Abstract

Include result summary!
This study explores the challenges inherent in variable selection for linear models in Bayesian statistics, particularly in high-dimensionality settings using R packages. The term ‘high-dimensionality’ here refers to contexts involving many predictors that are fewer, equal to, or greater than the number of data points. Frequentist penalised regression LASSO and Elastic Net methods and the Machine Learning XGBoost method are initially used to establish a benchmark. Subsequently, the Bayesian methods used include the Bayesian Lasso, Spike and Slab prior, Horseshoe priors, Laplace Approximation, and the recently revised Shotgun Stochastic Search with Screening. The packages that involve the same methodology are compared to establish user-friendliness and get insights into different approaches. The thesis compares the methodology using simulated scenarios and the socioeconomic crime dataset.

Contents

1 Acknowledgements

I am profoundly grateful to my advisor, Dr. Michail Papathomas, whose guidance was instrumental throughout my dissertation journey. His insights and constructive feedback on my draft greatly contributed to my work. My sincere appreciation goes to the Computer Science laboratory for providing a conducive environment that promoted privacy and cheerful atmosphere, free from the constraints of societal norms. A special mention goes to the new friends I have made over the last year - their support has been invaluable, and they will forever hold a special place in my heart. Finally, I would like to express my gratitude to the beaches of St Andrews, whose breathtaking beauty served as a constant source of inspiration and tranquility during this period.

2 Introduction

2.1 Personal Interest

The motivation to delve into the intricacies of the Bayesian statistical framework was kindled by Nobel laureate Daniel Kahneman's seminal and highly digestible book 'Thinking Fast and Slow' (Kahneman, 2011). Kahneman's dissection of human cognitive biases and flawed decision-making, especially in ambiguous situations, characterises System 1 thinking as heuristic-driven, low-cost, and low-effort, often resulting in pitfalls of emotionally induced biases. It is posited that the Bayesian methodology parallels System 2 thinking, where initial prior beliefs are systematically updated with new evidence, considering the strength of the information. The Bayesian approach provides an explicit mathematical framework for handling uncertainty, counterbalancing overconfidence and confirmation bias by quantifying uncertainties in light of new evidence.

2.2 Concept Comparison: Bayesian Against Frequentist Against Machine Learning

Bayesian statistics employs probability distributions, constructed using probability theory, to describe the 'degree of belief'. Its strengths and weaknesses simultaneously lie in the flexible incorporation of background information. Arguably, it presents a more organic approach to scientific reasoning than frequentist methods by continually updating probability distributions with new data. Focusing solely on the likelihood of data under specific hypotheses, without incorporating prior beliefs, leads to a rigid inference process that lacks the capability for iterative updating of beliefs. Prior beliefs are particularly advantageous when data is scarce or hard to acquire; using prior information might be the only option. Even with a non-informative prior, Bayesian methodology yields a distribution in contrast to the classical approach's point estimate, making it more intuitively understandable.

Bayesian approaches elegantly circumvent challenges encountered by classical methods. They sidestep issues of functional maximisation, which eliminates problems with algorithm convergence and the delicate task of choosing starting values close to the maximum (Train, 2012). Further, Bayesian methods alleviate the dilemma between local and global maxima, as convergence does not inherently imply the attainment of a global maximum. Moreover, they allow for desirable consistency and efficiency under more forgiving conditions: consistency is achievable with a fixed number of simulation draws, and any increase in draw numbers with sample size ensures efficiency.

2.3 MACHINE LEARNING

2.4 Motivation

In the realm of data-driven decision-making, substantial emphasis is placed on the abundance of data. This idea follows the mathematical principle that the number of observations should generally exceed the number of explanatory variables, which aids in preventing overfitting in models and enhances predictive power. Technological advancements have quickly fueled scientific progress, enabling us to amass large volumes of data, often with the number of variables greatly exceeding the number of data points. One of the critical challenges in this high-dimensional setting is to impose sparsity, a strategy that promotes model simplicity and interpretability without sacrificing significant predictive accuracy, thereby addressing the variance and overfitting issues inherent in high-dimensional data (Hastie et al., 2017). For instance, thousands of noisy pixel observations may exist in astronomy and image processing, but only a tiny subset is typically required to identify the objects of interest (Johnstone & Silverman, 2004). Meanwhile, in medical research involving rare diseases or novel treatments, data is often scarce, i.e. there are few data points. In those instances, the Central Limit Theorem (CLT) of normality sometimes needs to be more appropriately invoked to make assumptions about the underlying distribution of the data, despite insufficient sample sizes for the theorem to hold accurately.

A significant and well-known challenge in high-dimensional model selection is the issue of collinearity among predictors (Jianqing et al., 2010). This collinearity can often be misleading, especially in high-dimensional geometry (Fan & Lv, 2008), leading to the selection of an inaccurate model.

This thesis investigates high-dimensionality linear regression variable selection problems in both simulated data, where the number of predictors is smaller, equal, greater or even much greater than the number of data points, and the real data with a large number of predictors and a larger number of data points. Both will allow the exploration of issues in high dimensionality. The methods used encompass a selected handful of Bayesian approaches, classical penalised regression techniques, and a machine learning ensemble tree, all implemented using R software. It should be noted that some methodologies deployed in this study are not the original versions but extensions found within the implemented packages. This approach is intended to simplify the reader's narrative and illuminate the adaptations required to overcome computational limitations, enhance methodological efficiency and improve performance outcomes. In doing so, this study provides insights into which adjustments have proved most effective. Furthermore, comparative analyses of different packages for some methodologies are undertaken.

The aims of this thesis extend beyond applying and comparing a variety of Bayesian variable selection methods in linear regression problems. Equally important is the personal journey into the depth of this statistical framework. Prior to my Master's degree in Applied Statistics and Data Mining, my academic foundation was rooted in a creative discipline. Hence, this exploration of statistical frameworks stands not only as a scholarly endeavour but also as a pivotal chapter in my academic transition and growth. Hence, this document delivers a somewhat more extensive description of the methods, blending theory with application, to foster a deep understanding of the methodology and its practical testing.

2.5 Structure of Thesis

This thesis is structured as follows: Chapter 1 introduced a succinct overview of dissertation coupled with a reflection on personal interests. Chapter 2 delves into a more detailed and formally-structured motivation for variable selection in high-dimensional setting. Chapter 3 lays the foundation of Bayesian inference theory. Chapter 4 specifically overviews the variable selection methods within the Bayesian framework providing mathematical scaffolding and application-motivated package overview in R. Chapters 5 and 6 explore the application of variable and model selection in simulated and real high-dimensional data, with focus on methods that utilise shrinkage priors and penalised regression techniques. Finally, Chapter 7 provides the discussion of the analysis and findings.

3 Building Blocks

This thesis focuses on parametric models, characterised by a finite number of parameters independent of the sample size, belonging to the parameterised family of distributions. In contrast, non-parametric models allow for an adaptable number of parameters as the sample size expands. Model complexity is reflected by the number of parameters.

To begin with, the steps involved in Bayesian inference methodology are outlined to facilitate familiarity with the concepts:

1. *Development of a full Probability Model*: A joint probability distribution that encompasses all observable and latent variables is formulated. Ensuring the model is consistent with the prevailing understanding of the scientific problem and the data collection procedure is crucial.

2. *Conditioning on Observed Data*: The posterior distribution is computed and analysed. The most common approach to posterior distribution computation is Markov Chain Monte Carlo (MCMC) and Gibbs Sampling. Given the observed data, this distribution represents the conditional probability of the latent variables of interest.

3. *Assessment of Model Fit and Posterior Distribution Implications*: The model is evaluated, as are the plausibility of the substantive conclusions derived from the posterior distribution. The assessment also includes checking the conclusions' robustness and the results' sensitivity to the initial modelling assumptions. If necessary, the model is then modified or expanded, and the process is iterated.

3.1 Prior Distribution

The following three sub-chapters are heavily based on Watkins 'Theory of Statistics' (2010) lecture notes from the University of Arizona.

Firstly, a realisation of random variables on a sample space X is observed, represented as

$$X(s) = (X_1(s), ..., X_n(s)), \tag{1}$$

where each X_i shares the same distribution. The allowable distributions are typically restricted to a class P. If these distributions can be indexed by a set $\Omega \subset \mathbb{R}^d$, then P is termed a parametric family. The indexing is usually set up to ensure identifiability, i.e., the mapping from Ω to P is bijective. For a chosen parameter $\theta \in \Omega$, the distribution of the observations and the expectation are denoted by P_θ and E_θ, respectively.

As mentioned in the Chapter XX, *prior distribution* can be informed using additional data, expert knowledge, elicitation techniques, or sometimes, it may be challenging to define. (mention uninformative priors)

In Bayesian statistics, (X, Θ) is considered as a pair of random variables with an associated state space $X \times \Omega$. The distribution, μ of Θ over Ω, is called the *prior distribution.* The joint distribution of (X, Θ) is determined by the prior distribution in conjunction with the family $\{P_\theta : \theta \in \Omega\}$:

$$\Pr\{(X, \Theta) \in B\} = \int \int I_B(x, \theta) \mu_{X|\Theta}(dx|\theta) \mu_\Theta(d\theta). \tag{2}$$

Here, $\mu_{X|\Theta}(\cdot|\theta)$ represents the distribution of X under P_θ.

3.2 Bayes' Theorem and Posterior Distribution

Consider the scenario where μ_Θ has density f_Θ and $\mu_{X|\Theta}(\cdot|\theta)$ has density $f_{X|\Theta}$ with respect to the Lebesgue measure. The probability is then:

$$\Pr\{(X, \Theta) \in B\} = \int \int I_B(x, \theta) f_{X|\Theta}(x|\theta) f_\Theta(\theta)\, dx\, d\theta. \tag{3}$$

The Lebesgue measure, here, serves as a standard way of assigning a length, area, or volume to subsets of a Euclidean space and is fundamental in integration theory.

After observing $X = x$, the conditional density of Θ given $X = x$ using *Bayes' theorem*, the *posterior distribution* $f_{\Theta|X}(\theta|x)$ is given as:

$$f_{\Theta|X}(\theta|x) = \frac{f_{X|\Theta}(x|\theta) \cdot f_\Theta(\theta)}{\int_\Omega f_{X|\Theta}(x|t) \cdot f_\Theta(t)\, dt}. \tag{4}$$

The term $f_{X|\Theta}(x|\theta)$ is the likelihood of observing $X = x$ given $\Theta = \theta$, and $f_\Theta(\theta)$ is the prior density of Θ. The denominator represents the marginal likelihood of $X = x$, a normalising constant to ensure that the posterior density integrates to 1.

The likelihood function $f_{X|\Theta}(x|\theta)$ evaluates how probable the observed data x is under various parameter values θ. Unlike the prior distribution, the likelihood function depends solely on the data and quantifies the support it provides for various parameter values. It is important to note that the likelihood function is not a probability distribution over θ. That is, it does not provide probabilities for different parameter values but rather gives a measure of how well each parameter value θ is supported by the data.

Here, it is also important to mention that Bayesian inference obeys The Likelihood principle, according to which, different probability models that produce the same likelihood for the data should result in the same inference for the parameter θ. The data only influence the posterior through the likelihood function $f(x|\theta)$, while the prior remains independent of the data. Experimental variations are irrelevant for inference about θ.

The *posterior distribution* synthesises all available information regarding the parameter of interest. However, deriving analytical summaries, such as the *posterior distribution's* mean or variance, often requires evaluating complex integrals. The evaluation can be incredibly challenging for high-dimensional posterior distributions. Monte Carlo integration, a simulation technique, offers an effective solution for estimating these integrals. Within Monte Carlo integration, Markov Chain Monte Carlo (MCMC) methodology is a powerful tool for approximating posterior summary statistics, the application of this methodology is defined in the following Chapter.

3.3 MCMC Algorithm

Markov Chain Monte Carlo (MCMC) involves generating samples of θ from approximate distributions and iteratively refining these samples to converge to the desired posterior distribution, $p(\theta|y)$. The essence of *MCMC* is not the Markov property per se but the progressive improvement of the approximation towards the target distribution with each iteration. A Markov chain is defined as a stochastic sequence $\{\theta^0, \theta^1, \theta^2, \ldots, \theta^n\}$, where each state θ^n depends only on its immediate predecessor θ^{n-1}, with the initial state θ^0 is set to an arbitrary value. The Markov chain evolves according to a transition kernel P, dependent only on θ^n:

$$\theta^{n+1} \sim P(\theta^n, \theta^{n+1}) \, (\equiv P(\theta^{n+1}|\theta^n)). \tag{5}$$

Under the conditions of aperiodicity and irreducibility, the Markov chain attains a stationary distribution, independent of initial values. After discarding initial states, the remaining states serve as dependent samples from the target posterior distribution. It is essential to ensure sufficient iterations for convergence and an adequate post-convergence sample size for minimal Monte Carlo errors. The complexity of the posterior distribution does not significantly affect the simplicity of state updates in the chain.

Determining the number of initial states to discard, known as burn-in, is crucial to ensure that the Markov Chain has converged to the stationary distribution. Two common approaches to assess burn-in are through trace plots and the Brooks-Gelman-Rubin (BGR) diagnostic.

The number of iterations after burn-in is essential to estimate the summary statistics accurately, and it is important to evaluate the Monte Carlo error. A common approach to estimating the Monte Carlo error involves batching. The chain is divided into m batches, each of length T, so that $n = mT$. Let $\theta_1, \ldots, \theta_m$ be the sample means for each batch, and θ denote the mean overall n samples. The batch means estimate of σ^2 is then,

$$\hat{\sigma}^2 = \frac{T}{m-1} \sum_{i=1}^{m} (\bar{\theta}_i - \bar{\theta})^2. \tag{6}$$

An estimate of the Monte Carlo error is $\sqrt{\frac{\hat{\sigma}^2}{n}}$. The efficiency of the Markov chain in exploring the parameter space can be evaluated using the autocorrelation function (ACF). The ACF is the correlation of a parameter's value in the Markov chain with itself at a lag j, defined as $\text{cor}(\theta^t, \theta^{t+j})$.

Efficient chains show a fast decrease in ACF as the lag increases, indicating low dependency between chain values within a few iterations.

An alternative estimate of the Monte Carlo error uses the effective sample size M, defined as

$$M = \frac{n}{1 + 2\sum_{k=1}^{\infty} \rho_k}, \tag{7}$$

where ρ_k is the autocorrelation at lag k. Practically, M is estimated through an alternative method accounting for autocorrelations. The Monte Carlo error can be estimated as $\sqrt{\frac{\hat{\sigma}^2}{\hat{M}}}$.

Thinning is a process of selecting every k-th realisation to reduce the autocorrelation of the *MCMC* sample. It is used primarily to alleviate storage or memory issues but should be used judiciously as information is lost in the discarded samples.

The description above lays the foundation of *MCMC* algorithms, of which three are particularly prominent: Gibbs sampling, Metropolis-Hastings, and Importance/Rejection sampling.

Deciding the number of iterations for a Markov chain involves two key considerations: the chain's convergence to the stationary distribution and the required sample size for small Monte Carlo errors post-convergence.The convergence process, referred to as 'burn-in', discards the initial iterations of the chain until a stable mean is achieved. Various techniques, including trace plots and multiple replications, help ascertain the burn-in length. The Brooks-Gelman-Rubin method uses an analysis of variance to assess the similarity of estimates from different starting points. After convergence, the subsequent sample size is determined to minimise Monte Carlo errors. Batching is a standard method for dividing the chain into distinct batches to yield reliable sample mean estimates. The autocorrelation function (ACF) can also be employed to analyse the chain's performance by exploring the parameter space. Finally, 'thinning', or selectively storing every k_{th} realisation of the chain, helps reduce the autocorrelation of the Markov Chain Monte Carlo (MCMC) sample and manage memory allocation effectively.

Chapter XX will introduce an advanced extension to the MCMC, namely the Reversible Jump MCMC method involving Gibbs sampling, which is the most relevant to this paper.

4 The Setting

4.1 Linear Regression in High-Dimensional Setting

The thesis explores the multivariate linear regression model, described as follows:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n) \tag{8}$$

In this setup, $\mathbf{Y} \in \mathbb{R}^n$ is a response vector, $\mathbf{X} = [\mathbf{X}_1, ..., \mathbf{X}_p] \in \mathbb{R}^{n \times p}$ is the given design matrix comprising of p potential predictors. The vector $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T \in \mathbb{R}^p$ represents the set of regression coefficients that will be estimated. Lastly, $\boldsymbol{\epsilon} \in \mathbb{R}^n$ is the noise vector, constituted by independently distributed normal random variables, each sharing a common but unknown variance σ^2. This thesis investigates instances where the number of predictors is less than, equal to, exceeds, or greatly exceed the number of data points.

4.2 Variable Selection Problem

Variable selection seeks the optimal subset of predictors and coefficients to drive the most fitting model for the data. However, it is essential to remember that the “best” model does not claim to uncover the absolute truth about the underlying natural processes, which are far too intricate to be fully captured in mathematical terms (Steyerberg, 2019). With unseen or undiscovered predictors, and potential effects too minuscule to empirically detect, statistical models remain valuable approximations, drawing from the limited palette of known predictors to paint a feasible picture of the complex reality. In situations where the vector of regression coefficients β is large and sparse, i.e., most elements are zero or negligible, the task of identifying the significant elements of β becomes particularly important (Moran et al., 2019). In high-dimensional scenarios where observations are limited, the crucial task is variable selection. The goal is to identify a sparse subset of predictors that adequately capture the true signals within the data, allowing for the construction of parsimonious, interpretable models that effectively mitigate overfitting (Fan & Lv, 2008).

In Bayesian analysis, a Gaussian likelihood for $\mathbf{y}$ is commonly employed, predicated on the normal distribution of errors:

$$\mathbf{y} | (\mathbf{X}, \boldsymbol{\beta}) \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}). \tag{9}$$

For variable selection, binary indicators γ_j are defined to identify the non-zero elements of $\boldsymbol{\beta}$.

WRITE EXPRESSION FOR BAYESIAN MODEL EXPLICITLY? WRITE ABOUT NOISE IN DATA

5 Model Selection Methodology

This thesis discusses methods to compute the relative quality of statistical models, including Akaike information criterion (AIC), deviance information criterion (DIC), Watanabe–Akaike information criterion (WAIC), and Bayesian Information Criterion (BIC). While this thesis focuses on variable selection, the task of model selection is closely related, and some of the applied R packages use these criteria (f.e. *'spikeslab'* calculates AIC criteria, see Chapter XX), warranting their brief explanation for methodological completeness and personal understanding.

5.1 AIC

The following three sections rely on 'Bayesian Data Analysis' by Gelman et al. (2020).

In statistical literature, inference for θ is typically summarised using a point estimate, $\hat{\theta}$, rather than the full posterior distribution. Often, the maximum likelihood estimate (MLE) is used as this point estimate. A common approach for calculating out-of-sample predictive accuracy involves using the log posterior density of the observed data y given the point estimate, $\log p(y|\hat{\theta})$, and correcting for overfitting bias. When k represents the number of estimated parameters, the bias correction is performed by subtracting k from the log predictive density based on the MLE, according to the formula:

$$\widehat{elpd}_{\text{AIC}} = \log p(y|\hat{\theta}_{\text{mle}}) - k. \tag{10}$$

AIC is then defined as twice the negative of this quantity:

$$\text{AIC} = -2\log p(y|\hat{\theta}_{\text{mle}}) + 2k. \tag{11}$$

AIC favours models with good predictive capabilities, penalising those with more parameters, thereby discouraging overfitting. Though *AIC's* bias correction is applicable in normal linear models with known variance and uniform priors, it is not adequate in Bayesian models. In such cases, the penalty of k simply does not accurately represent the effective number of parameters. Hence, other criteria was introduced.

5.2 DIC

DIC is, to some degree, a Bayesian version of AIC, two changes are applied: the maximum likelihood estimate $\hat{\theta}$ is replaced with the posterior mean $\hat{\theta}_{\text{Bayes}} = E(\theta|y)$, while k is replaced with a data-based bias correction. The measure of predictive accuracy is then:

$$\widehat{elpd}_{\text{DIC}} = \log p(y|\hat{\theta}_{\text{Bayes}}) - p_{\text{DIC}}, \tag{12}$$

where p_{DIC} is the effective number of parameters, calculated as,

$$p_{\text{DIC}} = 2\left(\log p(y|\hat{\theta}_{\text{Bayes}}) - E_{\text{post}}\left(\log p(y|\theta)\right)\right), \tag{13}$$

where the expectation in the second term is an average of θ over its posterior distribution. The above expression can therefore be evaluated using simulation, specifically by generating a sequence of S samples θ^s for $s = 1, ..., S$, as

$$p_{\text{DIC, computed}} = 2\left(\log p(y|\hat{\theta}_{\text{Bayes}}) - \frac{1}{S}\sum_{s=1}^{S}\log p(y|\theta^s)\right). \tag{14}$$

The measure *DIC* is then defined in terms of the deviance rather than the log predictive density, as,

$$DIC = -2\log p(y|\hat{\theta}_{\text{Bayes}}) + 2p_{\text{DIC}}. \tag{15}$$

DIC is only valid when the posterior distribution is approximately multivariate normal. In some cases, a negative effective number of parameters can be obtained. *DIC*, besides being heavily criticised (please refer to detailed criticisms in Spiegelhalter et al.(2014)), it also cannot be used with discrete parameters since $E_\pi(\theta|x)$ typically does not coincide with one of the discrete values under consideration.

When considering a vast array of possible models, conducting a fit for each one against the data may not be feasible due to computational constraints. Furthermore, the *DIC* does not yield a readily interpretable quantitative comparison, limiting its practicality in such situations. An alternative approach for Bayesian model discrimination that offers more intuitive results is using Bayes Factors (please refer to description xx sections above) or posterior model probabilities. These methodologies not only offer quantitative comparisons of contending models but also facilitate the incorporation of model averaging concepts, potentially enhancing the overall predictive strength and robustness of the modelling process.

5.3 WAIC

The popularity of DIC is mainly due to its simplicity and inclusion in standard software, while *WAIC*, which requires Monte Carlo estimation of predictive densities, can be significantly more challenging to implement robustly (Spiegelhalter et al., 2014). *WAIC* represents a more fully Bayesian approach designed to estimate the out-of-sample expectation. This approach begins with the computation of the log pointwise posterior predictive density, followed by adding a corrective measure for the effective number of parameters, mitigating potential overfitting.

In this thesis, the robust modification is considered. The effective number of parameters is computed using the variance of individual terms in the log predictive density summed across all the n data points:

$$p_{WAIC2} = \sum_{i=1}^{n} \text{var}_{\text{post}}(\log p(y_i|\theta)). \tag{16}$$

Formula XX is then computed by evaluating the posterior variance of the log predictive density for each data point y_i, that is $V_{s=1}^{S} \log p(y_i|\theta^s)$, where $V_{s=1}^{S}$ represents the sample variance given by

$$V_{s=1}^{S} a_s = \frac{1}{S-1} \sum_{s=1}^{S} (a_s - \bar{a})^2. \tag{17}$$

The total sum across all data points y_i gives the effective number of parameters, as,

$$\text{computed } p_{WAIC2} = \sum_{i=1}^{n} V_{s=1}^{S} (\log p(y_i|\theta^s)). \tag{18}$$

Then use it for bias correction, as,

$$\widehat{\text{elppd}}_{\text{WAIC}} = \text{lppd} - \text{p}_{\text{WAIC}}. \tag{19}$$

Similar to AIC and DIC, the *WAIC* is determined by multiplying the above-mentioned expression by negative two to align it with the deviance scale:

$$WAIC = -2\text{lppd} + 2\text{p}_{WAIC2}, \tag{20}$$

with *lppd* computed as in XX and p_{WAIC2} in XX.

Unlike *AIC* and *DIC*, which gauge the plug-in predictive density's performance, *WAIC* assesses predictions for new data in a Bayesian context by averaging over the posterior distribution. While *AIC*, *DIC*, and *WAIC* aim to estimate the expected out-of-sample deviance of a model, akin to versions of cross-validation, BIC focuses on approximating the marginal probability density of the data under the model, pertinent in the context of discrete model comparison.

5.4 Bayes' Factor

The *Bayes factor*, introduced by Jeffrey (1935), is a key tool in traditional Bayesian model comparison, widely discussed in Bayesian literature. It quantifies the relative evidence for two models, M_i and M_j, given data y,

$$B_{ij} = \frac{p(y|M_i)}{p(y|M_j)} = \frac{\int p(y|\theta_i, M_i)p(\theta_i|M_i)d\theta_i}{\int p(y|\theta_j, M_j)p(\theta_j|M_j)d\theta_j}, \tag{21}$$

where $p(y|\theta_k, M_k)$ denotes the likelihood under model k, and $p(\theta_k|M_k)$ represents the prior distribution of θ_k. The variable M denotes the model, taking values from a finite set of K models. *Bayes factors* can be derived from posterior model probabilities with known prior model probabilities (Kass & Raftery, 1995), enabling Bayesian model averaging, which accounts for model uncertainty in posterior estimates (Hoeting et al., 1999).

Evaluating *Bayes factors* is challenging due to the difficulty in computing marginal likelihoods for each model. MCMC methods offer a solution but necessitate careful handling of varying parameter numbers to ensure ergodicity in the Markov chain. Carlin and Chib (1995) presented a method involving a global model with 'pseudo-priors' for all parameters. Green's (1995) Reversible Jump MCMC (RJMCMC) employs auxiliary variables to manage model dimension changes. However, defining pseudo-priors or auxiliary variables can be challenging and is an area of ongoing research. Barker & Link (2013) introduced a simplified RJMCMC approach compatible with Gibbs sampling. Chapter XX discusses Barker & Link's approach, its integration with prior MCMC results, and introduces the RJMCMC algorithm along with the *'rjmcmc'* package that facilitates *Bayes factor* and posterior model probability calculations using RJMCMC outputs.

5.5 RJMCMC

The Reversible Jump Markov Chain Monte Carlo (RJMCMC) method, suggested as a potential methodology within the scope of this dissertation, was explored. The RJMCMC refines the Metropolis-Hastings algorithm by facilitating the constructed Markov chain traversing varying dimensions.

As this methodology is still very much under development, the outline is based on the work of Gelling, Schofield and Barker for the package *'rjmcmc'* (Gelling et al., 2019). Barker and Link (2013) introduced a modified version of Green's(1995) RJMCMC algorithm which is compatible with Gibbs sampling, specifically suited for high-dimensional models.

Given data y with models indexed $1, ..., K$, and a set of model-specific parameters θ_k for each model k, along with prior model probabilities $p(M_k)$, the posterior model probabilities are related to Bayes factors as:

$$\frac{p(M_i|y)}{p(M_j|y)} = B_{ij} \times \frac{p(M_i)}{p(M_j)} \tag{22}$$

RJMCMC, as proposed by Green (1995), enables sampling across models by considering the model indicator as a latent variable that is sampled using MCMC. Transition between models i and j necessitates that: both models have an equal number of parameters, and a bijective mapping exists between the parameters of the two models. Auxiliary variables u_i are introduced to ensure the dimensions match, i.e., $\dim(\theta_i, u_i) = \dim(\theta_j, u_j)$. With the freedom to transition between any pair of models, $K(K-1)/2$ bijections must be defined. The choice of auxiliary variables and bijections

does not alter the posterior distribution, but affects computational efficiency. Limiting transitions between models can reduce the number of required bijections.
Insert a graph here
During the t iteration of the Markov chain, a proposed model M_j^* is introduced, while the current value is denoted as $M_i^{(t-1)}$. For model M_j^*, the proposed parameter values are determined using the bijection $f_{ij}(\cdot)$ as

$$(\theta_j^*, u_j^*) = f_{ij}(\theta_i^{(t-1)}, u_i^{(t-1)}). \tag{23}$$

The joint proposal is accepted or rejected using a Metropolis step. The selection of the bijection is crucial, as it affects the efficiency and convergence rate of the chain. Barker and Link (2013) introduced a restricted version of Green's *RJMCMC* algorithm, suitable for implementation via Gibbs sampling. This method involves introducing a universal parameter vector ψ, whose dimension is at least the maximum dimension of the model-specific parameters, i.e.,

$$\dim(\psi) \geq \max_k \{\dim(\theta_k)\}. \tag{24}$$

Model-specific parameters θ_i and auxiliary variables u_i are derived from ψ using a bijection $g_i(\cdot)$:

$$(\theta_i, u_i) = g_i(\psi), \tag{25}$$

$$\psi = g_i^{-1}(\theta_i, u_i). \tag{26}$$

Model parameters in model i are mapped to model j through the universal parameter vector ψ,

$$\begin{aligned}(\theta_j, u_j) &= g_j(\psi) \\ &= g_j(g_i^{-1}(\theta_i, u_i)).\end{aligned} \tag{27}$$

This method necessitates K bijections to move among K models. The joint distribution is expressed as:

$$p(y, \psi, M_k) = p(y|\psi, M_k)p(\psi|M_k)p(M_k), \tag{28}$$

where $p(y|\psi, M_k) = p(y|\theta_k, M_k)$ is the joint probability density for the data under model k, $p(\psi|M_k)$ is the prior for ψ given model k, and $p(M_k)$ is the prior model probability for model k.

Since priors are typically in the form $p(\theta_k|M_k)$, $p(\psi|M_k)$ is found as:

$$p(\psi|M_k) = p(g_k(\psi)|M_k) \left| \frac{\partial g_k(\psi)}{\partial \psi} \right|, \tag{29}$$

where $\left|\frac{\partial g_k(\psi)}{\partial \psi}\right|$is the determinant of the Jacobian for the bijection g_k, later denoted as $|J_k|$. The algorithm employs a Gibbs sampler that alternates between updating M and ψ. The full-conditional distribution for M is categorical, with probabilities:

$$p(M_k|\cdot) = \frac{p(y, \psi, M_k)}{\sum_j p(y, \psi, M_j)}. \tag{30}$$

A sample from the full-conditional for ψ is obtained by drawing θ_k and u_k from their respective distributions and computing $\psi = g_k^{-1}(\theta_k, u_k)$. Barker and Link (2013) also detailed a Rao-Blackwellized estimator for posterior model probabilities, based on estimating a transition matrix whose (i, j) entry represents the probability of transitioning from model M_i to M_j. The posterior model probabilities are derived by normalising the left eigenvector of this estimated transition matrix.

An essential feature of the *'rjmcmc'* package is the automatic computation of $|J_k|$ through automatic differentiation, which greatly simplifies implementation, especially when dealing with a large number of parameters.

Bayes factors computation is challenging, often limiting Bayesian multimodel inference. The *'rjmcmc'* package facilitates precise estimation of Bayes factors and posterior model probabilities for a predefined set of models. While RJMCMC is a versatile algorithm, allowing parameter changes in MCMC simulations, the *'rjmcmc'* package is tailored to refine posterior distributions and facilitate model comparison. Notably, it only permits transitions between models with an equal number of parameters.

It remains a developing technique, which suggests the need for further refinement. Comprehensive domain knowledge of the socio-economic data analysed in later chapters was necessary, including the ability to pre-select variables, a requirement that could only partially be met. Prioritising established methodologies ensured reliability in the findings of this dissertation.

6 Variable Selection Methodology

To provide a comprehensive perspective in this thesis, a comparison will be drawn between the classical statistical methods of variable selection, one machine learning method, and some traditional and more modern Bayesian techniques. The output of Bayesian analysis can often be more intuitive to interpret than the output of a frequentist analysis. Posterior distributions provide a probability distribution for each parameter, which tells what values are plausible given the data and the model. In the classical approach to statistical analysis, variable selection typically hinges on point estimates of parameters, which are then subjected to hypothesis testing to determine their significance (Krantz, 1999). Notably, a p-value, standing alone, does not constitute a robust measure of evidence in support of a particular model or hypothesis, which may lead to misinterpretation (often among non-statisticians) around direct evidence about the null hypothesis (Tanha et al., 2017).

6.1 Penalised Regression Methods

In penalised regression methods, a penalty term is added to the log-likelihood function to enforce a trade-off between bias and variance in regression coefficients, consequently optimising prediction error.

Least Absolute Shrinkage and Selection Operator (Lasso) incorporates the *L1-norm*, originally proposed by Tibshirani (1996), of regression coefficients (excluding the intercept) as the penalty term:

$$-\log L + \lambda \sum_{j=1}^{p} |\beta_j|, \quad \lambda > 0. \tag{31}$$

This penalisation shrinks coefficients toward zero and sets those with a negligible predictive contribution to zero, serving as an embedded feature selection method. When the number of predictors, p, exceeds the number of observations, n, *Lasso* selects at most n variables. *Lasso* does not group predictors, often arbitrarily selecting one from a group of highly correlated predictors.

Ridge regression employs the L2-norm of regression coefficients (excluding the intercept) as the penalty:

$$-\log L + \lambda \sum_{j=1}^{p} \beta_j^2, \quad \lambda > 0. \tag{32}$$

Ridge regression shrinks coefficients towards zero but retains all predictors in the model. When $n > p$ and there is high multicollinearity, *Ridge regression* often offers superior predictions. Note that since the penalty term is the sum of squared coefficients, shrinkage would not be fair across predictors with different scales. Hence, they need to be standardised.

Elastic Net combines the *L1-norm* and *L2-norm* penalties, the method first proposed by Zou and Hastie (2005):

$$-\log L + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2, \quad \lambda_1, \lambda_2 > 0. \tag{33}$$

The combination of penalties can also be expressed as:

$$\lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + \frac{1-\alpha}{2} \sum_{j=1}^{p} \beta_j^2 \right), \quad \lambda > 0,\ 0 \leq \alpha \leq 1. \tag{34}$$

Here, α controls the mixing of *Lasso* and *Ridge* penalties, and λ regulates the overall strength of regularisation.

The *'glmnet'* package in R provides efficient procedures for fitting generalised linear and similar models using penalised maximum likelihood, allowing for Lasso or elastic-net regularisation with a spectrum of lambda values, and includes capabilities for prediction, plotting, and cross-validation, even for sparse datasets.

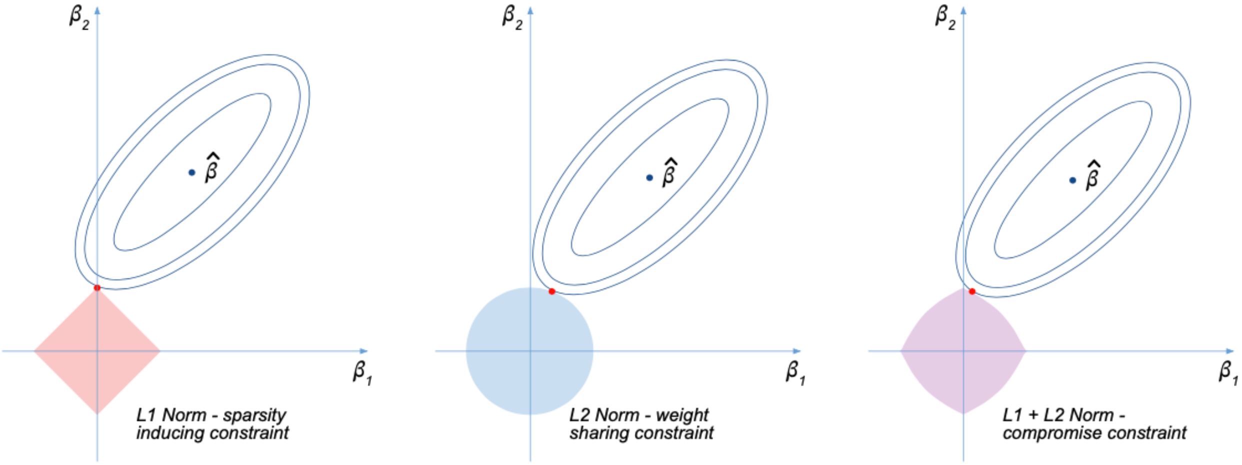
β_2
$\hat{\beta}$
β_1
L1 Norm - sparsity inducing constraint
β_2
$\hat{\beta}$
β_1
L2 Norm - weight sharing constraint
β_2
$\hat{\beta}$
β_1
L1 + L2 Norm - compromise constraint

Figure 1: Penalised Regression

Despite its proven efficacy, the original *Lasso* method has certain constraints. Tibshirani (1996) noted that ridge regression surpasses *Lasso* when dealing with multicollinearity among predictors. In situations with more predictors than observations, namely $p > n$, Lasso's convex optimisation limits it to selecting no more than n variables. Moreover, it disregards any meaningful feature ordering and struggles to effectively select highly correlated grouped variables, tending to choose individual ones instead. Later, Meier et al. (**Merier2008?**) introduced algorithms designed for extremely high-dimensional problems to solve convex optimisation issues. They demonstrated that the group lasso estimator for logistic regression remains statistically consistent with a sparse true underlying structure, even when the number of predictors significantly outnumbers the observations $p >> n$.

6.2 Machine Learning

In exploring variable selection methods, comparing frequentist and Bayesian inference approaches with a renowned machine learning method, Extreme Gradient Boosting (*XGBoost*), would offer valuable insights. *XGBoost* introduced by Chen and Guestrin (2016) is often cited for its outstanding performance in Kaggle competitions/ The method incorporates a feature importance mechanism, which, in simple terms, quantifies the contribution of individual attributes to the construction of decision trees within the ensemble (Chen & Guestrin, 2016).

XGBoost boasts exceptional scalability, enabling rapid processing on single machines and adept scaling to billions of examples in memory-constrained environments. As a comprehensive tree-boosting system, it introduces innovations such as a sparsity-aware algorithm for sparse data and a theoretically grounded weighted quantile sketch for handling instance weights, effectively streamlining resource employment in processing large datasets (Chen & Guestrin, 2016).

The following methodology is based on Wang, Xu, Zhao, Peng and Wang's definition (2019).

In this thesis, the *XGBoost* model is first classified based on all features. Secondly, all the importances of feature variables are computed and then sorted in descending order based on their information in the generated model process. Lastly, the features are filtered using a THRESHOLD XX and are inputted into the final model. The *XGBoost* model has the advantage of high accuracy and is not easy to overfit. It supports weak classification algorithm and weak regression model and is suitable for establishing regression model. The model is presented as follows:

$$\hat{y_i} = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F} \tag{35}$$

where $\hat{y}_i$ is the predicted value for the i-th instance, K denotes the number of trees, $\mathcal{F}$ denotes the set of all possible regression trees, and f_k represents a specific regression tree.

The goal of *XGBoost* is to build a K regression tree such that the predictions of the tree group are as close as possible to the true values while ensuring the greatest generalisation ability. The prediction process is achieved by minimising an objective function, given by:

$$\text{obj}(\theta) = \sum_{i}^{n} l(y_i, \hat{y_i}) + \sum_{k=1}^{K} \Omega(f_k), \tag{36}$$

where the first component, $\sum_{i=1}^{n} l(y_i, \hat{y}i)$, is a loss function that measures the deviation of the predicted values from the true values; the second part, $\sum_{k=1}^{K} \Omega(f_k)$, acts as a regularisation term that controls the complexity of the model, as:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|w\|^2, \tag{37}$$

where T represents the number of leave nodes in the tree, and $\|w\|^2$ is the weight of the corresponding leaf nodes.

During the t-th iteration of training, the objective function is defined as:

$$\text{obj}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_k) + \sum_{i=1}^{t-1} \Omega(f_i) \tag{38}$$

This formulation encapsulates each tree's training error and complexity, steering the algorithm toward building an ensemble of trees that effectively balances accuracy and generalisation.

Figure XX illustrates the tree-boosting process, based on Guo et al. (2020).

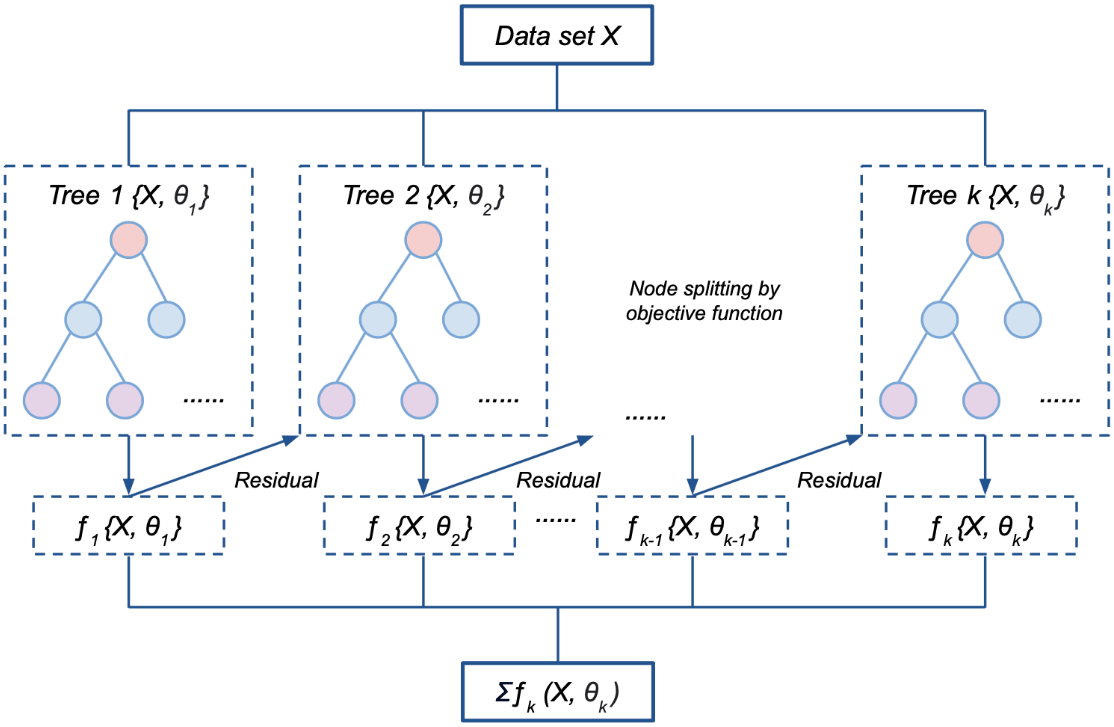

Figure 2: XGBoost Algorithm: Iterative Tree Building Process

While *XGBoost* is acknowledged for its efficacy in classification tasks, its application for regression with continuous outcomes in high-dimensionality settings is debated among the data analytics community as per various informal yet esteemed sources (Li, 2020). A caveat associated with *XGBoost* is its suboptimal performance when the number of features surpasses the number of observations in the training data. In this study, the real dataset being investigated comprises a scenario where the number of data points exceeds the number of features, albeit with a high dimensional feature space. It was deemed essential to incorporate an additional augmented simulated dataset configuration defined in Chapter XX.

6.3 Bayesian Framework

6.3.1 Bayesian Lasso

Building on the frequentist penalised regression Lasso, which minimises the Residual Sum of Squares (RSS) subject to the non-differentiable constraint expressed in terms of earlier defined L1-norm of the coefficients, achieve:

$$\min_{\beta} (\tilde{y} - X\beta)^T (\tilde{y} - X\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \tag{39}$$

where $\tilde{y} = y - \bar{y}1_n$ for some $\lambda \geq 0$.

Tibshirani [-@ Tibshirani1996] interpreted the lasso estimates as Bayes posterior mode under individual Laplace priors for each predictor. The Laplace distribution's ability to manifest as a scale mixture of normal distributions with independently exponentially distributed variances offers advantages. It prompted several researchers to adopt Laplace priors within a hierarchical Bayesian framework. This thesis discusses the one that the *'monomvn'* package implements. Park and Casella (2008) proposed Gibbs sampling for the Lasso, incorporating a Laplace prior within the hierarchical model. They considered a fully Bayesian analysis using a conditional Laplace prior, as:

$$\pi(\beta|\sigma^2) = \prod_{j=1}^{p} \frac{\lambda}{2\sigma} e^{-\lambda|\beta_j|/\sigma} \tag{40}$$

where the noninformative, scale-invariant marginal prior is $\pi(\sigma^2) = \frac{1}{\sigma^2}$. The conditioning on σ^2 asserts that it secures a unimodal full posterior. A lack of unimodality can hinder the convergence of the Gibbs sampler, thus rendering point estimates less reliable. Park and Casella (2008) argue that the *Bayesian Lasso* offers a middle ground between Lasso and Ridge regression by providing smooth paths similar to Ridge but with a tendency to push less significant parameters towards zero faster, akin to Lasso. This behaviour suggests an edge of the Laplace prior over Gaussian or Student-t priors in rapidly diminishing weakly related parameters.

The *'monomvn'* package implements the *Bayesian Lasso* model using the Gibbs Sampling algorithm, as described by Park and Casella (**Park2008?**). It introduced a feature to use a Rao-Blackwellized sample of σ^2, with β integrated out, to improve the mixing of the sampling algorithm. A unique feature of this package is the inclusion of Reversible Jump (RJ) MCMC for Bayesian model selection and averaging, which allows users to determine the best model based on the columns of the design matrix and their corresponding β parameters. Unlike Hans (2009) and Geweke (1996) methods, which require a specific prior on each β_i and individual conditional sampling, this implementation maintains joint sampling from the full β vector of non-zero entries, thus facilitating better Markov chain mixing. It also allows RJ proposals to alter the count of non-zero entries on a component-wise basis, with high acceptance rates due to marginalised between-model moves.

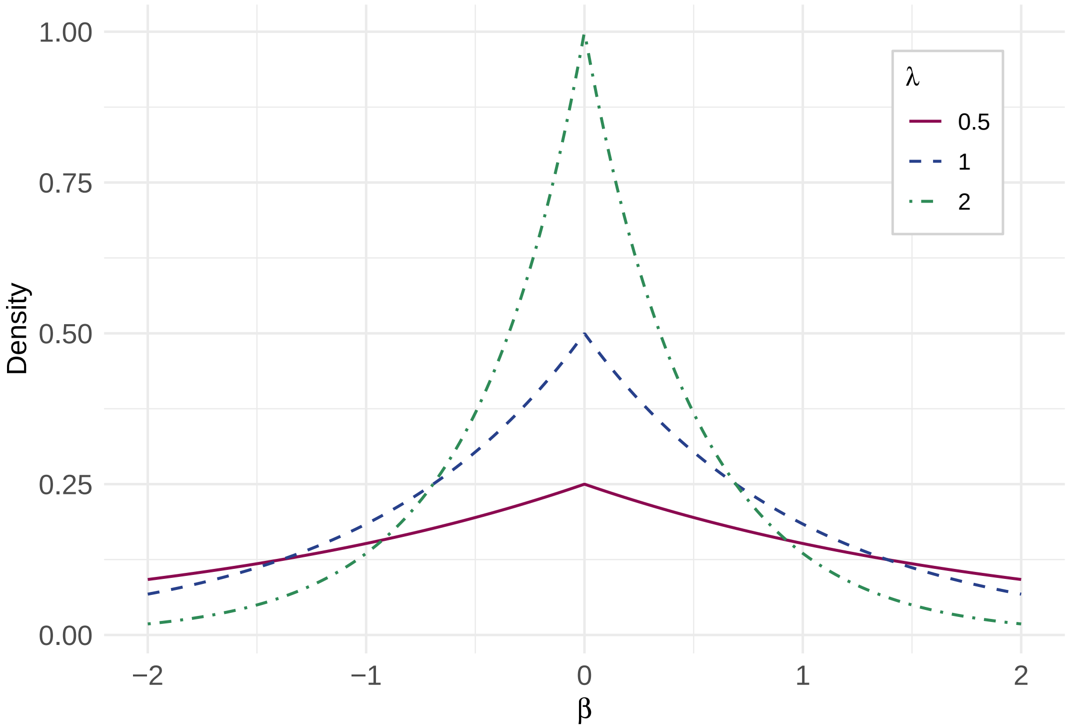

Figure 3: Density Plots of Laplace Prior Distributions for Bayesian Lasso: Impact of tau

6.3.2 Spike and Slab Prior

The *spike-and-slab* approach was initially pioneered by Lempers (1971), Mitchell, and Beauchamp (Mitchell & Beauchamp, 1988). The expression 'spike and slab' was characterised by a two-component mixture prior for β. This prior was set such that the β_k elements were mutually independent, consisting of a flat uniform distribution (the slab) and a zero degenerate distribution (the spike). See Figure XX that illustrates two samples drawn from normal distributions, representing the 'spike' and the 'slab'. The 'spike' is a sample drawn from a distribution with a small standard deviation, representing the zero degenerate distribution in the spike-and-slab prior. The 'slab' is a sample drawn from a distribution with a large standard deviation, representing the flat uniform distribution in the prior.

Ishwaran and Rao (2005) proposed a departure from this design. Instead of a two-component mixture, they posited a multivariate normal scale mixture distribution for β, dictated by the prior π for the hypervariance γ. Despite the divergence in distribution choice, the core principle paralleled the original methodology, aiming to shrink truly zero β_k coefficients via small posterior mean values. The hypervariances played a vital role in this, with smaller values driving coefficient shrinkage and larger ones inflating coefficients for final model selection.

In a further development of the *spike-and-slab* model, Ishwaran and Rao (2005) introduced a continuous bimodal prior in a rescaled variant of the model. The use of such a flexible prior helps to alleviate calibration challenges. To prevent the diminishing influence of the prior on the posterior

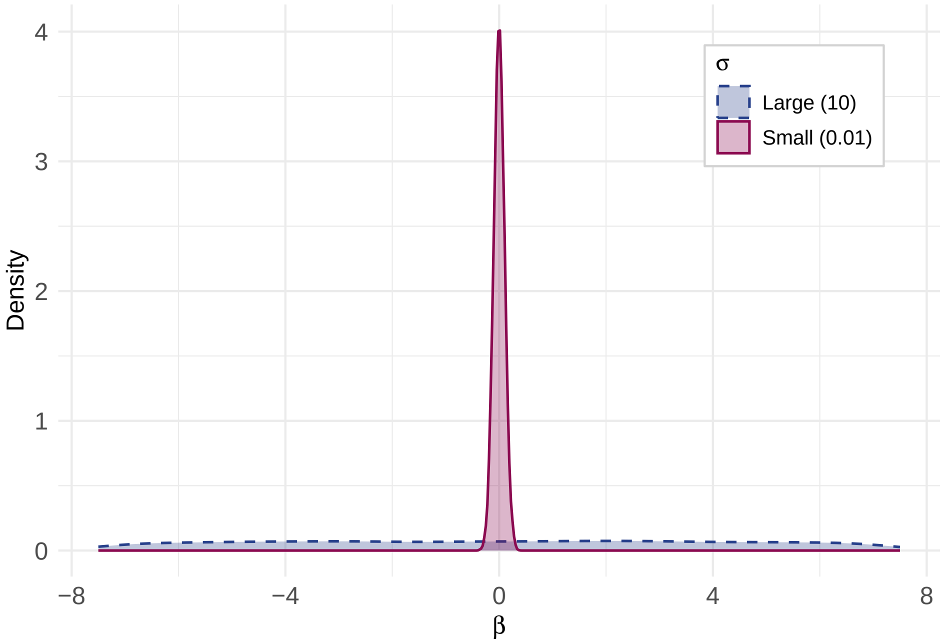

Figure 4: Density Plots of Three Spike-and-Slab Variants

as the sample size increases, they proposed a modification: a sample size invariant or ‘universal’ rescaling of the spike-and-slab model. This modification entails transforming the original Y_i values by a factor of $\sqrt{n}$ and incorporating a variance inflation factor to compensate for the altered variance of the transformed data. The chosen inflation factor can be viewed as a penalisation shrinkage effect of the posterior mean. They demonstrated that selecting n as the inflation factor ensures that the prior continues to influence the posterior, avoiding a vanishing effect. Coupled with a suitably chosen prior for γ, this ensures a robust model selection procedure based on the posterior mean, yielding superior performance over ordinary least squares (OLS) methods.

The rescaled *spike-and-slab* model is defined by a Bayesian hierarchical structure as follows (Ishwaran & Rao, 2005):

$$
\begin{aligned}
(Y_i^*|x_i, \boldsymbol{\beta}, \sigma^2) &\overset{ind}{\sim} \mathcal{N}(x_i^t\boldsymbol{\beta}, \sigma^2\lambda_n), \quad i = 1, \ldots, n, \\
(\boldsymbol{\beta}|\boldsymbol{\gamma}) &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}), \quad \boldsymbol{\Gamma} = \mathrm{diag}(\gamma_1, \ldots, \gamma_K), \\
\boldsymbol{\gamma} &\sim \pi(d\boldsymbol{\gamma}), \\
\sigma^2 &\sim \mu(d\sigma^2),
\end{aligned}
$$

where the values of $Y_i^* = \hat{\sigma}_n^{-1}n^{1/2}Y_i$ are scaled versions of the original response Y_i, where $\hat{\sigma}_n^2 = ||Y - X\hat{\beta}_n^\circ||^2/(n-K)$ serves as an unbiased estimate for σ_0^2 based on the full model, and $\hat{\beta}_n^\circ = (X^tX)^{-1}X^tY$ is the ordinary least squares estimate for β_0. Here, λ_n is a variance inflation factor introduced to account for the scaling of the Y_i's. While a natural choice for λ_n might be n, to match the

$\sqrt{n}$-scaling, λ_n also plays a critical role in controlling the increase in the variance of the data. In this context, $\lambda_n = n$ represents the amount of penalisation necessary to guarantee a significant shrinkage effect in the limit.

The *'spikeslab'* package, developed by Ishwaran, Kogalur and Rao (2010), implements a rescaled three-step spike and slab algorithm using the generalised elastic net (gnet) and Bayesian model averages (BMA) estimator. The BMA estimator effectively addresses correlation issues, particularly prevalent in high-dimensional datasets, leveraging the strengths of weighted generalised ridge regression (WGRR) - a fundamental advantage of the Bayesian approach. On the other hand, the gnet estimator applies the principle of soft-thresholding, a potent frequentist regularisation concept, to achieve sparse variable selection in complex, high-dimensional data settings.

The algorithm that underlines the package involved three main steps: First, variables are filtered down to the top nF, where n is the sample size and $F > 0$ is the user-specified fraction, and ordered based on absolute posterior mean coefficient value, computed via Gibbs sampling applied to an approximate rescaled spike and slab posterior. The user is provided with an option to engage this filtering step when $p \geq n$, but it should be noted that in such instances, the function should be passed a matrix rather than a data frame. Further, a rescaled spike and slab model is fitted to unfiltered variables from Step 1 using a Gibbs sampler, employing a blocking technique for computational efficiency, and the posterior mean of the regression coefficients BMA is computed and returned as an estimator for the regression coefficients.

Lastly, the gnet estimator is computed with its L2-regularisation parameters fixed, determined from the restricted BMA of Step 2, and its solution path for L1-regularisation is obtained using the *'lars'* package, selecting the model that minimises the AIC criterion, negating the need for cross-validation. Unlike the elastic net approach, this method simplifies optimisation and reduces computational time, especially in high-dimensional problems.

6.3.3 Spike and Slab Prior Meets Lasso

Within the frequentist statistics, sparse recovery for β is often achieved through the Lasso, whereas in the Bayesian domain, *spike-and-slab priors* are favoured for sparse modelling of β. In the Bayesian framework, the *spike-and-slab Lasso (SSL)*, introduced by Ročková & George (2018), bridges penalised likelihood Lasso method with the traditional *spike-and-slab prior* approach, capitalising on the strengths of both while mitigating their drawbacks. The package *'SSLASSO'* specifically implements *SSL*, which uses a dynamic penalty that adjusts based on the level of sparsity and performs selective shrinkage. It also supports fast algorithms for finding the most probable estimates, ensuring efficiency and scalability. Lastly, debiasing the posterior modal estimate or applying effective posterior sampling techniques can quantify uncertainty for the *SSL*. The package primarily focuses on settings where $p > n$.

The methodology definition is based on Tadesse and Vannucci (2022). The spike-and-slab Lasso prior is given by:

$$\begin{aligned}
\pi(\beta|\gamma) &= \prod_{i=1}^{p}[(1-\gamma_i)\psi(\beta_i|\lambda_0)+\gamma_i\psi(\beta_i|\lambda_1)], \\
\pi(\gamma|\theta) &= \prod_{i=1}^{p}[\theta^{\gamma_i}(1-\theta)^{1-\gamma_i}], \\
\theta &\sim Beta(a,b)
\end{aligned}$$

where $\psi(\beta|\lambda) = (\lambda/2)e^{-\lambda|\beta|}$ denotes the Laplace density with scale parameter λ. As depicted in Figure XX, larger values of λ result in a density peaked around zero (the 'spike'), while smaller λ values lead to a diffuse density (the 'slab'). The original model assumed a known variance $\sigma^2 = 1$. Later works extended this to handle unknown variance, placing an independent Jeffreys prior on σ^2 where $p(\sigma^2) \propto \sigma^{-2}$.

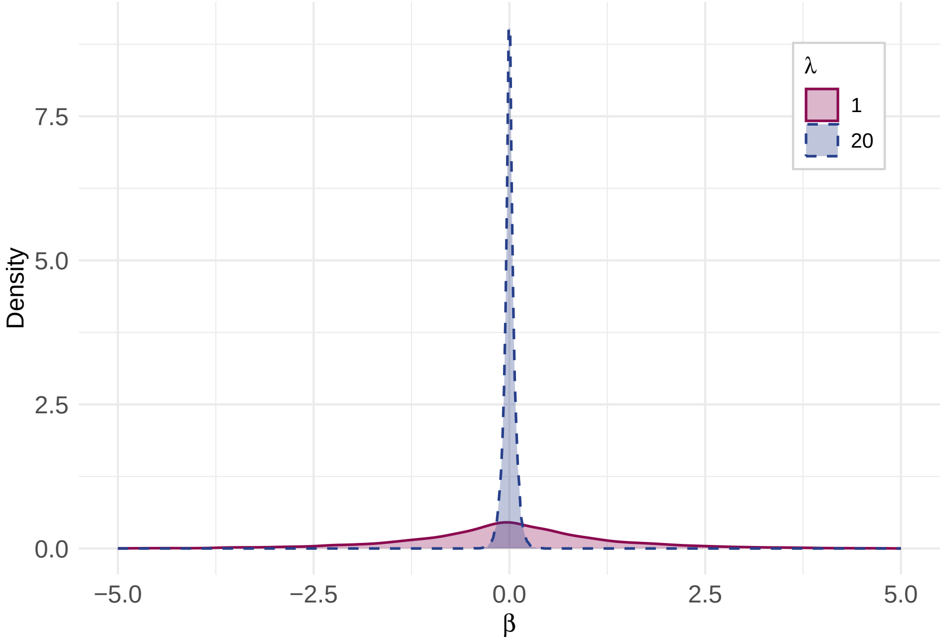

Figure 5: Density Plots of Laplace Distributions: Impact of Scale Parameter

Setting $\lambda_1 = \lambda_0$ results in the L1 penalty used in the Lasso. As $\lambda_0 \to \infty$, the spike-and-slab Lasso approaches the ideal point-mass model. Therefore, the SSL prior allows for a non-concave continuum between penalised likelihood and point-mass constructs.

The *SSL* prior, a mixture of two Laplace distributions, can be viewed as a two-group refinement of the L1 penalty in Lasso, leading to exactly sparse posterior modes for $p(\beta|y)$, enabling simultaneous variable selection and parameter estimation. This offers an advantage over traditional spike-and-slab formulations, which often require post hoc thresholding. Although the original Lasso is known to suffer from estimation bias, *SSL* offers two key advantages, as demonstrated by Tadesse and Vannucci (2022). First, it adaptively mixes two Lasso "bias" terms, applying either high shrinkage for small $|\beta_i|$ or low shrinkage for large $|\beta_i|$. Unlike the adaptive Lasso, which assigns fixed penalties, *SSL* adjusts the coefficient-specific penalties to extremes. Second, the prior on θ introduces dependency in the marginal prior $p(\beta)$ and non-separability in the SSL penalty, enabling *SSL* to borrow information across coordinates and adapt to sparsity information. The *'SSLASSO'* package fits a set of models, each distinguished by the regularisation parameter λ_0, using a coordinate descent algorithm. This algorithm utilises screening rules to exclude irrelevant predictors, adopting a similar approach to that proposed by Breheny and Huang (2011).

6.3.4 Horseshoe Priors

The **horseshoe** prior was introduced by Carvalho, Polson, and Scott (2010), who characterised it as multivariate-normal scale mixtures. They further modified the prior specification to set λ_i to be conditionally independent as further defined (Carvalho et al., 2010). In the context of a p-dimensional vector $y|\theta \sim N(\theta, \sigma^2 I)$, when sparsity is assumed for β, the Bayesian *horseshoe* prior, denoted as π_{HS}, is applied. The assumption here is that each β_i is conditionally independent, each having a density of $\pi_{HS}(\beta_i|\tau)$. The *horseshoe* prior is then formulated as follows:

$$\begin{aligned}
\beta_i|\lambda_i &\sim \mathcal{N}(0, \lambda_i^2), \quad \text{for } i = 1, \ldots, p, \\
\lambda_i|\tau &\sim \mathcal{C}^+(0, \tau), \\
\tau|\sigma &\sim \mathcal{C}^+(0, \sigma),
\end{aligned}$$

where $\mathcal{N}$ is the normal distribution and $\mathcal{C}^+$ is the half-Cauchy distribution, specifically over the positive reals with a scale parameter denoted by a. It is vital to note that each β_i is a mixture of its own λ_i and that all λ_i elements have a half-Cauchy prior with a common scale, τ. The λ_i is referred to as the local shrinkage parameter and τ as the global shrinkage parameter. Additionally, Jeffreys' prior is employed for the variance, denoted by $p(\sigma^2) \propto 1/\sigma^2$, which is non-informative. Similarly, the prior for τ also follows Jeffreys' treatment, scaled by σ, the standard deviation of the error model.

The *horseshoe* prior enforces sparsity on the regression coefficients $\boldsymbol{\beta}$. Specifically, the posterior mean of each coefficient β_j can be expressed as a linear function of the corresponding observation y_i:

$$E(\beta_i|y_i) = y_i(1 - E(k_i|y_i)), \tag{41}$$

where k_i represents the shrinkage coefficient. The half-Cauchy prior on λ_i induces a Beta$(\frac{1}{2}, \frac{1}{2})$ distribution for k_i, which has a *horseshoe* shape. For $k_i \approx 0$, there is negligible shrinkage representing signals, while for $k_i \approx 1$, there is substantial shrinkage representing noise.

The *horseshoe* prior has the property that it tends to shrink the majority of the coefficients β_j towards zero, enforcing sparsity, while allowing some coefficients to remain large if they are indeed associated with the response variable: its flat, Cauchy-like tails, ensures that significant signals maintain their magnitude, resulting in minimal post-hoc shrinkage. Simultaneously, its infinitely tall spike centred at the origin facilitates intense shrinkage for elements of β that are zero, effectively emphasising the sparse nature of the solution (Carvalho et al., 2010). A notable advantage of the *horseshoe* prior is that it does not require user-specified hyperparameters as the priors for λ_i, τ, and σ are fully defined. Figure XX showcases the horseshoe prior with three different magnitudes of the global shrinkage parameter. As visible from the plot, a smaller τ value tends to concentrate more mass around zero, thereby leading to an amplified global shrinkage effect.

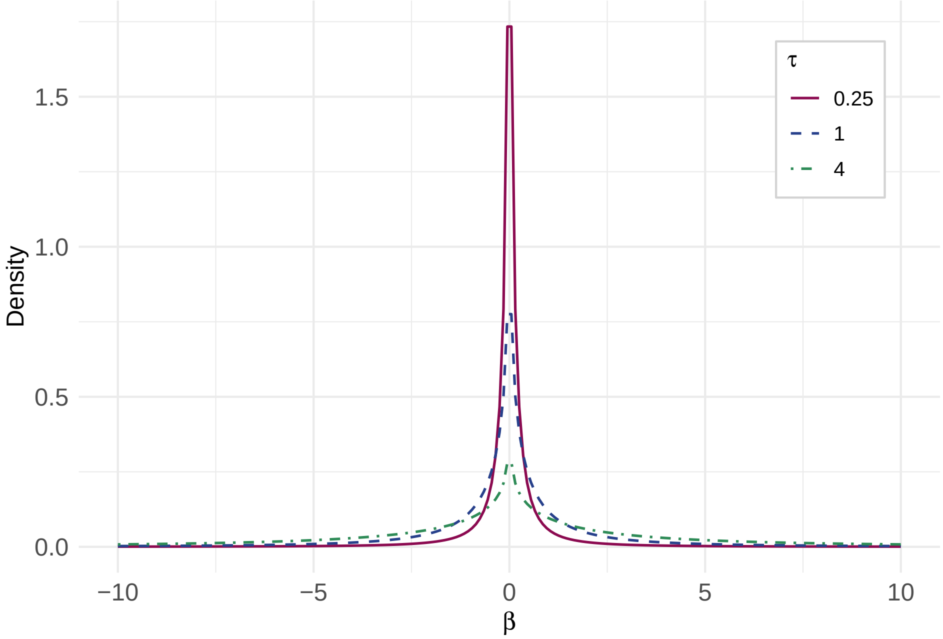

Figure 6: Density Plots of Horseshoe Distributions: Impact of tau

The **horseshoe+** estimator (Bhadra et al., 2016), an extension of the horseshoe estimator, excels in ultra-sparse problems. In contrast to the horseshoe estimator, the *horseshoe+* estimator has a lower posterior mean squared error and faster posterior concentration rates in terms of the Kullback–Leibler divergence metric. The prior distribution π_{HS+} for local shrinkage hyperparameters $(\lambda_1, ..., \lambda_p)$ retains the zero-mean half-Cauchy form and an additional layer of hyperparameters $(\eta_1, ..., \eta_p)$ is applied. Each η_i relates to the prior variance of the corresponding hyperparameter λ_i, creating an expanded hierarchy, as per (Makalic & Schmidt, 2016):

$$\beta_i|\lambda_i, \eta_i, \tau \sim \mathcal{N}(0, \lambda_i^2), \tag{42}$$
$$\lambda_i|\eta_i, \tau \sim \mathcal{C}^+(0, \tau\eta_i), \tag{43}$$
$$\eta_i \sim \mathcal{C}^+(0, 1). \tag{44}$$
$$\tag{45}$$

In both horseshoe and *horseshoe+* models, the local shrinkage random effects λ_i are not marginally independent after the global shrinkage parameter τ is considered. The *horseshoe+* model further develops the concept by introducing an additional level of local shrinkage parameters η_i alongside τ, yielding conditionally independent λ_i. Integrating over η_i yields λ_i's density:

$$p(\lambda_i|\tau) = \frac{4\log(\lambda_i/\tau)}{\pi^2\tau(\lambda_i/\tau)^2 - 1} \tag{7}$$

The introduction of the additional $\log(\lambda_i/\tau)$ term in the numerator leads to unique properties for the proposed estimator. Global shrinkage parameter τ can be handled in various ways. A full Bayesian approach might involve assigning a standard half-Cauchy or Uniform(0,1) prior on τ. An alternative approach could appeal to an asymptotic argument, suggesting τ's empirical Bayes estimator be set to $\hat{\tau} = p_n/n$, where p_n is the count of non-zero entries in θ.

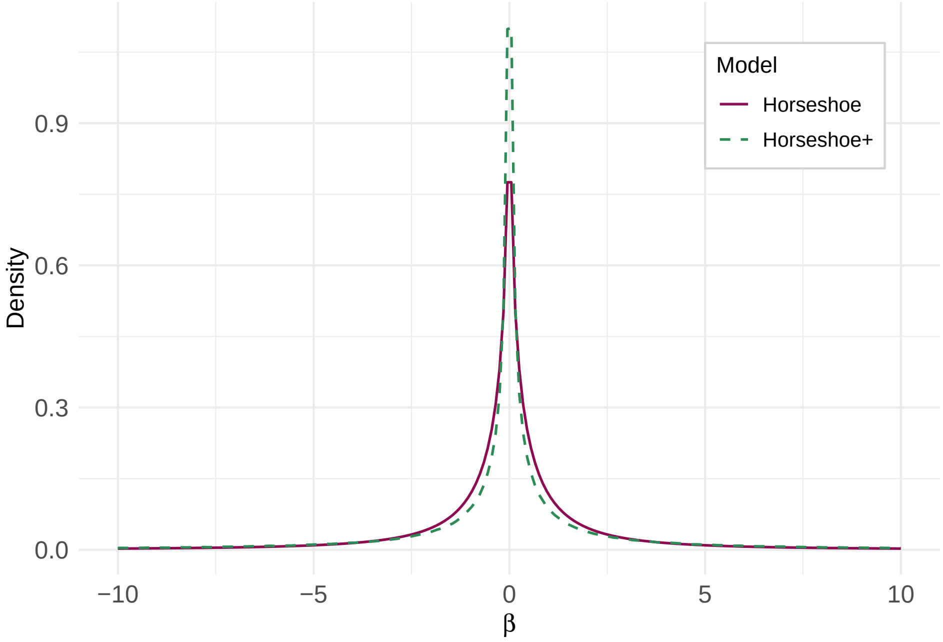

Figure 7: Density Plots of Horseshoe and Horseshoe+ Prior Distributions

As the usage of Horseshoe prior in Bayesian variable methodology is prominent, this thesis explores two R packages that employ it, namely *'bayesreg'* and *' horseshoe'* with some distinct features.

The *'bayesreg'* package, introduced by Bhattacharya, Chakraborty and Mallick (2016), fits linear or generalised linear regression models using Bayesian global-local shrinkage prior hierarchies, described by Polson and Scott (Polson & Scott, 2010). This thesis narrows its focus specifically on the horseshoe and horseshoe+, which are adept at handling high-dimensional datasets. The package automatically groups factor variables together and applies an additional level of shrinkage to the set of dummy variables that these factor variables are expanded into, this is a way to control the complexity of the model. It also provides both variable ranking and importance, credible intervals, and diagnostics, which include earlier described WAIC to assist with prior selection. The *'bayesreg'* package employs Gibbs sampling for its implementation and strategically selects between two algorithms for the efficient sampling of regression coefficients based on the ratio of predictors to sample size: Rue's algorithm when p/n < 2 (Rue, 2001), and Bhattacharya et al.'s algorithm otherwise (Shin et al., 2015). This approach circumvents the computational and numerical accuracy challenges inherent in directly computing matrix inverses, particularly in high-dimensional settings.

The *'horseshoe'* package allows for conducting sparse linear regression using the horseshoe prior, providing results such as posterior means and credible intervals. It is also grounded on the work of Bhattacharya (2016). The package's underlying algorithm updates the global-local scale parameters through a slice sampling scheme, with the regression coefficients' posterior samples computed differently depending on whether the number of predictors is greater or less than/equal to the number of observations. For the case where $p > n$, the method proposed by Bhattacharya et al. (Bhattacharya et al., 2016) is used, while for $p <= n$, the approach from Rue (Rue, 2001) is used. Despite not offering the option to specify the horseshoe+ prior, the package allows users to select methodological choices for handling the tau and error variance parameters. For tau, options include "truncatedCauchy" for full Bayes with a truncated Cauchy prior, "halfCauchy" for full Bayes with a half-Cauchy prior, or "fixed" for a fixed value approach, typically based on an empirical Bayes estimate. In this thesis, the first two are tested. Regarding the error variance σ^2, options include 'Jeffreys' for full Bayes with Jeffrey's prior or 'fixed' for a fixed value, again typically based on an empirical Bayes estimate.

6.3.5 Simplified Shotgun Stochastic Search Algorithm with Screening

On a quest to find a more modern variable selection method within the Bayesian framework, here is presented the somewhat more recently adapted version of Shotgun Stochastic Search (*SSS*). First, it is important to understand the original *SSS* algorithm. The *SSS*, introduced by Hans et al. (2007), is designed to efficiently navigate high-dimensional model spaces in regression settings with a large number of candidate predictors, where $p \gg n$. Its primary objective is to swiftly pinpoint regions with high posterior probabilities and ascertain the maximum a posteriori (MAP) model. To achieve this, the algorithm amalgamates sparsity-inducing priors promoting parsimony, temperature control akin to that used in global optimisation algorithms like simulated annealing (Kirkpatrick et al., 1983), and screening techniques resembling Iterative Sure Independence Screening (Fan & Lv, 2008). Furthermore, SSS exploits parallel computation to enhance performance on cluster computers.

The MAP model, denoted $\hat{k}$, is formally defined as:

$$\hat{k} = \underset{k \in \Gamma^*}{\arg\max} \{\pi(k|y)\}, \tag{46}$$

where Γ^* represents the set of models that are assigned non-zero prior probability.

In its quest to traverse large model spaces and pinpoint global maxima efficiently, SSS algorithm defines $\text{nbd}(k) = \{\Gamma^+, \Gamma^-, \Gamma^0\}$, where $\Gamma^+ = \{k \cup \{j\} : j \in k^c\}$, $\Gamma^- = \{k \setminus \{j\} : j \in k\}$, and $\Gamma^0 = \{[k \setminus \{j\}] \cup \{l\} : l \in k^c, j \in k\}$. The *SSS* algorithm proceeds as follows:

1. Select an initial model $k^{(1)}$.
2. For $i = 1$ to $i = N - 1$:
    - Compute $\pi(k|y)$ for all $k \in \text{nbd}(k^{(i)})$.
    - Sample k^+, k^-, and k^0 from Γ^+, Γ^-, and Γ^0, respectively, with probabilities proportional to $\pi(k|y)$.
    - Sample $k^{(i+1)}$ from $\{k^+, k^-, k^0\}$, with probability proportional to $\{\pi(k^+|y), \pi(k^-|y), \pi(k^0|y)\}$.

The MAP model is determined as the model with the highest unnormalised posterior probability among those models searched by SSS.

Simplified Shotgun Stochastic Search with Screening

As the objective of this thesis is to blend statistical methodology with application, it is imperative to dissect the proposed computational tools. Over the years, the SSS algorithm has evolved, and a streamlined version incorporating screening has been developed and made available through the package *'Bayes5'* (Shin et al., 2015).

The Simplified Shotgun Stochastic Search with Screening (*S5*) algorithm is a modified version of the SSS designed to further enhance computational efficiency. *S5* restricts its search to models in Γ^+ and Γ^-, thereby omitting the computationally intensive evaluation of marginal probabilities for models in Γ^0. However, this focused search might lead the algorithm to overlook certain high-posterior probability regions and risk settling in local maxima. To mitigate this, S5 introduces a temperature parameter, akin to simulated annealing, enabling broader exploration.

Furthermore, *S5* incorporates an Iterative Sure Independence Screening strategy to focus on variables highly correlated with the residuals of the current model. Specifically, it assesses $|r_k^T X_j|$, where r_k

is the residual of model k, for $j = 1, \ldots, p$, and prioritises variables for which this product is large.

In S5, S_k represents the union of variables in k and the top M_n variables obtained through residual-based screening. The screened neighborhood, denoted as $\text{nbd}_{scr}(k) = \{\Gamma^+_{\text{scr}}, \Gamma^-\}$, is defined with $\Gamma^+_{\text{scr}} = \{k \cup \{j\} : j \in k^c \cap S_k\}$. This results in a scalable algorithm, particularly beneficial when the number of variables p is large.

S5 algorithm employs a temperature schedule and utilises a screened set of variables to improve efficiency in identifying the MAP model. It approximates the posterior model probability and assesses model space uncertainty by approximating the normalising constant from the unnormalised posterior probabilities.

The computational complexity of the original SSS algorithm is proportional to the product of the number of models explored and the complexity of evaluating the unnormalised posterior probability for the largest model, denoted as E_n, and is given by $[O\{Np\} + O\{Nq_n\} + O\{N(p - q_n)q_n\}] \times E_n$, where q_n is the maximum size of model among searched models and $q_n < n << p$.

In contrast, S5 dramatically reduces the number of models considered by focusing on M_n variables post-screening. This leads to a computational complexity of $[O\{JL(M_n - q_n)\} + O(JLM_n)] \times E_n + O(JLnp)$, where $q_n < M_n$. The algorithm is scalable since its complexity is relatively insensitive to the size of p.

Shin et al. (2015) demonstrated that S5 is significantly faster than SSS in identifying the MAP model and requires fewer model evaluations.

7 Simulated Data Study

7.1 Simulation Overview

As noted before, variable selection methods are applied within the framework of linear regression, denoted by $Y = X\beta + e$, where $e \sim \mathcal{N}(0, \sigma^2)$. The thesis provides an analytical contrast between Bayesian techniques and their frequentist counterparts, specifically, Lasso and Elastic Net penalisations, which were rigorously studied in semesters 1 and 2. Additionally, a contemporary machine learning technique, *XGBoost*, is also applied to provide a comprehensive comparative.

Each of the datasets *Type 1*, *Type 2*, *Type 3*, and *Type 4* that follow are designed to be adaptable across various dimensionality settings. To further note, the selection of true signals, the predetermined magnitude of error variance, and the inclusion of interaction terms, polynomials, and other features are strategically chosen to present varying levels of complexity for the models being tested, thereby examining their robustness and adaptability under distinct circumstances.

The **Type 1** datasets consist of uncorrelated continuous covariates with a moderate level of noise. The covariates are simulated from a multivariate normal distribution:

$$\mathbf{X} \sim \mathrm{MVN}(\mathbf{u}_x, \sigma_x^2 \mathbf{\Sigma}_x), \tag{47}$$

where $\mathbf{u}_x$ is a $1 \times p$ mean vector, $\mathbf{\Sigma}_x$ is a $p \times p$ correlation matrix, and σ_x^2 is the common variance of all covariates. Each x_i is normally distributed with a mean of 5, i.e., $\mathbf{u}_x = (5, 5, \ldots, 5)$. $\mathbf{\Sigma}_x$ is a $p \times p$ identity matrix and $\sigma_x^2 = 1$. The response variable y is generated as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{48}$$

where $\boldsymbol{\epsilon_i} \sim N(0, \sigma_e^2)$ for $i = 1, 2, \ldots, n$. $\boldsymbol{\beta}$ is a $p \times 1$ vector of true coefficients, $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of error terms, σ_e^2 is the error variance, and $\mathbf{I_n}$ is an $n \times n$ identity matrix. The regression coefficients are enforced as $\beta_1, \ldots, \beta_{10} = 3$; $\beta_{11}, \ldots, \beta_{20} = 5$, and $\beta_{p-20} = 0$. The intercept term is zero, and the errors (unexplained variability in the response variable) are normally distributed with $\sigma_e^2 = 15$.

The **Type 2** dataset comprises continuous covariates with temporal correlation and moderate noise. The generation of the **Type 2** dataset parallels the method utilised for the *Type 1* dataset with certain distinctions. Specifically, the mean vector for the covariates $\mathbf{u}_x$ is constructed such that the first 30 variables each have a mean of 3, and the rest have a mean of 7; that is, $\mathbf{u}_x = (3, 3, \ldots, 3, 7, 7, \ldots, 7)$.

The covariance matrix of the covariates $\mathbf{\Sigma}_x$ adheres to an autoregressive order 1, commonly refered to as AR(1), structure, with the correlation coefficient $\rho = 0.8$, σ_x^2 is held constant at 1. For the response variable, 20 covariates are true signals. The vector of true regression coefficients, $\boldsymbol{\beta}$, is defined with non-zero values for the first 20 entries (e.g., 5 for each), while the remaining entries are set to zero; thus, $\boldsymbol{\beta} = (5, 5, \ldots, 5, 0, 0, \ldots, 0)^T$. The error term variance $\sigma_e^2 = 10$.

The **Type 3** dataset is a rich blend of continuous and categorical covariates, including interaction terms and polynomial features, with moderate noise. In this setup, the mean vector for continuous covariates $\mathbf{u}_x$ is segmented into three groups: the first 20 variables have a mean of 2, the next 30 have a mean of 5, and the remaining variables have a mean of 8, resulting in $\mathbf{u}_x = (2, 2, \ldots, 2, 5, 5, \ldots, 5, 8, 8, \ldots, 8)$. The covariance matrix $\mathbf{\Sigma}_x$ adheres to an AR(1) structure with a correlation coefficient of $\rho = 0.6$, while σ_x^2 remains constant at 1. The vector of true regression coefficients for continuous predictors, $\boldsymbol{\beta}$, takes on values such as $\boldsymbol{\beta} = (6, 6, 6, 6, 6, 4, 4, 4, 4, 4, 3, 3, 3, 3, 3, 0, 0, \ldots, 0)^T$.

First categorical variable is binary (akin to having or not having a certain illness) and is set to have $\beta = 4$. Second categorical variable is treated as ordinal, with categories 1 through 5 (serves as ordered category, f.e., progressive education levels from middle school to higher degree) and is set to have $\beta = 0$.

Interaction terms are created by multiplying selected pairs of covariates: $X1$ and $X2$ continuous, $X3$ and $X4$ continuous, $X11$ and $X22$ continuous, $binarycategorical$ and $X22$ continuous. Polynomial features are generated by elevating $X5$ and $X23$ covariates to the power of 2, and $X6$, $X23$ to the power of 3. From interaction terms and polynomials, only $X1 : X2$ was enforced to have $\beta = 3$ and $X23^2$ to have $\beta = 6$, the rest have $\boldsymbol{beta} = (0, \ldots, 0)$

The intercept, in this context, has set to a of value 2, serving as the expected response value when all covariates are at zero. Lastly, the error term variance is set at $\sigma_e^2 = 12$.

The **Type 4** dataset is characterised by grouping structures among continuous covariates, where covariates within each group are highly correlated, while covariates between different groups are independent. The mean vector for continuous covariates $\mathbf{u}_x$ is segmented into five groups, i.e. $\mathbf{u}_x = (2, \ldots, 2, 4, \ldots, 4, 6, \ldots, 6, 8 \ldots, 8, 10, \ldots, 10)$. The vector of true regression coefficients corresponding to true signals within each group $\boldsymbol{\beta} = (6, 6, 6, 6, 6, 4, 4, 4, 4, 4, 3, 3, 3, 3, 3, 0, \ldots, 0)^T$, where the number of groups $T = 5$.

The covariance matrix $\mathbf{\Sigma}_x$ is constructed as a block-diagonal matrix, where each block corresponds to a group, and within each block, the elements are highly correlated. The diagonal blocks can have a specific structure, the AR(1), with correlation coefficients $\rho = 0.6$, while the off-diagonal blocks are matrices of zeros, indicating independence between groups.

First categorical variable is binary and is set to have $\beta = 4$. Second categorical variable is treated as ordinal, with categories 1 through 5 and is set to have $\beta = 0$. Interaction terms are created by multiplying selected covariates: the continuous $X1$, $X2$ and $X3$, $X3$ and $X4$, $X16$ and $X17$. $X1 : X2 : X3$ was enforced to have $\beta = 3$, $X4 : X5$ and $X16 : X17$ to have $\beta = 0$. The error term variance is set at $\sigma_e^2 = 10$.

Four different dimensionality settings are considered under all four data types: $p(50) < n(200)$, reflecting a traditional setting with more observations than variables; $p(100) = n(100)$, representing a balanced case; $p(200) > n(150)$, indicative of high-dimensional scenarios such as in genomics; and $p(200) \gg n(50)$, where the number of variables substantially surpasses the number of observations. An additional configuration is proposed to better suit XGBoost's strengths and establish a baseline performance capacity. While retaining the data generation methodology outlined earlier, a subsequent dataset is simulated with a modified proportion of data points to features, namely $p = 50, n = 500$. These settings enable comparisons within data types and offer insights into practical challenges regarding data availability and computation. The choice of p and n values is guided by computation time and an approximation to real-world data, where obtaining data can be costly. This structure facilitates a comprehensive analysis of various scenarios. For reference, for all data generation the reproducibility seed was set to 42.

The packages *'glmnet'*, *'caret'*, *'spikeslab'*, *'horseshoe'*, *'SSLASSO'*, *'monomvn'*, *'BayesS5'* necessitate a matrix specification, where each column denotes a variable in the dataset, and each row represents an observation. If the data frame comprises only continuous variables, these are directly converted into a matrix. However, in the presence of categorical variables, these undergo a transformation. The constructed matrix integrates both the continuous and dummy variables representing each level of the categorical variables, excluding one level to avert the dummy variable trap, which could lead to multicollinearity. In contrast, the *'bayesreg'* package requires data to be specified as a data frame. Consequently, categorical variables are retained in their original format, i.e., as factors, to be appropriately fitted within the model.

7.2 Results

Three metrics are used to evaluate the performance of the introduced methodologies for the variable selection task: total signals (TS), false positives (FP), and false negatives (FN). TS is the count of covariates chosen by the method, FP quantifies the noise covariates wrongly identified as signals, while FN represents the missed true signal covariates.

7.3 Type 1

Add results of ceofficients and list which methods do not draw coefficients to exactly zero. Which methods include confidence intervals and where is 0 included should be disregarded.

Table 1: Summary of Type 1 Data Results

		p < n			p = n			p > n			p » n		
Package	Method	TS	FP	FN	TS	FP	FN	TS	FP	FN	TS	FP	FN
Frequentist Methods													
glmnet	Lasso	20	19	0	20	50	0	20	37	0	10	22	10
glmnet	Elastic Net	20	12	0	20	37	0	20	10	0	12	18	8
Bayesian Methods													
spikeslab	Spike-and-Slab Prior	20	20	0	20	79	0	20	117	0	12	12	8
SSLASSO	Spike-and-Slab Lasso	20	1	0	15	5	5	6	5	14	20	0	0
horseshoe	Horseshoe Prior, TC	20	0	0	20	1	0	20	0	0	1	0	19
horseshoe	Horseshoe Prior, HC	20	0	0	20	1	0	20	0	0	0	0	20
bayesreg	Horseshoe Prior	20	0	0	20	1	0	20	0	0	7	5	13
bayesreg	Horseshoe+ Prior	20	0	0	20	1	0	20	0	0	8	5	12
BayesS5	S5 Method												
monomvn	Bayesian Lasso												

[a] TS=True Signal, FP=False Positive, FN=False Negative
[b] p(50) < n(200), p(100) = n(100), p(200) > n(150), p(200) » n(50)

7.4 Type 2

Bla bla

Table 2: Summary of Type 2 Data Results

		p < n			p = n			p > n			p » n		
Package	Method	TS	FP	FN	TS	FP	FN	TS	FP	FN	TS	FP	FN
Frequentist Methods													
glmnet	Lasso	20	0	0	20	9	0	20	6	0	20	7	0
glmnet	Elastic Net	20	0	0	20	0	0	20	0	0	20	2	0
Bayesian Methods													
spikeslab	Spike-and-Slab Prior	20	2	0	20	75	0	20	8	0	20	25	0
SSLASSO	Spike-and-Slab Lasso	17	0	3	8	0	12	1	0	19	13	0	7
horseshoe	Horseshoe Prior, TC	20	0	0	20	2	0	20	0	0	3	0	17
horseshoe	Horseshoe Prior, HC	20	0	0	20	1	0	20	0	0	2	0	18
bayesreg	Horseshoe Prior	20	0	0	20	2	0	20	0	0	18	0	2
bayesreg	Horseshoe+ Prior	20	0	0	20	2	0	20	0	0	20	0	0
BayesS5	S5 Method												
monomvn	Bayesian Lasso												

[a] TS=True Signal, FP=False Positive, FN=False Negative
[b] p(50) < n(200), p(100) = n(100), p(200) > n(150), p(200) » n(50)

7.5 Type 3

Bla bla

Table 3: Summary of Type 3 Data Results

		p < n			p = n			p > n			p » n		
Package	Method	TS	FP	FN	TS	FP	FN	TS	FP	FN	TS	FP	FN
Frequentist Methods													
glmnet	Lasso	18	11	0	18	13	0	18	10	0	17	14	1
glmnet	Elastic Net	18	6	0	17	6	1	18	2	0	15	6	3
Bayesian Methods													
spikeslab	Spike-and-Slab Prior	17	10	1	18	82	0	18	19	0	9	36	9
SSLASSO	Spike-and-Slab Lasso	18	4	0	12	2	6	16	1	2	4	1	14
horseshoe	Horseshoe Prior, TC	17	0	1	17	0	1	17	0	1	5	2	12
horseshoe	Horseshoe Prior, HC	17	0	1	17	0	1	17	0	1	5	1	12
bayesreg	Horseshoe Prior	18	1	0	18	1	0	18	1	0	0	0	0
bayesreg	Horseshoe+ Prior	18	0	0	18	2	0	18	0	0	10	5	8
BayesS5	S5 Method												
monomvn	Bayesian Lasso												

[a] TS=True Signal, FP=False Positive, FN=False Negative

[b] p(50) < n(200), p(100) = n(100), p(200) > n(150), p(200) » n(50)

7.6 Type 4

Bla bla

Table 4: Summary of Type 4 Data Results

		p < n			p = n			p > n			p » n		
Package	Method	TS	FP	FN	TS	FP	FN	TS	FP	FN	TS	FP	FN
Frequentist Methods													
glmnet	Lasso	17	10	0	17	3	0	17	6	0	16	9	1
glmnet	Elastic Net	17	3	0	17	2	0	16	5	1	16	10	1
Bayesian Methods													
spikeslab	Spike-and-Slab Prior	17	7	0	17	77	0	17	47	0	16	27	1
SSLASSO	Spike-and-Slab Lasso	15	5	2	8	2	9	12	4	5	1	0	16
horseshoe	Horseshoe Prior, TC	16	0	1	16	0	1	16	0	1	2	0	15
horseshoe	Horseshoe Prior, HC	16	0	1	16	0	1	16	0	1	2	0	15
bayesreg	Horseshoe Prior	17	0	0	17	0	0	17	0	0	14	6	3
bayesreg	Horseshoe+ Prior	17	0	0	17	0	0	17	0	0	13	4	4
BayesS5	S5 Method												
monomvn	Bayesian Lasso												

[a] TS=True Signal, FP=False Positive, FN=False Negative

[b] p(50) < n(200), p(100) = n(100), p(200) > n(150), p(200) » n(50)

7.7 XGBoost

XGBoost selection of variables requires to set a threshold. For feature importance list with weights, see Appendix XX.

7.8 Discussion

Bla bla bla

Table 5: Summary of XGBoost Results for All Simulated Data

	p « n			p < n			p = n			p > n			p » n		
Data	TS	FP	FN	TS	FP	FN	TS	FP	FN	TS	FP	FN	TS	FP	FN
Type 1															
Type 2															
Type 3															
Type 4															
Type 5															

[a] TS=True Signal, FP=False Positive, FN=False Negative

[b] p(50) « n(500), p(50) < n(200), p(100) = n(100), p(200) > n(150), p(200) » n(50)

8 Crime Data

While the analysis of simulated data establishes a foundational understanding of various methodologies in a controlled environment, it is vital to extend this analysis to a real data set. Real data often present a more complex set of challenges and intricacies. Due to personal interest, the data set chosen is regarding sociological issues.

The dataset under study amalgamates socio-economic data from the 1990 Census, law enforcement data from the 1990 Law Enforcement Management and Admin Stats survey, and crime data from the 1995 Federal Bureau of Investigation Uniform Crime Reports (FBI UCR) for various communities in the U.S. Comprising 2215 instances and 147 attributes, the dataset includes factors related to community characteristics, law enforcement, and crime rates, focusing on 125 predictive attributes and 18 potential target variables (crime attributes). It is essential to recognise that the dataset has limitations due to discrepancies in population values, the omission of some communities, and the absence of certain data, particularly regarding rapes. The FBI cautions against using this dataset as the sole criterion for evaluating communities, as it does not encompass all relevant factors. For a comprehensive list of variables included in the dataset, please refer to Appendix Table XX.

The wrangled data consist of 99 variables: 98 predictors and the target variable “Violent Crimes per 100k Population” with 1992 instances. The other variables were disregarded as community names were non-predictive. Additionally, the analysis disregarded variables with over 80% missing values to maintain data integrity and reliable outputs. The data set was donated to UC Irvine Machine Learning Repository and is accessible online. For a more comprehensive understanding, refer to UC Irvine Machine Learning Repository (Redmond, 2011).

8.1 Ethical Considerations

Algorithmic decision-making mechanisms permeate many sectors of modern life, from spam classification in emails to credit scoring and employment candidate assessment. However, concerns have emerged regarding transparency, accountability, and fairness, specifically when these systems predicate their decisions on historical data. There exists a risk of perpetuating biases against certain demographic groups identified by ‘sensitive attributes’, such as gender, age, race, or religion, should these groups have been historically correlated with higher risk factors (Komiyama et al., 2018). Such variables refer to data that could be used to predict attributes protected by anti-discrimination laws, where the prejudiced actions are directed towards individuals based on their membership in certain groups, rather than assessing them on their individual merits. The caution around discriminatory impacts can manifest in two significant forms: disparate treatment and disparate impact (Zafar et al., 2017). The former describes intentional discrimination against groups with evidence of explicit reference to group membership. The latter examines the unintentional yet potentially harmful consequences that decision-making processes can have on specific groups, and despite it being facially neutral, it can still contribute to unintentional discrimination.

Decision-making entities such as banks, consultancies, and universities must strive to build classifiers free from discrimination, even if their historical data might inherently contain discriminatory elements. Žliobaitė et al. (2011) highlight a legal case where a leading consultancy firm faced allegations of indirect racial discrimination. They used existing criminal records for pre-employment screening, inadvertently creating bias because of the data's historical correlation between race and criminality. Even though the firm did not intend to discriminate, its use of criminal records resulted in racial discrimination. This case underscores that discrimination can inadvertently occur, even when sensitive information is not explicitly employed in the model, and such indirect discrimination is also legally unacceptable.

Likewise, Komiyama et al. (2018) has pointed out that the mere exclusion of the 'sensitive variables' is not sufficient. The publication further proposed the fairness of an algorithm through a coefficient of determination (CoD) of the sensitive attributes as a constraint. The CoD measures the predictable proportion of the variance of an estimator from sensitive attributes, effectively extending the correlation coefficient for multiple sensitive characteristics. For a deeper exploration of this topic, particularly in the realms of linear and logistic regression, readers are directed to the works of (Scutari et al., 2022), (Komiyama et al., 2018), and (Žliobaitė et al., 2011).

In this thesis, the pre-selection of variables included a thoughtful consideration of data sensitivity. The crime data used here include the variables describing the percentage of the African American population and the percentage of foreign-born individuals. Though a more in-depth exploration of sensitivity could be a progressive step beyond this work, it was deemed pertinent to acknowledge ongoing developments in data decision methodologies in ethically charged contexts. Thus, this thesis pivots back to the exclusion of the aforementioned variables, particularly those that could raise potential legal and ethical concerns in the application of the final model. Furthermore, the analysis explores interactions between variables to uncover potential underlying complex relationships. This approach aims to commit to bias prevention and the promotion of ethical data analysis.

8.2 Exploratory Data Analysis

Before proceeding with model fitting, conducting exploratory data analysis is standard statistical practice, it is important to spot any unwanted data characteristics that could adversely impact the models.

Firstly, to apply linear regression, it is vital to assess the normality assumption underlying it. The Shapiro-Wilk Normality test yields p-values well below 0.05 for all variables, providing evidence to reject the normality hypothesis. Given the skewness in most variables' distribution, illustrated by the sample histogram in $FigureXX$, a $log(X + 1)$ transformation is applied across all variables. While this aids in normalising the data, it also complicates the interpretation of variable importance later. Post-transformation, the Normality test shows marginal improvement but still falls short of confirming normal distribution for all variables, see $FigureXX$.

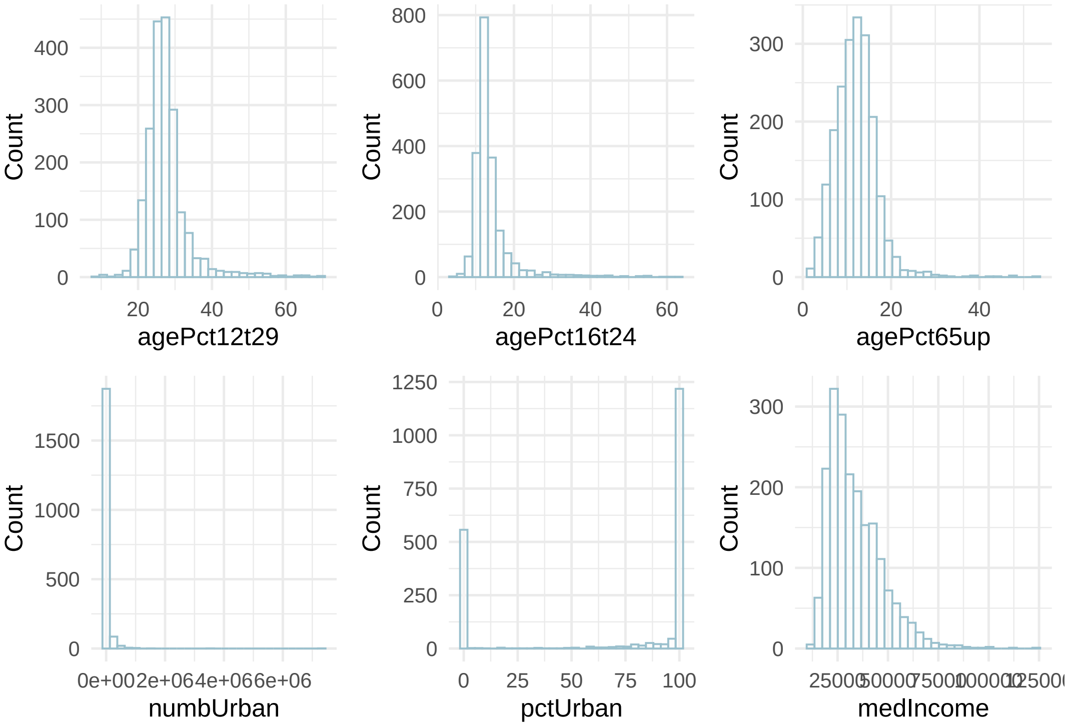

Figure 8: Crime Data: Sample Histograms Illustrating Variable Distribution Prior to Transformation

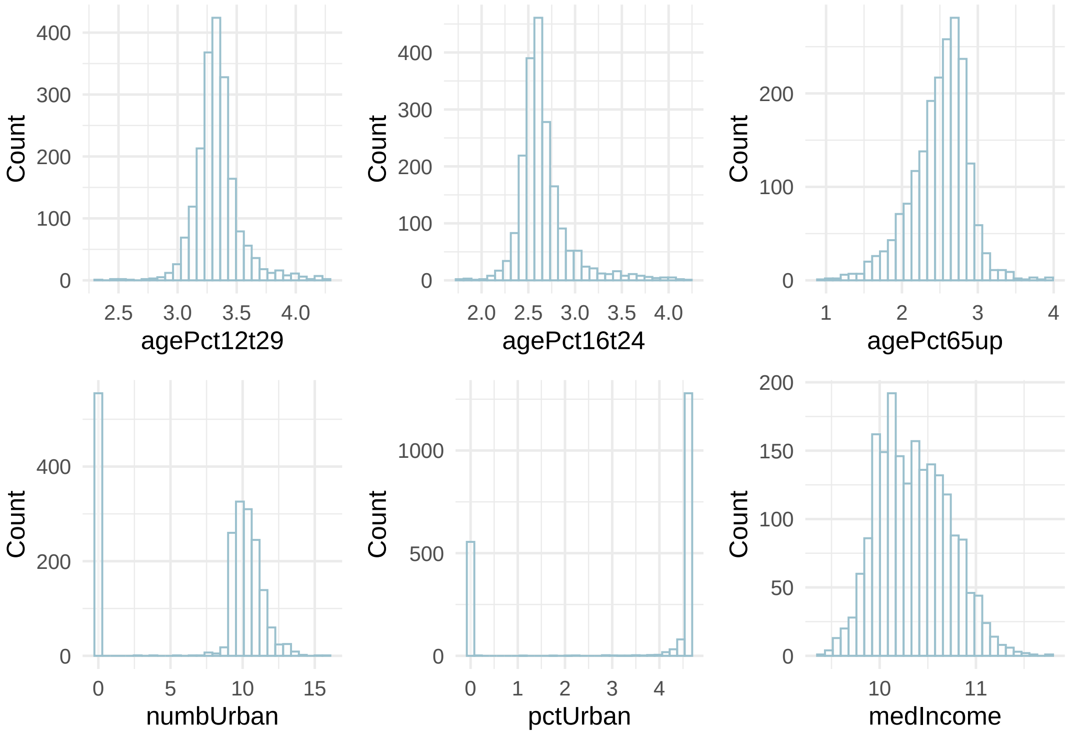

Figure 9: Crime Data: Sample Histograms Demonstrating Variable Distribution Post Logarithmic Transformation

Addressing outliers is typically critical since they can influence model performance and alter the correlations between variables. Tukey's approach to spotting high outliers in data variables, as described in Kannan et al. (2015), entails determining the interquartile range IQR, which represents the difference between the third ($Q3$) and first ($Q1$) quartiles. Then the lower and upper bounds are then calculated as $Q1 - factorIQR$ and $Q3 + factorIQR$, respectively. After identifying 1300 high outliers, a significant portion of the data, the number reduces to 1216 following the transformation. Although this number is concerning, it also mirrors the complexity of real-world data. Consequently, these outliers will not be eliminated. Refer to $FigureXX$ for a representative selection of boxplots signifying the outliers.

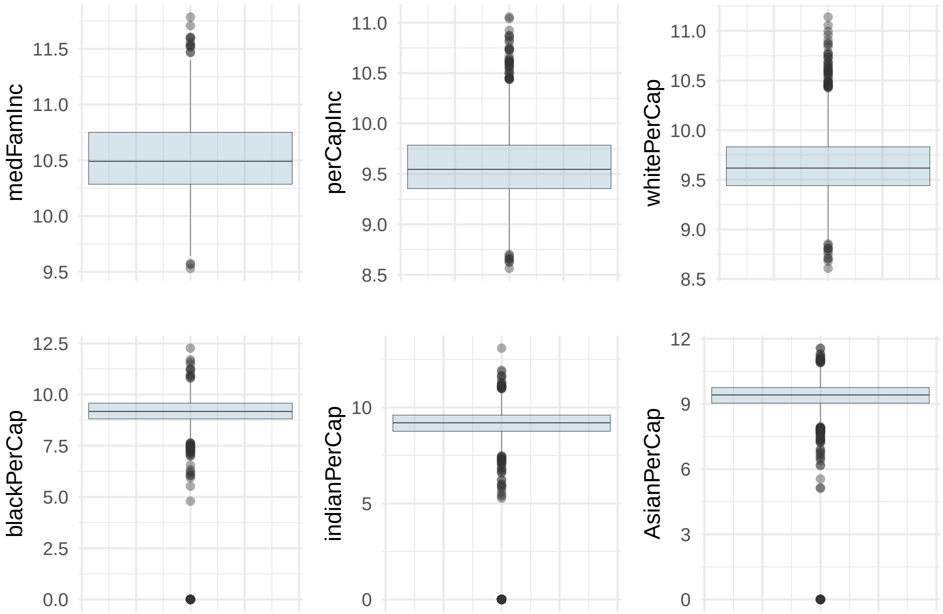

Figure 10: Crime Data: Sample of Boxplots for Visible Outliers

The linear correlations among variables have been examined, as shown in $FigureXX$. Given the high number of variables, a summary of correlations is deemed sufficient at this stage. The correlations vary from negligible to near absolute 1, signifying diverse relationships among the variables and underscoring the necessity for precise variable selection.

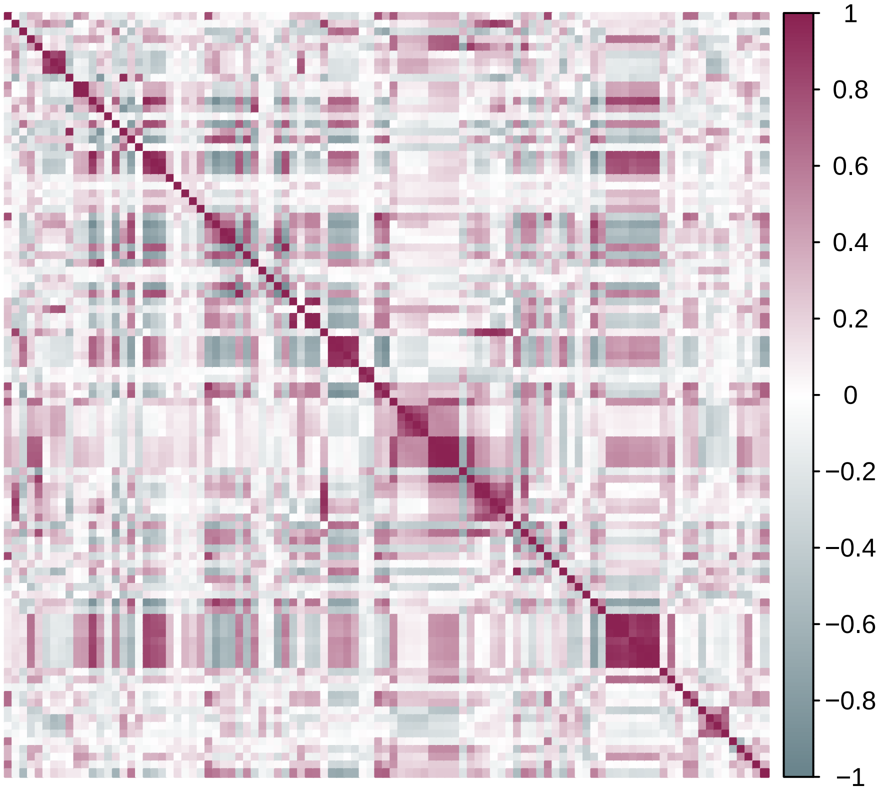

Figure 11: Crime Data: Linear Relationship Strength Among Variables

The scatter plot, see $FigureXX$, shows a sample of the analysis of linear correlations between predictors and the target. The correlations vary, some predictors show a linear relationship with the target.

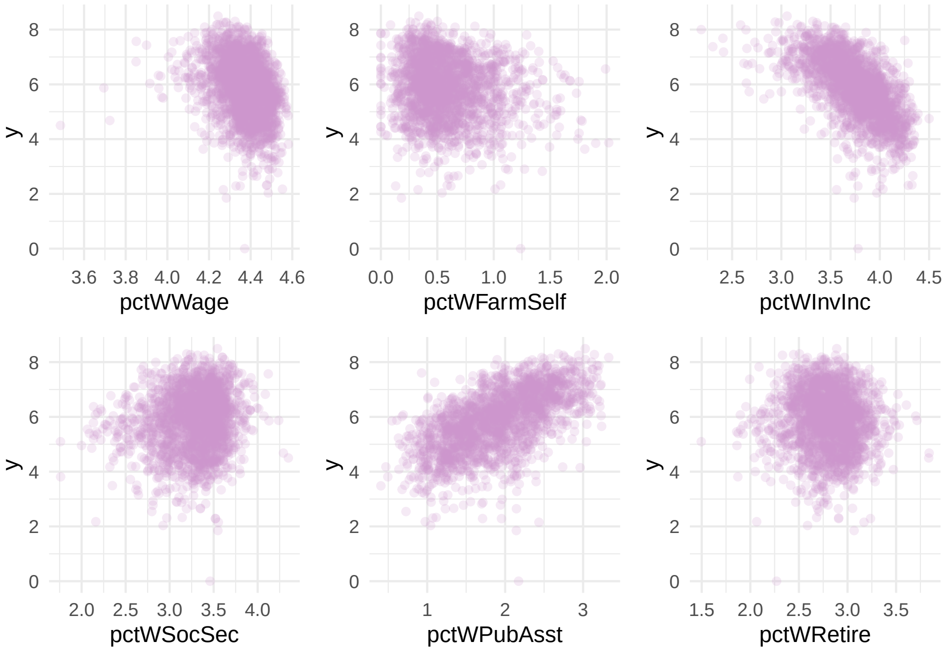

Figure 12: Crime Data: Sample Scatter Plot Analysis of Linear Correlations Between Predictors and Target Variable

With a sizeable sample and apparent relationships between variables and the target, the next step is applying the methodology that was tested on simulated data.

In the fitted models, interaction terms are included to examine the potential joint effects of certain pairs of variables:

The combined influence of the percentage of families with two parents (PctFam2Par) and the percentage of kids in two-parent households (PctKids2Par); the percentage of families with two parents (PctFam2Par) and the total percentage of divorces (TotalPctDiv); vacant boarded houses (PctVacantBoarded) and the percentage of households without a phone (PctHousNoPhone) are examined.

Other pairs investigated include the percentage of people in owner-occupied households (PctPersOwnOccup) with the percentage of people in densely populated houses (PctPersDenseHous); the percentage of houses with fewer than three bedrooms (PctHousLess3BR) with median number of bedrooms per house (MedNumBR); the number of homeless people in shelters (NumInShelters) and the number of homeless people on the streets (NumStreet); and the number of vacant houses (HousVacant) with the percentage of houses that are owner-occupied (PctHousOwnOcc).

Lastly, specific socioeconomic factors are compared with median income, such as the percentage of families with investment income (pctWInvInc), the percentage of the population under poverty (PctPopUnderPov), unemployment rates (PctUnemployed), public assistance rates (pctWPubAsst), population under poverty (PctPopUnderPov), percentage of individuals with less than 9th-grade education (PctLess9thGrade), high school graduation rates (PctNotHSGrad), divorce rates (TotalPctDiv), families with two parents (PctFam2Par), and the number of kids born to never married (NumKidsBornNeverMar). Each interaction could provide a more nuanced understanding of the complex factors influencing crime rates.

9 Crime Data Study Results

10 Conclusions

11 Appendix

11.1 Tables

Table 6: Crime Data: Summary of Selected Variables and Their Characteristics for Model Fitting

Variable	Included	Type
US state	No	Nominal
numeric code for county	No	Categorical
numeric code for community	No	Categorical
community name	No	Text
fold number	No	Categorical
population of community	Yes	Continuous
mean people per household	Yes	Continuous
percentage of population that is african american	No	Continuous
percentage of population that is caucasian	Yes	Continuous
percentage of population that is of asian heritage	Yes	Continuous
percentage of population that is of hispanic heritage	Yes	Continuous
percentage of population that is 12-21 in age	Yes	Continuous
percentage of population that is 12-29 in age	Yes	Continuous
percentage of population that is 16-24 in age	Yes	Continuous
percentage of population that is 65 and over in age	Yes	Continuous
number of people living in areas classified as urban	Yes	Continuous
percentage of people living in areas classified as urban	Yes	Continuous
median household income	Yes	Continuous
percentage of households with wage or salary income in 1989	Yes	Continuous
percentage of households with farm or self employment income in 1989	Yes	Continuous
percentage of households with investment / rent income in 1989	Yes	Continuous
percentage of households with social security income in 1989	Yes	Continuous
percentage of households with public assistance income in 1989	Yes	Continuous
percentage of households with retirement income in 1989	Yes	Continuous
median family income	Yes	Continuous
per capita income	Yes	Continuous
per capita income for caucasians	Yes	Continuous
per capita income for african americans	Yes	Continuous
per capita income for native americans	Yes	Continuous
per capita income for people with asian heritage	Yes	Continuous

Table 6: Crime Data: Summary of Selected Variables and Their Characteristics for Model Fitting *(continued)*

Variable	Included	Type
per capita income for people with other heritage	Yes	Continuous
per capita income for people with hispanic heritage	Yes	Continuous
number of people under the poverty level	Yes	Continuous
percentage of people under the poverty level	Yes	Continuous
percentage of people 25 and over with less than a 9th grade education	Yes	Continuous
percentage of people 25 and over that are not high school graduates	Yes	Continuous
percentage of people 25 and over with a bachelors degree or higher education	Yes	Continuous
percentage of people 16 and over, in the labor force, and unemployed	Yes	Continuous
percentage of people 16 and over who are employed	Yes	Continuous
percentage of people 16 and over who are employed in manufacturing	Yes	Continuous
percentage of people 16 and over who are employed in professional services	Yes	Continuous
percentage of people 16 and over who are employed in management	Yes	Continuous
percentage of people 16 and over who are employed in professional occup.	Yes	Continuous
percentage of males who are divorced	Yes	Continuous
percentage of males who have never married	Yes	Continuous
percentage of females who are divorced	Yes	Continuous
percentage of population who are divorced	Yes	Continuous
mean number of people per family	Yes	Continuous
percentage of families (with kids) that are headed by two parents	Yes	Continuous
percentage of kids in family housing with two parents	Yes	Continuous
percent of kids 4 and under in two parent households	Yes	Continuous
percent of kids age 12-17 in two parent households	Yes	Continuous
percentage of moms of kids 6 and under in labor force	Yes	Continuous
percentage of moms of kids under 18 in labor force	Yes	Continuous
number of kids born to never married	Yes	Continuous
percentage of kids born to never married	Yes	Continuous
total number of people known to be foreign born	Yes	Continuous
percentage of immigrants who immigated within last 3 years	Yes	Continuous
percentage of immigrants who immigated within last 5 years	Yes	Continuous
percentage of immigrants who immigated within last 8 years	Yes	Continuous
percentage of immigrants who immigated within last 10 years	Yes	Continuous
percent of population who have immigrated within the last 3 years	Yes	Continuous

Table 6: Crime Data: Summary of Selected Variables and Their Characteristics for Model Fitting *(continued)*

Variable	Included	Type
percent of population who have immigrated within the last 5 years	Yes	Continuous
percent of population who have immigrated within the last 8 years	Yes	Continuous
percent of population who have immigrated within the last 10 years	Yes	Continuous
percent of people who speak only English	Yes	Continuous
percent of people who do not speak English well	Yes	Continuous
percent of family households that are large (6 or more)	Yes	Continuous
percent of all occupied households that are large (6 or more people)	Yes	Continuous
mean persons per household (numeric - decimal)	Yes	Continuous
mean persons per owner occupied household	Yes	Continuous
mean persons per rental household (numeric - decimal)	Yes	Continuous
percent of people in owner occupied households (numeric - decimal)	Yes	Continuous
percent of persons in dense housing (more than 1 person per room)	Yes	Continuous
percent of housing units with less than 3 bedrooms	Yes	Continuous
median number of bedrooms	Yes	Continuous
number of vacant households	Yes	Continuous
percent of housing occupied	Yes	Continuous
percent of households owner occupied	Yes	Continuous
percent of vacant housing that is boarded up	Yes	Continuous
percent of vacant housing that has been vacant more than 6 months	Yes	Continuous
median year housing units built	Yes	Continuous
percent of occupied housing units without phone (in 1990, this was rare!)	Yes	Continuous
percent of housing without complete plumbing facilities	Yes	Continuous
owner occupied housing - lower quartile value	Yes	Continuous
owner occupied housing - median value	Yes	Continuous
owner occupied housing - upper quartile value	Yes	Continuous
rental housing - lower quartile rent	Yes	Continuous
rental housing - median rent (Census variable H32B from file STF1A)	Yes	Continuous
rental housing - upper quartile rent	Yes	Continuous
median gross rent (Census H43A from STF3A - with utilities)	Yes	Continuous
median gross rent as a percentage of household income	Yes	Continuous
median owners cost as a pct of household income (mortgage)	Yes	Continuous
median owners cost as a pct of household income (without mortgage)	Yes	Continuous
number of people in homeless shelters	Yes	Continuous

Table 6: Crime Data: Summary of Selected Variables and Their Characteristics for Model Fitting *(continued)*

Variable	Included	Type
number of homeless people counted in the street	Yes	Continuous
percent of people foreign born	No	Continuous
percent of people born in the same state as currently living	Yes	Continuous
percent of people living in the same house as in 1985 (5 years before)	Yes	Continuous
percent of people living in the same city as in 1985 (5 years before)	Yes	Continuous
percent of people living in the same state as in 1985 (5 years before)	Yes	Continuous
number of sworn full time police officers	No	Continuous
sworn full time police officers per 100K population	No	Continuous
number of sworn full time police officers in field operations	No	Continuous
sworn full time police officers in field operations	No	Continuous
total requests for police	No	Continuous
total requests for police per 100K popuation	No	Continuous
total requests for police per police officer	No	Continuous
police officers per 100K population	No	Continuous
a measure of the racial match between the community and the police force	No	Continuous
percent of police that are caucasian	No	Continuous
percent of police that are african american	No	Continuous
percent of police that are hispanic	No	Continuous
percent of police that are asian	No	Continuous
percent of police that are minority of any kind	No	Continuous
number of officers assigned to special drug units	No	Continuous
number of different kinds of drugs seized	No	Continuous
police average overtime worked	No	Continuous
land area in square miles	Yes	Continuous
population density in persons per square mile	Yes	Continuous
percent of people using public transit for commuting	Yes	Continuous
number of police cars	No	Continuous
police operating budget	No	Continuous
percent of sworn full time police officers on patrol	No	Continuous
gang unit deployed	No	Categorical
percent of officers assigned to drug units	Yes	Continuous
police operating budget per population	No	Continuous

Table 6: Crime Data: Summary of Selected Variables and Their Characteristics for Model Fitting *(continued)*

Variable	Included	Type
total number of violent crimes per 100K popuation	Target	Continuous

11.2 Plots

Simulated Data

Crime Data

Figure 13: Crime Data: Histograms Illustrating Variable Distribution Prior to Transformation

11.3 Code

For the full code, which includes the entire project, please access the publicly available GitHub repository. The files included are as listed:

1. data_crime_raw.R - handling of the Crime data
2. simulate_data.R - Type 1 through Type 4 data simulation
3. functions.R - functions for fitting the all methodology
4. main.R - main working file
5. read.me - text file describing the files contained in the repository

6. crime_raw.csv - copy of the file containing the Crime data
7. dissertation.rmd - the dissertation RMarkdown file

Snippets of code follow:

Data Simulation

Crime Data

Functions

Main Working File

12 List of Figures and Tables

List of Figures

List of Tables

13 Bibliography

Barker, R. J., & Link, W. A. (2013). Bayesian multimodel inference by RJMCMC: A gibbs sampling approach. *The American Statistician*, *67*(3), 150–156. https://doi.org/10.1080/00031305.2013.791644

Bhadra, A., Datta, J., Polson, N. G., & Willard, B. T. (2016). *Default bayesian analysis with global-local shrinkage priors*. https://doi.org/10.48550/arXiv.1510.03516

Bhattacharya, A., Chakraborty, A., & Mallick, B. K. (2016). *Fast sampling with gaussian scale-mixture priors in high-dimensional regression*. http://arxiv.org/abs/1506.04778

Bradley, P. C., & Chib, S. (1995). Bayesian model choice via markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*, 473–484. https://doi.org/10.1111/j.2517-6161.1995.tb02042.x

Breheny, P., & Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics*, *5*(1), 232–253. https://doi.org/10.1214/10-AOAS388

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, *97*, 465–480. https://doi.org/10.1093/biomet/asq017

Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. https://doi.org/10.1145/2939672.2939785

Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *70*, 849–911. https://doi.org/10.1111/j.1467-9868.2008.00674.x

Gelling, N., Schofield, M. R., & Barker, R. J. (2019). R package rjmcmc: Reversible mump MCMC using post-processing. *Australian and New Zealand Journal of Statistics*, *61*, 189–212. https://doi.org/10.1111/ANZS.12263

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2020). *Bayesian data analysis. Third edition* (pp. 165–175).

Geweke, J. (1996). *Variable selection and model comparison in regression* (Working Papers 539). Federal Reserve Bank of Minneapolis. https://ideas.repec.org/p/fip/fedmwp/539.html

Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, *82*, 711–732. https://doi.org/10.2307/2337340

Guo, R., Zhao, Z., Wang, T., Liu, G., Zhao, J., & Gao, D. (2020). Degradation state recognition of piston pump based on ICEEMDAN and XGBoost. *Applied Sciences*, *10*, 6593. https://doi.org/10.3390/app10186593

Hans, C. (2009). Bayesian lasso regression. *Biometrika*, *96*(4), 835–845. http://www.jstor.org/stable/27798870

Hans, C., Dobra, A., & West, M. (2007). Shotgun stochastic search for "large p" regression. *Journal of* *the American Statistical Association*, *102*, 507–516. https://doi.org/10.1198/016214507000000121

Hastie, T., Tibshirani, R., & Friedman, J. H. (2017). *The elements of statistical learning: Data mining, inference, and prediction* (2nd Edition). Springer. http://www.springer.com/series/692

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*(4), 382–401. http://www.jstor.org/stable/2676803

Ishwaran, H., Kogalur, U. B., & Rao, J. S. (2010). Spikeslab: Prediction and variable selection using spike and slab regression. *R Journal*, *2*, 68–73. https://doi.org/10.32614/rj-2010-018

Ishwaran, H., & Rao, J. S. (2005). Spike and slab variable selection: Frequentist and bayesian strategies. In *Annals of Statistics* (Vol. 33, pp. 730–773). https://doi.org/10.1214/009053604000001147

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, *31*(2), 203–222. https://doi.org/10.1017/S030500410001330X

Jianqing, F., Moore, F. L., & Jinchi, L. (2010). A selective overview of variable selection in high dimensional feature space. In *Statistica Sinica*.

Johnstone, I. M., & Silverman, B. W. (2004). Needles and straw in haystacks: Empirical BAYES estimates of possibly sparse sequences. *Annals of Statistics*, *32*, 1594–1649. https://doi.org/10.1214/009053604000000030

Kahneman, D. (2011). *Thinking, fast and slow.* Farrar, Straus; Giroux.

Kaliyaperumal, S. K., Kuppusamy, M., & Arumugam, S. (2015). Labeling methods for identifying outliers. In *International Journal of Statistics and Systems* (Vol. 10, pp. 231–238). http://www.ripublication.com

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773. https://doi.org/10.2307/2291091

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*(4598), 671–680. https://doi.org/10.1126/science.220.4598.671

Komiyama, J., Takeda, A., Honda, J., & Shimao, H. (2018). Nonconvex optimization for regression with fairness constraints. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th international conference on machine learning* (Vol. 80, pp. 2737–2746). PMLR. https://proceedings.mlr.press/v80/komiyama18a.html

Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. In *Source: Journal of the American Statistical Association* (Vol. 94, pp. 1372–1381).

Lempers, F. B. (1971). *Posterior probabilities of alternative linear models: Some theoretical considerations and empirical experiments.* Rotterdam University Press.

Li, J. (2020). *When to not use XGBoost?* https://www.kaggle.com/discussions/general/196542

Makalic, E., & Schmidt, D. F. (2016). *High-dimensional bayesian regularised regression with the BayesReg package.*

Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, *83*, 1023. https://doi.org/10.2307/2290129

Moran, G. E., Ročková, V., & George, E. I. (2019). Variance prior forms for high-dimensional bayesian variable selection. *Bayesian Analysis*, *14*. https://doi.org/10.1214/19-BA1149

Park, T., & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, *103*(482), 681–686. https://doi.org/10.1198/016214508000000337

Polson, N. G., & Scott, J. G. (2010). Shrink globally, act locally: Sparse bayesian regularization and prediction*. In *Bayesian Statistics 9* (pp. 501–538). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199694587.003.0017

Redmond, M. (2011). *Communities and Crime Unnormalized*. UCI Machine Learning Repository.

Ročková, V., & George, E. I. (2018). The spike-and-slab LASSO. *Journal of the American Statistical Association*, *113*, 431–444. https://doi.org/10.1080/01621459.2016.1260469

Rue, H. (2001). Fast sampling of gaussian markov random fields. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *63*, 325–338. https://doi.org/10.1111/1467-9868.00288

Scutari, M., Panero, F., & Proissl, M. (2022). Achieving fairness with a simple ridge penalty. *Statistics and Computing*, *32*, 77. https://doi.org/10.1007/s11222-022-10143-w

Shin, M., Bhattacharya, A., & Johnson, V. E. (2015). *Scalable bayesian variable selection using nonlocal prior densities in ultrahigh-dimensional settings*. http://arxiv.org/abs/1507.07106

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Linde, A. van der. (2014). The deviance information criterion: 12 years on. In *Journal of the Royal Statistical Society. Series B (Statistical Methodology): Vols. 76. No. 3* (pp. 485–493).

Steyerberg, E. W. (2019). *Clinical prediction models: Statistics for biology and health clinical prediction models a practical approach to development, validation, and updating second edition*. Springer Cham. https://doi.org/10.1007/978-3-030-16399-0

Tadesse, M. G., & Vannucci, M. (2022). *Handbook of bayesian variable selection*. Chapman & Hall. https://doi.org/10.1080/00031305.2013.791644

Tanha, K., Mohammadi, N., & Janani, L. (2017). P-value: What is and what is not. *Medical Journal of the Islamic Republic of Iran*, *31*, 377–378. https://doi.org/10.14196/mjiri.31.65

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288. http://www.jstor.org/stable/2346178

Train, K. E. (2012). *Discrete choice methods with simulation* (2nd Edition, pp. 284–314). Cambridge University Press. https://doi.org/10.1017/CBO9780511805271

Wang, J., Xu, J., Zhao, C., Peng, Y., & Wang, H. (2019). An ensemble feature selection method for high-dimensional data based on sort aggregation. *Systems Science and Control Engineering*, *7*, 32–39. https://doi.org/10.1080/21642583.2019.1620658

Watkins, J. C. (2010). *Theory of statistics contents*. University of Arizona.

Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. *26th International World Wide Web Conference, WWW 2017*, 1171–1180. https://doi.org/10.1145/3038912.3052660

Žliobaitė, I., Kamiran, F., & Calders, T. (2011). Handling conditional discrimination. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 992–1001. https://doi.org/10.1109/ICDM.2011.72

Zou, H., & Hastie, T. (2005). Zou h, hastie t. Regularization and variable selection via the elastic net. J r statist soc b. 2005;67(2):301-20. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x