_2^2 = (y - \boldsymbol{x}\beta)^T (y - \boldsymbol{x}\beta) = \|y - \boldsymbol{x}\hat{\beta}_{\mathrm{OLS}}\|_2^2 + \|\boldsymbol{x}\hat{\beta}_{\mathrm{OLS}} - \boldsymbol{x}\beta\|_2^2. \tag{6.16}$$

Then, from Eq. (6.14), then we can write following:

$$P(\beta, \sigma^2 \mid y) \propto \frac{1}{\sigma^{n-p+2}} \cdot \frac{1}{\sigma^p} \exp\left(-\frac{\|y - \boldsymbol{x}\hat{\beta}_{\mathrm{OLS}}\|_2^2}{2\sigma^2}\right) \exp\left(-\frac{\|\boldsymbol{x}\hat{\beta}_{\mathrm{OLS}} - \boldsymbol{x}\beta\|_2^2}{2\sigma^2}\right) \tag{6.17}$$

Now, for some suitable integration constant K, we get the posterior probability of σ^2 such that,

$$P(\sigma^2 \mid y) = K \cdot \frac{1}{\sigma^{n-p+2}} \exp\left(-\frac{\|y - \boldsymbol{x}\hat{\beta}_{\mathrm{OLS}}\|_2^2}{2\sigma^2}\right) \tag{6.18}$$

Therefore, σ^2 follows an inverse gamma distribution such that,

$$\sigma^2 \mid y \sim \text{InvGamma}\left(\frac{n-p}{2}, \frac{\|y - \boldsymbol{x}\hat{\beta}_{\text{OLS}}\|_2^2}{2}\right). \tag{6.19}$$

Then, the posterior expectation of σ^2 is given by:

$$E(\sigma^2 \mid y) = \frac{\|y - \boldsymbol{x}\hat{\beta}_{\text{OLS}}\|_2^2}{n-p-2}. \tag{6.20}$$

6.2.3 Informative Priors

Another possible approach for Bayesian linear regression is to use a normal prior to specify β and inverse gamma prior for σ^2. We can therefore, use a normal distribution with large variance to specify β such that,

$$P(\beta \mid \sigma_\beta^2) \sim \mathcal{N}(\mu_\beta, \sigma_\beta^2 \mathbf{I}_p). \tag{6.21}$$

We usually choose a large value for σ_β^2, which acts as a regularisation weight on the regression coefficients. However, we are certain about our prior information then we may consider smaller values. We can also use the same variance parameter σ^2 as of Eq. (6.11) for easier interpretation. Therefore another way to define priors for both β and σ^2 is given by:

$$P(\beta \mid \sigma^2) \sim \mathcal{N}(\mu_\beta, \sigma^2 \mathbf{I}_p) \tag{6.22}$$

$$P(\sigma^2) \sim \text{InvGamma}(a, b) \tag{6.23}$$

where $a, b > 0$ are fixed constants. Therefore, the joint posterior of β and σ^2 is given by:

$$\begin{aligned} &P(\beta, \sigma^2 \mid y) \\ &\propto P(y \mid \beta, \sigma^2) P(\beta \mid \sigma^2) P(\sigma^2) && (6.24) \\ &\propto \frac{1}{\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\|y - \boldsymbol{x}\beta\|_2^2\right) \frac{1}{\sigma^p} \exp\left(-\frac{\|\mu_\beta - \beta\|_2^2}{2\sigma^2}\right) \frac{1}{\sigma^{2(a+1)}} \exp\left(-\frac{b}{\sigma^2}\right). && (6.25) \end{aligned}$$

Now, when $\boldsymbol{x}^T\boldsymbol{x}$ is invertible, we can apply the identity from Eq. (6.16) and write the joint posterior of β conditional on σ^2 as

$$P(\beta \mid \sigma^2, y) \overset{\beta}{\propto} \exp\left(-\frac{(\hat{\beta}_{\text{OLS}} - \beta)^T (\boldsymbol{x}^T\boldsymbol{x})(\hat{\beta}_{\text{OLS}} - \beta)}{2\sigma^2}\right) \exp\left(-\frac{\|\mu_\beta - \beta\|_2^2}{2\sigma^2}\right) \tag{6.26}$$

Now,

$$\frac{(\hat{\beta}_{\text{OLS}}-\beta)^T(\boldsymbol{x}^T\boldsymbol{x})(\hat{\beta}_{\text{OLS}}-\beta)}{2\sigma^2}+\frac{\|\mu_\beta-\beta\|_2^2}{2\sigma^2}$$

$$=\frac{\beta^T\boldsymbol{x}^T\boldsymbol{x}\beta-2\beta^T\boldsymbol{x}^T\boldsymbol{x}\hat{\beta}_{\text{OLS}}-2\beta^T\mu_\beta+\beta^T\beta+R_1}{2\sigma^2} \tag{6.27}$$

where R_1 denote additional terms that are independent of β. Then,

$$=\frac{\beta^T(\boldsymbol{x}^T\boldsymbol{x}+\mathbf{I}_p)\beta-2\beta^T(\boldsymbol{x}^T\boldsymbol{x}\hat{\beta}_{\text{OLS}}+\mu_\beta)+R_1}{2\sigma^2} \tag{6.28}$$

$$=\frac{\beta^T(\boldsymbol{x}^T\boldsymbol{x}+\mathbf{I}_p)\beta-2\beta^T(\boldsymbol{x}^T\boldsymbol{x}+\mathbf{I}_p)(\boldsymbol{x}^T\boldsymbol{x})(\boldsymbol{x}^T\boldsymbol{x}+\mathbf{I}_p)^{-1}\left(\hat{\beta}_{\text{OLS}}+(\boldsymbol{x}^T\boldsymbol{x})^{-1}\mu_\beta\right)+R_1}{2\sigma^2}. \tag{6.29}$$

Now, it can be shown that,

$$\boldsymbol{x}^T\boldsymbol{x}(\boldsymbol{x}^T\boldsymbol{x}+\mathbf{I}_p)^{-1}=\left(\mathbf{I}_p+(\boldsymbol{x}^T\boldsymbol{x})^{-1}\right) \tag{6.30}$$

Then from Eq. (6.29) we have

$$\frac{(\hat{\beta}_{\text{OLS}}-\beta)^T(\boldsymbol{x}^T\boldsymbol{x})(\hat{\beta}_{\text{OLS}}-\beta)}{2\sigma^2}+\frac{\|\mu_\beta-\beta\|_2^2}{2\sigma^2}$$

$$=\frac{\beta^T(\boldsymbol{x}^T\boldsymbol{x}+\mathbf{I}_p)\beta-2\beta^T(\boldsymbol{x}^T\boldsymbol{x}+\mathbf{I}_p)\left(\mathbf{I}_p+(\boldsymbol{x}^T\boldsymbol{x})^{-1}\right)\left(\hat{\beta}_{\text{OLS}}+(\boldsymbol{x}^T\boldsymbol{x})^{-1}\mu_\beta\right)+R_1}{2\sigma^2} \tag{6.31}$$

$$=\frac{(\beta-\beta')^T(\boldsymbol{x}^T\boldsymbol{x}+\mathbf{I}_p)(\beta-\beta')+R_2}{2\sigma^2} \tag{6.32}$$

where $\beta' := \left(\mathbf{I}_p+(\boldsymbol{x}^T\boldsymbol{x})^{-1}\right)\left(\hat{\beta}_{\text{OLS}}+(\boldsymbol{x}^T\boldsymbol{x})^{-1}\mu_\beta\right)$ and R_2 denote additional terms independent of β. Therefore, from Eq. (6.26) and Eq. (6.32), we get the following posterior of β conditional on σ^2:

$$\beta \mid \sigma^2, y \sim \mathcal{N}\left(\left(\mathbf{I}_p+(\boldsymbol{x}^T\boldsymbol{x})^{-1}\right)^{-1}(\hat{\beta}_{\text{OLS}}+(\boldsymbol{x}^T\boldsymbol{x})^{-1}\mu_\beta), \sigma^2\left(\mathbf{I}_p+\boldsymbol{x}^T\boldsymbol{x}\right)^{-1}\right). \tag{6.33}$$

Then the posterior expectation of β conditional on σ^2 is given by:

$$E(\beta \mid \sigma^2, y)=\left(\mathbf{I}_p+(\boldsymbol{x}^T\boldsymbol{x})^{-1}\right)^{-1}(\hat{\beta}_{\text{OLS}}+(\boldsymbol{x}^T\boldsymbol{x})^{-1}\mu_\beta). \tag{6.34}$$

Now, if we set our prior information on β around zero then we can write Eq. (6.34) as

$$E(\beta \mid \sigma^2, y)=\left(\mathbf{I}_p+(\boldsymbol{x}^T\boldsymbol{x})^{-1}\right)^{-1}\hat{\beta}_{\text{OLS}} \tag{6.35}$$

$$=\left(\mathbf{I}_p+(\boldsymbol{x}^T\boldsymbol{x})^{-1}\right)^{-1}(\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^Ty \tag{6.36}$$

$$=\left(\boldsymbol{x}^T\boldsymbol{x}+\mathbf{I}_p\right)^{-1}\boldsymbol{x}^Ty. \tag{6.37}$$

Eq. (6.37) corresponds to the ridge estimates in Eq. (3.15) for fixed $\lambda = 1$. Alternatively, we can say that using a normal prior on β gives us Bayesian alternative for ridge regression.

6.3 Bayesian Variable Selection

One of the major issues in high dimensional statistical modelling is to achieve variable selection in a model. This led to several likelihood-based approaches, which we discussed in Chapter 3. In this section, we discuss the Bayesian alternatives for variable selection. One of the earlier methods for Bayesian variable selection was proposed by Mitchell and Beauchamp [56]. They proposed a hierarchical model for Bayesian variable selection. Later, George and McCulloch [41] proposed the use of latent variables to attain sparsity. They suggested the use of the Gibbs sampling algorithm [16, 38] to obtain the posterior.

6.3.1 Gibbs Sampling Algorithm

Gibbs sampling is an iterative Markov Chain Monte Carlo (MCMC) algorithm for sampling from the posterior. Initially, the algorithm was proposed by Geman and Geman [40] as a special case of the Metropolis-Hastings algorithm [47]. Later Gelfand and Smith [38] proposed a generalised framework to sample from multivariate probability distributions.

Let $\theta \coloneqq (\theta_1, \theta_2, \cdots, \theta_r)$ be r modelling parameters. Then the Gibbs sampling algorithm can be performed using the following algorithm:

Algorithm

- Initial guess: $\theta^{(0)}$

- Updating:

$$\begin{aligned}
&\text{draw } \theta_1^{(k+1)} \text{ from } p(\theta_1 \mid \theta_2^{(k)}, \theta_3^{(k)}, \cdots, \theta_r^{(k)}) \\
&\text{draw } \theta_2^{(k+1)} \text{ from } p(\theta_2 \mid \theta_1^{(k+1)}, \theta_3^{(k)}, \cdots, \theta_r^{(k)}) \\
&\qquad\qquad \vdots \\
&\text{draw } \theta_r^{(k+1)} \text{ from } p(\theta_r \mid \theta_1^{(k+1)}, \theta_2^{(k+1)}, \cdots, \theta_{r-1}^{(k+1)})
\end{aligned}$$

where $p(\theta_i \mid \theta_1, \cdots, \theta_{i-1}, \theta_{i+1}, \cdots, \theta_r)$ denotes the known conditional distribution of θ_i. The simple framework also allows us to perform block Gibbs sampling, where we can sample from a multivariate conditional distribution by exploiting conditional independence.

6.3.2 Variable Selection Via Gibbs Sampling

Variable selection via Gibbs sampling was first suggested by George and McCulloch [41]. They suggested the use of latent variables to specify active and inactive co-variates. The Gibbs sampling framework avoids the computationally expensive search of the whole model space of dimension 2^p. Their suggested hierarchical model is given by:

$$y \mid \boldsymbol{x}, \beta, \sigma^2 \sim \mathcal{N}\left(\boldsymbol{x}\beta, \sigma^2 \mathbf{I}_n\right) \tag{6.38}$$

$$\beta_j \mid z_j = 0, \sigma^2_{\beta_j} \sim \mathcal{N}(0, \sigma^2_{\beta_j}) \tag{6.39}$$

$$\beta_j \mid z_j = 1, \sigma^2_{\beta_j} \sim \mathcal{N}(0, c_j^2 \sigma^2_{\beta_j}) \tag{6.40}$$

$$z_j \sim \text{Ber}(q_j) \tag{6.41}$$

$$\sigma^2 \sim \text{InvGamma}(a, b). \tag{6.42}$$

George and McCulloch [41] suggested sufficiently small values for $\sigma^2_{\beta_j}$'s such that β_j can be safely replaced with zero and choice of $c_j > 1$ should be sufficiently large so that a true active co-variate is included in the model. They also discussed the possibility of choosing the value of c_j based on the intersection points of the densities $\mathcal{N}(0, c_j\sigma^2_{\beta_j})$ and $\mathcal{N}(0, \sigma^2_{\beta_j})$. This allows us to interpret c_j's as prior odds.

The choice of q_j is based on expert opinion. A special case is $q_j = 1/2$, this corresponds to the indifference prior or uniform prior on the selection probability.

The model selection is performed by inspecting the posterior of z. The authors suggested that inspection of individual sub model can be considered to perform variable selection, that is if $P(z_j \mid y) > 0.5$ then we can conclude that $\boldsymbol{x}_j$ is included in the model.

6.4 Bayesian LASSO

The Bayesian LASSO provides a natural way to quantify the model uncertainty in a LASSO-fitted model. To motivate this approach, recall firstly that, under the assumption $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$, we can write the likelihood of a linear model $y = \boldsymbol{x}\beta + \epsilon$ in the following way,

$$\begin{aligned} p(y \mid \boldsymbol{x}, \beta) &\propto e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\epsilon_i^2} \\ &\propto e^{-\frac{1}{2\sigma^2}\|y-\boldsymbol{x}\beta\|_2^2}. \end{aligned} \tag{6.43}$$

Tibshirani [69] suggested the use of a Laplace prior

$$p(\beta) \propto e^{-\lambda\|\beta\|_1} \tag{6.44}$$

for the model parameters, yielding the following posterior,

$$\begin{aligned} p(\beta \mid \boldsymbol{x}, y) &\propto p(y \mid X, \beta)p(\beta) \\ &\propto e^{-\left(\frac{1}{2\sigma^2}\|y-\boldsymbol{x}\beta\|_2^2+\lambda\|\beta\|_1\right)} \end{aligned} \tag{6.45}$$

It is a well-established result that the mode of Eq. (6.45), that is the posterior mode of β under Laplace priors, corresponds just to the frequentist LASSO estimate [54, 60, 69]. Draws from this posterior are not necessarily sparse, but still can be used to assess uncertainty of model parameters through checking the sample variances of the modelling parameters[44].

The Bayesian LASSO has been implemented in several different ways, which differ essentially in how sparsity is induced, and how the regularisation parameter is handled.

Let, $\tau = (\tau_1, \cdots, \tau_p)$. Then for $j = 1, \cdots, p$, Park and Casella [60] proposed the following hierarchical mixture model for parameter estimation:

$$y \mid \mu, \beta, \sigma^2 \sim \mathcal{N}(\mu + \boldsymbol{x}\beta, \sigma^2 \mathbf{I}_n), \tag{6.46}$$

$$\beta \mid \sigma^2, \tau \sim \mathcal{N}(\mathbf{0}_p, \sigma^2 \mathbf{D}_\tau) \tag{6.47}$$

$$\tau_j^2 \sim \frac{\lambda^2}{2} e^{-\lambda^2 \tau_j^2/2} \tag{6.48}$$

$$\sigma^2 \sim \pi_{\sigma^2}(\sigma^2). \tag{6.49}$$

where $\mathbf{D}_\tau = \operatorname{diag}\left(\tau_1^2, \cdots, \tau_p^2\right)$ and $\pi_{\sigma^2}(\sigma^2)$ denotes the improper prior. In this formulation τ_j^2 acts as a scale for the regression coefficients and after marginalising these regression coefficients over all τ_j^2 we get the conditional prior on β of the following form

$$\pi(\beta \mid \sigma^2) = \prod_{j=1}^{p} \frac{\lambda}{2\sigma} e^{-\lambda|\beta_j|\sigma}. \tag{6.50}$$

Choice of λ

Unlike the variable selection via Gibbs sampling, we don't have any natural way of co-variate selection fo Bayesian LASSO. Therefore choice of λ plays an important role in this context as λ forces β to give sparse posterior medians. For this, Park and Casella suggested two different techniques.

Firstly, they suggested the possibility of using marginal maximum likelihood estimates for the choice of λ. They considered a Monte Carlo EM algorithm [52], which in iteration k, updates the parameter λ using the iterative scheme

$$\lambda_k = \sqrt{\frac{2p}{\sum_{j=1}^{p} E_{\lambda_{k-1}}[\tau_j^2|y]}}, \tag{6.51}$$

where y is assumed to be centred, and the conditional expectation is estimated via averages of Gibbs samples. For $p < n$, the initial value λ_0 was suggested to be

$$\lambda_0 = \frac{p\sqrt{\hat{\sigma}^2_{\text{OLS}}}}{\sum_{j=1}^{p} |\hat{\beta}_{\text{OLS}_j}|}, \tag{6.52}$$

where $\hat{\sigma}^2_{\text{OLS}}$ and $\hat{\beta}_{\text{OLS}}$ are ordinary least squares estimates.

In another approach, they discussed the possibility of using gamma priors on λ^2:

$$\pi(\lambda^2) = \frac{\delta^r}{z(r)} (\lambda^2)^{r-1} e^{-\delta\lambda^2}; \quad \lambda^2 > 0\, (r > 0, \delta > 0), \tag{6.53}$$

where r is the shape parameter and δ the rate parameter. Lykou and Ntzoufras [54] used gamma priors for λ, and developed a concept for specification of the hyperparameters based on Bayes factors, which evaluate the evidence for inclusion of the respective predictor variables.

6.5 Spike and Slab Priors

Spike and slab priors belong to a family of distributions which are widely used in Bayesian variable selection methods. As the name suggests, these types of prior consists of a spike component and a slab component. Ishwaran and Rao [48] proposed the following compact form to describe spike and slab models.

$$y \mid \beta, \sigma^2 \sim \mathcal{N}(\boldsymbol{x}\beta, \sigma^2 \mathbf{I}_n), \tag{6.54}$$

$$\beta \mid \sigma^2, \tau \sim \mathcal{N}(\mathbf{0}_p, \mathbf{D}_\tau) \tag{6.55}$$

$$\sigma^2 \sim \pi_{\sigma^2}(\sigma^2) \tag{6.56}$$

$$\tau_j^2 \sim \pi_{\tau_j^2}(\tau_j^2). \tag{6.57}$$

where $\mathbf{D}_\tau = \operatorname{diag}\left(\tau_1^2, \cdots, \tau_p^2\right)$ works as a scale for the regression coefficients similar to what we see for the Bayesian LASSO in Section 6.4. Both π_{σ^2} and $\pi_{\tau_j^2}$ are chosen to ensure that these excludes values of zero with probability 1. Ishwaran and Rao [48] classified these types of priors in two broad categories, one with two component indifference priors and the other with continuous bimodal priors.

Two component Indifference priors

A popular example of two component indifference prior is the hierarchical model proposed by George and McCulloch [41], which we discuss in Section 6.3. The prior specification using the 0, 1 latent variables can be easily translated into formal spike and slab specification so that,

$$\tau_j^2 \mid c_j, \sigma_{\beta_j}^2, z_j \sim (1 - z_j)\delta_{\sigma_{\beta_j}^2}(\cdot) + z_j \delta_{c_j \sigma_{\beta_j}^2}(\cdot) \tag{6.58}$$

$$z_j \mid q_j \sim (1 - q_j)\delta_0(\cdot) + q_j \delta_1(\cdot). \tag{6.59}$$

Here, $\delta_\eta(\cdot)$ denotes the discrete mass concentrated at η. Usually, we set the value of $q_j = 1/2$ which is referred as indifference towards the selection of a co-variate.

Continuous bimodal priors

The choice of c_j, $\sigma^2_{\beta_j}$ and q_j can be difficult and therefore improperly chosen values may perform poorly in variable selection. To overcome this issue, Ishwaran and Rao [48] proposed a continuous model based on their previous unpublished work on variable selection. They suggested the following hierarchical model:

$$\beta_j \mid z_j, \sigma^2_{\beta_j} \sim \mathcal{N}(0, z_j \sigma^2_{\beta_j}) \tag{6.60}$$

$$z_j \mid q, \eta_0 \sim (1-q)\delta_{\eta_0}(\cdot) + q\delta_1(\cdot) \tag{6.61}$$

$$\sigma^{-2}_{\beta_j} \mid a_{\beta_j}, b_{\beta_j} \sim \text{Gamma}(a_{\beta_j}, b_{\beta_j}) \tag{6.62}$$

$$q \sim \text{Uniform}[0,1]. \tag{6.63}$$

Alternatively we can write $\tau^2_j = z_j \sigma^2_{\beta_j}$ and therefore, integrating over q gives us the regular spike and slab model specification. The uniform prior on q ensures the continuity of the model and specifies how likely a co-variate β to be selected.

Spike and Slab LASSO

Several other hierarchical models have been proposed based on the spike and slab specification. Ročková and George [64] suggested the use of Laplace priors on the regression coefficients β and coined the term spike and slab LASSO as the log posterior resembles weighted LASSO. They suggested the following specification for the prior on β

$$P(\beta_j \mid z_j) = z_j \psi_1(\beta_j) + (1 - z_j)\psi_0(\beta_j), \tag{6.64}$$

where $\psi_1(\beta_j) = \frac{\lambda_1}{2} e^{-\lambda_1 |\beta|}$ with small λ_1 to express the slab component and $\psi_0(\beta_j) = \frac{\lambda_0}{2} e^{-\lambda_0 |\beta|}$ with large λ_0 to specify the spike component.

Chapter 7

Robust Bayesian Analysis

In the previous chapters, we saw how likelihood based approaches can be used for regularisation in high dimensional problems. We investigated the use of sensitivity analysis to understand the variability of the LASSO estimates. However, these likelihood based approaches do not allow us to incorporate our prior belief in the model for which we need to do a Bayesian analysis. In Chapter 6, we discussed Bayesian analysis from a regressional point of view and showed how the choice of priors can play an important role in the analysis. In this chapter we explore a robust Bayesian framework to perform Bayesian analysis to obtain an efficient model under limited information.

Section 7.1 is focused on why a robust Bayesian analysis is important and how this is efficient in processing prior information. In Section 7.2 we discuss the Imprecise Beta Model (or IBM) and its usage in robust Bayesian analysis. We use the framework of IBM to specify Bernoulli distributed random variables. In Section 7.3, we investigate different sources of uncertainty in high dimensional models and how these uncertainties can be treated using robust Bayesian analysis and propose a stepping stone for our novel hierarchical model for robust Bayesian variable selection.

7.1 Motivation for Robust Bayesian Analysis

In high dimensional models, we require an efficient parameter estimation routine to determine the regression coefficients as well as select co-variates to attain sparsity. LASSO or other likelihood based approaches rely on the optimisation methods to achieve sparsity and we do not have any straightforward expression to explain the level of sparsity. This motivates us to perform a Bayesian analysis based on spike and slab prior specification to understand our modelling parameters β as well as the level of sparsity in the model. However, choosing a suitable prior for β is particularly difficult for high dimensional models for variety of reasons. Firstly, high dimensional models come with very limited information as the number of predictors are much more than the observation. Therefore, it is hard to extract information to specify our priors for the modelling parameters and usually dealt with the assumption that number of true active covariates are less than the total number of observations.

Another issue, that occurs in high dimensional problems, is choosing a prior to specify the selection of a predictor. The choice of selection indicator plays an important role in understanding the level of sparsity in the model. However, the prior specification of the selection indicators has not been explored much in the literature, the previous works on the spike and slab priors mostly relied on the use of a uniform prior to specify the selection probability. This can be problematic as the model allows to learn from the data only and doesn't incorporate any expert opinion on the inclusion of the predictors. Another conventional approach for specifying this prior probability is to fix a beta distribution with prior expectation 1/2. This is also considered as an indifference prior among the researchers. It has been argued that setting prior with prior expectation 1/2 is useful to show our lack of evidence. However, this can increase the chance of selecting a non-important predictor.

Moreover, one common issue in statistical modelling is to incorporate expert opinions. The Bayesian paradigm allows us to incorporate expert opinion through suitable prior specification and we would like to exploit this in all possible ways. However, as we discussed earlier, extracting information for these kind of problems is very hard and therefore expert opinions may vary based on their analyses. This motivates us to perform a robust Bayesian analysis for high dimensional models. In robust Bayesian analysis, we specify a set of priors instead of a single prior. This modification allows us to incorporate all these expert opinions in a more convenient manner. In this case, we get a set of posteriors instead of a single posterior.

7.2 Imprecise Beta Model

The imprecise beta model is a robust Bayesian approach to analyse binomial data. This is a special case of Imprecise Dirichlet model for multinomial data [77]. We formulate the imprecise beta model using an alternative parametrisation using mean (α) and concentration (s). Let q be the probability of a binomial distribution then we define the imprecise beta distribution in the following way:

$$f(q;\alpha,s)=\frac{1}{B(s\alpha,s-s\alpha)}q^{s\alpha-1}(1-q)^{s(1-\alpha)-1} \tag{7.1}$$

where $\alpha\in[\underline{\alpha},\overline{\alpha}]$ and $s>0$ is fixed constant. The choice of s can also be imprecise and we may use an interval instead. For a fixed value of the concentration parameter s, we can compute the prior lower and upper expectation in the following way:

$$\underline{E}(q\mid\alpha,s):=\inf_{\alpha\in[\underline{\alpha},\overline{\alpha}]}\alpha=\underline{\alpha} \tag{7.2}$$

$$\overline{E}(q\mid\alpha,s):=\sup_{\alpha\in[\underline{\alpha},\overline{\alpha}]}\alpha=\overline{\alpha}. \tag{7.3}$$

Similarly, we can get a set of variances such that

$$\operatorname{Var}(q\mid\alpha,s):=\left\{\frac{\alpha(1-\alpha)}{s+1}:\alpha\in[\underline{\alpha},\overline{\alpha}]\right\}. \tag{7.4}$$

Therefore, if $[\underline{\alpha},\overline{\alpha}]$ contains $1/2$ then we have the maximum variance given by:

$$\overline{\operatorname{Var}}(q\mid\alpha,s)=\frac{1}{4(s+1)} \tag{7.5}$$

and minimum variance occurs in one of the bounds of α. Larger values of s lead to smaller variances and vice-versa.

This particular property of imprecise beta distribution allows us to represent the imprecision in terms of variance as well. Based on the range of the variance, we can specify a set for the concentration parameter s. A further discussion on the properties of imprecise beta model can be found in [78].

Now, since the Bernoulli distribution can be interpreted as binomial distribution, therefore, we can apply this imprecise beta distribution to specify the selection indicators in Bayesian variable selection. Let z be a Bernoulli distributed variable so that

$$P(z \mid q) = q^z(1-q)^{1-z}. \tag{7.6}$$

Then for $q \sim \text{Beta}(s\alpha, s-s\alpha)$, we have the following

$$\begin{aligned} P(q \mid z, \alpha, s) &\propto q^z(1-q)^{1-z}q^{s\alpha-1}(1-q)^{s(1-\alpha)-1} && (7.7)\\ &\propto q^{z+s\alpha-1}(1-q)^{1-z+s(1-\alpha)-1}. && (7.8) \end{aligned}$$

That is, $q \mid z, \alpha, s$ follows a beta distribution such that

$$q \mid z, \alpha, s \sim \text{Beta}\left(z+s\alpha, 1-z+s(1-\alpha)\right). \tag{7.9}$$

7.3 Uncertainty Treatment in Variable Selection

In Section 7.1, we discussed why want to perform a robust Bayesian analysis to tackle different issues. In this section we provide a basic framework, which we will use to perform a robust Bayesian variable selection. We adapt the framework of Narisetty and He [57] to propose our spike and slab model given by:

$$\begin{aligned} \beta_j \mid z_j = 1, \sigma^2 &\sim \mathcal{N}(0, \sigma^2\tau_1^2) && (7.10)\\ \beta_j \mid z_j = 0, \sigma^2 &\sim \mathcal{N}(0, \sigma^2\tau_0^2). && (7.11) \end{aligned}$$

This is slightly different to the model proposed by George and McCulloch [41]. In this specification, we fix τ_0 close to zero and our choice of τ_0 does not contribute in the slab component. For τ_1, we can perform a sensitivity analysis. We notice that setting a very large τ_1 for an orthogonal design case (that is when $\boldsymbol{x}^T\boldsymbol{x} = n\mathbf{I}_p$) will force non-zero components to be non-important. Therefore, it is reasonable to fix $\tau_1 \geq 1$, but not too large.

To overcome the uncertainty around the selection indicators z_j, we use the imprecise beta model introduced in Section 7.2. We specify our selection indicators in the following way:

$$z_j \mid q_j \sim \text{Ber}(q_j) \tag{7.12}$$

$$q_j \sim \text{Beta}(s\alpha_j, s(1-\alpha_j)). \tag{7.13}$$

Here $s > 0$ is fixed constant and $\alpha \in \mathcal{P}$, where $\mathcal{P}$ is any subset of p-dimensional unit hypercube. The use of the set $\mathcal{P}$ allows us to incorporate prior information about the co-variates. This can be done in two different ways, which we will explain in Chapter 8.

Chapter 8

Robust Bayesian Variable Selection

In Chapter 6, we discussed different variable selection methods in the Bayesian paradigm and introduced the notion of spike and slab priors in Section 6.5 which are efficient in achieving sparsity. However, we saw that choosing priors can be difficult in spike and slab models. This motivates us to perform a robust Bayesian analysis on spike and slab priors. We therefore introduced the notion of robust Bayesian analysis in Chapter 7 along with its applicability. Moreover, we addressed different sources of uncertainty in high-dimensional models and possible treatment of these uncertainties to obtain a robust model.

In this chapter, we use the frameworks discussed in Chapter 6 and Chapter 7 to give a formal description of our novel robust Bayesian model. We introduce our model in Section 8.1, followed by a discussion on the choice of different prior parameters. In Section 8.2, we investigate different properties of the posterior distributions by using an orthogonal design case. The orthogonal design case allows us to decompose the joint posterior in an efficient manner and obtain closed-form expressions for posterior distributions. We use these closed-form expressions to provide a connection between selection indicators and regression coefficients, both of which are important in Bayesian variable selection. Section 8.3 is focused on the general high-dimensional case, where we do not have analytical expressions and therefore we need numerical tools to perform statistical analysis. We show that our choice of priors allows us to obtain closed-form full conditional distributions and we can sample from our posteriors through a Gibbs sampling framework. Finally in Section 8.5, we illustrate our results using synthetic datasets to show our method's performance in variable selection.

8.1 A Hierarchical Model

We follow the discussion in Section 7.3 to propose the following hierarchical model for variable selection, such that for $\beta \coloneqq (\beta_1, \cdots, \beta_p)^T$ and $1 \leq j \leq p$,

$$y \mid \beta, \sigma^2 \sim \mathcal{N}\left(\boldsymbol{x}\beta, \sigma^2 \mathbf{I}_n\right) \tag{8.1}$$

$$\beta_j \mid z_j = 1, \sigma^2 \sim \mathcal{N}(0, \sigma^2 \tau_1^2) \tag{8.2}$$

$$\beta_j \mid z_j = 0, \sigma^2 \sim \mathcal{N}(0, \sigma^2 \tau_0^2) \tag{8.3}$$

$$z_j \mid q_j \sim \mathrm{Ber}(q_j) \tag{8.4}$$

$$q_j \sim \mathrm{Beta}(s\alpha_j, s(1-\alpha_j)) \tag{8.5}$$

$$\sigma^2 \sim \mathrm{InvGamma}(a, b), \tag{8.6}$$

where $s, a, b > 0$ are fixed constants.

The latent variables $z \coloneqq (z_1, \cdots, z_p)$ in the model correspond to spike and slab prior specification routine where z_j represents the selection of the co-variate $\boldsymbol{x}_j$. We consider normal distributions for both spike and slab components to exploit the continuity of our prior specification. We fix a sufficiently small τ_0 $(1 \gg \tau_0^2 > 0)$ so that $\beta_j | z_j = 0$ has its prior probability mass concentrated around zero. Therefore probability distribution of $\beta_j | z_j = 0$ represents the spike component of our prior specification. We can also specify the spike component by using Dirac measure at zero. However, we loose the continuity of the prior at zero, which is undesired for computation purposes. To construct the slab component, we consider τ_1^2 to be large so that $\tau_1 \gg \tau_0$. This allows the prior for $\beta_j \mid z_j = 1$ to be flat. We can also use log-normal distributions for the spike and slab specifications, which we have not explored as log-normal distributions do not allow us to obtain interesting analytical results which we will discuss in Section 8.2.

We use imprecise beta priors to specify the selection probabilities $q \coloneqq (q_1,\ldots,q_p)$.

We use $\alpha := (\alpha_1, \ldots, \alpha_p)$ to represent our prior expectation of the selection probabilities (q) and s to represent concentration parameter. We consider $\alpha \in \mathcal{P}$, where $\mathcal{P}$ is any subset of p-dimensional unit hypercube, that is $\mathcal{P} \subseteq [0,1]^p$. This setting allow us to incorporate prior information in two different ways. We can set individual α_j based on our prior information about the j-th co-variate or, we consider an equiprobable setting where we assume $\alpha_1 = \alpha_2 = \cdots = \alpha_p$ and α_j belong to any subset of $[0,1]$ for $j = 1, \cdots, p$. Therefore, if we have no prior information about the problem, then we consider a near-vacuous set for the elicitation of each α_j. That is, for $1 \gg \epsilon_1, \epsilon_2 > 0$, $\alpha_j \in [\epsilon_1, 1-\epsilon_2]$. This is equivalent to saying that the prior expectation of the total number of active co-variates lies between $p\epsilon_1$ to $p(1-\epsilon_2)$.

To show the importance of α_j, let $f_{z_j}(\beta_j)$ be the density of $\beta_j \mid z_j$ as mentioned in Eq. (8.2) and Eq. (8.3). So that,

$$f_{z_j}(\beta_j) := \frac{1}{\sqrt{2\pi}\sigma\tau_{z_j}} \exp\left(-\frac{\beta_j^2}{2\sigma^2\tau_{z_j}^2}\right). \tag{8.7}$$

Then the hierarchical model implies the following:

$$\begin{aligned} P(\beta_j \mid \sigma^2) &= \sum_{z_j} P(\beta_j \mid z_j, \sigma^2)\left(\int P(z_j \mid q_j)P(q_j)dq_j\right) && (8.8)\\ &= \sum_{z_j}[f_1(\beta_j)]^{z_j}[f_0(\beta_j)]^{1-z_j}\left(\int q_j^{z_j}(1-q_j)^{1-z_j}P(q_j)dq_j\right) && (8.9)\\ &= \sum_{z_j}[\alpha_j f_1(\beta_j)]^{z_j}[(1-\alpha_j)f_0(\beta_j)]^{1-z_j} && (8.10)\\ &= \alpha_j f_1(\beta_j) + (1-\alpha_j)f_0(\beta_j). && (8.11) \end{aligned}$$

That is, we can express our prior on β_j as a mixture of normal distributions where the weights are the prior expectation of the selection probability. In Fig. 8.1 we show the effect of α_j on the prior specification of β for fixed $\tau_0 = 10^{-4}, \tau_1 = 10$ and $\sigma = 1$. We notice that smaller values of α_j forces the prior to be more concentrated around 0 whereas higher values of α_j result to a flatter prior. This also suggests that we can impose our prior belief on β_j through α_j. We can assign a sufficiently large value for τ_1 to capture the prior expected range of β_j and vary α_j to control the tail of the marginalised probability distribution.

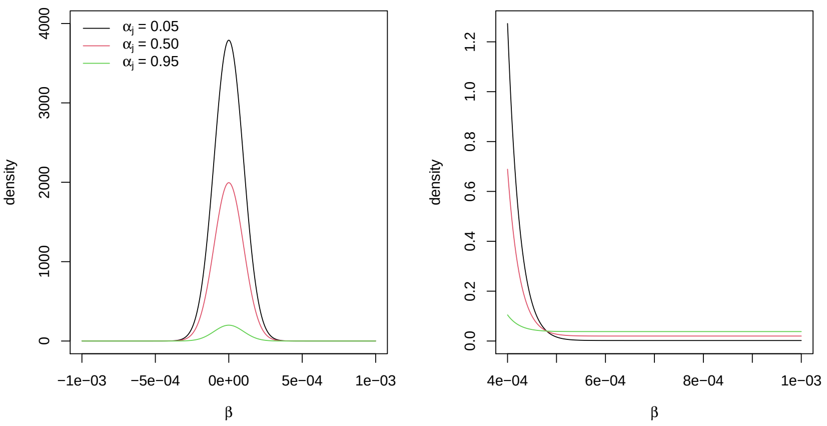

Figure 8.1: Marginalised densities of β_j (Eq. (8.11)) for different values of α_j. The figure on the right side shows the tails of the distributions.

8.2 Posterior for Orthogonal Design

For investigating different analytical properties of the posterior, it is useful to have the different modelling parameters a posteriori independent. In general it is not possible to have such parameters. However, orthogonal design case allows to obtain parameters which are a posteriori independent. As described earlier in Section 7.3, we consider a case to be orthogonal design when $\boldsymbol{x}^T\boldsymbol{x} = n\mathbf{I}_p$. Clearly, for orthogonal design, we have $\hat{\beta}_{\text{OLS}} = (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T y = \boldsymbol{x}^T y/n$, where $\hat{\beta}_{\text{OLS}} \coloneqq (\hat{\beta}_{\text{OLS},\,1}, \ldots, \hat{\beta}_{\text{OLS},\,p})^T$ are the ordinary least squares estimates. Then,

$$P(y \mid \beta) \tag{8.12}$$

$$= \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp\left(-\frac{1}{2\sigma^2}\|y - \boldsymbol{x}\beta\|_2^2\right) \tag{8.13}$$

$$= \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp\left(-\frac{1}{2\sigma^2}\left(n\beta^T\beta - 2n\beta^T\hat{\beta}_{\text{OLS}} + y^T y\right)\right) \tag{8.14}$$

$$= \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp\left(-\frac{1}{2\sigma^2}\left(n\left(\beta^T\beta - 2\beta^T\hat{\beta}_{\text{OLS}} + \hat{\beta}_{\text{OLS}}^T\hat{\beta}_{\text{OLS}}\right) + y^T y - n\hat{\beta}_{\text{OLS}}^T\hat{\beta}_{\text{OLS}}\right)\right) \tag{8.15}$$

$$= \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp\left(-\frac{n\|\beta - \hat{\beta}_{\text{OLS}}\|_2^2}{2\sigma^2}\right) \exp\left(\frac{y^t y - n\hat{\beta}_{\text{OLS}}^t\hat{\beta}_{\text{OLS}}}{2\sigma^2}\right) \tag{8.16}$$

$$= \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp\left(\frac{y^t y - n\hat{\beta}^t_{\text{OLS}}\hat{\beta}_{\text{OLS}}}{2\sigma^2}\right) \prod_j \exp\left(-\frac{n\left(\beta_j - \hat{\beta}_{\text{OLS},\, j}\right)^2}{2\sigma^2}\right) \tag{8.17}$$

$$\propto \prod_j \exp\left(-\frac{n\left(\beta_j - \hat{\beta}_{\text{OLS},\, j}\right)^2}{2\sigma^2}\right). \tag{8.18}$$

The above expression shows that for orthogonal design case, the likelihood is proportional to the product of the functions of each component of β. This allows us to decompose the joint posterior and show that the modelling parameters are a posteriori independent.

Let $z \coloneqq (z_1, \ldots, z_p)$ and $q \coloneqq (q_1, \ldots, q_p)$, then the joint posterior of the proposed hierarchical model can be computed in the following way:

$$P(\beta, \sigma^2, z, q \mid y) \propto P(y \mid \beta, \sigma^2) P(\beta \mid z, \sigma^2) P(z \mid q) P(q) P(\sigma^2). \tag{8.19}$$

To show the analytical properties of our model we will assume that σ^2 is known and fixed. First, we will discuss the posterior of selection indicators and then regression coefficients.

8.2.1 Selection indicators

To examine selection indicators or z, we marginalise the joint posterior in Eq. (8.19), we write the posterior of z as

$$P(z \mid y) = \iint P(\beta, z, q \mid y) dq d\beta \tag{8.20}$$

$$\propto \int P(y \mid \beta) \left(P(\beta \mid z) \int P(z \mid q) P(q) dq\right) d\beta. \tag{8.21}$$

Since $P(z_j \mid q_j) = q_j^{z_j}(1-q_j)^{1-z_j}$ and q_j follows Beta distribution,

$$P(\beta \mid z) \int P(z \mid q) P(q) dq$$

$$= \prod_j \left([f_1(\beta_j)]^{z_j} [f_0(\beta_j)]^{1-z_j} \int q_j^{z_j}(1-q_j)^{1-z_j} P(q_j) dq_j\right) \tag{8.22}$$

$$= \prod_j \left([\alpha_j f_1(\beta_j)]^{z_j} [(1-\alpha_j) f_0(\beta_j)]^{1-z_j}\right). \tag{8.23}$$

Now combining Eq. (8.18), Eq. (8.21) and Eq. (8.23) we get

$$P(z \mid y) \propto \int \prod_j \exp\left(-\frac{n\left(\beta_j - \hat{\beta}_{\text{OLS, }j}\right)_2^2}{2\sigma^2}\right)\left([\alpha_j f_1(\beta_j)]^{z_j}[(1-\alpha_j)f_0(\beta_j)]^{1-z_j}\right) d\beta \tag{8.24}$$

$$\propto \prod_j \int \exp\left(-\frac{n\left(\beta_j - \hat{\beta}_{\text{OLS, }j}\right)_2^2}{2\sigma^2}\right)\left([\alpha_j f_1(\beta_j)]^{z_j}[(1-\alpha_j)f_0(\beta_j)]^{1-z_j}\right) d\beta_j \tag{8.25}$$

Note that in Eq. (8.24), $d\beta$ has not been changed as the integration operator is outside of the product. Now, we have the decomposed posterior of z_j such that

$$P(z_j \mid y) = M_j \int \exp\left(-\frac{n\left(\beta_j - \hat{\beta}_{\text{OLS, }j}\right)^2}{2\sigma^2}\right)[\alpha_j f_1(\beta_j)]^{z_j}[(1-\alpha_j)f_0(\beta_j)]^{1-z_j} d\beta_j, \tag{8.26}$$

where M_j is a normalisation constant independent of z_j. Then we have,

$$P(z_j = 1 \mid y) = M_j \alpha_j \int \exp\left(-\frac{n\left(\beta_j - \hat{\beta}_{\text{OLS, }j}\right)^2}{2\sigma^2}\right) f_1(\beta_j) d\beta_j. \tag{8.27}$$

Now, for $k \in \{0, 1\}$ and $j \in \{1, \cdots, p\}$ we have

$$\exp\left(-\frac{n\left(\beta_j - \hat{\beta}_{\text{OLS, }j}\right)^2}{2\sigma^2}\right) f_k(\beta_j)$$

$$= \exp\left(-\frac{n\left(\beta_j - \hat{\beta}_{\text{OLS, }j}\right)^2}{2\sigma^2}\right) \frac{1}{\sqrt{2\pi}\sigma\tau_k} \exp\left(-\frac{\beta_j^2}{2\sigma^2\tau_k^2}\right) \tag{8.28}$$

$$= \frac{1}{\sqrt{2\pi}\sigma\tau_k} \exp\left(-\frac{n\left(\beta_j - \hat{\beta}_{\text{OLS, }j}\right)^2}{2\sigma^2} - \frac{\beta_j^2}{2\sigma^2\tau_k^2}\right) \tag{8.29}$$

$$= \frac{1}{\sqrt{2\pi}\sigma\tau_k} \exp\left(-\frac{n\tau_k^2\left(\beta_j - \hat{\beta}_{\text{OLS, }j}\right)^2 + \beta_j^2}{2\sigma^2\tau_k^2}\right) \tag{8.30}$$

$$= \frac{1}{\sqrt{2\pi}\sigma\tau_k} \exp\left(-\frac{(n\tau_k^2+1)\beta_j^2 - 2n\tau_k^2\beta_j\hat{\beta}_{\text{OLS, }j} + n\tau_k^2\hat{\beta}_{\text{OLS, }j}^2}{2\sigma^2\tau_k^2}\right) \tag{8.31}$$

$$= \frac{1}{\sqrt{2\pi}\sigma\tau_k} \exp\left(-\frac{(n\tau_k^2+1)\beta_j^2 - 2n\tau_k^2\beta_j\hat{\beta}_{\text{OLS, }j} + \frac{n^2\tau_k^4}{n\tau_k^2+1}\hat{\beta}_{\text{OLS, }j}^2 + \frac{n\tau_k^2}{n\tau_k^2+1}\hat{\beta}_{\text{OLS, }j}^2}{2\sigma^2\tau_k^2}\right) \tag{8.32}$$

$$= \frac{1}{\sqrt{2\pi}\sigma\tau_k} \exp\left(-\frac{(n\tau_k^2+1)\left(\beta_j - \frac{n\tau_k^2}{n\tau_k^2+1}\hat{\beta}_{\text{OLS, }j}\right)^2 + \frac{n\tau_k^2}{n\tau_k^2+1}\hat{\beta}_{\text{OLS, }j}^2}{2\sigma^2\tau_k^2}\right) \tag{8.33}$$

$$= \frac{\sqrt{n\tau_k^2+1}}{\sqrt{2\pi}\sigma\tau_k\sqrt{n\tau_k^2+1}} \exp\left(-\frac{n\hat{\beta}_{\text{OLS, }j}^2}{2\sigma^2(n\tau_k^2+1)}\right) \exp\left(-\frac{\left(\beta_j - \frac{n\tau_k^2}{n\tau_k^2+1}\hat{\beta}_{\text{OLS, }j}\right)^2}{\frac{2\sigma^2\tau_k^2}{n\tau_k^2+1}}\right) \tag{8.34}$$

$$= w_{k,j}\frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{\left(\beta_j - \hat{\beta}_{k,j}\right)^2}{2\sigma_k^2}\right), \tag{8.35}$$

where $\hat{\beta}_{k,j} := \frac{n\tau_k^2\hat{\beta}_{\text{OLS, }j}}{n\tau_k^2+1}$, $\sigma_k^2 := \frac{\sigma^2\tau_k^2}{n\tau_k^2+1}$ and $w_{k,j} := \frac{1}{\sqrt{n\tau_k^2+1}} \exp\left(-\frac{n\hat{\beta}_{\text{OLS, }j}^2}{2(n\sigma^2\tau_k^2+\sigma^2)}\right)$. Then using Eq. (8.35) we have

$$P(z_j = 1 \mid y) = M_j\alpha_j w_{1,j} \tag{8.36}$$

and

$$P(z_j = 0 \mid y) = M_j(1-\alpha_j)w_{0,j}. \tag{8.37}$$

Therefore,

$$z_j \mid y \sim \text{Ber}\left(\frac{\alpha_j w_{1,j}}{\alpha_j w_{1,j} + (1-\alpha_j)w_{0,j}}\right). \tag{8.38}$$

Co-variate selection

For the co-variate selection we investigate the posterior odds of each z_j. We assign a co-variate to be non-active when

$$\sup_{\alpha_j\in\mathcal{P}}\left\{\frac{P(z_j = 1 \mid y)}{P(z_j = 0 \mid y)}\right\} < 1, \tag{8.39}$$

for $j = 1, \cdots, p$. Or equivalently when

$$\sup_{\alpha_j\in\mathcal{P}}\left\{\frac{w_{1,j}\alpha_j}{w_{0,j}(1-\alpha_j)}\right\} < 1. \tag{8.40}$$

Similarly, we assign a co-variate to be active if,

$$\inf_{\alpha_j \in \mathcal{P}} \left\{ \frac{w_{1,j}\alpha_j}{w_{0,j}(1-\alpha_j)} \right\} > 1. \tag{8.41}$$

We define the rest to be indeterminate or indecisive.

Property of the posterior odds:

For $1 \leq j \leq p$, the posterior odds of the selection indicators are given by:

$$\frac{w_{1,j}\alpha_j}{w_{0,j}(1-\alpha_j)} = \frac{w_{1,j}}{w_{0,j}} \left(\frac{1}{1-\alpha_j} - 1 \right). \tag{8.42}$$

Now, the first derivatives of the posterior odds are given by:

$$\frac{w_{1,j}}{w_{0,j}} \frac{1}{(1-\alpha_j)^2} > 0. \tag{8.43}$$

Therefore, we see that the posterior odds are monotone increasing with respect to the prior selection probability α_j.

Now, recall the near-vacuous set defined in Section 8.1. Because of the monotonicity property of the posterior odds, we only need to compute the posterior odds on the lower and upper bounds of the set instead of the whole interval. That is

$$\sup_{\alpha_j \in [\epsilon_1, 1-\epsilon_2]} \left\{ \frac{w_{1,j}\alpha_j}{w_{0,j}(1-\alpha_j)} \right\} = \frac{(1-\epsilon_2)}{\epsilon_2} \cdot \frac{w_{1,j}}{w_{0,j}} \tag{8.44}$$

and,

$$\inf_{\alpha_j \in [\epsilon_1, 1-\epsilon_2]} \left\{ \frac{w_{1,j}\alpha_j}{w_{0,j}(1-\alpha_j)} \right\} = \frac{\epsilon_1}{(1-\epsilon_1)} \cdot \frac{w_{1,j}}{w_{0,j}}. \tag{8.45}$$

Therefore, we say that a co-variate is considered to be active if $\frac{\epsilon_1}{(1-\epsilon_1)} \cdot \frac{w_{1,j}}{w_{0,j}} > 1$ and a co-variate is considered to be inactive if $\frac{(1-\epsilon_2)}{\epsilon_2} \cdot \frac{w_{1,j}}{w_{0,j}} < 1$.

8.2.2 Regression coefficients

Similar to the selection indicators, the joint posterior of the regression coefficients is given by:

$$P(\beta \mid y) = \sum_z \int P(\beta, z, q \mid y) dq \tag{8.46}$$

$$\propto \sum_z \int P(y \mid \beta)P(\beta \mid z)P(z \mid q)P(q)dq \tag{8.47}$$

$$\propto P(y \mid \beta) \sum_z \left(P(\beta \mid z) \int P(z \mid q)P(q)dq \right). \tag{8.48}$$

From Eq. (8.23) we have

$$P(\beta \mid z) \int P(z \mid q)P(q)dq = \prod_j \left([\alpha_j f_1(\beta_j)]^{z_j} [(1-\alpha_j)f_0(\beta_j)]^{1-z_j} \right). \tag{8.49}$$

Then we can write Eq. (8.48) as

$$P(\beta \mid y) \propto P(y \mid \beta) \sum_z \left(\prod_j \left([\alpha_j f_1(\beta_j)]^{z_j} [(1-\alpha_j)f_0(\beta_j)]^{1-z_j} \right) \right). \tag{8.50}$$

Now,

$$\sum_z \left(\prod_j \left([\alpha_j f_1(\beta_j)]^{z_j} [(1-\alpha_j)f_0(\beta_j)]^{1-z_j} \right) \right)$$

$$= \sum_{z_1} \cdots \sum_{z_p} \left(\prod_j \left([\alpha_j f_1(\beta_j)]^{z_j} [(1-\alpha_j)f_0(\beta_j)]^{1-z_j} \right) \right) \tag{8.51}$$

$$= \sum_{z_1} \cdots \sum_{z_p} [\alpha_1 f_1(\beta_1)]^{z_1} [(1-\alpha_1)f_0(\beta_1)]^{1-z_1} \cdots [\alpha_p f_1(\beta_p)]^{z_p} [(1-\alpha_p)f_0(\beta_p)]^{1-z_p} \tag{8.52}$$

$$= \prod_j \sum_{z_j} [\alpha_j f_1(\beta_j)]^{z_j} [(1-\alpha_j)f_0(\beta_j)]^{1-z_j} \tag{8.53}$$

$$= \prod_j [\alpha_j f_1(\beta_j) + (1-\alpha_j)f_0(\beta_j)]. \tag{8.54}$$

Therefore we get,

$$P(\beta \mid y) \propto P(y \mid \beta) \prod_j [\alpha_j f_1(\beta_j) + (1-\alpha_j)f_0(\beta_j)]. \tag{8.55}$$

Now combining Eq. (8.18) and Eq. (8.55) we have

$$P(\beta \mid y) \propto \exp\left(-\frac{1}{2\sigma^2} \left(n\beta^T\beta - 2n\beta^T\hat{\beta}_{\text{OLS}} \right) \right) \prod_j [\alpha_j f_1(\beta_j) + (1-\alpha_j)f_0(\beta_j)]$$

$$\propto \exp\left(-\frac{n}{2\sigma^2} \|\beta - \hat{\beta}_{\text{OLS}}\|_2^2 \right) \prod_j [\alpha_j f_1(\beta_j) + (1-\alpha_j)f_0(\beta_j)] \tag{8.56}$$

$$\propto \prod_j \exp\left(-\frac{n(\beta_j - \hat{\beta}_{\text{OLS},\, j})^2}{2\sigma^2} \right) [\alpha_j f_1(\beta_j) + (1-\alpha_j)f_0(\beta_j)]. \tag{8.57}$$

Therefore, the β_j's are a posteriori independent and for each $1 \leq j \leq p$, we have,

$$P(\beta_j \mid y) \propto \exp\left(-\frac{n(\beta_j - \hat{\beta}_{\text{OLS, }j})^2}{2\sigma^2}\right)[\alpha_j f_1(\beta_j) + (1-\alpha_j) f_0(\beta_j)]. \tag{8.58}$$

Let $W_j := \alpha_j w_{1,j} + (1-\alpha_j) w_{0,j}$. Then combining Eq. (8.35) and Eq. (8.58) we show,

$$\beta_j \mid y \sim \frac{\alpha_j w_{1,j}}{W_j}\mathcal{N}\left(\hat{\beta}_{1,j}, \sigma_1^2\right) + \frac{(1-\alpha_j) w_{0,j}}{W_j}\mathcal{N}\left(\hat{\beta}_{0,j}, \sigma_0^2\right), \tag{8.59}$$

where $\hat{\beta}_{k,j} := \frac{n\tau_k^2 \hat{\beta}_{\text{OLS, }j}}{n\tau_k^2+1}$, $\sigma_k^2 := \frac{\sigma^2\tau_k^2}{n\tau_k^2+1}$ and $w_{k,j} := \frac{1}{\sqrt{n\tau_k^2+1}} \exp\left(-\frac{n\hat{\beta}^2_{\text{OLS, }j}}{2(n\sigma^2\tau_k^2+\sigma^2)}\right)$.

Eq. (8.59) shows that the posteriors of the regression coefficients are mixtures of two normal distributions. Clearly, the posteriors are bimodal when $\hat{\beta}_{\text{OLS, }j} \neq 0$. We illustrate the posteriors in Fig. 8.2 for fixed $\sigma^2 = 1$, $n = 100$, $\tau_0 = 10^{-4}$ and τ_1=10. In Fig. 8.2, the left column shows the density functions and in the right column shows the posterior cumulative distribution functions (CDF). We show these posteriors for four different values of $\hat{\beta}_{\text{OLS, }j}$ ($\hat{\beta}$ in the figure) over equispaced grids of α_j so that $\alpha_j \in [0.05, 0.95]$. We observe that the posterior densities are bimodal except for the top row and each of the posteriors has a spike component at zero. We also notice that for smaller values of $\hat{\beta}_{\text{OLS, }j}$, the posterior CDFs are more concentrated at zero. However, as we increase the value of $\hat{\beta}_{\text{OLS, }j}$, the posterior CDFs shift towards $\hat{\beta}_{\text{OLS, }j}$. For a sufficiently large value of $\hat{\beta}_{\text{OLS, }j}$, the posterior CDFs are concentrated at $\hat{\beta}_{\text{OLS, }j}$.

Properties of the posterior:

To analyse the properties of the posterior, we first consider the ratio of the weights in Eq. (8.59). For $1 \leq j \leq p$, ratios of the weights are given by:

$$\frac{\alpha_j w_{1,j}}{(1-\alpha_j) w_{0,j}}. \tag{8.60}$$

This corresponds to posterior selection probability of selection indicators. Therefore, for active co-variates this ratio becomes greater than 1 for all $\alpha_j \in [\epsilon_1, 1-\epsilon_2]$ and $\mathcal{N}\left(\hat{\beta}_{1,j}, \sigma_1^2\right)$ dominates the posterior. Similarly, for non-active co-variates this ratio becomes less than 1 for all values of α_j and $\mathcal{N}\left(\hat{\beta}_{0,j}, \sigma_0^2\right)$ dominates the posterior.

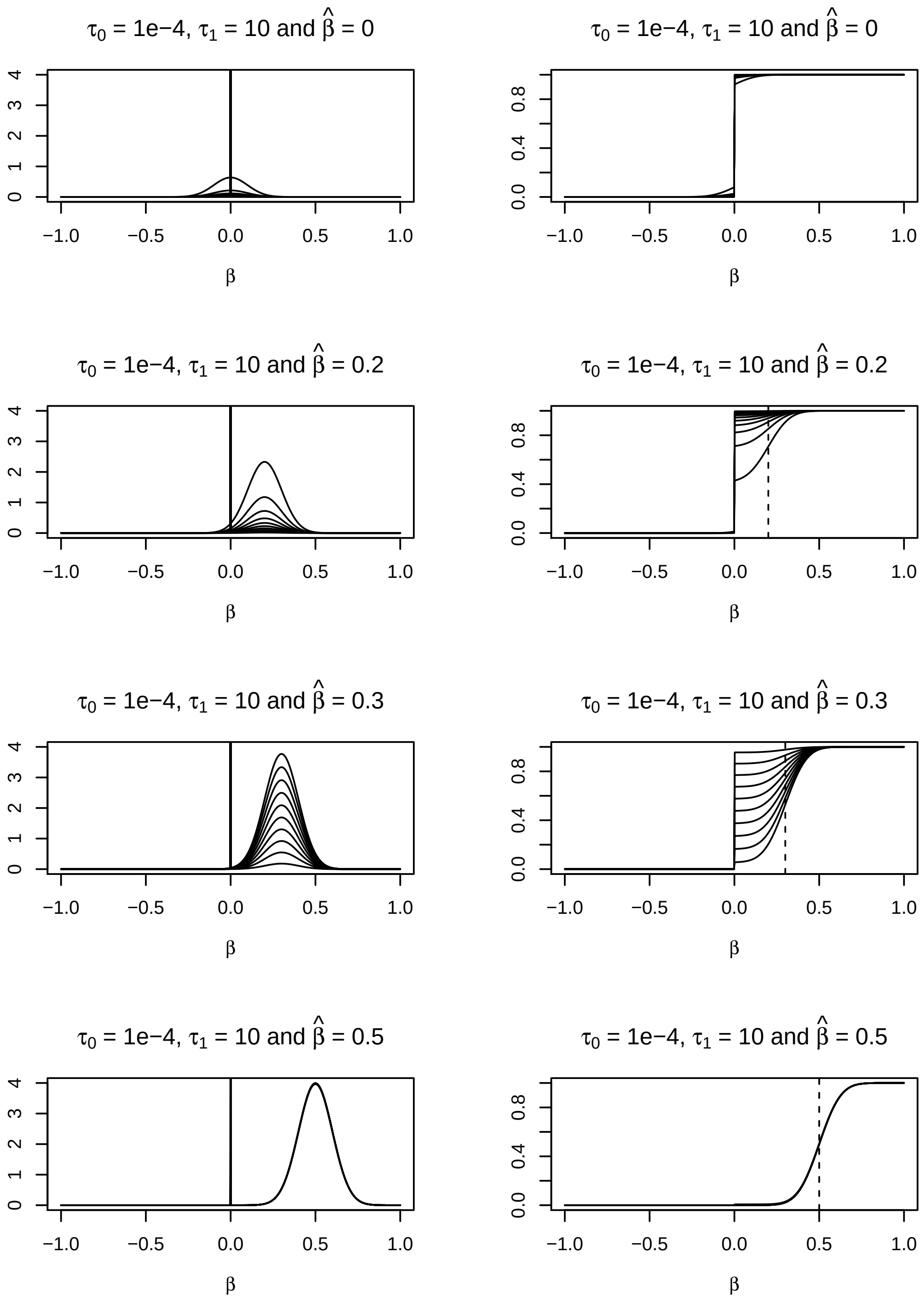

Figure 8.2: Posterior density function and corresponding cumulative distribution function of β_j for different values of $\hat{\beta}_{\text{OLS},\ j}$ over a set of α_j such that $\alpha_j \in [0.05, 0.95]$.

Alternatively, exploiting the monotonicity property of the posterior odds, we can say that $\mathcal{N}\left(\hat{\beta}_{1,j}, \sigma_1^2\right)$ dominates the posterior if $\frac{\epsilon_1}{(1-\epsilon_1)} \cdot \frac{w_{1,j}}{w_{0,j}} > 1$. That is, if

$$\exp\left(-\frac{n\hat{\beta}^2_{\text{OLS, }j}}{2(n\sigma^2\tau_1^2+\sigma^2)} + \frac{n\hat{\beta}^2_{\text{OLS, }j}}{2(n\sigma^2\tau_0^2+\sigma^2)}\right) > \frac{(1-\epsilon_1)\sqrt{n\tau_1^2+1}}{\epsilon_1\sqrt{n\tau_0^2+1}} \tag{8.61}$$

$$-\frac{n\hat{\beta}^2_{\text{OLS, }j}}{2(n\sigma^2\tau_1^2+\sigma^2)} + \frac{n\hat{\beta}^2_{\text{OLS, }j}}{2(n\sigma^2\tau_0^2+\sigma^2)} > \ln\left(\frac{(1-\epsilon_1)\sqrt{n\tau_1^2+1}}{\epsilon_1\sqrt{n\tau_0^2+1}}\right) \tag{8.62}$$

$$\frac{n\hat{\beta}^2_{\text{OLS, }j}}{2\sigma^2}\left[-\frac{1}{(n\tau_1^2+1)} + \frac{1}{(n\tau_0^2+1)}\right] > \ln\left(\frac{(1-\epsilon_1)\sqrt{n\tau_1^2+1}}{\epsilon_1\sqrt{n\tau_0^2+1}}\right) \tag{8.63}$$

$$\frac{n\hat{\beta}^2_{\text{OLS, }j}}{2\sigma^2}\frac{n\tau_1^2-n\tau_0^2}{(n\tau_1^2+1)(n\tau_0^2+1)} > \ln\left(\frac{(1-\epsilon_1)\sqrt{n\tau_1^2+1}}{\epsilon_1\sqrt{n\tau_0^2+1}}\right). \tag{8.64}$$

Then after rearranging the terms on both sides, we get:

$$\hat{\beta}^2_{\text{OLS, }j} > \frac{\sigma^2}{n}\frac{(n\tau_1^2+1)(n\tau_0^2+1)}{n\tau_1^2-n\tau_0^2}\left[2\ln\left(\frac{1-\epsilon_1}{\epsilon_1}\right) + \ln\left(\frac{n\tau_1^2+1}{n\tau_0^2+1}\right)\right]. \tag{8.65}$$

Similarly, we say that, $\mathcal{N}\left(\hat{\beta}_{0,j}, \sigma_0^2\right)$ dominates the posterior if,

$$\hat{\beta}^2_{\text{OLS, }j} < \frac{\sigma^2}{n}\frac{(n\tau_1^2+1)(n\tau_0^2+1)}{n\tau_1^2-n\tau_0^2}\left[2\ln\left(\frac{\epsilon_2}{1-\epsilon_2}\right) + \ln\left(\frac{n\tau_1^2+1}{n\tau_0^2+1}\right)\right]. \tag{8.66}$$

We can further simplify this for $\epsilon_1 = \epsilon_2 = \epsilon$, that is when $\alpha_j \in [\epsilon, 1-\epsilon]$. Let $\tau_0 \ll 1/n$, then $\mathcal{N}\left(\hat{\beta}_{1,j}, \sigma_1^2\right)$ dominates the posterior if,

$$\hat{\beta}^2_{\text{OLS, }j} > \frac{\sigma^2}{n}\frac{(n\tau_1^2+1)}{n\tau_1^2}\left[\ln\left(n\tau_1^2+1\right) + 2\ln\left(\frac{1-\epsilon}{\epsilon}\right)\right], \tag{8.67}$$

and similarly, $\mathcal{N}\left(\hat{\beta}_{0,j}, \sigma_0^2\right)$ dominates the posterior if,

$$\hat{\beta}^2_{\text{OLS, }j} < \frac{\sigma^2}{n}\frac{(n\tau_1^2+1)}{n\tau_1^2}\left[\ln\left(n\tau_1^2+1\right) - 2\ln\left(\frac{1-\epsilon}{\epsilon}\right)\right]. \tag{8.68}$$

We can compute a region of indeterminacy using Eq. (8.67) and Eq. (8.68). If the value of $\hat{\beta}^2_{\text{OLS, }j}$ lies in between these bounds then we consider the j-th co-variate as indeterminate. We illustrate this in Fig. 8.3 for fixed $\sigma^2 = 1$, $n = 100$ and $\tau_0 = 10^{-4}$. The shaded area shows the region of indeterminacy for different values of $\alpha_j \in [\epsilon, 1-\epsilon]$. Clearly, the region of indeterminacy depends on the values of ϵ. Higher values of $\epsilon(< 0.5)$ shrink the region of indeterminacy. We also notice that higher values of τ_1 force the bounds to be higher. Therefore, extreme values of τ_1 may lead to poor results in variable selection. A very small value of τ_1 will force some non-active co-variates to be indeterminate whereas a very high value of τ_1 will force some non-zero small effects to be inactive.

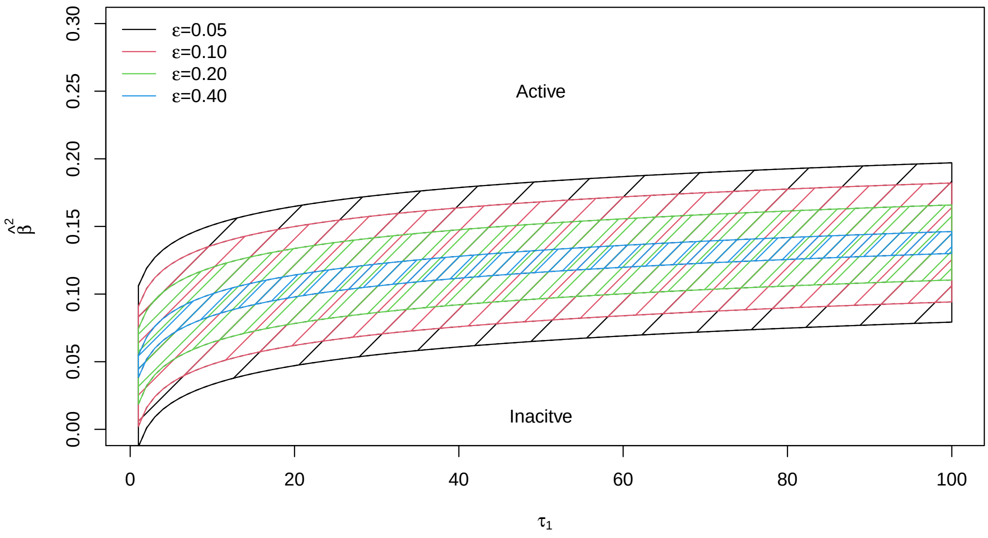

Figure 8.3: Effect of τ_1 in specifying the region of indeterminacy for different values of ϵ.

Posterior mean and variance:

The posterior mean of β_j is given by:

$$E(\beta_j \mid y) = \frac{\alpha_j w_{1,j}}{W_j}\hat{\beta}_{1,j} + \frac{(1-\alpha_j)w_{0,j}}{W_j}\hat{\beta}_{0,j}. \tag{8.69}$$

We show illustrate these posterior means in Fig. 8.4. We fix $\tau_0 = 10^{-4}, \tau_1 = 10$, $n = 100$ and $\sigma^2 = 1$. We check posterior means for six different possible values of $\hat{\beta}_{\text{OLS},\, j}$ ($\hat{\beta}$ in the figure). In the top row we show our results for $\hat{\beta}_{\text{OLS},\, j} > 0$. We see that in the first two cases, the posterior means are monotonically increasing and in the third case it is close to constant. Similarly in the bottom row, we show our result for $\hat{\beta}_{\text{OLS},\, j} < 0$. We see that the posterior means are decreasing in the first two cases, and remains close to constant in the third case.

We also get a closed-form expression for the posterior variance. By Lemma A.4, we have

$$\begin{aligned}&\text{Var}(\beta_j \mid y)\\ &= \frac{\alpha_j w_{1,j}}{W_j}\left(\sigma_1^2 + \hat{\beta}_{1,j}^2\right) + \frac{(1-\alpha_j)w_{0,j}}{W_j}\left(\sigma_0^2 + \hat{\beta}_{0,j}^2\right) - \left[\frac{\alpha_j w_{1,j}\hat{\beta}_{1,j} + (1-\alpha_j)w_{0,j}\hat{\beta}_{0,j}}{W_j}\right]^2\end{aligned} \tag{8.70}$$

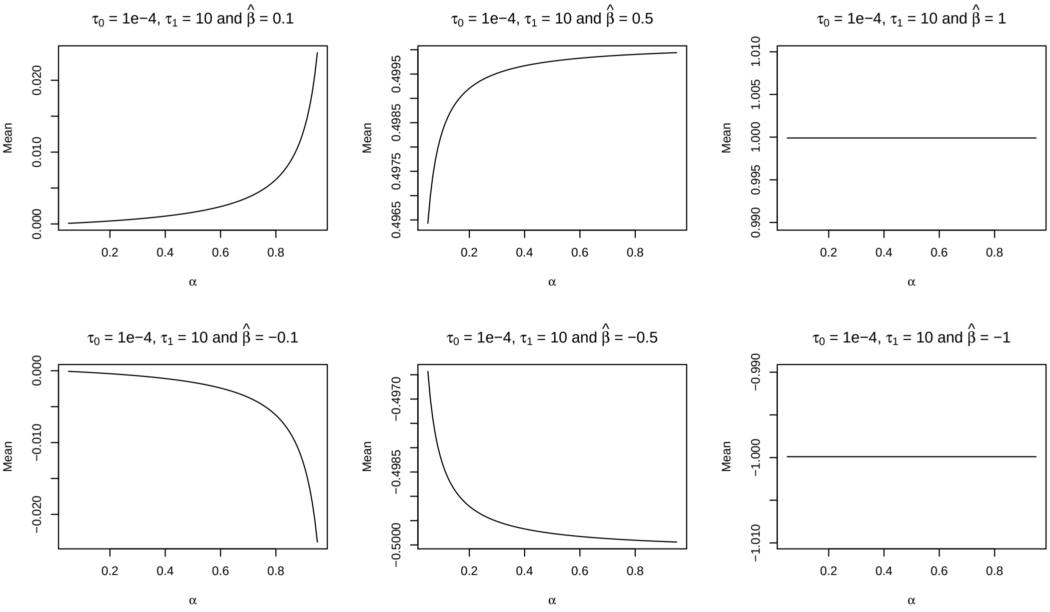

Figure 8.4: Relation between posterior expectation of β and prior selection probability α for different values of $\hat{\beta}$.

$$= \frac{\alpha_j w_{1,j}\sigma_1^2 + (1-\alpha_j)w_{0,j}\sigma_0^2}{W_j} + \frac{\alpha_j w_{1,j}\hat{\beta}_{1,j}^2 + (1-\alpha_j)w_{0,j}\hat{\beta}_{1,j}^2}{W_j} - \left[\frac{\alpha_j w_{1,j}\hat{\beta}_{1,j} + (1-\alpha_j)w_{0,j}\hat{\beta}_{0,j}}{W_j}\right]^2 \tag{8.71}$$

$$= \frac{\alpha_j w_{1,j}\sigma_1^2 + (1-\alpha_j)w_{0,j}\sigma_0^2}{W_j} + \frac{\alpha(1-\alpha)w_{1,j}w_{0,j}(\hat{\beta}_{1,j} - \hat{\beta}_{0,j})^2}{W_j^2}. \tag{8.72}$$

Therefore, we get a set of posterior variances $\mathcal{V}_j$ such that:

$$\mathcal{V}_j := \left\{ \frac{\alpha_j w_{1,j}\sigma_1^2 + (1-\alpha_j)w_{0,j}\sigma_0^2}{W_j} + \frac{\alpha(1-\alpha)w_{1,j}w_{0,j}(\hat{\beta}_{1,j} - \hat{\beta}_{0,j})^2}{W_j^2} : \alpha_j \in (0,1) \right\}, \tag{8.73}$$

where $w_{k,j}$ and σ_k are as defined before.

The posterior variance of β_j does not show a monotone trend like the posterior mean. In Fig. 8.5, we show the effect of α_j on the posterior variance for six different values $\hat{\beta}_{\text{OLS},\, j}$. We fix $\tau_0 = 10^{-4}, \tau_1 = 10$, $n = 100$ and $\sigma^2 = 1$ to obtain these posterior variances. In the top row, we show variances for $\hat{\beta}_{\text{OLS},\, j} > 0$ and in the bottom row we show the case for $\hat{\beta}_{\text{OLS},\, j} < 0$. We notice that for extreme values of $\hat{\beta}_{\text{OLS},\, j}$, the posterior variance behaves like a constant and is close to $\frac{\sigma^2}{n} = 0.01$.

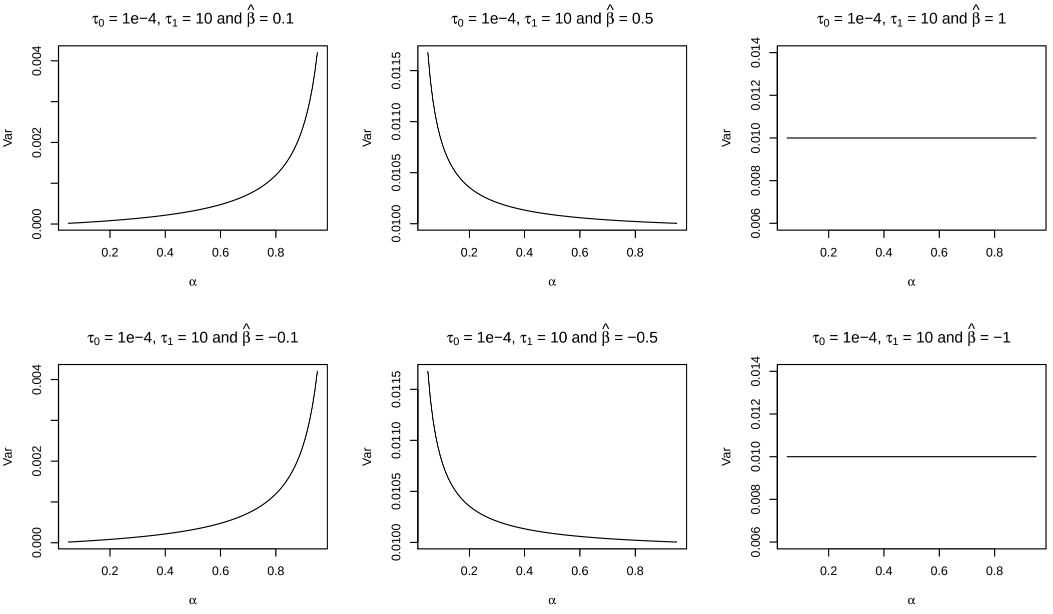

Figure 8.5: Relation between posterior variance of β and prior selection probability α for different values of $\hat{\beta}$.

8.3 Posterior Computation for the General Case

The orthogonal case allows us to decompose the joint density function in a convenient way for known variance σ^2. However, this can be non-trivial when the variance is unknown. Moreover, variable selection is generally applied for correlated datasets or high-dimensional problems. As a consequence, it is not possible to have an orthogonal design in many cases. Therefore, we need a suitable computation scheme for general cases, that is, when the datasets are non-orthogonal or we don't have any information about the variance. Interestingly, our choice of priors allows us to obtain full conditional distributions of the modelling parameters and therefore for the general case, we follow a Gibbs sampling routine (Section 6.3.1) to compute posterior distributions and hence perform variable selection [57]. To avoid confusion we use a special notation $\overset{t}{\propto}$, which means that left hand side is proportional to the terms which are function of t and the rest are considered as constants. Now, recall the joint posterior in Eq. (8.19). Then the joint conditional distribution of the regression coefficients is given by:

$$P(\beta \mid z, \sigma^2, q, y) \overset{\beta}{\propto} P(\beta, z, \sigma^2, q \mid y) \tag{8.74}$$

$$\overset{\beta}{\propto} P(y \mid \beta, \sigma^2) P(\beta \mid z, \sigma^2) \tag{8.75}$$

$$\overset{\beta}{\propto} \exp\left(-\frac{1}{2\sigma^2}\|y - \boldsymbol{x}\beta\|_2^2\right) \prod_{j=1}^{p} f_{z_j}(\beta_j) \tag{8.76}$$

$$\overset{\beta}{\propto} \exp\left(-\frac{\beta^T \boldsymbol{x}^T \boldsymbol{x} \beta - 2\beta^T \boldsymbol{x}^T y}{2\sigma^2}\right) \prod_{j=1}^{p} f_{z_j}(\beta_j) \tag{8.77}$$

$$\overset{\beta}{\propto} \exp\left(-\frac{\beta^T \boldsymbol{x}^T \boldsymbol{x} \beta - 2\beta^T \boldsymbol{x}^T y}{2\sigma^2}\right) \prod_{j=1}^{p} \exp\left(-\frac{\beta_j^2}{2\sigma^2 \tau_{z_j}^2}\right). \tag{8.78}$$

Let $D_z := \text{diag}(\tau_{z_j}^{-2})$, then we rewrite Eq. (8.78) as

$$P(\beta \mid z, \sigma^2, q, y) \overset{\beta}{\propto} \exp\left(-\frac{\beta^T \boldsymbol{x}^T \boldsymbol{x} \beta - 2\beta^T \boldsymbol{x}^T y}{2\sigma^2}\right) \exp\left(-\frac{\beta^T D_z \beta}{2\sigma^2}\right) \tag{8.79}$$

$$\overset{\beta}{\propto} \exp\left(-\frac{\beta^T \boldsymbol{x}^T \boldsymbol{x} \beta - 2\beta^T \boldsymbol{x}^T y + \beta^T D_z \beta}{2\sigma^2}\right) \tag{8.80}$$

$$\overset{\beta}{\propto} \exp\left(-\frac{(\beta - \mu^*)^T V^{-1} (\beta - \mu^*)}{2\sigma^2}\right), \tag{8.81}$$

where $\mu^* := V\boldsymbol{x}^T y$ and $V := (\boldsymbol{x}^T \boldsymbol{x} + D_z)^{-1}$. Therefore the full conditional of β follows a multivariate normal distribution such that:

$$\beta \mid z, \sigma^2, q, y \sim \mathcal{N}(\mu^*, \sigma^2 L). \tag{8.82}$$

For the selection indicators, we only need to compute the probability of z_j conditional on β, σ^2 and q_j. Therefore, we can compute these posteriors component-wise, such that:

$$P(z_j \mid \beta_j, \sigma^2, q_j) \overset{z_j}{\propto} P(\beta_j \mid z_j, \sigma^2) P(z_j \mid q_j) \tag{8.83}$$

$$\overset{z_j}{\propto} q_j^{z_j} (1 - q_j)^{1 - z_j} f_{z_j}(\beta_j) \tag{8.84}$$

$$\overset{z_j}{\propto} [q_j f_{z_j}(\beta_j)]^{z_j} [(1 - q_j) f_{z_j}(\beta_j)]^{1 - z_j}. \tag{8.85}$$

Therefore, $z_j \mid \beta_j, \sigma^2$ follows a Bernoulli distribution such that

$$P(z_j = 1 \mid \beta_j, \sigma^2) = \frac{q_j f_1(\beta_j)}{q_j f_1(\beta_j) + (1 - q_j) f_0(\beta_j)}. \tag{8.86}$$

Unlike the orthogonal design case, the choice of concentration parameter plays an important role on the conditional distributions of q_j's for the Gibbs sampling algorithm. We know that,

$$P(q_j \mid z_j) \sim P(z_j \mid q_j) P(z_j). \tag{8.87}$$

Then the conditional distribution of the q_j follows a beta distribution such that:

$$q_j \mid z_j \sim \text{Beta}(s\alpha_j + z_j, s(1-\alpha_j) + 1 - z_j), \tag{8.88}$$

where $\alpha_j \in \mathcal{P}$.

For the general case, we are also interested in the posterior of σ^2. The conditional distribution of σ^2 is given by:

$$\begin{aligned}
&P(\sigma^2 \mid \beta, z, y) \\
&\overset{\sigma^2}{\propto} P(y \mid \beta, \sigma^2) P(\beta \mid z, \sigma^2) P(\sigma^2) && (8.89) \\
&\overset{\sigma^2}{\propto} \frac{1}{\sigma^n} \exp\left(-\frac{\|y - \boldsymbol{x}\beta\|_2^2}{2\sigma^2}\right) \frac{1}{\sigma^p} \exp\left(-\frac{\beta^T D_z \beta}{2\sigma^2}\right) \frac{1}{\sigma^{2(a+1)}} \exp\left(-\frac{b}{\sigma^2}\right) && (8.90) \\
&\overset{\sigma^2}{\propto} \frac{1}{\sigma^{2(p/2+n/2+a+1)}} \exp\left\{-\frac{1}{\sigma^2}\left(\frac{\|y - \boldsymbol{x}\beta\|_2^2}{2} + \frac{\beta^T D_z \beta}{2} + b\right)\right\} && (8.91)
\end{aligned}$$

Therefore,

$$\sigma^2 \mid \beta, z, y \sim \text{InvGamma}\left(a + \frac{p}{2} + \frac{n}{2}, b + \frac{\|y - \boldsymbol{x}\beta\|_2^2}{2} + \frac{\beta^T D_z \beta}{2}\right) \tag{8.92}$$

8.4 Measures for Prediction

A robust Bayesian routine needs different measures of accuracy as we don't have a single posterior for prediction. We introduce a new measure which can be considered to evaluate prediction accuracy and call it minimum squared error. Let

$$\mathcal{A}(\alpha) := \left\{ j : \left\{ \frac{w_{1,j}\alpha_j}{w_{0,j}(1-\alpha_j)} \right\} > 1 \right\}. \tag{8.93}$$

Therefore, $\mathcal{A}(\alpha)$ or simply, $\mathcal{A}$ denotes the set of active variables for each value of α. Let, $\boldsymbol{x}_{\mathcal{A}} := [x_j]_{j \in \mathcal{A}}$ and $\beta_{\mathcal{A}} := [\beta_j]_{j \in \mathcal{A}}$. We define minimum squared error by:

$$\text{Minimum Squared Error} = \min_{\alpha \in \mathcal{P}} \|y - \boldsymbol{x}_{\mathcal{A}} \hat{\beta}_{\mathcal{A}}^{\text{post}}\|_2^2, \tag{8.94}$$

where $\hat{\beta}_{\mathcal{A}}^{\text{post}} := E(\beta_{\mathcal{A}} \mid y)$ is the posterior mean of $\beta_{\mathcal{A}}$.

The sensitivity analysis also creates an indeterminacy in prediction. Therefore, we define a similar measure called maximum squared error over the set of $\alpha \in \mathcal{P}$. We use both minimum and maximum squared error to introduce a new measure to capture the indeterminacy such that:

$$\text{Indeterminacy} = \frac{\text{Maximum Squared Error} - \text{Minimum Squared Error}}{\text{Maximum Squared Error}}. \tag{8.95}$$

Therefore, indeterminacy gives us a relative difference between the best fitted model and worst fitted model obtained from the robust Bayesian analysis. Clearly, we will aim to reduce the indeterminacy for our robust Bayesian model.

8.5 Simulation Studies

In this section we will show the accuracy of our method in terms of variable selection. We construct four different synthetic datasets to investigate different aspects of variable selection problems.

Example 8.1. *In this example, we construct an orthogonal design matrix* $x_{i,j}$ *with 50 predictors and 100 observations. We assign the regression coefficients so that* $\beta_j \sim$ *Uniform*$([-200, -80] \cup [80, 200])$ *for* $1 \leq j \leq 6$ *and* $\beta_j = 0$ *for* $j > 6$*. We consider standard normal noise to construct the response vector* $y_i = \sum_{j=1}^{6} x_{i,j}\beta_j + \epsilon_i$*, where* $\epsilon_i \sim N(0, 1)$ *for* $i = 1, \cdots, 100$*. This setting allows us to evaluate the performance of our method with only strong non-zero effects.*

Example 8.2. *In this case, we construct an orthogonal design matrix as of Example 8.1. We assign the regression coefficients such that the first 12* β_j*'s represent a strong effect and the next 20* β_j*'s represent a mild effect. To do so, we consider* $\beta_j \sim$ *Uniform*$([-200, -80] \cup [80, 200])$ *for* $1 \leq j \leq 12$*,* $\beta_j \sim$ *Uniform*$([-20, -10] \cup [10, 20])$ *for* $13 \leq j \leq 32$ *and* $\beta_j = 0$ *for* $j > 32$*. We construct the response vector in the following way:* $y_i = \sum_{j=1}^{32} x_{i,j}\beta_j + \epsilon_i$*, where* $\epsilon_i \sim N(0, 1)$ *for* $i = 1, \cdots, 100$*. This type of coefficient assignment allows us to investigate both medium and large effects within the model.*

Example 8.3. *We use this example to illustrate the high-dimensional case. We construct the design matrix with 100 observations and 200 predictors from a multivariate normal distribution so that* $x_i \sim \mathcal{N}(0, \mathbf{I}_{200})$*, where* $1 \leq i \leq 100$*. We set regression coefficients so that* $\beta_j \sim$ *Uniform*$([-200, -80] \cup [80, 200])$ *for* $1 \leq j \leq 12$*,* $\beta_j \sim$ *Uniform*$([-20, -10] \cup [10, 20])$ *for* $13 \leq j \leq 32$ *and* $\beta_j = 0$ *for* $j > 32$*. We construct the response vector in a similar fashion as of Example 8.1 and Example 8.2. Clearly, in this case the design matrix can not be constructed as an orthogonal design matrix as the total number of observations is less than the total number of predictors.*

Example 8.4. *We use this dataset to show the performance of our method for high dimensional problems with small effects. We generate the predictors from a multivariate normal distributions so that $x_i \sim \mathcal{N}(0, \mathbf{I}_{100})$, where $1 \leq i \leq 50$. To show the small effects, we set $\beta_j \sim$ Uniform$([-4,-1] \cup [1,4])$ for $1 \leq j \leq 60$ and $\beta_j = 0$ for $j > 60$. For the random noise we consider smaller variance unlike the previous examples where we take 1 as the variance of the error term. We use smaller variance as we consider small effects only and higher variance of random noise may contribute more in the response than the predictors. Therefore, we construct the response vector so that $y_i = \sum_{j=1}^{60} x_{i,j}\beta_j + \epsilon_i$, where $\epsilon_i \sim N(0, 0.01)$ for $i = 1, \cdots, 50$. This way, we get a problem with small effects only. For this dataset, the number of true active regression coefficients is more than the total number of observations unlike the previous examples.*

Results

To investigate our method's accuracy in variable selection, we consider two different sets for α so that one set represents a near-vacuous case and the other set represents prior information. To specify the near-vacuous case, we consider $\alpha_j \in [0.1, 0.9]$ for the j-th co-variate. For the sake of simplicity, we drop this subscript j and write α_j as α. The choice of the elicitation-based set is dependent on the example. For instance, for Example 8.1, we set $\alpha \in [0.1, 0.12]$ based on the true values of the regression coefficients. Similarly, we consider $\alpha \in [0.1, 0.64]$ for Example 8.2, $\alpha \in [0.06, 0.2]$ for Example 8.3 and $\alpha \in [0.1, 0.6]$ for Example 8.4. We fix $\tau_0 = 10^{-6}$ for all of the experiments to specify the spike component of our prior. For σ^2, we use an inverse-gamma distribution with both scale and shape parameters being equal to 10^{-5}. Experiments suggest that higher values of τ_1 give us poor results for variable selection. To show that, we consider three different values of τ_1 for Example 8.1 and Example 8.2, which are 5, 10 and 50. This way, we get the effect of τ_1 on variable selection. However, for Example 8.3 and Example 8.4, we notice that setting $\tau_1 > 1$ gives us poor results. Therefore, for these two datasets, we consider $\tau_1 = 1$ for the illustration, along with 10 and 50. We provide the summary of our results in Table 8.1. The left-most column shows the method of variable selection followed by three columns which represent the status of the true active variables after variable selection which are 'active', 'inactive' and 'indeterminate'. Similarly we show the status of true inactive variables in the next three columns. We also perform variable selection with three other Bayesian methods for comparison. For this, we use `basad` [57], `blasso` [60] and `SSLASSO` [64].

We observe that for the first dataset, all methods are in good agreement except for `SSLASSO` which identifies only 4 co-variates as active. It can be seen from the Table 8.1 that the choice of τ_1 or α have no effect on the variable selection and our method identifies all the true important co-variates correctly.

The analyses using Example 8.2 is particularly interesting. In this case, the effect of τ_1 is more prominent. We see that increasing value of τ_1 results in fewer active variables, which follows our result for orthogonal design case. We also notice that choice of α can be crucial in identifying the active co-variates. The elicitation-based choice underperforms when $\tau_1 = 50$, α puts less weights on some of the mild effects and higher τ_1 reduces the posterior odds of their corresponding selection indicators. However, this is not the case for near-vacuous case and all of the mild effects remain indeterminate. We also see that our results are somewhat in agreement with `basad` and `SSLASSO` for higher values of α and selects less variables as active. This is not the case for `blasso` which identifies all the 32 true important coefficients as active.

The Example 8.3 is used to illustrate the high-dimensional problem. In this case, our method is in good agreement with `basad` for $\tau_1 = 1$. However, for higher values of τ_1 it tends to select fewer covariates as active and gives a similar result to that of `SSLASSO`. However, unlike the previous cases, there is not a single choice of τ_1, for which our method identifies every true important covariate correctly. In this particular example, `blasso` outperforms other methods in terms of variable selection.

We use Example 8.4 to show the performance of our method for high dimensional models with small effects. We see that `blasso` performs poorly in terms of variable selection and considers many of the true active effects as inactive. This also the case for `SSLASSO`, which gives similar result. We observe that our method also tends to select fewer covariates as active. However, it does not assign all of those variable as inactive and classifies most of them as indeterminate. As a result, our method reduces the risk of producing too many false inactive covariates. We also see that our method gives least number of false active covariates among the four methods. However, our method is not a clear winner, as `basad` performs the best when it comes to identification of active covariates despite giving more false inactive covariates than our method.

We see that for the first two datasets, both `blasso` and our method identify all the true active covariates, especially when $\tau_1 = 1$. This is not the case for the third dataset, where our method fails to identify every true active covariate and `blasso` is the clear winner. However, for the fourth dataset `blasso` performs the worst in terms of the variable selection. This happens as the double exponential prior do not have enough mass at zero relative to the tail (check Castillo et al. [17] for further discussion on this). Therefore, `blasso` can not allow near zero active effects and inactive effects simultaneously and it tends to overshrink the small effects and the median estimates become sparser than desired. Our method do not experience such issues and the robust Bayesian approach makes sure that we do not produce too many false inactive covariates. As a result, our method performs well for every dataset irrespective of the construction.

	True Active			True Inactive		
Parameter Setting/ Method	Act	Inact	Indet	Act	Inact	Indet
Dataset 1, 6 active and 44 inactive						
$\alpha \in [0.1, 0.9]$, $\tau_0 = 10^{-6}$, $\tau_1 = 5$	6	0	0	0	44	0
$\alpha \in [0.1, 0.9]$, $\tau_0 = 10^{-6}$, $\tau_1 = 10$	6	0	0	0	44	0
$\alpha \in [0.1, 0.9]$, $\tau_0 = 10^{-6}$, $\tau_1 = 50$	6	0	0	0	44	0
$\alpha \in [0.1, 0.12]$, $\tau_0 = 10^{-6}$, $\tau_1 = 5$	6	0	0	0	44	0
$\alpha \in [0.1, 0.12]$, $\tau_0 = 10^{-6}$, $\tau_1 = 10$	6	0	0	0	44	0
$\alpha \in [0.1, 0.12]$, $\tau_0 = 10^{-6}$, $\tau_1 = 50$	6	0	0	0	44	0
BASAD	6	0	–	0	44	–
BLASSO (Median)	6	0	–	0	44	–
SSLASSO (Double Exponential)	4	2	–	0	44	–
Dataset 2, 32 active and 18 inactive						
$\alpha \in [0.1, 0.9]$, $\tau_0 = 10^{-6}$, $\tau_1 = 5$	32	0	0	0	18	0
$\alpha \in [0.1, 0.9]$, $\tau_0 = 10^{-6}$, $\tau_1 = 10$	17	0	15	0	18	0
$\alpha \in [0.1, 0.9]$, $\tau_0 = 10^{-6}$, $\tau_1 = 50$	4	0	28	0	18	0
$\alpha \in [0.1, 0.64]$, $\tau_0 = 10^{-6}$, $\tau_1 = 5$	32	0	0	0	18	0
$\alpha \in [0.1, 0.64]$, $\tau_0 = 10^{-6}$, $\tau_1 = 10$	18	0	14	0	18	0
$\alpha \in [0.1, 0.64]$, $\tau_0 = 10^{-6}$, $\tau_1 = 50$	7	5	20	0	18	0
BASAD	16	16	–	0	18	–
BLASSO (Median)	32	0	–	0	18	–
SSLASSO (Double Exponential)	4	28	–	0	18	–
Dataset 3, 40 active and 160 inactive						
$\alpha \in [0.1, 0.9]$, $\tau_0 = 10^{-6}$, $\tau_1 = 1$	14	0	26	0	0	160
$\alpha \in [0.1, 0.9]$, $\tau_0 = 10^{-6}$, $\tau_1 = 10$	3	0	37	0	0	160
$\alpha \in [0.1, 0.9]$, $\tau_0 = 10^{-6}$, $\tau_1 = 50$	4	1	35	0	1	159
$\alpha \in [0.06, 0.2]$, $\tau_0 = 10^{-6}$, $\tau_1 = 1$	14	1	25	0	160	0
$\alpha \in [0.06, 0.2]$, $\tau_0 = 10^{-6}$, $\tau_1 = 10$	14	15	11	0	154	6
$\alpha \in [0.06, 0.2]$, $\tau_0 = 10^{-6}$, $\tau_1 = 50$	5	10	25	0	160	0
BASAD	12	28	–	0	160	–
BLASSO (Median)	40	0	–	0	160	–
SSLASSO (Double Exponential)	3	37	–	0	160	–
Dataset 4, 60 active and 40 inactive						
$\alpha \in [0.1, 0.9]$, $\tau_0 = 10^{-6}$, $\tau_1 = 1$	6	0	54	1	0	39
$\alpha \in [0.1, 0.9]$, $\tau_0 = 10^{-6}$, $\tau_1 = 10$	0	0	60	0	0	40
$\alpha \in [0.1, 0.9]$, $\tau_0 = 10^{-6}$, $\tau_1 = 50$	0	1	59	0	0	40
$\alpha \in [0.1, 0.6]$, $\tau_0 = 10^{-6}$, $\tau_1 = 1$	5	10	45	1	12	27
$\alpha \in [0.1, 0.6]$, $\tau_0 = 10^{-6}$, $\tau_1 = 10$	0	1	59	0	5	35
$\alpha \in [0.1, 0.6]$, $\tau_0 = 10^{-6}$, $\tau_1 = 50$	0	16	44	0	8	32
BASAD	34	26	–	11	29	–
BLASSO (Median)	16	44	–	4	36	–
SSLASSO (Double Exponential)	17	43	–	3	37	–

Table 8.1: Summary of variable selection for four different synthetic datasets.

Chapter 9

Data Analysis

We have discussed different Bayesian modelling approaches for linear regression in Chapter 6 and our novel robust Bayesian approach in Chapter 8. In these two chapters, we showed how our choice of priors contributes to parameter estimation. In Chapter 8, we illustrated these Bayesian variable selection techniques along with our novel approach using synthetic dataset. However, we are also interested in model fitting which is an important part of statistical modelling with real datasets.

In this chapter, we perform robust Bayesian analysis using different real datasets. These datasets are carefully chosen so that we can perform our analysis for different cases which may occur in variable selection problems. We start our analysis using the Diabetes dataset in Section 9.1 followed by Gaia dataset in Section 9.2. These two dataset are not high-dimensional in nature, however, these are correlated in nature. Especially for Gaia, we can observe collinearity within the dataset which is an important problem in Bayesian variable selection. In Section 9.3, we investigate an ultra high-dimensional dataset for which use a preliminary screening before performing our robust Bayesian analysis.

9.1 Diabetes Dataset

The Diabetes dataset[1] [29] concerns 10 predictors which are age, sex, body mass index, average blood pressure and six blood serum measurements. The response denotes the disease progression in one year. Here the 'sex' predictor is not Gaussian and we use dummy variables to work with this. We show the correlation plot in Fig. 9.1, where we can see a mild correlation between the predictors.

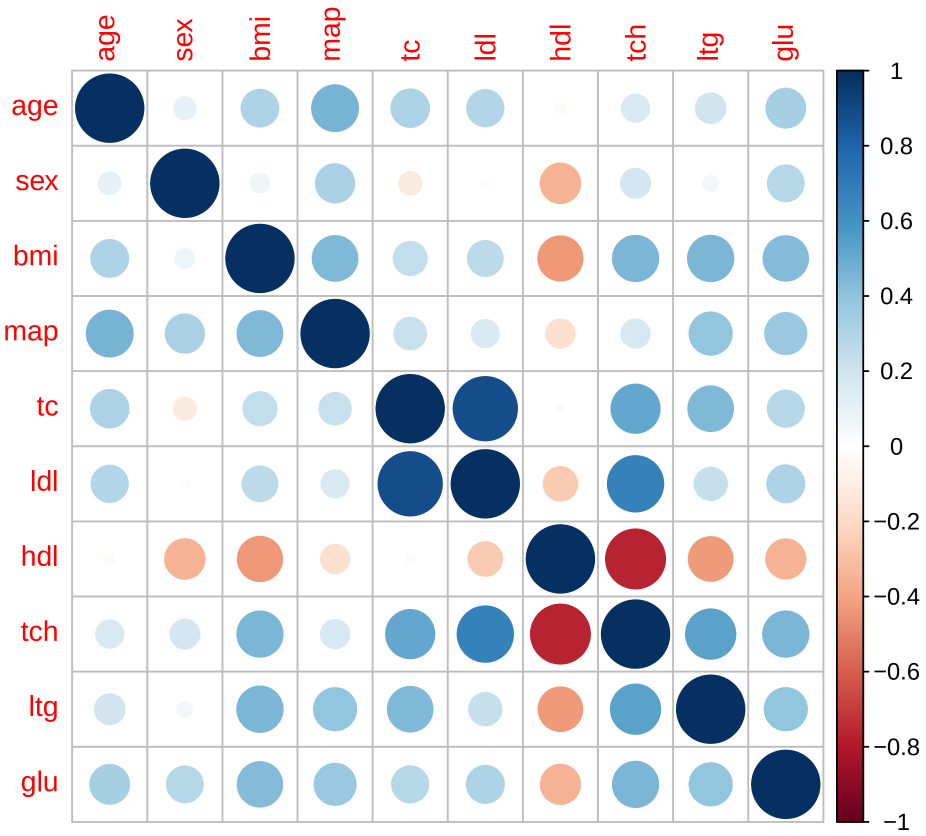

Figure 9.1: Correlation plot matrix of Diabetes dataset

We perform a preliminary analysis to get an idea about the number of active covariates in the dataset. We randomly sample 100 observations from the dataset and fit ordinary least squares. We provide the summary of ordinary least squares in Table 9.1. As we discussed earlier for ridge estimates in Section 3.3, we can simply check the p-values to get an idea about the importance of the covariates. We see

[1] This dataset is openly available and has been loaded from the `R` package `lars` [45] for illustration.

from Table 9.1, p-values of ‘sex’ and ‘ltg’ are less than 0.01. So we can safely assume that there are at least two active covariates in the model. Similarly we can consider 3 other variables to be active based on our threshold for the p-values. Therefore, we can expect to have 2 to 5 active variables in the dataset. Now, based on this preliminary analysis, we consider two different sets to specify our prior expectation of the selection probabilities denoted by $\alpha := (\alpha_1, \ldots, \alpha_p)$. We first specify a near-vacuous set so that, $\alpha_j \in [0.1, 0.9]$ and we choose the other set so that $\alpha_j \in [0.2, 0.5]$. Therefore, our second choice of α_j's a direct representation of our prior information on the selection probability of variables.

	Estimate	Std. Error	t-value	p-value
(Intercept)	155	5	27	<2e-16
age	62	138	0.5	6.5e-01
sex	-381	136	-2.8	6.3e-03
bmi	381	164	2.3	2.3e-02
map	365	167	2.2	3.2e-02
tc	-1149	887	-1.3	1.9e-01
ldl	924	759	1.2	2.2e-01
hdl	334	447	0.7	4.6e-01
tch	244	334	0.7	4.7e-01
ltg	971	361	2.7	8.6e-03
glu	182	143	1.3	2.1e-01

Table 9.1: Summary of ordinary least squares estimates for the Diabetes dataset.

Analysis

To perform variable selection and model fitting, we randomly sample 100 observations from the dataset. We perform our analysis with two different choices of α_j as we mentioned earlier. We fix $\tau_0 = 10^{-2}$, $\tau_1 = 5$ and use inverse-gamma distribution to specify σ^2 so that the scale and shape parameters are equal to 10^{-5}. We also consider four other methods for comparison. Three of these methods are based on spike and slab priors which are `spikeslab` [48], `SSLASSO` [64] and `basad` [79]. The other method we explore for illustration is Bayesian LASSO [60] using the package `blasso` [43]. We also randomly sample 20 new observations for investigating prediction accuracy and posterior predictive checking.

Method	Act	Inact	Indet	Min. Sq. Err	Indeterminacy
RBVS; $\alpha_j \in [0.1, 0.9]$	2	5	3	2.4e+04	0.29
RBVS; $\alpha_j \in [0.2, 0.5]$	2	6	2	2.4e+04	0.28
SSLASSO	2	8	–	3.3e+04	–
Spike & Slab	8	2	–	2.6e+04	–
BASAD	2	8	–	3.3e+04	–
BLASSO	5	5	–	2.5e+04	–

Table 9.2: Summary of variable selection and model fitting for the Diabetes dataset.

We show the summary of our analysis in Table 9.2. In the left-most column we provide different methods followed by three columns which represent the number of active covariates, inactive covariates and indeterminate covariates. From Table 9.2, we notice that both choices of α_j give us 2 active co-variates which are 'bmi' and 'ltg'. This is also the case for `SSL` and `basad`. However, `blasso` and `spikeslab` include more variables in the model. Our method also identifies some indeterminate variables in the Diabetes dataset. For the near-vacuous case, we have 3 indeterminate variables whereas 2 indeterminate variables for the second case. We show the cumulative distributions of the selected covariates in Fig. 9.2, which are obtained from 1000 MCMC samples of the posteriors.

We provide the minimum squared error and indeterminacy in last two columns of Table 9.2. We observe that our method outperforms other methods in terms of minimum squared error. We also see that the near-vacuous case and the elicitation-based case are in good agreement in terms of the indeterminacy. We also show the posterior predictive distributions in Fig. 9.3. We use 20 newly sampled responses to construct the reference distribution, which we show by the black bold line. The red shaded lines on the left denotes the posterior predictive distributions obtained from near-vacuous case and green shaded lines on the right shows the posterior predictive distributions obtained from the elicitation-based case. To construct this distributions we randomly choose 100 MCMC samples from the posteriors. We see that both are in good agreement and the shaded areas cover the reference distribution.

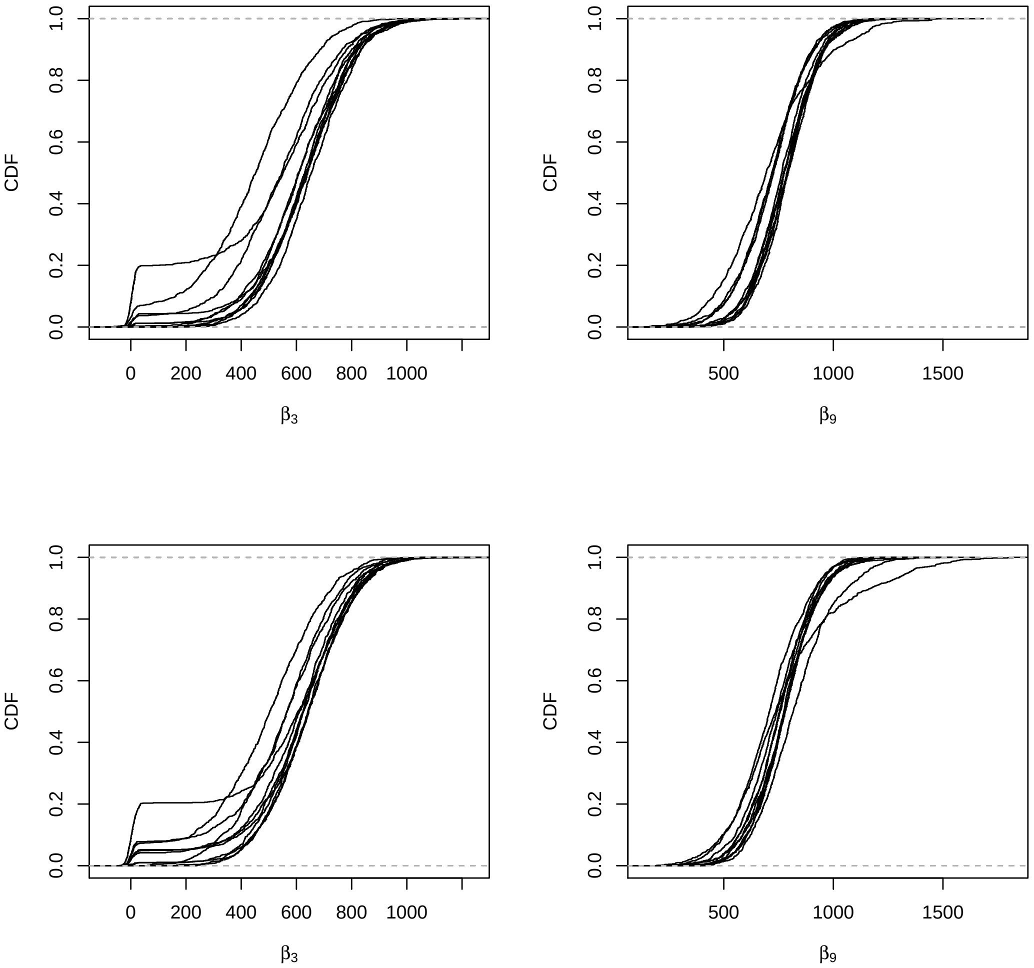

Figure 9.2: Empirical cumulative distribution functions of the selected covariates for near-vacuous set (top) and elicitation-based set (bottom).

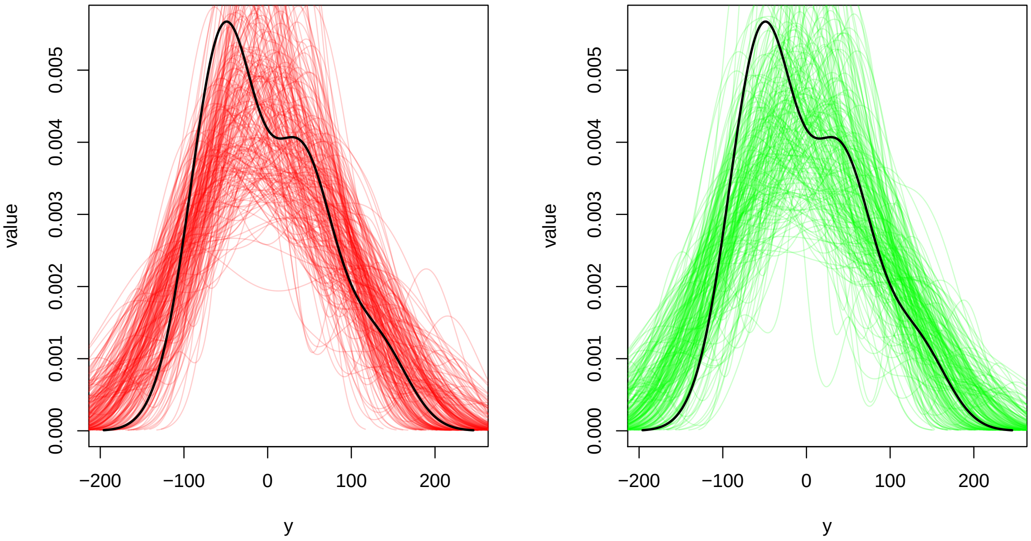

Figure 9.3: Posterior predictive distributions obtained from the Diabetes dataset for near-vacuous set (left) and elicitation-based set (right).

9.2 Gaia Dataset

The Gaia dataset[2] was used for computer experiments [4, 31] prior to the launch of European Space Agency's Gaia mission [32]. The data contains spectral information of 16 (p) wavelength bands, and four different stellar parameters. In this example, we take stellar-temperature (in Kelvin scale) as the response variable. This dataset contains 8286 observations which are highly correlated. We show the correlation between the co-variates in Fig. 9.4.

Previous work by Einbeck et al. [31] suggests that this dataset contains 1-3 main contributory variables. Based on this information, we take two sets for α_j similar to our choice of α_j for Diabetes dataset in Section 9.1. We specify our first set as near-vacuous set and choose $\alpha_j \in [0.1, 0.9]$. The second set is based on our prior information on the contributory variables and therefore a natural choice of α_j is $[1/16, 3/16]$.

[2] This dataset is openly available and has been loaded from the `R` package `LPCM` [30] for illustration.

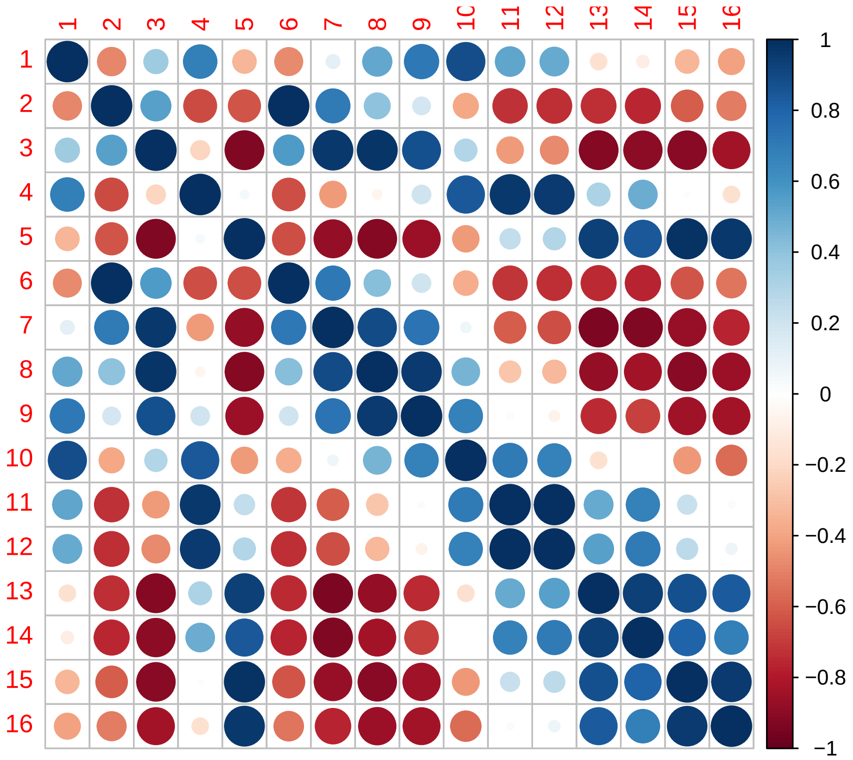

Figure 9.4: Correlation plot matrix of the Gaia dataset

Analysis

Similar to our example in Section 9.1, we consider 100 observations to fit our model. We also use same parameter values to perform our robust Bayesian analysis. We compare our method with four other methods and provide the summary of our comparison in Table 9.3. For both choices of α_j, we notice that our method considers 'band 6' to be the active co-variate in the model. However, the choice of α_j is more significant in identifying the inactive variables. We observe that for the near-vacuous set, our method remain indecisive in terms of rejecting a variable and produces 15 indeterminate variables. For the elicitation-based set we see that there are only two indeterminate variables unlike for the near-vacuous set. We notice that our method is in good agreement with `SSLASSO` and `blasso` in terms of variable selection. The other two methods include more variables in the model. We show the empirical cumulative distribution functions of the 6-th co-variate in Fig. 9.5. It can be seen from the figure that the CDFs obtained for near-vacuous set have larger variance than that of elicitation-based set.

Method	Act	Inact	Indet	Min. Sq. Err	Indeterminacy
RBVS; $\alpha_j \in [0.1, 0.9]$	1	0	15	6.2e+07	0.45
RBVS; $\alpha_j \in [1/16, 3/16]$	1	13	2	6.1e+07	0.21
SSLASSO	1	15	–	6.4e+07	–
Spike & Slab	4	12	–	6.6e+07	–
BASAD	3	13	–	7.9e+07	–
BLASSO	1	15	–	6.5e+07	–

Table 9.3: Summary of variable selection and model fitting for the Gaia dataset.

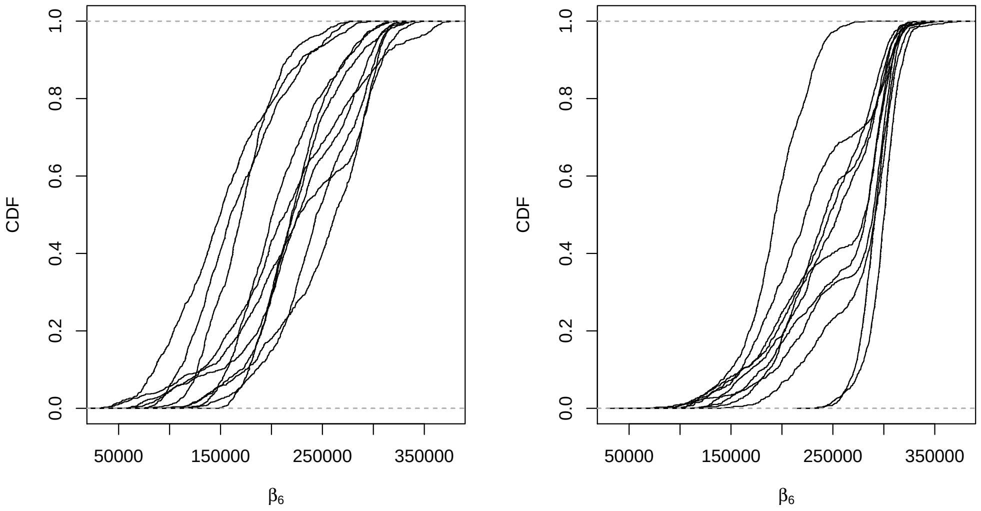

Figure 9.5: Empirical cumulative distribution functions of the selected covariate for near-vacuous set (left) and elicitation-based set (right).

We also investigate accuracy of our method in terms of prediction which we show in 4-th and 5-th columns. We sample 20 new observations from the Gaia dataset to evaluate prediction accuracy. We notice that our method outperforms other methods in terms of minimum squared error. It can be seen that the indetermincay is higher for near-vacuous set than the case where an elicitation-based set is used for α_j. We also use these observations to obtain posterior predictive distributions, which we show in Fig. 9.6, similar to our illustration with the Diabetes dataset. We see that the posterior predictive distributions are in good agreement with the reference distribution denoted by the black bold line and covers the reference distribution denoted by the black bold line.

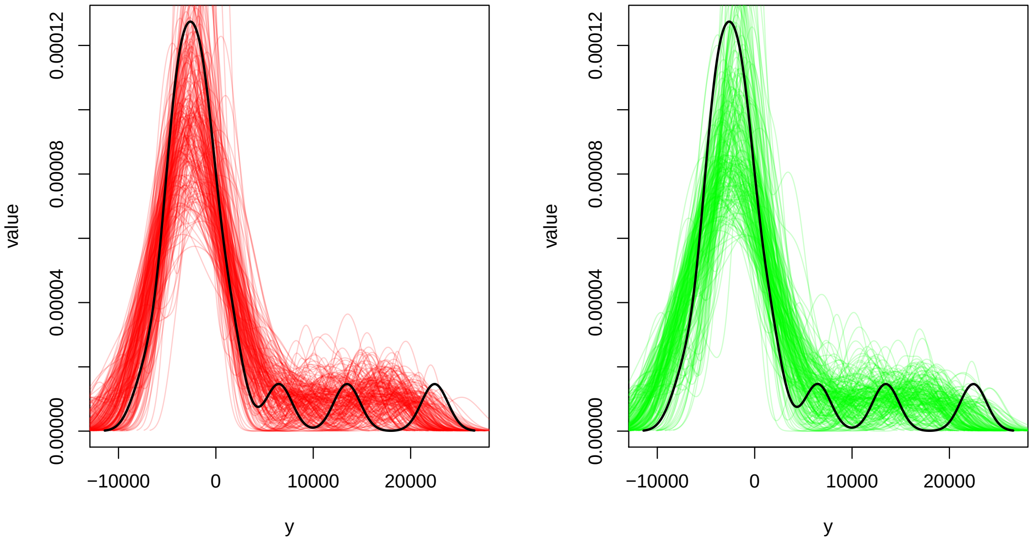

Figure 9.6: Posterior predictive distributions obtained from the Gaia dataset for near-vacuous set(left) and elicitation-based set (right).

9.3 Lymphoma Dataset

We investigate the Lymphoma dataset[3] [3] to illustrate our result for a high-dimensional problem. In this dataset, there are 7399 genes related to B-cell Lymphoma along with the response which denote censored survival times. There are only 240 observation in this dataset which makes the problem ultra-high-dimensional, that is $p \gg n$. Performing Bayesian analysis in this type of dataset is extremely difficult and we use a variable screening method to identify 200 important co-variates. We use the package `VariableScreening` [53] to obtain the first 200 co-variates based on the correlation distance. We provide the correlation plot of these co-variates in Fig. 9.7. It can be observed that the dataset is highly correlated and forms several cluster along the diagonal.

[3] This dataset is openly available and has been collected from the following website: `https://web.stanford.edu/~hastie/StatLearnSparsity/data.html`

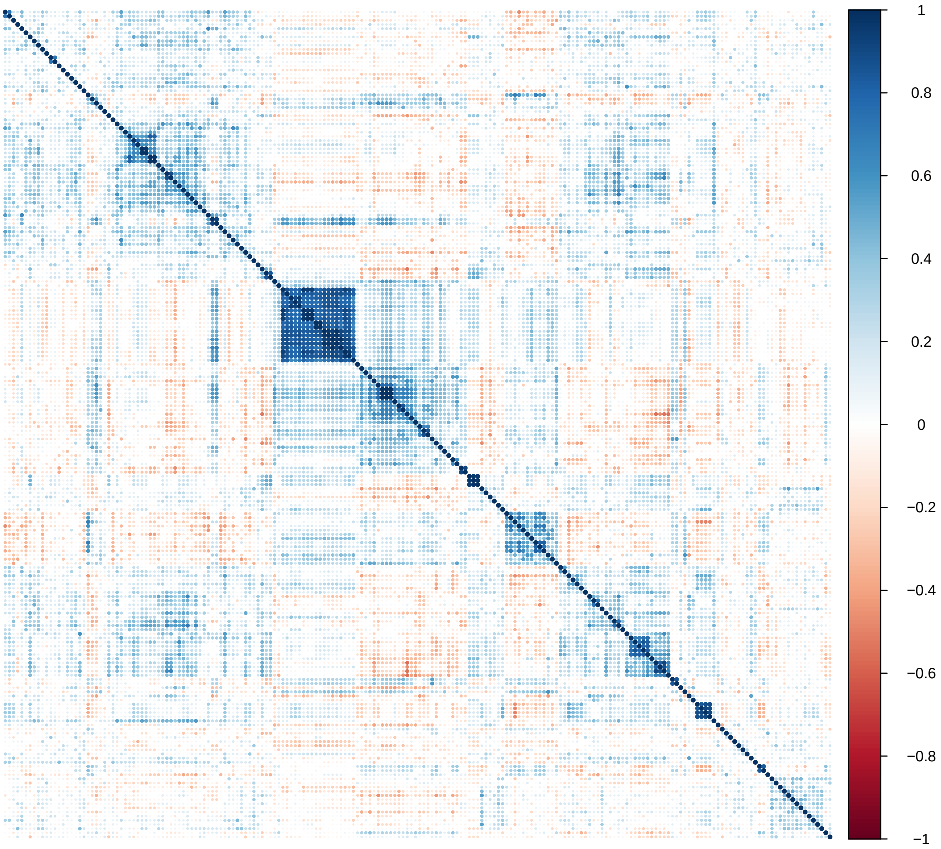

Figure 9.7: Correlation plot matrix of the Lymphoma dataset

The choice of α_j for this dataset is difficult and we choose α_j based on the selected co-variates after the variable screening. We fit a ridge regression model to examine the p-values. This preliminary analysis suggests that we may consider 20 to 30 variables based on our tolerance for p-values. Therefore, we specify our elicitation-based as $\alpha_j \in [0.1, 0.15]$. For the near-vacuous case, we stick to our previous examples and choose $\alpha_j \in [0.1, 0.9]$.

Analysis

Similar to our previous examples, we sample 100 observations for variable selection model fitting. For this dataset, we fix $\tau_0 = 10^{-3}$, $\tau_1 = 1$ and use an inverse-gamma distribution with shape and scale parameters being equal to 1. We provide the summary of our Bayesian analysis in Table 9.4. Similar to our analysis of Gaia dataset in Section 9.2, we observe that our method does not identify any inactive

Method	Act	Inact	Indet	Min. Sq. Err	Indeterminacy
RBVS; $\alpha_j \in [0.1, 0.9]$	1	0	199	1.2e+02	0.89
RBVS; $\alpha_j \in [0.1, 0.15]$	1	191	8	1.1e+02	0.54
SSLASSO	0	200	–	1.7e+02	–
Spike & Slab	10	190	–	1.6e+02	–
BASAD	3	197	–	1.4e+02	–
BLASSO	3	197	–	1.7e+02	–

Table 9.4: Summary of variable selection and model fitting for the Lymphoma dataset.

variable for the near-vacuous set of α_j. This is not the case for the elicitation-based set and identifies 191 inactive variables and only 8 as indeterminate variables. Another, interesting thing happens where, `SSLASSO`, selects the null model unlike other methods used for comparison. It can be seen that both `basad` and `blasso` identify three active co-variates, whereas `spikeslab` selects 10 co-variates. Our method in this case identifies only the 7251-th predictor as active, irrespective to the choice of α_j. We show the empirical CDFs in Fig. 9.8, it can be noticed that the variances are higher for the near-vacuous case than the elicitation-based case. We also notice that for the Lymphoma dataset, the estimates are close to 0, which results to the bimodal nature of the CDFs.

For the prediction accuracy, we sample 20 new observations similar to our analysis using the other two datasets. We observe that our method performs better than the other methods in terms of minimum squared error. However, the indeterminacy is higher than the previous examples and for the near-vacuous set, the indeterminacy is 0.89, which is undesirable. High indeterminacy suggests that we must incorporate some prior information on this dataset, which we do with the second choice of α_j. This is slightly better for the second choice of α_j which is based on elicitation. High indeterminacy for both choices of α_j is also an indication that we don't have a best method for the Lymphoma dataset. We also show the posterior predictive distributions in Fig. 9.9. In the figure, the left hand side shows the plots for the near-vacuous case and right hand side shows the plots for the elicitation-based case.

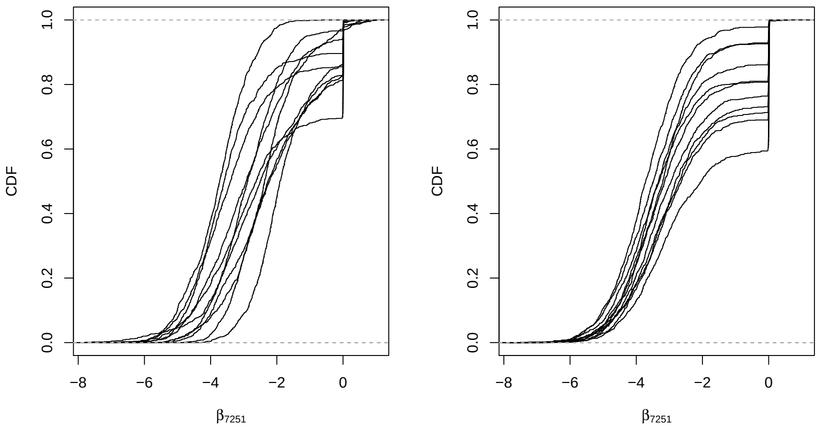

Figure 9.8: Empirical cumulative distribution functions of the selected covariates for near-vacuous set (left) and elicitation-based set (right).

We observe that elicitation-based set gives much better result than the near-vacuous set and posterior predictive distributions obtained from elicitation-based case covers the reference distribution. However, this is not the case for near-vacuous case.

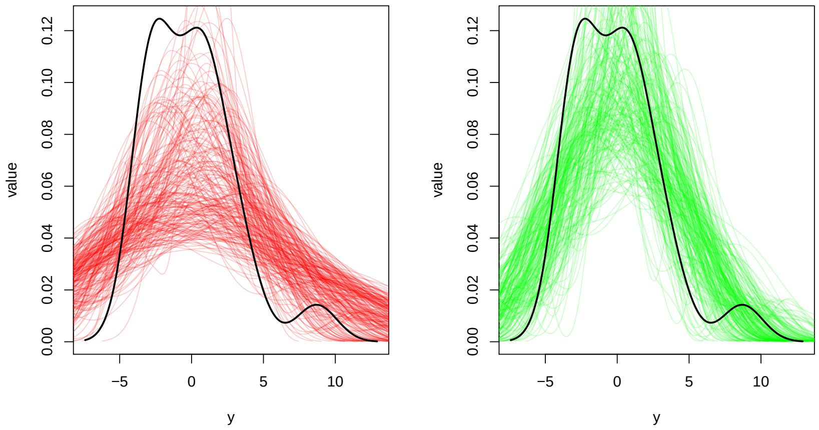

Figure 9.9: Posterior predictive distributions obtained from the Lymphoma dataset for near-vacuous set(left) and elicitation-based set (right).

Chapter 10

Conclusion

This chapter summarises all of the work presented in the thesis followed by discussions and potential areas of future work.

10.1 Summary of the thesis

The thesis was focused on investigating the imprecision in high-dimensional statistical modelling and building a robust variable selection routine for problems with limited information. We investigated this in two different ways. First, we examined the use of the weights in adaptive LASSO to perform a sensitivity analysis and check the resulting variation in variable selection and model fitting. The other approach we considered was robust Bayesian analysis. We specified the selection probabilities of the co-variates using an imprecise beta distribution to obtain a robust Bayesian variable selection routine. We applied our method to both synthetic datasets and real life datasets to check the efficiency of our method.

In Chapter 2, we introduced the notion of statistical modelling from a regressional point of view. We investigated linear regression and discussed its different properties followed by a general framework of uncertainty quantification. We briefly introduced the likelihood-based approach and the Bayesian approach for statistical inference to build the foundation of this thesis.

Chapter 3 was focused on the theoretical framework of the likelihood-based approaches for linear regression. We discussed the maximum likelihood estimation as a tool to perform parameter estimation, which laid the foundation for ordinary least squares and several other regularisation methods. We set up the idea of regularisation through ridge regression, which is one of the foremost works in high-dimensional statistical modelling. Despite being a popular method because of its simple implementation and closed form expressions, it is not particularly useful for achieving sparsity. To achieve sparsity, we require variable selection methods, which we discussed later on. We emphasised LASSO and other LASSO type problems because of their easy implementations and fast computation. The regularisation methods described in the thesis depend on the use of additional penalty term, which are often solved using a regularisation parameter. This regularisation parameter gives us different models based on its value and we require a model selection technique to find the best fit. Besides this, regularisation methods often require numerical optimisation and we may not have closed form expressions to obtain variance formulas. To overcome these issues, we need suitable model selection techniques and inference methods which have been discussed to assist the readers.

Chapter 4 was focused on theoretical aspects of numerical optimisation. We discussed the basic mathematics behind numerical optimisation and presented three different optimisation techniques for LASSO along with the convergence of these methods.

In Chapter 5, we discussed the theoretical aspects of variable selection methods for high-dimensional problems. For a consistent variable selection an estimator needs to satisfy several asymptotic conditions. These asymptotic conditions are termed as the oracle properties. LASSO, despite being a popular variable selection method, fails to satisfy the oracle properties. This led to several other works on variable selection, adaptive LASSO being one of them. We introduced the notion of the adaptive LASSO, which is a modification of LASSO and satisfies these oracle properties. The formulation of adaptive LASSO allows us to perform a sensitivity analysis over a set of weights. We discussed this sensitivity analysis along with novel error bounds, which help us to understand the effect of data driven weights in variable selection. The other important topic we discussed in this chapter is the binary credal classification under sparsity constraints. This is an extension of the adaptive penalised logistic regression. In binary credal classification, we obtain a robust classification routine through sensitivity analysis.

Chapter 6 was focused on the Bayesian approaches for linear modelling. We first discussed different types of prior for Bayesian analysis followed by the estimation techniques in Bayesian paradigm. Later on, we focused on different Bayesian modelling strategies for linear regression. We investigated the use of both informative priors and improper priors for linear models. In this thesis, we are particularly interested in high-dimensional models and variable selection strategies. To achieve sparsity, we need special types of hierarchical models to specify the regression coefficients. We discussed these hierarchical models in this chapter. At first, stochastic search variable selection was introduced, which achieves sparsity by introducing selection indicators in the model. This is a special version of spike and slab priors, which have been discussed as well. Besides this, we discussed the Bayesian LASSO, which is the direct Bayesian representation of LASSO.

The Bayesian paradigm allows us to incorporate our prior beliefs in an efficient way. However, in high-dimensional models, the severe uncertainty often results in different models based on different prior specifications. Therefore, we are interested in a robust Bayesian analysis which is performed through specifying a set of priors instead of a single prior. This robust Bayesian analysis was introduced in Chapter 7. We discussed the philosophy behind robust Bayesian analysis and its relevance in high-dimensional statistical modelling. After that, we introduced the imprecise beta model, which is an integral part of our robust Bayesian variable selection. Finally, we laid the foundation stone of our method by pointing out different sources of uncertainty in high-dimensional models and their remedies through robust Bayesian analysis.

Chapter 8 was dedicated to our novel robust Bayesian variable selection framework. We discussed our hierarchical model and the motivation behind our prior specifications. The choice of our prior selection probability plays an important role in robust variable selection, for which we use the imprecise beta prior. The robustness is achieved through a sensitivity analysis over a set of α. We used this α to specify our prior expectation of the selection probability. Our choice of conjugate priors gives us a nice framework for conditional analysis. We considered an orthogonal design case, which allowed us to decompose the joint posterior in a convenient way to perform conditional analysis. We discussed the posteriors for regression parameters and selection indicators through our analysis. We also provided a general framework for robust variable selection which can be achieved efficiently through a Gibbs sampling algorithm. Our robust Bayesian methodology gives us a set of posterior distributions instead of a single posterior, for which two measures have been introduced, which were used to capture the indeterminacy in model fitting.

Finally, an application of our methodology was shown in Chapter 9. In this chapter, we investigated three different real datasets to capture different aspects of high-dimensional modelling.

10.2 Discussion and future works

An important aspect of research work is to put light on the issues, which have not been tackled yet. This thesis investigates a novel robust Bayesian approach for variable selection, where we face some limitations on the modelling strategies and need to be improved. For instance, we need to find a suitable measure to evaluate prediction accuracy as well as the underlying imprecision. In this section, we briefly discuss some of these limitations of our work along with other promising areas of our research, which could be further investigated.

An important part of our variable selection method is to have a decision criterion for the posterior odds of the selection indicators. This is an interesting area of research, which we can develop further. In this thesis, we adapt the approach of George and McCulloch [41] to specify the active co-variates. George and McCulloch [41] considered the median probability of the selection indicators and checked if these are greater than 1/2. We adapt this decision rule in our robust Bayesian analysis by checking the posterior odds over the set of α. However, we may also consider this co-variate selection as a decision making problem and the notion of median probability can replaced by a more generalised utility-based decision rule.

Our methodology is based on the sensitivity analysis over the sets of prior selection probability but another important aspect of our method is the spike and slab prior specification. The choice of scale parameters in slab component of the model is very crucial. For the orthogonal design case, we can evaluate the effect of this scale parameter in variable selection through a closed form expression, however, this is not the case for general design. Extreme choices of this scale parameter contribute to more indeterminate variables. We were able to explain part of this effect but it is not fully understood yet. This remains as an open question where we would like to understand the behaviour of our model based on the specification of the slab component.

Our robust Bayesian method also raises research questions around prediction and model fitting. One aspect of linear regression is model fitting, where we are interested in the goodness of fit. In robust Bayesian analysis, we have a set of posteriors, which makes model fitting non-trivial. We introduced two different measures for this purpose. However, these are very crude ways of explaining goodness of fit as well as indeterminacy in model fitting. We would like to have a more sophisticated way of explaining these measures of accuracy, which can be compared with other methods as well. This is a very interesting aspect of robust Bayesian methodology and not just our variable selection routine. A unified measure of accuracy for goodness of fit will be beneficial for robust Bayesian analysis and will open the door of a more explicable comparison with other methods, where we don't have a set of posterior distributions or posterior estimates.

Moreover, we can exploit our robust Bayesian variable selection method to introduce modelling strategies for other types of regression models, especially when our regressors are continuous. Our hierarchical model with conjugate priors can be easily extended to other problems, which involve a likelihood from the exponential family of distributions. This opens the door for several future works in robust Bayesian variable selection, which we would like to explore in the future.

Appendix A

Proofs of Lemmas

Here, we provide proof of different lemmas and results, which we use in the thesis. The lemmas are well known and often can be found in the literature as statements. We aim to provide the proof for convenience and continuity of our proofs in the main text.

A.1 Likelihood-based Approaches

A.1.1 Invertible covariance matrices are postive definite

Let $\boldsymbol{x}$ be a $n \times p$ design matrix that is the matrix of predictors.

Lemma A.1. *If $\boldsymbol{x}^T\boldsymbol{x}$ is invertible then it is positive definite.*

Proof. Let, $v \in \mathbb{R}^p$ be a non zero vector. Then,

$$v^T\boldsymbol{x}^T\boldsymbol{x}v = (\boldsymbol{x}v)^T\boldsymbol{x}v = \|\boldsymbol{x}v\|_2^2 \geq 0. \tag{A.1}$$

That is $\boldsymbol{x}^T\boldsymbol{x}$ is positive semi definite. Now, since v is non zero vector therefore $\|\boldsymbol{x}v\|_2^2 = 0$ implies that columns of $\boldsymbol{x}$ are not linearly independent and $\boldsymbol{x}^T\boldsymbol{x}$ is not invertible.

This contradicts our assumption and therefore, $\|\boldsymbol{x}v\|_2^2 > 0$. That is $\boldsymbol{x}^T\boldsymbol{x}$ is positive definite when $\boldsymbol{x}^T\boldsymbol{x}$ is invertible.

□

A.1.2 Ridge estimates are root-n-consistent

Let $\| \cdot \|$ denote the matrix norm in the space $\mathbb{R}^{p \times p}$ such that, for any matrix A

$$\|A\| := \sup_{\|x\|_2=1} \{\|Ax\|_2 : x \in \mathbb{R}^p\}, \tag{A.2}$$

where $\| \cdot \|_2$ denotes the usual Euclidean norm in $\mathbb{R}^p$. Note that, $\|A\|$ is the largest eigenvalue of A.

Let $\{A_n\}_n$ be the sequence of matrices

$$A_n = \frac{1}{n} \left(\boldsymbol{x}^T \boldsymbol{x} + \lambda_n \mathbf{I}_p \right), \tag{A.3}$$

where $0 < \lambda_n < \infty$.

Lemma A.2. *The* $\lim_{n\to\infty} A_n^{-1}$ *exists and it equals to* Σ^{-1}.

Proof. To prove Lem. A.2, we first show that, $\lim_{n\to\infty} A_n$ exists and is equal to Σ.

$$\|A_n - \Sigma\| = \left\| \frac{1}{n} \left(\boldsymbol{x}^T \boldsymbol{x} + \lambda_n \mathbf{I}_p \right) - \Sigma \right\| \tag{A.4}$$

$$= \left\| \frac{1}{n} \boldsymbol{x}^T \boldsymbol{x} - \Sigma + \frac{\lambda_n}{n} \mathbf{I}_p \right\| \tag{A.5}$$

by applying triangle inequality in Eq. (A.5), ie. $\|a + b\| \leq \|a\| + \|b\|$, we get,

$$\|A_n - \Sigma\| \leq \left\| \frac{1}{n} \boldsymbol{x}^T \boldsymbol{x} - \Sigma \right\| + \left\| \frac{\lambda_n}{n} \mathbf{I}_p \right\| \tag{A.6}$$

$$= \left\| \frac{1}{n} \boldsymbol{x}^T \boldsymbol{x} - \Sigma \right\| + \frac{\lambda_n}{n}. \tag{A.7}$$

Now, as $n \to \infty$, $\lim_{n\to\infty} \frac{1}{n} \boldsymbol{x}^T \boldsymbol{x} = \Sigma$ Therefore,

$$\|A_n - \Sigma\| \to 0 \tag{A.8}$$

$$\implies \lim_{n\to\infty} A_n = \Sigma. \tag{A.9}$$

Since, $\{A_n\}_n$ is convergent, therefore it is a Cauchy sequence, that is, for every $\delta > 0$ there exists a positive natural number N such that for all natural numbers $m_1, m_2 > N$

$$\|A_{m_1} - A_{m_2}\| < \delta. \tag{A.10}$$

Now, since, A_n is sum of a positive semi-definite matrix $(\frac{1}{n} \boldsymbol{x}^T \boldsymbol{x})$ and a diagonal matrix with positive entries $(\lambda_n \mathbf{I}_p)$, it is easy to see that A_n is positive definite.

Then, the inverse A_n^{-1} exists. Let, $A_n = U_n D_n U_n^T$ where, D_n is a diagonal matrix and U_n is orthogonal. Now,

$$\|A_n^{-1}\| = \left\|(U_n D_n U_n^T)^{-1}\right\| \tag{A.11}$$

$$= \left\|(U_n D_n^{-1} U_n^T)\right\| \tag{A.12}$$

since, U_n is orthogonal and D_n is diagonal, we get,

$$\|A_n^{-1}\| = \sup_{1\leq j\leq p} \{[D_n^{-1}]_{jj}\} \tag{A.13}$$

$$= \frac{1}{\inf_{1\leq j\leq p}\{[D_n]_{jj}\}}. \tag{A.14}$$

As, $A_n = \frac{1}{n}\left(\boldsymbol{x}^T\boldsymbol{x} + \lambda_n \mathbf{I}_p\right)$ is positive definite, therefore all of its eigen values are greater than or equal to λ_n. Therefore,

$$\|A_n^{-1}\| \leq \frac{1}{\lambda_n}. \tag{A.15}$$

Then,

$$A_{m_1}^{-1} - A_{m_2}^{-1} = A_{m_1}^{-1} A_{m_2} A_{m_2}^{-1} - A_{m_1}^{-1} A_{m_1} A_{m_2}^{-1} \tag{A.16}$$

$$= A_{m_1}^{-1}\left(A_{m_2} - A_{m_1}\right) A_{m_2}^{-1} \tag{A.17}$$

$$\left\|A_{m_1}^{-1} - A_{m_2}^{-1}\right\| = \left\|A_{m_1}^{-1}\left(A_{m_2} - A_{m_1}\right) A_{m_2}^{-1}\right\| \tag{A.18}$$

applying the Cauchy-Schwartz inequality we get,

$$\left\|A_{m_1}^{-1} - A_{m_2}^{-1}\right\| \leq \left\|A_{m_1}^{-1}\right\| \left\|A_{m_2} - A_{m_1}\right\| \left\|A_{m_2}^{-1}\right\| \tag{A.19}$$

using Eq. (A.10),

$$\left\|A_{m_1}^{-1} - A_{m_2}^{-1}\right\| \leq \delta \left\|A_{m_1}^{-1}\right\| \left\|A_{m_2}^{-1}\right\| \tag{A.20}$$

$$\leq \frac{\delta}{\lambda_n^2}. \tag{A.21}$$

Therefore, for every $\frac{\delta}{\lambda_n^2} > 0$, we can find a positive natural number N, such that for every $m_1, m_2 > 0$, $\left\|A_{m_1}^{-1} - A_{m_2}^{-1}\right\| \leq \frac{\delta}{\lambda_n^2}$. Hence, $\{A_n^{-1}\}_n$ is a Cauchy sequence. Since, $\mathbb{R}^p$ is a Banach space under the Euclidean norm $\|\cdot\|_2$, therefore every Cauchy sequence is convergent. Then there exist L such that, $\lim_{n\to\infty} A_n^{-1} = L$. Now,

$$A_n A_n^{-1} = \mathbf{I}_p = A_n^{-1} A_n \tag{A.22}$$

$$\lim_{n\to\infty} A_n A_n^{-1} = \mathbf{I}_p = \lim_{n\to\infty} A_n^{-1} A_n \tag{A.23}$$

since both A_n and A_n^{-1} is convergent,

$$\lim_{n\to\infty} A_n \cdot \lim_{n\to\infty} A_n^{-1} = \mathbf{I}_p = \lim_{n\to\infty} A_n^{-1} \cdot \lim_{n\to\infty} A_n \tag{A.24}$$

$$\Sigma \cdot L = \mathbf{I}_p = L \cdot \Sigma \tag{A.25}$$

Therefore, $\lim_{n\to\infty} A_n^{-1} = \Sigma^{-1}$. □

Recall the Ridge estimates in Eq. (3.12) given by:

$$\hat{\beta}_{\mathrm{R}}(\lambda_n) := \arg\min_{\beta} \left(\frac{1}{2} \|Y - \boldsymbol{x}\beta\|_2^2 + \lambda_n \|\beta\|_2^2 \right). \tag{A.26}$$

Lemma A.3. *Let λ_n be sequence of regularisation parameters such that $\frac{\lambda_n}{\sqrt{n}} \to 0$ as $n \to \infty$, then the ridge estimates are root n-consistent.*

Proof. Let $A_n = \frac{1}{n}\left(\boldsymbol{x}^T\boldsymbol{x} + \lambda_n \mathbf{I}_p\right)$

$$\hat{\beta}_{\mathrm{R}}(\lambda_n) = (nA_n)^{-1} \boldsymbol{x}^T(\boldsymbol{x}\beta + \epsilon) \tag{A.27}$$

$$= (nA_n)^{-1} \boldsymbol{x}^T\boldsymbol{x}\beta + (nA_n)^{-1} \boldsymbol{x}^T\epsilon \tag{A.28}$$

We know that, $\mathbb{E}[\boldsymbol{x}^T\epsilon \mid \boldsymbol{x}] = 0$. Therefore, conditioning on $\boldsymbol{x}$, we get

$$\mathbb{E}[\hat{\beta}_{\mathrm{R}}(\lambda_n) - \beta \mid \boldsymbol{x}] = (nA_n)^{-1} \boldsymbol{x}^T\boldsymbol{x}\beta - \beta \tag{A.29}$$

$$= (nA_n)^{-1} (nA_n - \lambda_n \mathbf{I}_p) \beta - \beta \tag{A.30}$$

$$= \beta - (nA_n)^{-1} \lambda_n \beta - \beta \tag{A.31}$$

$$= -\lambda_n (nA_n)^{-1} \beta. \tag{A.32}$$

Multiplying $\sqrt{n}$ on both sides,

$$\mathbb{E}[\sqrt{n}(\hat{\beta}_{\mathrm{R}}(\lambda_n) - \beta) \mid \boldsymbol{x}] = -\sqrt{n}\lambda_n (nA_n)^{-1} \beta \tag{A.33}$$

$$= -\frac{\sqrt{n}}{n}\lambda_n A_n^{-1}\beta \tag{A.34}$$

$$= -\frac{\lambda_n}{\sqrt{n}} A_n^{-1}\beta \tag{A.35}$$

Now, as $n \to \infty$, from Eq. (A.35), we get:

$$\lim_{n\to\infty} \mathbb{E}[\sqrt{n}(\hat{\beta}_{\mathrm{R}}(\lambda_n) - \beta) \mid \boldsymbol{x}] = \lim_{n\to\infty} -\frac{\lambda_n}{\sqrt{n}} A_n^{-1}\beta \tag{A.36}$$

since, by Lem. A.2, $\lim_{n\to\infty} A_n^{-1}$ exists and β is independent of n, therefore using product rule of limits we get

$$= -\beta \lim_{n\to\infty} \frac{\lambda_n}{\sqrt{n}} \lim_{n\to\infty} A_n^{-1} \tag{A.37}$$

$$= -\beta \cdot 0 \cdot \Sigma^{-1} \tag{A.38}$$

$$= 0. \tag{A.39}$$

This proves $\sqrt{n}\left(\hat{\beta}_{\mathrm{R}}(\lambda_n) - \beta\right)$ is asymptotically unbiased. Along the same lines it would be possible to show that $n^s\left(\hat{\beta}_{\mathrm{R}}(\lambda_n) - \beta\right)$ is asymptotically unbiased for any $0 \le s < 1$ under suitable convergence criterion for λ_n.

As before, conditioning on $\boldsymbol{x}$, we get:

$$\begin{aligned}
\mathrm{Var}\left[\sqrt{n}\left(\hat{\beta}_{\mathrm{R}}(\lambda_n) - \beta\right) \mid \boldsymbol{x}\right] &= \mathrm{Var}\left[\sqrt{n}\left((nA_n)^{-1}\boldsymbol{x}^T\epsilon\right)\right] && \text{(A.40)}\\
&= \mathrm{Var}\left[\frac{\sqrt{n}}{n} A_n^{-1}\boldsymbol{x}^T\epsilon\right] && \text{(A.41)}\\
&= \mathrm{Var}\left[A_n^{-1}\left(\frac{1}{\sqrt{n}}\boldsymbol{x}^T\epsilon\right)\right] && \text{(A.42)}\\
&= A_n^{-1}\mathrm{Var}\left[\left(\frac{1}{\sqrt{n}}\boldsymbol{x}^T\epsilon\right)\right] \cdot \left(A_n^{-1}\right)^T && \text{(A.43)}
\end{aligned}$$

since, $A_n^{-1} = \left(\frac{1}{n}\boldsymbol{x}^T\boldsymbol{x} + \frac{1}{n}\lambda_n\mathbf{I}_p\right)^{-1}$ is symmetric

$$= A_n^{-1} \cdot \frac{1}{n}\boldsymbol{x}^T\boldsymbol{x} \cdot \mathrm{Var}[\epsilon] \cdot A_n^{-1}. \tag{A.44}$$

Now, as $n \to \infty$,

$$\lim_{n\to\infty} \mathrm{Var}\left[\sqrt{n}\left(\hat{\beta}_{\mathrm{R}}(\lambda_n) - \beta\right) \mid \boldsymbol{x}\right] = \lim_{n\to\infty} A_n^{-1} \cdot \frac{1}{n}\boldsymbol{x}^T\boldsymbol{x} \cdot \mathrm{Var}[\epsilon] \cdot A_n^{-1}. \tag{A.45}$$

Since, by Lem. A.2, $\lim_{n\to\infty} A_n^{-1}$ exists and $\lim_{n\to\infty} \frac{1}{n}\boldsymbol{x}^T\boldsymbol{x}$ exists by assumption, therefore applying product rule of limits, we get:

$$= \mathrm{Var}[\epsilon] \cdot \lim_{n\to\infty} A_n^{-1} \cdot \lim_{n\to\infty} \frac{1}{n}\boldsymbol{x}^T\boldsymbol{x} \cdot \lim_{n\to\infty} A_n^{-1} \tag{A.46}$$

$$= \sigma^2\Sigma^{-1}\Sigma\Sigma^{-1} \tag{A.47}$$

$$= \sigma^2\Sigma^{-1}. \tag{A.48}$$

Now, by the central limit theorem we know,

$$\sqrt{n}\left(\hat{\beta}_{\mathrm{R}}(\lambda_n) - \beta\right) \xrightarrow{d} \mathcal{N}\left(\mathbb{E}\left[\sqrt{n}\left(\hat{\beta}_{\mathrm{R}}(\lambda_n) - \beta\right)\right], \mathrm{Var}\left[\sqrt{n}\left(\hat{\beta}_{\mathrm{R}}(\lambda_n) - \beta\right)\right]\right). \tag{A.49}$$

Now, applying Eq. (A.28), Eq. (A.35) and Eq. (A.48) we get,

$$\sqrt{n}\left(\hat{\beta}_{\mathrm{R}}(\lambda_n) - \beta\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\Sigma^{-1}\right). \tag{A.50}$$

□

A.2 Robust Bayesian Variable Selection

A.2.1 Variance formula for mixture of distributions

Lemma A.4. *Let*

$$X \sim w_1 f_1 + w_2 f_2 \tag{A.51}$$

where f_i denotes a normal density with mean μ_i and variance σ_i^2 for $i = 1, 2$. Then,

$$Var(X) = \sum_{i=1}^{2} w_i(\sigma_i^2 + \mu_i^2) - \left(\sum_{i=1}^{2} w_i\mu_i\right)^2 \tag{A.52}$$

Proof. First, note that

$$E(X^2) = \int x^2[w_1 f_1(x) + w_2 f_2(x)]dx \tag{A.53}$$

$$= w_1 \int x^2 f_1(x)dx + w_2 \int x^2 f_2(x)dx \tag{A.54}$$

$$= w_1(\sigma_1^2 + \mu_1^2) + w_2(\sigma_2^2 + \mu_2^2). \tag{A.55}$$

Consequently, Then, the variance of X is given by:

$$\mathrm{Var}(X) = E(X^2) - [E(X)]^2 \tag{A.56}$$

$$= \sum_{i=1}^{2} w_i(\sigma_i^2 + \mu_i^2) - \left(\sum_{i=1}^{2} w_i\mu_i\right)^2. \tag{A.57}$$

□

Bibliography

[1] A. Agresti. *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley, 2013. ISBN 9780470463635. URL `https://books.google.co.uk/books?id=UOrr47-2oisC`.

[2] Zakariya Yahya Algamal and Muhammad Hisyam Lee. Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification. *Expert Systems with Applications*, 42(23):9326 – 9332, 2015. ISSN 0957-4174. doi: 10.1016/j.eswa.2015.08.016.

[3] Ash A. Alizadeh, Michael B. Eisen, R. Eric Davis, Chi Ma, Izidore S. Lossos, Andreas Rosenwald, Jennifer C. Boldrick, Hajeer Sabet, Truc Tran, Xin Yu, John I. Powell, Liming Yang, Gerald E. Marti, Troy Moore, James Hudson, Lisheng Lu, David B. Lewis, Robert Tibshirani, Gavin Sherlock, Wing C. Chan, Timothy C. Greiner, Dennis D. Weisenburger, James O. Armitage, Roger Warnke, Ronald Levy, Wyndham Wilson, Michael R. Grever, John C. Byrd, David Botstein, Patrick O. Brown, and Louis M. Staudt. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, Feb 2000. ISSN 1476-4687. doi: 10.1038/35000501. URL `10.1038/35000501`.

[4] C. A. L. Bailer-Jones. The ILIUM forward modelling algorithm for multivariate parameter estimation and its application to derive stellar parameters from Gaia spectrophotometry. *Monthly Notices of the Royal Astronomical Society*, 403(1): 96–116, 2010. doi: 10.1111/j.1365-2966.2009.16125.x.

[5] Tathagata Basu, Jochen Einbeck, and Matthias Troffaes. A sensitivity analysis

and error bounds for the adaptive lasso. In I. Irigoien, D.-J. Lee, J. Martinez-Minaya, and M.X. Rodriguez-Alvarez, editors, *Proceedings of the 35th International Workshop on Statistical Modelling.*, pages 278–281. Universidad del Pais Vasco, 2020. URL `http://dro.dur.ac.uk/31805/`.

[6] Tathagata Basu, Matthias C. M. Troffaes, and Jochen Einbeck. Binary credal classification under sparsity constraints. In Marie-Jeanne Lesot, Susana Vieira, Marek Z. Reformat, João Paulo Carvalho, Anna Wilbik, Bernadette Bouchon-Meunier, and Ronald R. Yager, editors, *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 82–95, Cham, 2020. Springer International Publishing. ISBN 978-3-030-50143-3. doi: 10.1007/978-3-030-50143-3_7.

[7] Thomas Bayes and Richard Price. LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1763. doi: 10.1098/rstl.1763.0053.

[8] A. Beck and M. Teboulle. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009. doi: 10.1137/080716542.

[9] James O Berger. *Statistical Decision Theory and Bayesian Analysis; 2nd ed.* Springer Series in Statistics. Springer, New York, 1985. doi: 10.1007/978-1-4757-4286-2.

[10] James O. Berger. Robust Bayesian analysis: sensitivity to the prior. *Journal of Statistical Planning and Inference*, 25(3):303 – 328, 1990. ISSN 0378-3758. doi: 10.1016/0378-3758(90)90079-A.

[11] Anirban Bhattacharya, Debdeep Pati, Natesh S. Pillai, and David B. Dunson. Dirichlet-Laplace Priors for Optimal Shrinkage. *Journal of the American Statistical Association*, 110(512):1479–1490, 2015. doi: 10.1080/01621459.2014.960967. PMID: 27019543.

[12] Miķelis Bickis. The Imprecise Logit-Normal Model and its Application to Estimating Hazard Functions. *Journal of Statistical Theory and Practice*, 3(1): 183–195, Mar 2009. doi: 10.1080/15598608.2009.10411919.

[13] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004. URL `http://www.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf`.

[14] Leo Breiman. Better Subset Regression Using the Nonnegative Garrote. *Technometrics*, 37(4):373–384, 1995. ISSN 00401706. URL `http://www.jstor.org/stable/1269730`.

[15] G. Casella and R.L. Berger. *Statistical Inference.* Duxbury advanced series in statistics and decision sciences. Thomson Learning, 2002. ISBN 9780534243128. URL `https://books.google.co.in/books?id=0x_vAAAAMAAJ`.

[16] George Casella and Edward I. George. Explaining the Gibbs Sampler. *The American Statistician*, 46(3):167–174, 1992. ISSN 00031305. URL `http://www.jstor.org/stable/2685208`.

[17] Ismal Castillo, Johannes Schmidt-Hieber, and Aad van der Vaart. Bayesian linear regression with sparse priors. *Ann. Statist.*, 43(5):1986–2018, 10 2015. doi: 10.1214/15-AOS1334.

[18] Anthony Christidis, Stefan Van Aelst, and Ruben Zamar. *nnGarrote: Non-Negative Garrote Estimation with Penalized Initial Estimators*, 2020. URL `https://CRAN.R-project.org/package=nnGarrote`. R package version 1.0.3.

[19] G. Corani and A. Antonucci. Credal ensembles of classifiers. *Computational Statistics & Data Analysis*, 71:818 – 831, 2014. ISSN 0167-9473. doi: 10.1016/j.csda.2012.11.010.

[20] G. Corani and C. P. de Campos. A tree augmented classifier based on Extreme Imprecise Dirichlet Model. *International Journal of Approximate Reasoning*, 51(9):1053 – 1068, 2010. ISSN 0888-613X. doi: 10.1016/j.ijar.2010.08.007.

[21] Giorgio Corani and Marco Zaffalon. Learning Reliable Classifiers From Small or Incomplete Data Sets: The Naive Credal Classifier 2. *J. Mach. Learn. Res.*, 9:581–621, 2008.

[22] D.R. Cox. *Principles of Statistical Inference.* Cambridge University Press, 2006. ISBN 9780521685672. URL `https://books.google.co.in/books?id=EMYWpoVn7vcC`.

[23] Erika Cule and Maria De Iorio. A semi-automatic method to guide the choice of ridge parameter in ridge regression, 2012.

[24] Erika Cule, Steffen Moritz, and Dan Frankowski. *ridge: Ridge Regression with Automatic Selection of the Penalty Parameter*, 2020. URL `https://CRAN.R-project.org/package=ridge`. R package version 2.5.

[25] R. Davidson. *Econometric Theory and Methods: International Edition.* OUP Oxford, 2009. ISBN 9780195391053. URL `https://books.google.co.in/books?id=wTP_RAAACAAJ`.

[26] Norman R. Draper and Harry Smith. *Fitting a Straight Line by Least Squares: Applied Regression Analysis*, pages 15–46. John Wiley & Sons, Inc., 1998. ISBN 9781118625590. doi: 10.1002/9781118625590.ch1.

[27] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

[28] A.W.F. Edwards. *Likelihood.* Cambridge science classics. Cambridge University Press, 1984. ISBN 9780521318716. URL `https://books.google.co.in/books?id=LL08AAAAIAAJ`.

[29] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Ann. Statist.*, 32(2):407–499, April 2004. doi: 10.1214/009053604000000067.

[30] Jochen Einbeck and Ludger Evers. *LPCM: Local Principal Curve Methods*, 2019. URL `https://CRAN.R-project.org/package=LPCM`. R package version 0.46-3.

[31] Jochen Einbeck, Ludger Evers, and Coryn Bailer-Jones. Representing Complex Data Using Localized Principal Components with Application to Astronomical Data. In Alexander N. Gorban, Balázs Kégl, Donald C. Wunsch, and Andrei Y. Zinovyev, editors, *Principal Manifolds for Data Visualization and Dimension Reduction*, pages 178–201, Berlin, Heidelberg, 2008. Springer. ISBN 978-3-540-73750-6.

[32] ESA. Science & Technology: Gaia. http://sci.esa.int/gaia, 2013. Accessed: 2020-07-11.

[33] B.S. Everitt. *Cambridge Dictionary of Statistics*. Cambridge University Press, 2006. ISBN 9780521860390.

[34] Jianqing Fan and Runze Li. Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001. ISSN 01621459. URL http://www.jstor.org/stable/3085904.

[35] Jerome Friedman, Trevor Hastie, Holger Hofling, and Robert Tibshirani. Pathwise coordinate optimization. *Ann. Appl. Stat.*, 1(2):302–332, December 2007. doi: 10.1214/07-AOAS131.

[36] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL http://www.jstatsoft.org/v33/i01/.

[37] Paul H Garthwaite, Joseph B Kadane, and Anthony O'Hagan. Statistical Methods for Eliciting Probability Distributions. *Journal of the American Statistical Association*, 100(470):680–701, 2005. doi: 10.1198/016214505000000105.

[38] Alan E. Gelfand and Adrian F. M. Smith. Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, 85 (410):398–409, 1990. ISSN 01621459. URL http://www.jstor.org/stable/2289776.

[39] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

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

[41] Edward I. George and Robert E. McCulloch. Variable Selection Via Gibbs Sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993. ISSN 01621459. URL `http://www.jstor.org/stable/2290777`.

[42] R. Paul Gorman and Terrence J. Sejnowski. Analysis of hidden units in a layered network trained to classify sonar targets. *Neural Networks*, 1(1):75–89, 1988. ISSN 0893-6080. doi: 10.1016/0893-6080(88)90023-8.

[43] Robert B. Gramacy. *monomvn: Estimation for Multivariate Normal and Student-t Data with Monotone Missingness*, 2017. R package version 1.9-7.

[44] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, 2015. ISBN 9781498712170. URL `https://books.google.co.uk/books?id=f-A_CQAAQBAJ`.

[45] Trevor Hastie and Brad Efron. *lars: Least Angle Regression, Lasso and Forward Stagewise*, 2013. URL `https://CRAN.R-project.org/package=lars`. R package version 1.2.

[46] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[47] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 04 1970. ISSN 0006-3444. doi: 10.1093/biomet/57.1.97.

[48] Hemant Ishwaran and J. Sunil Rao. Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, 33(2):730773, Apr 2005. ISSN 0090-5364. doi: 10.1214/009053604000001147.

[49] Juan José del Coz, Jorge Díez, and Antonio Bahamonde. Learning Nondeterministic Classifiers. *J. Mach. Learn. Res.*, 10:22732293, December 2009. ISSN 1532-4435.

[50] E. L. Lehmann. On the history and use of some standard statistical models. In Deborah Nolan and Terry Speed, editors, *Probability and Statistics: Essays in Honor of David A. Freedman*, volume Volume 2 of *Collections*, pages 114–126. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2008. doi: 10.1214/193940307000000419.

[51] E.L. Lehmann and G. Casella. *Theory of Point Estimation.* Springer Texts in Statistics. Springer New York, 2011. ISBN 9781441931306. URL `https://books.google.co.in/books?id=WaBQcgAACAAJ`.

[52] Richard A. Levine and George Casella. Implementations of the Monte Carlo EM Algorithm. *Journal of Computational and Graphical Statistics*, 10(3):422–439, 2001. ISSN 10618600. URL `http://www.jstor.org/stable/1391097`.

[53] Runze Li, Liying Huang, and John Dziak. *VariableScreening: High-Dimensional Screening for Semiparametric Longitudinal Regression*, 2018. URL `https://CRAN.R-project.org/package=VariableScreening`. R package version 0.2.0.

[54] Anastasia Lykou and Ioannis Ntzoufras. On Bayesian lasso variable selection and the specification of the shrinkage parameter. *Statistics and Computing*, 23 (3):361–390, May 2013. ISSN 1573-1375. doi: 10.1007/s11222-012-9316-x.

[55] M. E. Maron. Automatic Indexing: An Experimental Inquiry. *J. ACM*, 8(3): 404417, July 1961. ISSN 0004-5411. doi: 10.1145/321075.321084.

[56] T. J. Mitchell and J. J. Beauchamp. Bayesian Variable Selection in Linear

Regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988. doi: 10.1080/01621459.1988.10478694.

[57] Naveen Naidu Narisetty and Xuming He. Bayesian variable selection with shrinking and diffusing priors. *Ann. Statist.*, 42(2):789–817, 04 2014. doi: 10.1214/14-AOS1207.

[58] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Publishing Company, Incorporated, 1st edition, 2014. ISBN 1461346916, 9781461346913.

[59] S original, from StatLib, and by Rob Tibshirani. R port by Friedrich Leisch. *bootstrap: Functions for the Book "An Introduction to the Bootstrap"*, 2019. URL `https://CRAN.R-project.org/package=bootstrap`. R package version 2019.6.

[60] Trevor Park and George Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008. doi: 10.1198/016214508000000337.

[61] Lewis Paton, Matthias C. M. Troffaes, Nigel Boatman, Mohamud Hussein, and Andy Hart. Multinomial Logistic Regression on Markov Chains for Crop Rotation Modelling. In Anne Laurent, Oliver Strauss, Bernadette Bouchon-Meunier, and Ronald R. Yager, editors, *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 476–485, Cham, 2014. Springer International Publishing. doi: 10.1007/978-3-319-08852-5_49.

[62] Lewis Paton, Matthias C. M. Troffaes, Nigel Boatman, Mohamud Hussein, and Andy Hart. A Robust Bayesian Analysis of the Impact of Policy Decisions on Crop Rotations. In T. Augustin, S. Doria, E. Miranda, and E. Quaeghebeur, editors, *ISIPTA '15: proceedings of the 9th International Symposium on Imprecise Probability : Theories and Applications, 20-24 July 2015, Pescara, Italy*, pages 217–226. SIPTA, July 2015. URL `http://dro.dur.ac.uk/15736/`.

[63] R Core Team. *R: A Language and Environment for Statistical Computing.* R

Foundation for Statistical Computing, Vienna, Austria, 2019. URL `https://www.R-project.org/`.

[64] Veronika Ročková and Edward I. George. The Spike-and-Slab LASSO. *Journal of the American Statistical Association*, 113(521):431–444, 2018. doi: 10.1080/01621459.2016.1260469.

[65] Ankan Saha and Ambuj Tewari. On the Finite Time Convergence of Cyclic Coordinate Descent Methods. *CoRR*, abs/1005.2146, 2010.

[66] K. Sauer and C. Bouman. A Local Update Strategy for Iterative Reconstruction from Projections. *Trans. Sig. Proc.*, 41(2):534–548, February 1993. ISSN 1053-587X. doi: 10.1109/78.193196.

[67] S.K. Shevade and S.S. Keerthi. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, 19(17):2246–2253, 2003. doi: 10.1093/bioinformatics/btg308.

[68] N. Z. Shor, Krzysztof C. Kiwiel, and Andrzej Ruszcayǹski. *Minimization Methods for Non-differentiable Functions*. Springer-Verlag, Berlin, Heidelberg, 1985. ISBN 0-387-12763-1.

[69] Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288, 1996. ISSN 00359246. URL `http://www.jstor.org/stable/2346178`.

[70] A. N. Tikhonov. On the solution of ill-posed problems and the method of regularization. *Dokl. Akad. Nauk SSSR*, 151(3):501–504, 1963. URL `http://www.ams.org/mathscinet-getitem?mr=0162377`.

[71] A. Tsanas, M. A. Little, C. Fox, and L. O. Ramig. Objective Automatic Assessment of Rehabilitative Speech Treatment in Parkinson's Disease. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 22(1):181–190, 2014.

[72] Athanasios Tsanas. UCI machine learning repository, 2014. URL `https://archive.ics.uci.edu/ml/datasets/LSVT+Voice+Rehabilitation`.

[73] P. Tseng. Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization. *J. Optim. Theory Appl.*, 109(3):475494, June 2001. ISSN 0022-3239. doi: 10.1023/A:1017501703105.

[74] Sara Van de Geer, Peter Bühlmann, and Shuheng Zhou. The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso). *Electron. J. Statist.*, 5:688–749, 2011. doi: 10.1214/11-EJS624.

[75] Sara A. Van De Geer and Peter Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat*, 2009.

[76] P. Walley. *Statistical Reasoning with Imprecise Probabilities.* Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1991. ISBN 9780412286605. URL `https://books.google.co.uk/books?id=-hbvAAAAMAAJ`.

[77] Peter Walley. Inferences from Multinomial Data: Learning about a Bag of Marbles. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1): 3–57, 1996. ISSN 00359246. URL `http://www.jstor.org/stable/2346164`.

[78] Gero Walter, Louis J.M. Aslett, and Frank P.A. Coolen. Bayesian nonparametric system reliability using sets of priors. *International Journal of Approximate Reasoning*, 80:67 – 88, 2017. ISSN 0888-613X. doi: 10.1016/j.ijar.2016.08.005.

[79] Qingyan Xiang and Naveen Narisetty. *basad: Bayesian Variable Selection with Shrinking and Diffusing Priors*, 2017. URL `https://CRAN.R-project.org/package=basad`. R package version 0.2.0.

[80] Ming Yuan and Yi Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007. ISSN 00063444. URL `http://www.jstor.org/stable/20441351`.

[81] Ming Yuan and Yi Lin. On the nonnegative garrotte estimator. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):143–161, 2007. doi: 10.1111/j.1467-9868.2007.00581.x.

[82] Marco Zaffalon. The naive credal classifier. *Journal of Statistical Planning and Inference*, 105(1):5 – 21, 2002. ISSN 0378-3758. doi: 10.1016/S0378-3758(01)00201-4. Imprecise Probability Models and their Applications.

[83] Marco Zaffalon, Giorgio Corani, and Denis Mau. Evaluating credal classifiers by utility–discounted predictive accuracy. *International Journal of Approximate Reasoning*, 53(8):1282 – 1301, 2012. ISSN 0888-613X. doi: 10.1016/j.ijar.2012.06.022. Imprecise Probability: Theories and Applications (ISIPTA'11).

[84] Peng Zhao and Bin Yu. On Model Selection Consistency of Lasso. *J. Mach. Learn. Res.*, 7:2541–2563, December 2006. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=1248547.1248637`.

[85] J. Zhu and T. Hastie. Classification of gene microarrays by penalized logistic regression. *Biostatistics*, 5(3):427–443, 2004. doi: 10.1093/biostatistics/kxg046.

[86] Hui Zou. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006. doi: 10.1198/016214506000000735.