Bayesian linear models, Bayesian ridge and Bayesian LASSO:

Parameter selection and predictive performance on the Boston housing data set

Oliver Chipperfield

30th August, 2019

School of mathematics and statistics
University of Glasgow

Contents

1 Introduction 1

1.1 Linear models and regularisation . . . 1

1.2 Bayesian hierachial models and regularisation . . . 2

1.3 Boston housing data . . . 3

2 Methods 4

2.1 Data and definitions . . . 4

2.2 The Bayesian linear model . . . 4

2.3 The Bayesian ridge model . . . 6

2.4 The Bayesian LASSO model . . . 7

2.5 Estimating posterior marginal densities . . . 8

2.6 Assessing predictive performance . . . 9

2.7 Assessing convergence . . . 9

2.8 Choice of priors . . . 9

2.9 Software and packages . . . 10

3 Results 11

3.1 ‘13 variable’ results . . . 11

3.2 ‘22 variable’ results . . . 11

3.3 Prediction performance . . . 13

3.4 Priors . . . 15

4 Discussion 16

4.1 Parameter estimation . . . 16

4.2 Interpretating the parameters . . . 17

4.3 Prediction performance . . . 18

4.4 Extensions to Bayesian regularisation . . . 19

5 Conclusions 19

6 Appendix 21

6.1 Model traceplots . . . 21

6.2 Bayesian linear model . . . 22

6.3 Bayesian ridge model . . . 23

6.4 Bayesian LASSO model . . . 25

6.5 Laplace as a mixture normal . . . 29

References 31

1 Introduction

Regression models are a broad and common class of statistical models that are used to both infer the importance of independent variables on the impact of a dependent variable, and to forecast future results given a set of independent variables

This project aims to compare and contrast Bayesian approaches to linear modelling and regularised regression, namely the Bayesian linear model (BLM), the Bayesian ridge and the Bayesian LASSO. To do this, these models will be defined, derived and applied to the Boston Housing data set, a data set describing median house prices and a set of covariates (Newman et al., 1998).

The parameter estimates of the these models will be compared and contrasted in the context of variable selection. In addition the predictive performance of these models will be assessed and compared. Parameter estimates will also be interpreted in relation to their role as drivers of median house prices in Boston.

1.1 Linear models and regularisation

Linear models are a simple class of statistical models where a univariate response vector ($\mathbf{y}$) is modelled using a design matrix ($\mathbf{X}$) of p covariates, and a vector β of p parameter estimates.

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \quad \epsilon \sim \mathsf{N}(0, \sigma^2) \tag{1}$$

The very simplest linear model is the ordinary least squares (OLS) model. Under the OLS model the maximum likelihood estimator of $\hat{\beta}$ is $(\mathbf{X^T X})^{-1}\mathbf{X^T y}$ (Freedman, 2009). Regularised regression are a class of regression models where the maximum likelihood / cost function includes a penalty λ which penalises the function for fitting larger regression coefficients (Abhijit, 2017). These approaches prove most effective with high-dimensional regression models where multi-collinearity between covariates can cause over fitting (Dormann et al., 2013). Regularised regression models have proven popular in applied machine-learning, where a forecaster can relatively naively include a large number of dependent variables in the design matrix without the fear of over-fitting.

Ridge regression is one type of regularised regression. Ridge regression has historical origins but has been applied in practice since the later half of the 20th century (e.g. Hoerl and Kennard (1970), Marquardt and Snee (1975)). Under this approach the likelihood / cost function for the estimation of β would be

$$L(\hat{\beta}) = (\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta}) + \lambda \sum_{i=1}^{p} \hat{\beta}_i^2 \tag{2}$$

The $\hat{\beta}$ parameters that best minimise this function would be the point estimates for β (Rifkin and Lippert, 2007). As $\lambda \to 0$ the cost function equates to the OLS estimate of β. The greater the value of λ the greater the penalty for larger β parameters, therefore causing less important parameters to shrink towards 0 (Abhijit, 2017).

Another class of regularised regression is the LASSO. The frequentist LASSO was first introduced by Tibshirani (1996) as an improvement to existing model reduction approaches such as ridge regression. Under this approach the likelihood / cost function for the estimation of β would be

$$L(\hat{\beta}) = (\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta}) + \lambda \sum_{i=1}^{p} |\hat{\beta}_i| \tag{3}$$

As λ increases the rate of parameter shrinkage is greater, therefore resulting in a greater propensity for parameters to be estimated as zero.

Under a frequentist framework, it's not possible to estimate the best value for λ analytically. In practice cross-validation is carried out to determine the value of λ which best minimises the mean-squared-error (MSE) (Friedman et al., 2010). One major drawback of frequentist regularised regression is that standard-errors are not easily constructed for the point estimates of the regression parameters (Kyung et al., 2010) therefore preventing effective inference or model testing (Lockhart et al., 2014).

1.2 Bayesian hierachial models and regularisation

All Bayesian statistics are fundamentally based on Bayes theorem, $p(a|b) \propto p(b|a)p(a)$. In Bayesian modelling, the data model, e.g. $\mathbf{y} \sim \text{N}(\mathbf{X}\beta, \sigma^2\mathbf{I_n})$ can be regarded as the likelihood function of the parameters given the data, i.e. $p(\mathbf{y}|\beta) = L(\mathbf{y}|\beta)$. In frequentist statistics, the parameters are assumed to be unknown but fixed, in Bayesian statistics the parameters are treated as random variables (Hoff, 2009). Therefore it's only necessary to define a prior probability for β to ascertain $p(\beta|\mathbf{y})$:

$$p(\beta|\mathbf{y}) \propto L(\mathbf{y}|\beta)p(\beta) \tag{4}$$

In hierachial Bayesian modelling this relationship is taken one step further. Different parameters can be layered on top of one another as long as the parameters are conditionally independent (Allenby et al., 2005). For example, using Bayes rule and through factorisation the joint posterior distribution for $\beta, \sigma^2|\mathbf{y}$ can be expressed as

$$p(\beta, \sigma^2|\mathbf{y}) \propto L(\mathbf{y}|\beta, \sigma^2)p(\beta|\sigma^2)p(\sigma^2) \tag{5}$$

The marginal posterior distribution can sometimes be derived given the posterior joint distribution (e.g $p(\beta_i|\mathbf{y})$) though integration, however in practice this is often not analytically possible. The full conditional distribution of a given parameter can also be derived by dropping irrelevant terms from the joint distribution (Hoff, 2009). Alternatively Monte-Carlo methods can be used to estimate the marginal posterior densities (see section 2.5).

The marginal posterior densities of the parameters estimate the probability of a parameter being a certain value. Therefore a 95% credible interval reflects the probability that a parameter occupies a range. This differs from the frequentist interpretation of a confidence interval which relates to the proportion of intervals containing the true parameter if the experiment was repeated an infinite number of times (Freedman, 2009).

Table 1: Descriptions of the response variable and the original 13 covariates of the Boston housing data (Harrison and Rubinfeld, 1978) after log and binary transformations

Variable	Description
medv	Median owner-occupied home value (1000's USD) (log-transformed).
age	Proportion of units built before 1940.
b	1000(B - 0.63)^2 where B is the proportion of people who are black.
chas	Bounds charles river (0 = False, 1 = True).
crim	Per capita crime rate (log-transformed).
dis	Weighted distance to five employment centres (log-transformed).
indus	Proportion non-retail business acres (log-transformed).
lstat	Proportion of population in low status occupations (log-transformed).
nox	Nitric oxide concentration (pp 10m) (log-transformed).
ptratio	Pupil-teacher ratio (log-transformed).
rad_over	Index of accessibility to radial highways (< 10 = 0, > 10 = 1).
rm	Average number of rooms per unit.
tax_over	Property-tax rate per 10,000 USD (< 500 = 0, > 500 = 1).
zn_over	Proportion of land zoned for lots over 25,000 sq. ft. (0 = 0, > 0 = 1).

It can be argued that the Bayesian credible interval is more intuitive and appropriate for inference (Grzenda, 2015).

Linear models are often easily extended to a Bayesian framework. The OLS extension is well established (Hoff, 2009) as is the ridge regression (Hsiang, 1975). Tibshirani (1996) noted that the LASSO regression could also be expressed as a Bayesian hierachial model using Laplace priors (see section 2).

Bayesian approaches to regularised regression have two major advantages. Given that parameter estimates are expressed as marginal probability densities, credible intervals are very easily obtained, making parameter inference much easier (Kyung et al., 2010). Additionally, the penalty parameter λ can be expressed as a weakly-informative hyper-prior, therefore reducing the need to carry out cross-validation to optimise λ.

1.3 Boston housing data

The data is sourced from the R package `mlbench` (Newman et al., 1998) and was originally compiled by (Harrison and Rubinfeld, 1978). The original aims of the Harrison and Rubinfeld (1978) paper focused on methodological problems associated with estimating how much people were willing to pay for cleaner air. To do this they analysed data from 506 census tracts in Boston recorded from the 1970 census. Median owner-occupied house prices were recorded along with 13 possible covariates (see table 1), amongst these nitric-oxide levels. The Harrison and Rubinfeld (1978) dataset since developed to be a common data set for testing and applying statistical methodologies, in particular those relating to variable selection and machine learning (e.g. Liang et al. (2019), Alhamzawi et al. (2019), Yue et al. (2017)).

2 Methods

2.1 Data and definitions

Given that median house prices displayed a strong skew and were bounded at 0, for the purposes of regression modelling, it was decided that median house price values should be log-transformed. In addition the variables, `crim`, `indus`, `nox`, `dis`, `lstat` and `ptratio` were log transformed in order to suppress the possible leverage effects of outliers (Freedman, 2009). The variables `rad`, `tax` and `zn` were recoded to binary variables (`rad_over`, `tax_over` and `zn_over`) determined by cut-off points. Figure 1 summarises the transformed covariates and their relationship to median house values.

It's argued that regularised regression is most applicable on data sets with a large number of collinear independent variables (Dormann et al., 2013). There's also often evidence of curvature in the relationship between `medv` and the covariates (see figure 1). In order to test the impact of regularisation on a higher dimensional collinear data set, and in order to more effectively model `medv`, an additional dataset was created where all continuous variables were squared to make nine additional covariates. Therefore there were two datasets tested, hereafter referred to as the **'13 variable'** data set and the **'22 variable'** data set.

Alternative design matrices were also considered. One approach was to add additional cubic terms for each continuous variable, results were similar to the '22 variable' regressor set. Alternatively a spline based approach was attempted where cubic basis functions were treated as regressors (using r function `bs` (R Core Team, 2018)). Prediction performance was moderately improved, however parameter interpretation became less clear.

Define $\mathbf{y}$ as a vector of length $n = 506$ of the log-transformed median house prices of the Boston census tracts (`medv`). $\mathbf{y}$ is centred and standardised to have a mean of 0 and standard deviation of 1. By centring the response variable about zero, we can define the intercept of the regression model as a flat invariant prior with $\mu = 0$. Define $\mathbf{X}$ to be the design matrix of length n and width p. All columns of $\mathbf{X}$ are standardised to have mean 0 and standard deviation of 1. By standardising the covariates, it's possible to use a common λ penalty value for all variables (Abhijit, 2017). Detmer and Slawski (2018) argue that it isn't necessary to standardise binary categorical variables, however this project will follow precedent (e.g. Abhijit (2017)). Define $\beta_1 ... \beta_p$ to be the p regression parameters of the regression model, β denotes the vector of regression parameters of length p.

2.2 The Bayesian linear model

The Bayesian linear model is hierachial conception of the OLS estimation of β (Hoff, 2009). The hierachial model for a Bayesian linear model can be defined as

$$\begin{aligned} \mathbf{y}|\beta, \sigma^2, \mathbf{X} &\sim \mathsf{N}(\mathbf{X}\beta, \sigma^2\mathbf{I_n}) \\ \beta|\sigma^2 &\propto 1 \\ \sigma^2 &\sim \text{Inv-Gamma}(a, b) \end{aligned} \tag{6}$$

Where $\mathbf{y}|\beta, \sigma^2, \mathbf{X}$ denotes the linear model and parameter σ^2 denotes the variance. The prior $\beta|\sigma^2$ is the

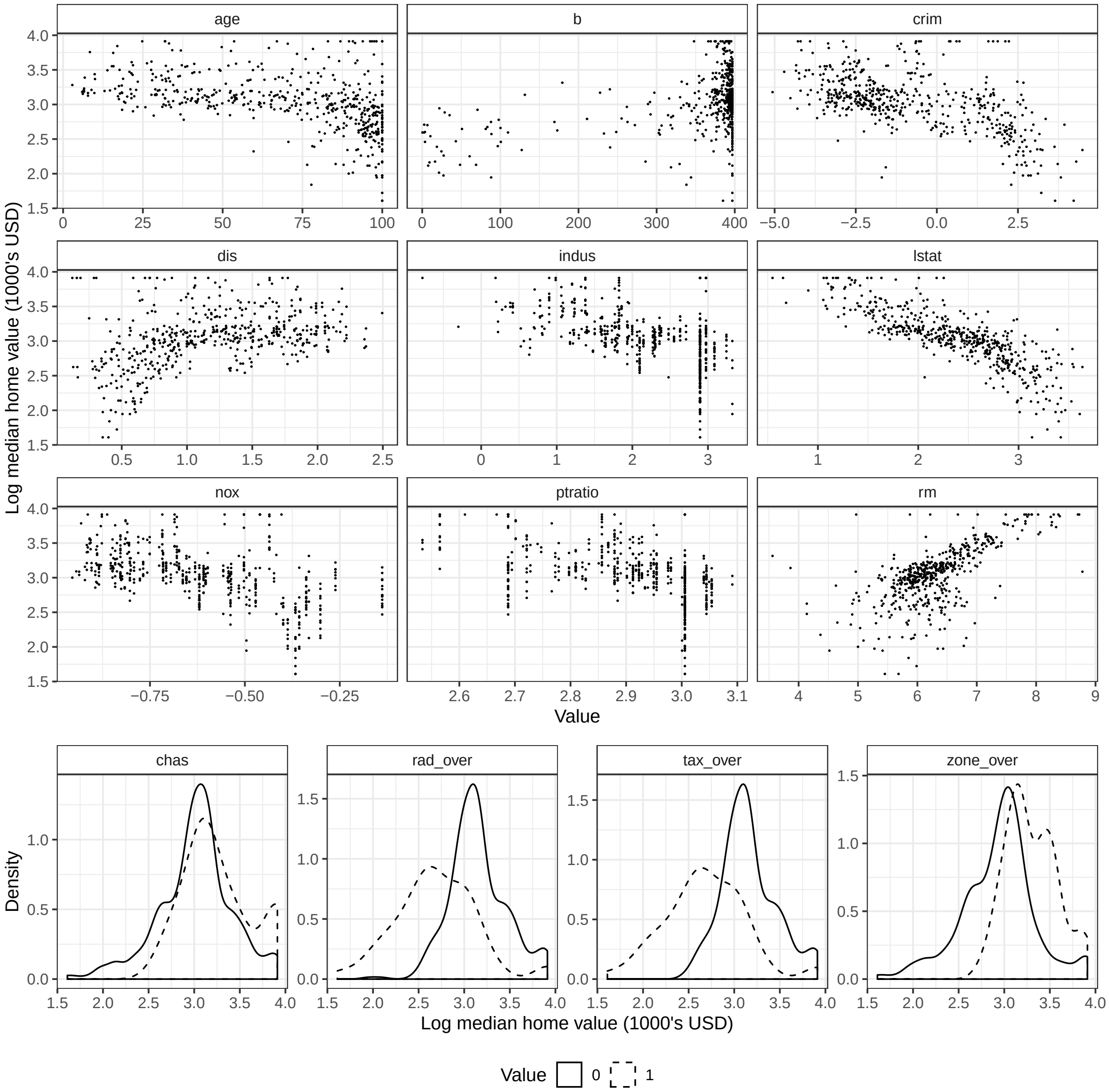

Figure 1: Scatterplots and histograms of Boston housing data (Harrison and Rubinfeld, 1978), after log and binary transformations

improper Jeffreys prior of a normal distribution with known variance. As noted in section **??**, the model is hierachial therefore the full conditional distribution of β and σ^2 can be derived without integrating through the joint distribution. Full details of the derivation of β and σ^2 can be seen in appendix 6.2. The conditional posterior distribution of β is

$$\beta|\sigma^2, \mathbf{y}, \mathbf{X} \sim \text{N}(\hat{\beta}, \sigma^2\mathbf{V}) \tag{7}$$

Where $\hat{\beta} = (\mathbf{X^TX})^{-1}\mathbf{X^Ty}$ and $\mathbf{V} = (\mathbf{X^TX})^{-1}$. Therefore the mean and maximum a posteriori estimation (MAP) of the posterior conditional distribution of β will be the same as the maximum likelihood estimator $\hat{\beta}$. The conditional posterior distribution of σ^2 is

$$\sigma^2|\mathbf{y}, \mathbf{X} \sim \text{Inv-Gamma}(\frac{n-p}{2} + a, \frac{1}{2}\text{SSE} + b) \tag{8}$$

Where SSE $= (\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta})$. Note that as $a \rightarrow 0$ and $b \rightarrow 0$, $E[\sigma^2|\mathbf{y}, \mathbf{X}] = (1/(n-p-2)) \times SSE$ which is slightly larger than the maximum likelihood estimator of $\hat{\sigma}^2 = (1/(n-p-1)) \times SSE$. Given that σ^2 depends on the data only, the Bayesian linear model is semi-conjugate and only requires simulation to estimate the posterior marginal distribution of β and the posterior predictive distribution.

2.3 The Bayesian ridge model

The Bayesian ridge model can also be formulated as hierachial model, it differs from a BLM in that the prior for β is no longer uninformative (Hsiang, 1975):

$$\begin{aligned} \mathbf{y}|\beta, \sigma^2, \mathbf{X} &\sim \text{N}(\mathbf{X}\beta, \sigma^2\mathbf{I_n}) \\ \beta|\sigma^2, \lambda &\sim \text{N}(\mathbf{0_p}, \mathbf{\Sigma}) \\ \sigma^2 &\sim \text{Inv-Gamma}(a, b) \\ \lambda &\sim \text{Gamma}(r, \delta) \end{aligned} \tag{9}$$

Where $\mathbf{y}|\beta, \sigma^2, \mathbf{X}$ is the linear model, $\mathbf{0_p}$ is a vector of zeros of length p and $\mathbf{\Sigma} = \frac{\sigma^2\mathbf{I_p}}{\lambda}$. The parameter σ^2 denotes the variance. Because the model is hierachial, it's possible to derive the full conditional distributions of β, λ and σ^2 without integrating through the joint distribution. Full details of the derivation can be found in appendix 6.3. The full conditional distribution of β is defined as

$$\beta|\sigma^2, \lambda, \mathbf{X} \sim \text{N}(\mathbf{E}, \mathbf{V^{-1}}) \tag{10}$$

Where $\mathbf{E} = \mathbf{V}\frac{\mathbf{X^Ty}}{\sigma^2}$ and $\mathbf{V} = (\mathbf{\Sigma^{-1}} + \frac{\mathbf{X^TX}}{\sigma^2})^{-1}$. It can be shown that the MAP of the $\beta = (\mathbf{X^TX} + \lambda\mathbf{I_p})^{-1}\mathbf{X^Ty}$ which is equal to the maximum likelihood estimator of the frequentist ridge cost-function (Rifkin and Lippert, 2007) (see appendix 6.3). The full conditional probability of σ^2 can be defined as

$$\sigma^2|\beta, \lambda, \mathbf{X} \sim \text{Inv-Gamma}(\frac{n+p}{2} + a, \frac{1}{2}\text{SSE} + b) \tag{11}$$

Where SSE $= (\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta})$. As $a \to 0$ and $b \to 0$, the expectation of $\sigma^2|\beta, \lambda, \mathbf{X}, \mathbf{y}$ is $(1/(n+p-2)) \times SSE$, which only equates to maximum likelihood estimate of σ^2 when $p = 1$, otherwise the estimator is biased and tends to underestimate σ^2. This is a known issue with Bayesian regularisation approaches (Moran et al., 2018). The posterior conditional distribution of λ can be defined as

$$\lambda|\beta, \sigma^2, \mathbf{X} \sim \text{Gamma}(\frac{p}{2} + r, \frac{1}{2\sigma^2}\beta^{\mathbf{T}}\beta + \delta) \tag{12}$$

Therefore, unlike the BLM, this model is not conjugate, and parameters cannot be derived analytically or by simulation. However as all conditional distributions can be defined as proper probability distributions it is possible to estimate the posterior densities through Gibbs sampling or other Monte-Carlo-Markov-chain (MCMC) methods.

2.4 The Bayesian LASSO model

There are several Bayesian forms of the LASSO, this project will use the Park and Casella (2008) formulation.

$$\begin{aligned} \mathbf{y}|\beta, \sigma^2, \mathbf{X} &\sim \text{N}(\mathbf{X}\beta, \sigma^2\mathbf{I_n}) \\ \beta_1 ... \beta_p|\sigma^2, \lambda^2, \mathbf{X} &\sim \text{Laplace}(0, \frac{\sigma^2}{\lambda^2}) \end{aligned} \tag{13}$$

Following Pericchi and Smith (1992), the vector of β coefficients can be redefined as a mixture multivariate normal with an exponential diagonal covariance matrix. See appendix 6.5 for proof.

$$\begin{aligned} \beta|\sigma^2, \tau_1^2 ... \tau_p^2, \lambda^2, \mathbf{X} &\sim \text{N}(\mathbf{0_p}, \sigma^2\mathbf{D}_\tau) \\ \tau_1^2 ... \tau_p^2|\lambda^2 &\sim \text{Exp}(\frac{1}{2}\lambda^2) \\ \sigma^2 &\sim \text{Inv-Gamma}(a, b) \\ \lambda^2 &\sim \text{Gamma}(r, \delta) \end{aligned} \tag{14}$$

Where $\mathbf{D}_\tau = diag(\tau_1^2 ... \tau_p^2)$, σ^2 denotes the variance, and λ^2 denotes the penalty parameter. Because the model is hierachial, it's possible to derive the full conditional distributions of β, $\tau_1^2 ... \tau_p^2$, σ^2 and λ^2 without integrating through the joint distribution. Full details can be seen in appendix 6.4. The conditional posterior distribution of β is

$$\beta|\sigma^2, \tau_1^2 ... \tau_p^2, \lambda^2, \mathbf{X} \sim \text{N}(\mathbf{A^{-1}X^Ty}, \sigma^2\mathbf{A^{-1}}) \tag{15}$$

Where $\mathbf{A} = \mathbf{X^TX} + \mathbf{D}_\tau^{\mathbf{-1}}$. It can be shown that the MAP of β is equal to arg.max$(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) - \lambda\sum_{i=1}^{p}|\beta_i|)$ which is equal to the minima of the frequentist LASSO cost function (see appendix 6.4). The posterior conditional distribution of σ^2 is defined as

$$\sigma^2|\beta, \lambda^2, \mathbf{X}, \mathbf{y} \sim \text{Inv-Gamma}(\frac{n+p}{2} + a, \frac{1}{2}\text{SSE} + \beta^{\mathbf{T}}\mathbf{D}_\tau^{\mathbf{-1}}\beta + b) \tag{16}$$

Where SSE $= (\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)$. Like ridge regression, as $a \to 0, b \to 0$ and $\lambda^2 \to 0$ the expectation of $\sigma^2|\beta, \lambda^2, \mathbf{X}, \mathbf{y}$ will tends towards a biased estimation of σ^2 of $(1/(n+p-2)) \times SSE$. The posterior conditional distributions of $\tau_1^2 ... \tau_p^2$ are defined in appendix 6.4. Finally, the posterior conditional distribution λ^2 is defined as

$$\lambda^2|\tau_i^2...\tau_p^2 \sim \text{Gamma}(p+r, \delta + \frac{1}{2}\sum_{i=1}^{p}\tau_i^2) \tag{17}$$

The Bayesian LASSO regression model is therefore not conjugate, however given that all posterior conditional distributions can be defined as proper distributions it is possible to estimate the posterior marginal densities using Gibbs Sampling or other Monte-Carlo techniques.

2.5 Estimating posterior marginal densities

As noted, the posterior densities of the parameters of interest can be simulated directly in the case of the BLM. For the Bayesian ridge and LASSO procedures however direct simulation is not possible. Instead Monte-Carlo-Markov-chains (MCMC) procedures must be used. MCMC procedures carry out random draws from the joint condition distribution of the parameters of interest. As long as the joint distribution is aperiodic, irreducible and recurrent, as the number of samples tends towards infinity the draws come to reflect the marginal posterior densities for each parameter (Brooks et al., 2011).

The Gibbs sampler is an MCMC procedure where samples are recursively drawn from the proper full conditional distributions of the parameters (Gelfand and Smith, 1990). The samples will come to represent samples from the marginal distributions for each parameter (Titterington, 1997). For both the ridge and LASSO implementations, the posterior conditional distributions can be derived and expressed as proper distributions. However Gibbs sampling is found to scale poorly on high-dimensional problems where the sample space becomes too large (Betancourt and Girolami, 2015). The sampler can also struggle to converge when distributions have high curvature (Betancourt, 2017).

For computational convenience, marginal posterior densities for all models were estimated using the R package `rstan` (Stan Development Team, 2019a). STAN uses the no-u-turn sampler to sample from the full joint distribution (Carpenter et al., 2017), the no U-turn sampler is a variant of Hamiltonian Monte-Carlo. Hamiltonian Monte-Carlo is a variant of the Metropolis-Hastings algorithm. The Metropolis-Hastings algorithm samples the parameter space using an acceptance/rejection criteria to move sampling to the highest density region of the parameter space (Brooks et al., 2011). Hamiltonian Monte-Carlo much improves the efficiency of the algorithm by using momentum parameters to inform the direction of the sampling. This approach reduces the number of rejected samples thus convergence occurs much more rapidly (Betancourt, 2017). `rstan` implements a further variant of Hamiltonian Monte-Carlo known as no-u-turn (NUTS) sampling. This self-optimises step-size and step-number parameters that are required in the Hamiltonian function (Hoffman et al., 2014). Marginal densities were estimated using four chains with 3000 draws per chain. The first 1000 draws were discarded. Following recommendations by Stan Development Team (2019b) thinning was not carried out.

The MAP of the marginal posterior distribution can be estimated from the sampler. Following `rjags` (Plummer, 2018) the MAP is estimated by fitting a kernel density estimator over the marginal posterior samples

(Kruschke, 2015). Gelman et al. (2004) notes that this approach can be over-sensitive to complex multi-modal distributions. Visual inspection of the posterior densities suggests that they're mono-model (see figures 2 and 3).

2.6 Assessing predictive performance

The predictive performance of each model is determined by calculating the root-mean-squared-prediction-error (RMSPE) with k-fold cross-validation. Using a k-fold cross-validation process, the data is split into k sets of testing and training data where the RMSPE of the kth trained model is calculated based on the kth set of testing data (Abhijit, 2017). The RMSPE is defined as

$$\text{RMSPE} = \sqrt{\frac{\sum_{i=1}^{n}(\tilde{y}_i - y_i)^2}{n}} \tag{18}$$

Where y_i denotes the ith element of a hold-out set of testing data, $\tilde{y}_i$ denotes the predicted value for y_i and n denotes the total sample size of the hold-out set of testing data. $\tilde{y}_i$ was calculated in two ways, using draws from the posterior predictive distribution, or via the estimated MAPs of the marginal posterior distribution. The MAP approach enables more direct comparisons to OLS, frequentist ridge and frequentist LASSO methods.

For comparison; frequentist OLS, ridge and LASSO models were calculated using `glmnet` (Simon et al., 2011). Point estimates were used to estimate RMSPE.

2.7 Assessing convergence

Convergence is only guaranteed asymptotically (Stan Development Team, 2019b). Under a finite number of draws it's therefore necessary to assess whether the MCMC has reached convergence. Informally convergence can be assessed by examining the trace plot of the draws from the posterior distribution, if there are no apparent trends in the trace plots we can conclude that the sampler has converged (Gelman et al., 2004).

A more formal approach is to use the Gelman-Rubin diagnostic ($\hat{R}$) (Gelman and Rubin, 1992). The Gelman-Rubin diagnostic is derived by comparing the within-chain variance of the MCMC draws against the between-chain variance of the MCMC draws. Customarily Gelman-Rubin statistics of < 1.1 have been considered indicative of convergence (Gelman et al., 2004). Recent research suggests that the traditional $\hat{R}$ is too high (Vats and Knudson, 2018) and instead a split-$\hat{R}$ is proposed with a more strict cut-off of < 1.05 (Vehtari et al., 2019). This project will use the split-$\hat{R}$ diagnostic.

2.8 Choice of priors

In all models the prior distribution of σ^2 is defined as Inv-Gamma(a, b) with $a = 0.001$ and $b = 0.001$. Spiegelhalter (2004) states that inverse-gamma parameters of $a = 0.001$ and and $b = 0.001$ are often suitable as proper weakly informative priors for the description of variance parameters. The Inv-Gamma$(0.001, 0.001)$ distribution is strongly skewed to 0, however the conditional distribution of σ^2 is little influenced by the prior parameters if other terms in the posterior distribution are sufficiently large.

Gelman (2006) argues that where a very low of σ^2 is plausible then the model can be very sensitive to the prior designations of a and b. This is not evident in equations 8, 11 or 16, where the influence of a and b is minimal. It can be seen that the non-prior components of the parametrisation of σ^2 will be well above 0.

Following the approach of Park and Casella (2008), λ is defined as a Gamma(r, δ) hyper-prior for both the ridge and LASSO model. Park and Casella (2008) chose $r = 1$ and $\delta = 1.78$, which they considered to be a weak prior. Hans (2009) used $r = 1$ and $\delta = 1$. Chung et al. (2013) suggests that where the expectation of a hierachial parameter is close to zero (a plausible occurrence in a LASSO model) priors of $r = 2$ and $\delta \to 0$ are uninfluential to the posterior distribution.

To assess if the choice of prior parameters on λ do have a significant impact on the results all 15 permutations of $r = \{0.0001, 0.001, 0.01, 1, 2\}$ and $\delta = \{0.0001, 0.001, 0.01\}$ were modelled. The resulting MAP for all parameters were calculated and 5-fold cross-validation was carried out to assess the RMSPE.

2.9 Software and packages

Analysis was carried out using R version 3.5.1 (R Core Team, 2018) and Rstudio version 1.1.456 (RStudio Team, 2019). In addition to the R packages mentioned elsewhere, the following packages were used in the manipulation and visualisation of the data. `data.table` (Dowle and Srinivasan, 2019), `caret` (from Jed Wing et al., 2018), `GGally` (Schloerke et al., 2018), `gridExtra` (Auguie, 2017), `MASS` (Venables and Ripley, 2002), and the 'tidyverse' suite of packages `dplyr`, `ggplot2`, `plyr`, `readr`, `reshape2`, `scales`, `stringr`, `tidyr` (Wickham, 2007).

3 Results

3.1 ‘13 variable’ results

Figure 2 displays the posterior marginal densities for the parameters of the BLM, ridge and LASSO models and σ^2 and λ for the ‘13 variables’ dataset. Table 2 displays the credible intervals and estimated MAP for all β parameters. The MAP for σ^2 was estimated to be 0.494, 0.495 and 0.496 for the BLM, ridge and LASSO models respectively. The ridge model MAP for λ was estimated to be 10.955, the LASSO model MAP for λ^2 was estimated to be 24.971. MAP estimates were found to be in line with frequentist point estimates for all classes of models.

Table 2: Maps and 95% credible intervals for all parameters of the ’13 variable’ model. Values highlighted in red denote credible intervals that include 0.

	BLM			Ridge			LASSO		
Variable	**Lower**	**MAP**	**Upper**	**Lower**	**MAP**	**Upper**	**Lower**	**MAP**	**Upper**
age	-0.023	0.057	0.137	-0.034	0.046	0.128	-0.029	0.042	0.119
b	0.062	0.114	0.164	0.064	0.114	0.163	0.06	0.111	0.162
chas	0.025	0.069	0.114	0.028	0.071	0.115	0.025	0.069	0.113
crim	-0.246	-0.127	-0.009	-0.215	-0.104	0.009	-0.2	-0.083	0.022
dis	-0.281	-0.183	-0.085	-0.264	-0.17	-0.074	-0.26	-0.162	-0.064
indus	-0.059	0.023	0.105	-0.069	0.01	0.089	-0.069	0.004	0.078
lstat	-0.664	-0.584	-0.505	-0.649	-0.569	-0.491	-0.663	-0.583	-0.504
nox	-0.297	-0.19	-0.088	-0.278	-0.177	-0.074	-0.267	-0.162	-0.056
ptratio	-0.232	-0.171	-0.109	-0.227	-0.167	-0.108	-0.221	-0.159	-0.098
rad_over	0.112	0.336	0.555	0.031	0.217	0.417	-0.026	0.166	0.391
rm	0.08	0.142	0.204	0.089	0.15	0.21	0.081	0.143	0.207
tax_over	-0.547	-0.338	-0.137	-0.419	-0.234	-0.059	-0.409	-0.197	-0.016
zone_over	-0.104	-0.041	0.022	-0.106	-0.041	0.022	-0.1	-0.036	0.024

Split-$\hat{R}$ diagnostic statistics were calculated for each parameter for all models. The maximum split-$\hat{R}$ for the BLM, Ridge and LASSO models were 1.0011, 1.001 and 1.0005 respectively. These values are well within the acceptable limit (Gelman et al. (2004) and Vehtari et al. (2019)). Visual assessment of the trace-plots also indicate that convergence has been reached. Appendix 6.1 displays the trace plots for the parameters with the highest split-$\hat{R}$ for each model.

3.2 ‘22 variable’ results

Figure 3 displays the posterior marginal densities for the parameters of the BLM, ridge and LASSO models, σ^2 and λ for the ‘22 variables’ dataset . Table 3 displays the credible intervals and estimated MAP for all β parameters. The MAP for σ^2 was estimated to be 0.424, 0.426 and 0.426 for the BLM, ridge and LASSO models respectively. The ridge model MAP for λ was estimated to be 3.823, the MAP for the LASSO model λ^2 was estimated to be 5.795. MAP estimates were found to be in line with frequentist point estimates for all classes of models.

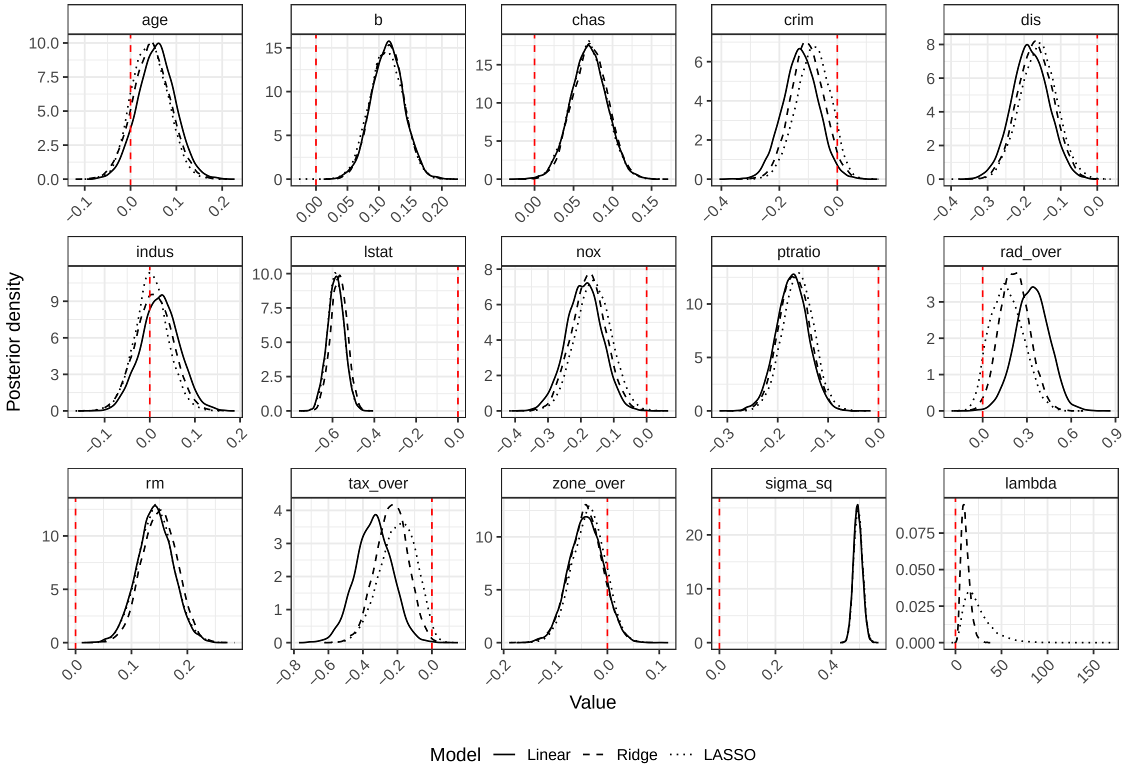

Figure 2: ’13 variable’ density plots of the posterior draws for the Bayesian linear model (solid-line), Bayesian ridge models (dashed-line) and Bayesian LASSO model (dotted-line). The dashed red-lines denotes 0. Note that lambda denotes lambda squared for the LASSO model.

Split-$\hat{R}$ diagnostic statistics were calculated for each parameter for all models. The maximum split-$\hat{R}$ for the BLM, Ridge and LASSO models were 1.0008, 1.0007 and 1.0013 respectively. These values are well within the acceptable limit (Gelman et al. (2004) and Vehtari et al. (2019)). Visual assessment of the trace-plots also indicate that convergence has been reached. Appendix 6.1 displays the trace plots for the parameters with the highest split-$\hat{R}$ for each model.

Table 3: Maps and 95% credible intervals for all parameters of the ’22 variable’ model. Values highlighted in red denote credible intervals that include 0.

	BLM			Ridge			LASSO		
Variable	**Lower**	**MAP**	**Upper**	**Lower**	**MAP**	**Upper**	**Lower**	**MAP**	**Upper**
age	-0.163	0.077	0.313	-0.104	0.101	0.31	-0.11	0.064	0.262
age_sq	-0.343	-0.095	0.149	-0.327	-0.11	0.1	-0.273	-0.067	0.111
b	0.101	0.317	0.53	0.089	0.284	0.484	0.049	0.248	0.464
b_sq	-0.48	-0.262	-0.04	-0.425	-0.223	-0.024	-0.406	-0.187	0.015
chas	0.024	0.064	0.103	0.024	0.063	0.102	0.023	0.062	0.101
crim	-0.59	-0.462	-0.33	-0.536	-0.412	-0.287	-0.548	-0.419	-0.293
crim_sq	-0.368	-0.305	-0.244	-0.355	-0.295	-0.236	-0.353	-0.293	-0.233
dis	-0.6	-0.336	-0.064	-0.537	-0.296	-0.069	-0.497	-0.26	-0.037
dis_sq	-0.151	0.101	0.343	-0.172	0.047	0.276	-0.188	0.017	0.234
indus	-0.535	-0.321	-0.115	-0.46	-0.269	-0.075	-0.449	-0.246	-0.044
indus_sq	0.092	0.303	0.522	0.033	0.23	0.432	0.003	0.211	0.418
lstat	-0.354	-0.063	0.231	-0.434	-0.189	0.055	-0.424	-0.145	0.102
lstat_sq	-0.763	-0.472	-0.181	-0.603	-0.358	-0.118	-0.649	-0.4	-0.125
nox	-0.833	-0.576	-0.321	-0.694	-0.444	-0.2	-0.68	-0.415	-0.18
nox_sq	-0.67	-0.377	-0.083	-0.503	-0.23	0.041	-0.5	-0.201	0.053
ptratio	-1.815	0.132	2.088	-0.573	-0.086	0.404	-0.552	-0.096	0.37
ptratio_sq	-2.261	-0.316	1.636	-0.589	-0.097	0.39	-0.561	-0.087	0.37
rad_over	0.518	0.73	0.931	0.394	0.604	0.815	0.404	0.616	0.829
rm	-1.421	-1.009	-0.604	-0.938	-0.549	-0.195	-1.125	-0.707	-0.262
rm_sq	0.742	1.154	1.573	0.322	0.686	1.079	0.395	0.847	1.274
tax_over	-0.63	-0.447	-0.263	-0.529	-0.342	-0.153	-0.54	-0.348	-0.161
zone_over	-0.048	0.016	0.08	-0.048	0.013	0.073	-0.047	0.01	0.068

3.3 Prediction performance

Figure 4 shows the RMSPE calculated using the MAP of β parameters from 20-fold cross-validation for both the ‘13 variable’ and ‘22 variable’ data sets. Figure 4 also includes RMSPE for the OLS, frequentist ridge and frequentist LASSO models. For the ‘13 variable’ data set the RMSPE from the posterior predictive distribution was 0.503, 0.503 and 0.503 for the BLM, ridge and LASSO models respectively. For the ‘22 variable’ data set the RMSPE from the posterior predictive distribution was 0.446, 0.445 and 0.446 for the BLM, ridge and LASSO models respectively.

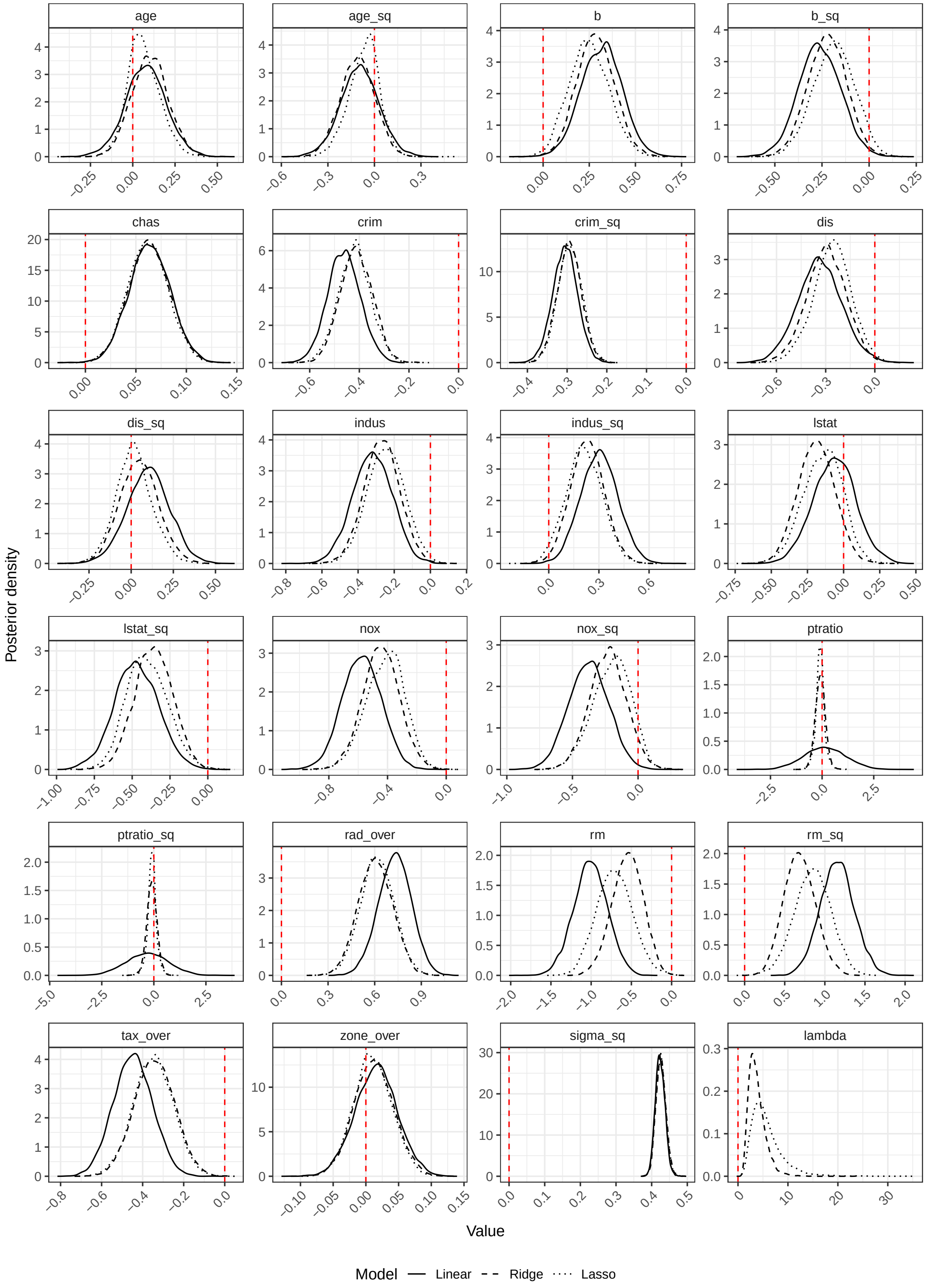

Figure 3: '22 variable' density plots of the posterior draws for the Bayesian linear model (solid-line), Bayesian ridge models (dashed-line) and Bayesian LASSO model (dotted-line). The dashed red-lines denotes 0. Note that lambda denotes lambda squared for the LASSO model.

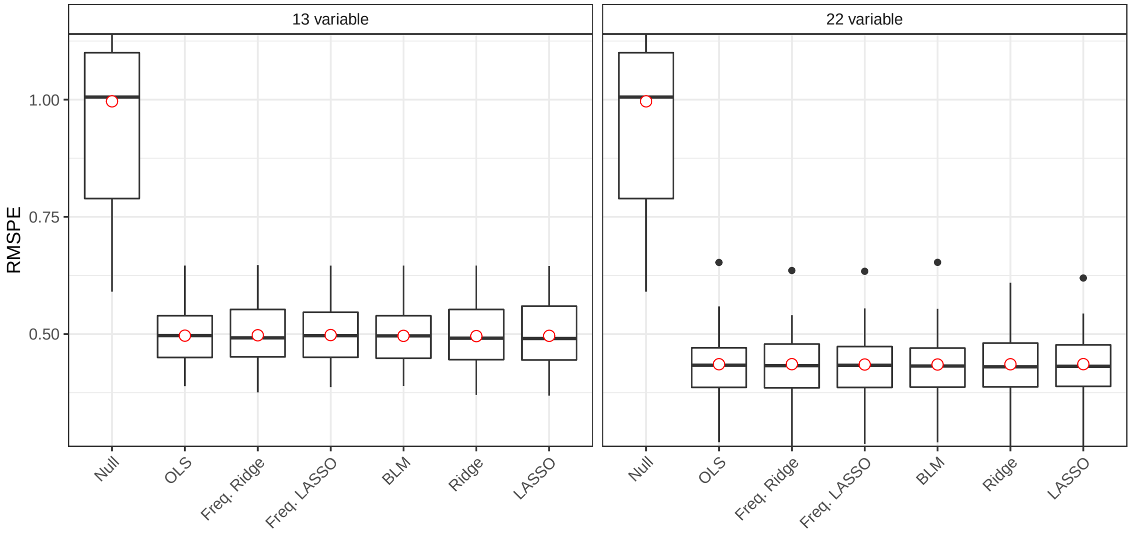

Figure 4: Box-plot of RMSPE values from 10 fold cross-validation for the Null (i.e. mean only), OLS, frequentist ridge, frequentist LASSO, Bayesian linear model, Bayesian ridge and the Bayesian LASSO for both the 13 variable and 22 variable models. Red dots indicate the weighted average RMSPE

3.4 Priors

As described in section 2.8, a range of alternative prior parameters for the λ and λ^2 parameters were applied for ridge and LASSO models for both the '13 variable' and '22 variable' data sets. It was found there was little variation in the mean RMSPE for any of the models and data sets. '13 variable' ridge RMSPE ranged from 0.452 to 0.455 and the LASSO RMSPE ranged from 0.452 to 0.455. For the '22 variable' ridge regression, RMSPE ranged from 0.452 to 0.455 and the LASSO RMSPE ranged from 0.452 to 0.455. Estimates of MAPs for the primary β parameters were similarly invariant.

4 Discussion

4.1 Parameter estimation

Referring to the ‘13 variable’ data set (figure 2 and table 2). When the credible intervals of the marginal densities did not include 0, it can be seen that the estimates for the BLM, ridge and LASSO models were broadly similar. For the ridge and LASSO models, parameter shrinkage can be observed for the `rad_over` and `tax_over` parameters. A look at the covariance of the posterior densities of these two variables shows why.

Figure 5 shows the covariance of the posterior marginal densities of the `tax_over` and `rad_over` parameters. Estimates for these two parameters are highly correlated, this is because of a very high correlation between the original two variables (r = 0.975). In the ridge and LASSO models the posterior distribution has been dragged closer to 0.

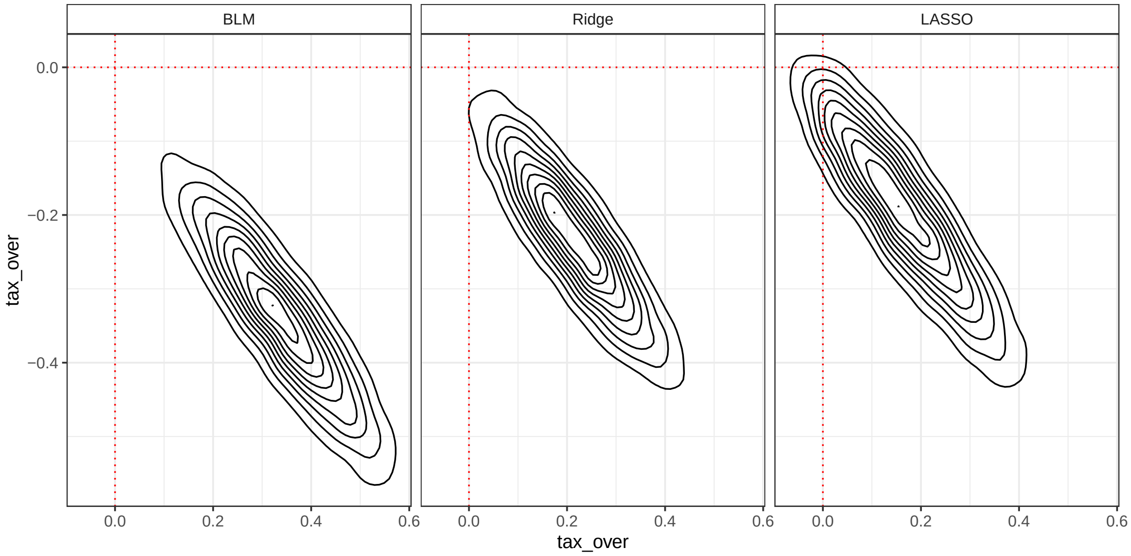

Figure 5: Covariance density plots of the posterior densities of the rad_over and tax_over parameters for the BLM, ridge and LASSO models of the 13 variable data set.

Referring to the ‘22 variable’ data set (figure 3 and table 3). In comparison to the ‘13 variable’ data set there’s much more evidence of parameter shrinkage in the ridge and LASSO models. The only β_i where BLM, ridge and LASSO posterior densities are very similar is the `chas` parameter. Parameter shrinkage can be seen for several variables such as `nox`, `nox_sq`, `rm` and `rm_sq` in addition to `tax_over` and `rad_over`. It can also be noticed that for some parameters where the MAP is close to 0, the variance range of the credible interval is much reduced in the ridge and LASSO regressions. The most dramatic examples are `ptratio` and `ptratio_sq`.

Figure 6 shows the covariance of the posterior densities of the `ptratio` and `ptratio_sq` parameters. It can be seen that estimates for these two parameters are highly correlated, this is due to the very high correlation of the original two variables (Pearson’s correlation coefficient = 0.99971). Notably it can be seen that the ridge and LASSO posterior parameters appear to have a much reduced covariance in comparison to the BLM posterior parameters.

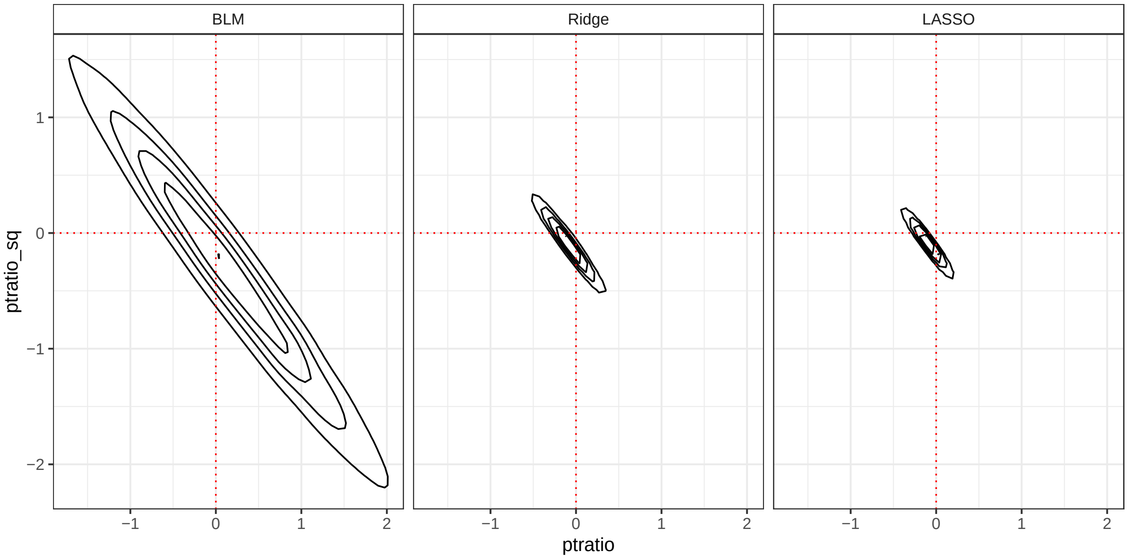

Figure 6: Covariance density plots of the posterior densities of the ptratio and ptratio_sq parameters for the BLM, ridge and LASSO models of the 22 variable data set.

As seen in figure 2 and figure 3, where parameters have an unambiguous effect on `medv`, parameter estimates were often similar. Where parameter shrinkage was observed in the ridge and LASSO regression it was most visible where the posterior densities and original variables were highly correlated. Regularisation methods have been shown to be effective at reducing the impacts of strong multi-collinearity (e.g. Herawati et al. (2018)) and this can be observed in these results. For the '22 variable' data set the square terms will exhibit greater multi-collinearity.

However, in the context of credible intervals, in only a few cases does the parameter shrinkage impact inference about which parameters significantly contribute to the model. For the '13 variable' model (table 2) it can be seen that only the `crim` and `rad_over` parameters displayed different interpretations depending on which type of model was used. `crim` is considered a significant parameter only under the BLM of formulation, and `rad_over` is no longer considered a significant parameter under the LASSO formulation. Similarly, for the '22 variable' model (table 3) model formulation makes little difference to the interpretation of the parameters, only `b_sq` is shrunk significantly in the LASSO model, and `nox_sq` is shrunk towards 0 under the ridge and LASSO model. Therefore using the credible interval as the criteria for model selection, Bayesian regularisation methods make little difference to the formulation of the most parsimonious model. Frequentest ridge and LASSO regressions displayed similar results.

4.2 Interpretating the parameters

Referring to the '13 variable' dataset, seven of the 13 variables had posterior credible intervals that did not include 0 under all model types, i.e seven variables can be regarded as having a significant effect on `medv`. These were `b`, `chas`, `dis`, `lstat`, `nox`, `ptratio`, `rm` and `tax_over`. These relationships with median house prices are largely intuitive. River side properties are more desirable and properties with more rooms were more valuable, therefore these variables show positive coefficients. High-levels of air pollution, a long distance to work and a high pupil-teacher ratio (however see below) are all undesirable and therefore these variables show positive correlations. In the context of the original research aims Harrison and Rubinfeld (1978), it does appear that nitric oxide levels (nox) do negatively impact house prices even when correcting for possible confounders such as industrial activity (indus) and car pollution (rad).

The proportion of the population who have have low status occupations (lstat) shows a significant negative correlation. In this case it seems unlikely that the presence of low-status composition of the population suppresses house-prices but rather low-status people are driven to low value housing, therefore the relationship is likely not causal. The relationship between low value housing and poverty is well established (e.g. Meen (2009)).

Harrison and Rubinfeld (1978) transformed racial composition data to variable b (see table 1) which is at it's highest when census tracts are racially homogeneous and at its lowest when tracts are racially heterogeneous. The lowest value for b (0) is when a tract is 63% black. The very highest values of b are where a tract is exclusively white (b = 396.6). Where a tract is exclusively black, b = 136.9. The estimated MAP for b is positive, given that the data is heavily weighted to majority white neighbourhoods and racially mixed areas are rare (see figure 1), it can be concluded that b better represents a measure of the 'white-ness' of a community. By this interpretation it must be concluded that 'white' neighbourhoods have significantly higher median property values.

The '22 variable' model provides some additional insights. indus_sq, and crim and crim_sq display credible intervals that do not include 0 under all model types. This suggests that once additional terms are introduced to describe the curvature of parameters, the relationship between crim and indus can be better described. Unsurprisingly criminal activity (crim) negatively impacts house prices. The indus and indus_sq parameters describe a relationship that is broadly positive suggesting that median house prices in fact increase around industrial areas. Notably when other terms are included, tax_over shows consistently significant parameters less than zero. Suggesting that high tax rates do significantly reduce median house prices.

4.3 Prediction performance

Figure 4 shows the RMSPE from 20-fold cross-validation for all classes of models for both the '13 variable' and '22 variable' sets off regressors. These include the RMSPE for the frequentist approaches for comparison. It can be see that the RMSPE for all classes of regression model are almost identical for both regressor sets. Therefore it seems that for these data sets, regularised approaches don't significantly improve prediction error and Bayesian approaches don't significantly improve prediction error in comparison to frequentist approaches.

Using real world data sets with 9 - 13 variables, Celeux et al. (2012) found that Bayesian regularisation approaches did not perform significantly better than their frequentist counterparts, and their frequentist counterparts did not perform significantly better than other traditional model selection approaches (e.g. BIC). Mallick (2015) found improved prediction using Bayesian regularisation when applying these approaches to very high dimension genetics data. It may be argued that the benefits of regularisation can only be observed with problems involving many more regressors with much higher levels of multi-collinearity than is observed in this data set.

Given these results. It can be argued, why bother with regularisation and Bayesian implementations? In a machine learning context, i.e. attempting to find a model which best predicts future data, there appears to be no benefit for this particular data using these particular sets of regressors. It may be found that using alternative sets of regressors with many more dimensions the benefits of regularisation at reducing multicollinearity (Dormann et al., 2013) will be more apparent. As proven in appendix (6), the formulations used in this project result in posterior MAPs which are equal to their frequentist counterparts, parameters for hyper-priors have been chosen to be as weakly informative as possible, therefore no prior information has been included in the modelling.

4.4 Extensions to Bayesian regularisation

The hierachial Bayesian framework allows for extensibility which would not be possible under a frequentist framework. Under a Bayesian hierachial conception it could be possible to specify a unique λ hyper-prior for each parameter, i.e.

$$\lambda_1 \ldots \lambda_p \sim \text{Gamma}(r, \delta) \tag{19}$$

Under the frequentist framework this would not be possible as the parameter space of $\lambda_1 \ldots \lambda_p$ would be too great for cross-validation (Friedman et al., 2010). This could enable greater flexibility where parameters can be penalised to a greater or lesser extent. The resulting posterior densities would however deviate from the frequentist approaches.

The elastic net attempts to combine the properties of a ridge regression and LASSO regression by including both LASSO and ridge parameters into the cost-function for β (Zou and Hastie, 2005) such that the cost function is

$$L(\hat{\beta}) = \text{arg.min}\left[(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta}) + \lambda_1 \sum_{i=1}^{p} \beta_i^2 + \lambda_2 \sum_{i=1}^{p} |\beta_i|\right] \tag{20}$$

Like frequentist ridge and frequentist LASSO approaches, λ_1 and λ_2 must be estimated by cross-validation over a grid of possible values, thus increasing the computational load. Li et al. (2010) have formulated a Bayesian hierarchical model for the elastic net where these problems have been overcome.

The census tracts of the Boston housing data have been assumed to be independent of one another. In practice this is not true, census tracts will be adjacent of to another and the characteristics of one tract will have an impact on the median house prices of its neighbours. Given that latitude and longitudinal data is available (Newman et al., 1998) it would be possible to include geo-spatial covariance properties into the hierarchical model (Ribeiro Jr and Diggle, 2018) or alternatively it would be possible to construct an areal model with a nearest neighbour covariance structure.

5 Conclusions

In conclusion, Bayesian approaches enabled easier and more intuitive interpretation of the regression coefficient parameter estimates, particularly so for the regularised approach. Bayesian regularised regression did impact marginal posterior parameter estimates, most notably where there were a greater number collinear regressors in the regression model. However regularised approaches rarely made a difference in the interpretation of the resulting parameter estimates. Bayesian approaches were no better at predicting house prices than frequentist approaches, and regularised regression was not better than non-regularised regression. It may be argued that the benefits of regularisation are more apparent where the number of parameters and the extent of multi-collinearity is much greater.

6 Appendix

6.1 Model traceplots

As seen in section 3, split-$\hat{R}$ statistics suggest that all MCMC sampling for all parameters have reached convergence. Figure 7 shows the trace-plots for the ‘13-variable’ regression. and figure 8 shows the trace-plots for the ‘22-variable’ regression for those parameters with the highest split-$\hat{R}$.

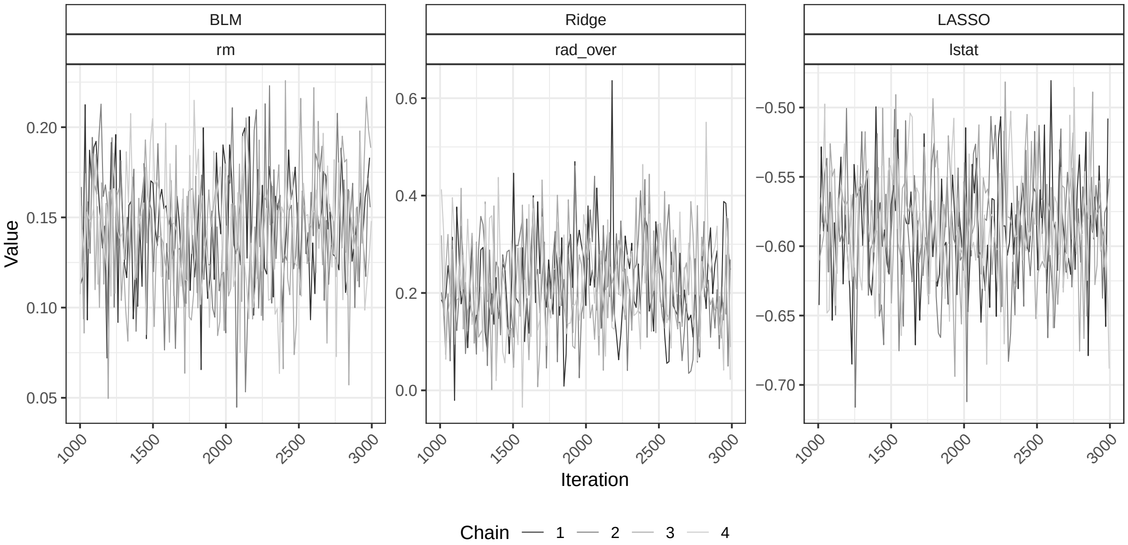

Figure 7: Trace-plots of the 13-variable regression analysis for the parameters showing the highest Gelman-Rubin statistic. Iterations have been thinned to every 10 points to aid interpretation

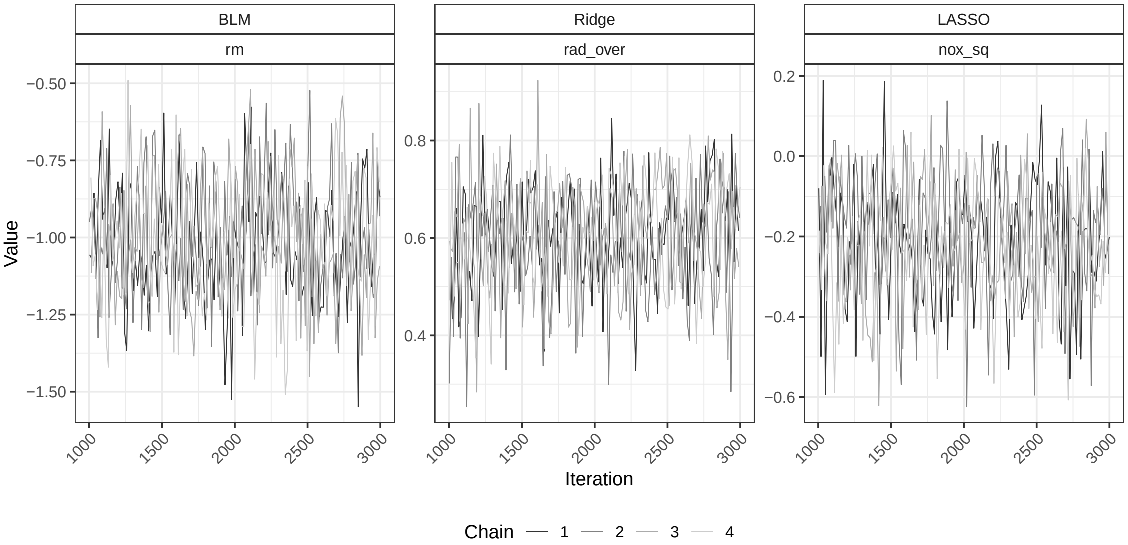

Figure 8: Trace-plots of the 22-variable regression analysis for the parameters showing the highest Gelman-Rubin statistic. Iterations have been thinned to every 10 points to aid interpretation

6.2 Bayesian linear model

Referring to the hierachial definition given in section 2.2. Using Bayes rule $p(a|b) \propto p(b|a)p(a)$ it can be seen that the joint distribution of $\mathbf{y}$, β and σ^2 can be stated as

$$\begin{aligned} p(\beta, \sigma^2|\mathbf{y}, \mathbf{X}) &\propto L(\mathbf{y}|\beta, \sigma^2, \mathbf{X}) \times p(\beta, \sigma^2|\mathbf{X}) \\ &\propto L(\mathbf{y}|\beta, \sigma^2, \mathbf{X}) \times p(\beta|\mathbf{y}, \sigma^2, \mathbf{X}) \times p(\sigma^2) \end{aligned} \tag{21}$$

Which make up the constituents of the hierachial model. Therefore the joint distribution $p(\beta, \sigma^2|\mathbf{y}, \mathbf{X})$ can be defined as

$$p(\beta, \sigma^2|\mathbf{y}, \mathbf{X}) \propto (\sigma^2)^{-n/2} \exp\Big[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)\Big] \times 1 \times (\frac{1}{\sigma^2})^{a+1} \exp(-\frac{b}{\sigma^2}) \tag{22}$$

By removing components of the joint distribution not dependent on β, the full conditional distribution of β can be derived as follows

$$\begin{aligned} p(\beta|\sigma^2, \mathbf{y}, \mathbf{X}) &\propto \exp\Big[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)\Big] \\ &= \exp\Big[-\frac{1}{2\sigma^2}(\mathbf{y^T y} + \beta^{\mathbf{T}}(\mathbf{X^T X})\beta - 2\mathbf{y^T X}\beta)\Big] \\ &\propto \exp\Big[-\frac{1}{2\sigma^2}(\beta^{\mathbf{T}}\mathbf{X^T X}\beta - 2\mathbf{y^T X}(\mathbf{X^T X})^{-1}(\mathbf{X^T X})\beta)\Big] \end{aligned} \tag{23}$$

Given that the maximum likelihood estimate of β is $\hat{\beta} = (\mathbf{X^T X})^{-1}\mathbf{X^T y}$, this can be rearranged to

$$\mathbf{y} = \hat{\beta}(\mathbf{X^T X})\mathbf{X^{-1}} \tag{24}$$

Substituting (24) into (23)

$$\begin{aligned} p(\beta|\sigma^2, \mathbf{y}, \mathbf{X}) &\propto \exp\Big[-\frac{1}{2\sigma^2}(\beta^{\mathbf{T}}(\mathbf{X^T X})\beta - 2[\hat{\beta}(\mathbf{X^T X})\mathbf{X^{-1}}]^{\mathbf{T}}(\mathbf{X^T X})^{-1}(\mathbf{X^T X}))\mathbf{X}\beta)\Big] \\ &\propto \exp\Big[-\frac{1}{2\sigma^2}(\beta^{\mathbf{T}}(\mathbf{X^T X})\beta - 2\hat{\beta}^{\mathbf{T}}(\mathbf{X^T X})\hat{\beta})\Big] \\ &\propto \exp\Big[-\frac{1}{2\sigma^2}(\beta^{\mathbf{T}}(\mathbf{X^T X})\beta - 2\hat{\beta}^{\mathbf{T}}(\mathbf{X^T X})\beta + \hat{\beta}^{\mathbf{T}}(\mathbf{X^T X})\hat{\beta})\Big] \end{aligned} \tag{25}$$

If we define $\mathbf{V} = (\mathbf{X^T X})^{-1}$ the conditional probability of β can be defined as

$$\begin{aligned} p(\beta|\sigma^2, \mathbf{y}, \mathbf{X}) &\propto \exp\Big[-\frac{1}{2\sigma^2}(\beta^{\mathbf{T}}\mathbf{V^{-1}}\beta - 2\hat{\beta}^{\mathbf{T}}\mathbf{V^{-1}}\beta + \hat{\beta}^{\mathbf{T}}\mathbf{V^{-1}}\hat{\beta})\Big] \\ &\propto \exp\Big[-\frac{1}{2\sigma^2}(\beta - \hat{\beta})^T\mathbf{V^{-1}}(\beta - \hat{\beta})\Big] \end{aligned} \tag{26}$$

Which is proportional to the multivariate normal distribution with mean $\hat{\beta}$ and variance $\sigma^2\mathbf{V}$, i.e.

$$\beta|\sigma^2, \mathbf{y}, \mathbf{X} \sim \mathsf{N}(\hat{\beta}, \sigma^2\mathbf{V}) \tag{27}$$

As we've already derived the conditional distribution of β we can derive $\sigma^2|\mathbf{y}, \mathbf{X}$ by rearranging the Bayesian formula.

$$\begin{aligned} p(\sigma^2|\mathbf{y}, \mathbf{X}) &= \frac{p(\beta, \sigma^2|\mathbf{y}, \mathbf{X})}{p(\beta|\sigma^2, \mathbf{y}, \mathbf{X})} \\ &\propto \frac{\Big[(\sigma^2)^{-n/2} \exp\Big(- \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta})\Big)(\sigma^2)^{-(a+1)} \exp(-\frac{b}{\sigma^2})\Big]}{\Big[(\sigma^2)^{-p/2}(|\mathbf{V}|)^{-1/2} \exp\Big(- \frac{1}{2\sigma^2}(\beta - \hat{\beta})^T\mathbf{V^{-1}}(\beta - \hat{\beta})\Big)\Big]} \\ &\propto (\sigma^2)^{-(\frac{n}{2} - \frac{p}{2} + a + 1)} \frac{\exp\Big[- \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta})\Big]}{\exp\Big[- \frac{1}{2\sigma^2}(\beta - \hat{\beta})^T\mathbf{V^{-1}}(\beta - \hat{\beta})\Big)\Big]} \exp(-\frac{b}{\sigma^2}) \end{aligned} \tag{28}$$

As $p(\sigma^2|\mathbf{y}, \mathbf{X})$ is not conditional on β we can set $\beta = \hat{\beta}$ therefore setting the denominator to 1. Therefore

$$\begin{aligned} p(\sigma^2|\mathbf{y}, \mathbf{X}) &\propto (\sigma^2)^{-(\frac{n}{2} - \frac{p}{2} + a + 1)} \exp\Big[- \frac{1}{2\sigma^2}\Big((\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta}) + b\Big)\Big] \\ \sigma^2|\mathbf{y}, \mathbf{X} &\sim \text{Inv-Gamma}(\frac{n-p}{2} + a, \frac{1}{2}\text{SSE} + b) \end{aligned} \tag{29}$$

Where SSE $= (\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta})$.

6.3 Bayesian ridge model

Referring to the hierachial definition given in section 2.3. Using Bayes theorem $p(a|b) \propto p(b|a)p(a)$ it can be seen that the joint distribution of β, σ^2 and λ can be stated as

$$\begin{aligned} p(\beta, \sigma^2, \lambda|\mathbf{y}, \mathbf{X}) &\propto L(\mathbf{y}|\beta, \sigma^2, \lambda, \mathbf{X}) \times p(\beta, \sigma^2, \lambda|\mathbf{X}) \\ &\propto L(\mathbf{y}|\beta, \sigma^2, \lambda, \mathbf{X}) \times p(\beta, \sigma^2|\lambda, \mathbf{X}) \times p(\lambda) \\ &\propto L(\mathbf{y}|\beta, \sigma^2, \lambda, \mathbf{X}) \times p(\beta|\sigma^2, \lambda, \mathbf{X}) \times p(\sigma^2) \times p(\lambda) \end{aligned} \tag{30}$$

Which make up the constituents of the hierachial model. Therefore the joint distribution of β, σ^2 and λ is

$$\begin{aligned} p(\beta, \sigma^2, \lambda|\mathbf{y}, \mathbf{X}) &\propto \frac{1}{(2\pi)^{n/2}(|\sigma^2\mathbf{I_n}|)^{1/2}} \exp\Big[-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)\Big] \\ &\times \frac{1}{(2\pi)^{p/2}(|\mathbf{\Sigma}|)^{1/2}} \exp\Big[-\frac{1}{2}\beta^{\mathbf{T}}\mathbf{\Sigma^{-1}}\beta\Big] \\ &\times (\sigma^2)^{-(a+1)} \exp(\frac{-b}{\sigma^2}) \times \lambda^{r-1}\exp(-\lambda\delta) \end{aligned} \tag{31}$$

As the model's hierachial, we can derive the full conditions of β, σ^2 and λ without integrating through the joint distribution. The parameter β can derived as such:

$$\begin{aligned} p(\beta|\sigma^2, \lambda, \mathbf{X}) &\propto \exp\Big[-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)\Big]\exp\Big[-\frac{1}{2}\beta^{\mathbf{T}}\mathbf{\Sigma^{-1}}\beta\Big] \\ &\propto \exp\Big[-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta) - \frac{1}{2}\beta^{\mathbf{T}}\mathbf{\Sigma^{-1}}\beta\Big] \\ &\propto \exp\Big[-\frac{1}{2}\Big(\frac{\beta^{\mathbf{T}}\mathbf{X^T X}\beta}{\sigma^2} - 2\frac{\beta\mathbf{X^T y}}{\sigma^2} + \beta^{\mathbf{T}}\mathbf{\Sigma^{-1}}\beta\Big)\Big] \\ &\propto \exp\Big[-\frac{1}{2}\Big(\beta^{\mathbf{T}}(\mathbf{\Sigma^{-1}} + \frac{\mathbf{X^T X}}{\sigma^2})\beta - 2\frac{\beta\mathbf{X^T y}}{\sigma^2}\Big)\Big] \end{aligned} \tag{32}$$

If we define $\mathbf{E} = \mathbf{V}\frac{\mathbf{X^T y}}{\sigma^2}$ and $\mathbf{V} = (\mathbf{\Sigma^{-1}} + \frac{\mathbf{X^T X}}{\sigma^2})^{-1}$ and we can redefine $p(\beta|\sigma^2, \lambda, \mathbf{X})$ as such:

$$\begin{aligned} p(\beta|\sigma^2, \lambda, \mathbf{X}) &\propto \exp\Big[-\frac{1}{2}\Big(\beta^{\mathbf{T}}\mathbf{V^{-1}}\beta - 2\beta\mathbf{V_r^{-1}}\mathbf{E}\Big)\Big] \\ &\propto \exp\Big[-\frac{1}{2}\Big(\beta^{\mathbf{T}}\mathbf{V^{-1}}\beta - 2\beta\mathbf{V^{-1}}\mathbf{E} + \mathbf{E^T}\mathbf{V^{-1}}\mathbf{E}\Big)\Big] \\ &\propto \exp\Big[-\frac{1}{2}(\beta - \mathbf{E})^T\mathbf{V^{-1}}(\beta - \mathbf{E})\Big] \end{aligned} \tag{33}$$

Which is equivalent to the multivariate normal distribution with mean $\mathbf{E}$ and variance $\mathbf{V^{-1}}$, i.e.

$$\beta|\sigma^2, \lambda, \mathbf{X} \sim \mathsf{N}(\mathbf{E}, \mathbf{V^{-1}}) \tag{34}$$

Under the frequentist ridge regression model, the cost / likelihood function is

$$L(\hat{\beta}) = (\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta}) + \lambda\sum_{i=1}^{p}\beta_i^2 \tag{35}$$

Therefore the maximum likelihood estimate of beta is

$$\hat{\beta}_{MLE} = (\mathbf{X^T X} + \lambda\mathbf{I_p})^{-1}\mathbf{X^T y} \tag{36}$$

The expectation and MAP of $\beta|\sigma^2, \lambda, \mathbf{X}$ is $\mathbf{E}$. With some manipulation it can be shown to be equal to the $\hat{\beta}_{MLE}$ of the frequentist solution.

$$\begin{aligned}\mathbf{E} = \mathbf{V}\frac{\mathbf{X^T y}}{\sigma^2} &= (\mathbf{\Sigma^{-1}} + \frac{\mathbf{X^T X}}{\sigma^2})^{-1}\frac{\mathbf{X^T y}}{\sigma^2} \\ &= (\frac{\lambda\mathbf{I_p}}{\sigma^2} + \frac{\mathbf{X^T X}}{\sigma^2})^{-1}\frac{\mathbf{X^T y}}{\sigma^2} \\ &= \sigma^2(\lambda\mathbf{I_p} + \mathbf{X^T X})^{-1}\frac{\mathbf{X^T y}}{\sigma^2} \\ &= (\mathbf{X^T X} + \lambda\mathbf{I_p})^{-1}\mathbf{X^T y}\end{aligned} \tag{37}$$

The derivation of the full conditional distribution of $\sigma^2|\beta, \lambda, \mathbf{X}, \mathbf{y}$ can also be derived through factorisation:

$$\begin{aligned}p(\sigma^2|\beta, \lambda, \mathbf{X}, \mathbf{y}) &\propto (\sigma^2)^{-n/2}(\sigma^2)^{-p/2}\exp\Big[-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)\Big](\sigma^2)^{-(a+1)}\exp(-\frac{b}{\sigma^2}) \\ &\propto (\frac{1}{\sigma^2})^{\frac{n+p}{2}+a+1}\exp\Big[-\frac{1}{\sigma^2}\Big(\frac{1}{2}(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)+b\Big)\Big]\end{aligned} \tag{38}$$

Therefore

$$\sigma^2|\beta, \lambda, \mathbf{X}, \mathbf{y} \sim \text{Inv-Gamma}(\frac{n+p}{2}+a, \frac{1}{2}\text{SSE}+b) \tag{39}$$

Where SSE $= (\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)$.

Finally the derivation of $\lambda|\beta, \sigma^2, \mathbf{X}, \mathbf{y}$:

$$\begin{aligned}p(\lambda|\beta, \sigma^2, \mathbf{X}, \mathbf{y}) &\propto \frac{1}{(|\mathbf{\Sigma}|)^{1/2}}\exp\Big[-\frac{1}{2}\beta^{\mathbf{T}}\mathbf{\Sigma^{-1}}\beta\Big]\lambda^{r-1}\exp(-\lambda\delta) \\ &\propto \frac{1}{\lambda^{-p/2}}\exp\Big[-\frac{1}{2}\Big(\beta^{\mathbf{T}}(\frac{\lambda}{\sigma^2}\mathbf{I_p})\beta\Big)-\lambda\delta\Big]\lambda^{r-1} \\ &\propto \lambda^{\frac{p}{2}+r-1}\exp\Big[-\lambda\Big(\frac{1}{2\sigma^2}\beta^{\mathbf{T}}\beta+\delta\Big)\Big] \\ \lambda|\beta, \sigma^2, \mathbf{X}, \mathbf{y} &\sim \text{Gamma}(\frac{p}{2}+r, \frac{1}{2\sigma^2}\beta^{\mathbf{T}}\beta+\delta)\end{aligned} \tag{40}$$

6.4 Bayesian LASSO model

Referring to the hierachial definition given in section 6.5. Using Bayes theorem $p(a|b) \propto p(b|a)p(a)$ it can be seen that the joint distribution of β, σ^2, $\tau_1^2...\tau_p^2$ and λ can be stated as

$$\begin{aligned}p(\beta, \sigma^2, \tau_1^2...\tau_p^2, \lambda|\mathbf{y}, \mathbf{X}) &\propto L(\mathbf{y}|\beta, \sigma^2, \tau_1^2...\tau_p^2, \lambda, \mathbf{X}) \times p(\beta, \sigma^2, \tau_1^2...\tau_p^2, \lambda|\mathbf{X}) \\ &\propto L(\mathbf{y}|\beta, \sigma^2, \tau_1^2...\tau_p^2, \lambda, \mathbf{X}) \times p(\beta|\sigma^2, \tau_1^2...\tau_p^2, \lambda^2, \mathbf{X}) \\ &\times p(\sigma^2|\mathbf{X}) \times p(\tau_1^2...\tau_p^2|\lambda^2, \mathbf{X}) \times p(\lambda^2|\mathbf{X})\end{aligned} \tag{41}$$

Which correspond to the constituents of the hierachial model. Therefore the joint probability of $\beta, \sigma^2, \tau_1^2...\tau_p^2, \lambda|\mathbf{y}, \mathbf{X}$ can be defined as

$$
\begin{aligned}
p(\beta, \sigma^2, \tau_1^2...\tau_p^2, \lambda|\mathbf{y}, \mathbf{X}) \propto\ & \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)\Big] \\
& \times \prod_{i=1}^{p} \frac{1}{(2\pi\sigma^2\tau_i^2)^{1/2}} \exp\big[-\frac{1}{2\sigma^2\tau_i^2}\beta_i^2\big] \\
& \times (\sigma^2)^{-(a+1)} \exp(-\frac{b}{\sigma^2}) \\
& \times \prod_{i=1}^{p} \frac{1}{2}\lambda^2 \exp(\frac{1}{2}\lambda^2\tau_i^2) \\
& \times (\lambda^2)^{r-1} \exp(-\delta\lambda^2)
\end{aligned}
\tag{42}
$$

Deriving the full conditional distribution of β does not require integration of the joint probability:

$$
\begin{aligned}
p(\beta|\sigma^2, \tau_1^2...\tau_p^2, \lambda^2, \mathbf{X}) &\propto \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)\Big] \Bigg[\prod_{i=1}^{p} \frac{1}{(2\pi\sigma^2\tau_i^2)^{1/2}} \exp\big[-\frac{1}{2\sigma^2\tau_i^2}\beta_i^2\big] \Bigg] \\
&\propto \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big[-\frac{1}{2}(\mathbf{y} - \mathbf{X}\beta)^T(\sigma^2\mathbf{I_n})^{-1}(\mathbf{y} - \mathbf{X}\beta)\Big] \\
&\times \frac{1}{(2\pi)^{p/2}} \frac{1}{(|\sigma^2\mathbf{D}_\tau|)^{1/2}} \exp\Big[\frac{1}{2}\beta^\mathbf{T}(\sigma^2\mathbf{D}_\tau)^{-1}\beta\Big] \\
&\propto \frac{1}{(2\pi)^{\frac{n+p}{2}}(|\sigma^2\mathbf{I_n}|)^{1/2}(|\sigma^2\mathbf{D}_\tau|)^{1/2}} \exp\Big[\frac{-1}{2}\Big((\mathbf{y} - \mathbf{X}\beta)^T(\sigma^2\mathbf{I_n})^{-1}(\mathbf{y} - \mathbf{X}\beta) - \beta^\mathbf{T}\mathbf{D}_\tau^\mathbf{-1}\beta\Big)\Big]
\end{aligned}
\tag{43}
$$

The expression $(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) - \beta^\mathbf{T}\mathbf{D}_\tau^\mathbf{-1}\beta$ can be stated as $\beta^\mathbf{T}(\mathbf{X^TX} + \mathbf{D}_\tau^\mathbf{-1})\beta - 2\mathbf{y^TX}\beta + \mathbf{y^Ty}$. Letting $\mathbf{A} = \mathbf{X^TX} + \mathbf{D}_\tau^\mathbf{-1}$ this equals

$$
\beta^\mathbf{T}\mathbf{A}\beta - 2\mathbf{y^TX}\beta + \mathbf{y^Ty}
\tag{44}
$$

By completing the square

$$
\begin{aligned}
\beta^\mathbf{T}\mathbf{A}\beta - 2\mathbf{y^TX}\beta + \mathbf{y^Ty} &= \mathbf{A}\Big[\beta^\mathbf{T}\beta - 2\mathbf{A^{-1}yX}\beta + \mathbf{A^{-1}y^Ty}\Big] \\
&= \mathbf{A}\Big[(\beta - \mathbf{A^{-1}X^Ty})^T(\beta - \mathbf{A^{-1}X^Ty}) - (\mathbf{A^{-1}})^T(\mathbf{A^{-1}})\mathbf{X^TXy^Ty} + \mathbf{A^{-1}y^Ty}\Big] \\
&= \mathbf{A}\Big[(\beta - \mathbf{A^{-1}X^Ty})^T(\beta - \mathbf{A^{-1}X^Ty}) + \mathbf{A^{-1}y^T}(\mathbf{I_n} - \mathbf{XA^{-1}X^T})\mathbf{y}\Big] \\
&= (\beta - \mathbf{A^{-1}X^Ty})^T\mathbf{A}(\beta - \mathbf{A^{-1}X^Ty}) + \mathbf{y^T}(\mathbf{I_n} - \mathbf{XA^{-1}X^T})\mathbf{y}
\end{aligned}
\tag{45}
$$

Therefore by removing the parts which are not conditional on β, $p(\beta|\sigma^2, \tau_1^2...\tau_p^2, \lambda^2, \mathbf{X})$ can be stated as

$$p(\beta|\sigma^2, \tau_1^2...\tau_p^2, \lambda^2, \mathbf{X}) \propto \frac{1}{(2\pi)^{\frac{n+p}{2}}(|\sigma_2\mathbf{I_n}|)^{1/2}(|\sigma^2\mathbf{D}_\tau|)^{1/2}} \exp\Big[\frac{1}{2\sigma^2}(\beta - \mathbf{A^{-1}X^Ty})^T\mathbf{A}(\beta - \mathbf{A^{-1}X^Ty})\Big] \tag{46}$$

Which is proportional to the multivariate normal with mean $\mathbf{A^{-1}X^Ty}$ and variance $\sigma^2\mathbf{A^{-1}}$, i.e.

$$\beta|\sigma^2, \tau_1^2...\tau_p^2, \lambda^2, \mathbf{X} \sim \mathsf{N}(\mathbf{A^{-1}X^Ty}, \sigma^2\mathbf{A^{-1}}) \tag{47}$$

Under the frequentist LASSO model (Tibshirani (1996)), the cost function is minimised at

$$\text{arg.min}\Big((\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta}) + \lambda\sum_{i=1}^{p}|\beta_i|\Big) \tag{48}$$

It can be shown that the MAP of the posterior conditional distribution of β is identical to the minima of the frequentist LASSO cost function by referring back to the Laplace formulation of the $\beta_1...\beta_p$ parameters and the likelihood for $\mathbf{y}$.

$$p(\beta_1...\beta_p|\sigma^2, \lambda^2, \mathbf{X}) = \prod_{i=1}^{p}\frac{\lambda}{2\sigma^2}\exp\Big(-\frac{\lambda}{\sigma^2}|\beta_i|\Big) \propto (\frac{\lambda}{2})^p\exp\Big(-\lambda\sum_{i=1}^{p}|\beta_i|\Big) \tag{49}$$

$$L(\mathbf{y}|\beta, \sigma^2, \mathbf{X}) \propto \frac{1}{(\sigma^2)^{n/2}}\exp\Big[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)\Big] \tag{50}$$

Once again using Bayes rule, $p(\beta|\sigma^2, \lambda^2, \mathbf{X})$ can be expressed as a combination of probabilities

$$\begin{aligned} p(\beta_1...\beta_p|\sigma^2, \lambda^2, \mathbf{X}, \mathbf{y}) &\propto L(\mathbf{y}|\beta_1...\beta_p, \sigma^2, \mathbf{X}) \times p(\beta_1\beta_p|\sigma^2, \lambda^2, \mathbf{X}) \\ &\propto \exp\Big[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) - \lambda\sum_{i=1}^{p}|\beta_i|\Big] \end{aligned} \tag{51}$$

The MAP of which will be the arg.max$(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) - \lambda\sum_{i=1}^{p}|\beta_i|)$ which is equivalent to the minima of the frequentist LASSO cost function. Derivation of the full conditional distribution for σ^2 is also performed through factorisation.

$$\begin{aligned} p(\sigma^2|\beta, \lambda^2, \mathbf{X}, \mathbf{y}) &\propto \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\Big[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)\Big] \\ &\times \frac{1}{(2\pi\sigma^2)^{p/2}}\exp\Big[-\frac{1}{2\sigma^2}\beta^\mathbf{T}\mathbf{D}_\tau^{\mathbf{-1}}\beta\Big] \\ &\times (\sigma_2)^{-(a+1)}\exp(-\frac{b}{\sigma_2}) \\ &\propto (\frac{1}{\sigma^2})^{\frac{n}{2}+\frac{p}{2}+a+1}\exp\Big[-\frac{1}{\sigma^2}\Big(\frac{1}{2}(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \frac{1}{2}\beta^\mathbf{T}\mathbf{D}_\tau^{\mathbf{-1}}\beta + b\Big)\Big] \end{aligned} \tag{52}$$

Which is conjugate to the inverse gamma distribution

$$\sigma^2|\beta, \lambda^2, \mathbf{X}, \mathbf{y} \sim \text{Inv-Gamma}(\frac{n}{2} + \frac{p}{2} + a, \frac{1}{2}\text{SSE} + \beta^{\mathbf{T}}\mathbf{D}_{\tau}^{-\mathbf{1}}\beta + b) \tag{53}$$

Where SSE $= (\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta})$. The full conditional distributions of $\tau_1^2 ... \tau_p^2$ requires some transformations:

$$\begin{aligned} p(\tau_i^2|\sigma^2, \lambda^2, \mathbf{X}, \mathbf{y}) &\propto \frac{1}{(2\pi\sigma^2\tau_i^2)^{1/2}} \exp\Big[-\frac{1}{2\sigma^2\tau_i^2}\beta_i^2 \Big] \exp\Big[-\frac{1}{2}\lambda^2\tau_i^2 \Big] \\ &\propto \frac{1}{(\tau_i^2)^{1/2}} \exp\Big[-\frac{1}{2}\Big(\frac{1}{\sigma^2\tau_i^2}\beta_i^2 + \lambda^2\tau_i^2 \Big) \Big] \end{aligned} \tag{54}$$

If we define $\eta_i = \frac{1}{\tau_i^2}$

$$\begin{aligned} p(\tau_i^2|\beta_i, \sigma^2, \lambda^2, \mathbf{X}, \mathbf{y}) &\propto (\eta_i)^{-3/2} \exp\Big[-\frac{1}{2}\Big(\frac{\beta_i^2}{\sigma^2}\eta_i + \frac{\lambda^2}{\eta_i} \Big) \Big] \\ &\propto (\eta_i)^{-3/2} \exp\Big[-\frac{\beta_i^2}{2\sigma^2\eta_i}\Big(\eta_i^2 - \frac{\sigma^2\lambda^2}{\beta_i} \Big) \Big] \\ &\propto (\eta_i)^{-3/2} \exp\Big[-\frac{\beta_i^2}{2\sigma^2\eta_i}\Big(\eta_i - (\frac{\sigma^2\lambda^2}{\beta_i}) \Big)^2 \Big] \end{aligned} \tag{55}$$

Which is of the form of the inverse-Gaussian distribution, with mean parameter $(\frac{\sigma^2\lambda^2}{\beta_i})^{1/2}$ and scale parameter λ^2.

$$\frac{1}{\tau_i^2}|\sigma^2, \lambda^2, \mathbf{X}, \mathbf{y} = \eta_i \sim \text{Inv-Gaussian}((\frac{\sigma^2\lambda^2}{\beta_i})^{1/2}, \lambda^2) \tag{56}$$

Finally, the full conditional distribution of λ^2 is derived from the joint probability distribution.

$$\begin{aligned} p(\lambda^2|\beta, \sigma^2, \tau_i^2 ... \tau_p^2, \lambda^2, \mathbf{X}, \mathbf{y}) &\propto \Big[\prod_{i=1}^{p} \frac{1}{2}\lambda^2 \exp(-\frac{1}{2}\lambda^2\tau_i^2) \Big] \times (\lambda^2)^{r-1} \exp(-\delta\lambda^2) \\ &\propto (\lambda^2)^{p+r-1} \exp\Big[-(\delta\lambda^2 + \frac{1}{2}\lambda^2 \sum_{i=1}^{p} \tau_i^2) \Big] \\ &\propto (\lambda^2)^{p+r-1} \exp\Big[-\lambda^2(\delta + \frac{1}{2}\sum_{i=1}^{p} \tau_i^2) \Big] \end{aligned} \tag{57}$$

Therefore

$$\lambda^2|\tau_i^2 ... \tau_p^2 \sim \text{Gamma}(p + r, \delta + \frac{1}{2}\sum_{i=1}^{p} \tau_i^2) \tag{58}$$

6.5 Laplace as a mixture normal

Referring to the hierachial conception of the LASSO model in section , where $\phi = \frac{\sigma^2}{\lambda^2}$ and $\beta_i \sim$ Laplace$(\mu = 0, \phi)$. The probability density function for β_i is therefore:

$$f_{\mathrm{B}_i}(\beta_i|\phi) = \frac{1}{2\phi}\exp(-\frac{|\beta_i|}{\phi}) \tag{59}$$

Let $A \sim$ Exp$(\frac{1}{2\phi^2})$ and $B|A \sim$ N$(0, A)$. The marginal distribution of B will therefore be

$$\begin{aligned} f_B(b) &= \int_0^\infty f_{AB}(a,b)\,\mathrm{d}a = \int_0^\infty f_B(b|a)f_A(a)\,\mathrm{d}a \\ &= \int_0^\infty \frac{1}{(2\pi a)^{1/2}}\exp(-\frac{b^2}{2a})\frac{1}{2\phi^2}\exp(-\frac{a}{2\phi^2})\,\mathrm{d}a \\ &= \frac{1}{2\phi^2}\int_0^\infty \frac{1}{(2\pi a)^{1/2}}\exp\Big(-\frac{b^2}{2a}-\frac{a}{2\phi^2}\Big)\,\mathrm{d}a \\ &= \frac{1}{2\phi^2}\int_0^\infty \frac{1}{(2\pi a)^{1/2}}\exp\Big(-\frac{1}{2}(\frac{\phi^2|b|^2+a^2}{a\phi^2})\Big)\,\mathrm{d}a \end{aligned} \tag{60}$$

By completing the square

$$\begin{aligned} f_B(b) &= \frac{1}{2\phi^2}\int_0^\infty \frac{1}{(2\pi a)^{1/2}}\exp\Big[-\frac{(a-|b|\phi)^2+2a|b|\phi}{2\phi^2 a}\Big]\,\mathrm{d}a \\ &= \frac{1}{2\phi^2}\int_0^\infty \frac{1}{(2\pi a)^{1/2}}\exp\Big[-\frac{|b|}{\phi}\Big]\exp\Big[-\frac{(a-|b|\phi)^2}{2\phi^2 a}\Big]\,\mathrm{d}a \end{aligned} \tag{61}$$

Let $C \sim$ Inv-Gaussian(γ, δ) where. The probability density function for C is therefore

$$f_C(c|\gamma,\delta) = \Big(\frac{\delta}{2\pi c^3}\Big)^{1/2}\exp\Big[-\frac{\delta(c-\gamma)^2}{2\gamma^2 c}\Big] \tag{62}$$

Let $\gamma = |b|\phi$ and $\delta = |b|^2$. With some manipulation.

$$\begin{aligned} f_B(b) &= \frac{1}{2\phi^2}\exp(-\frac{|b|}{\phi})\int_0^\infty \frac{1}{(2\pi a)^{1/2}}\Big[\frac{\delta^{1/2}}{\delta^{1/2}}\Big]\Big[\frac{(a^2)^{1/2}}{(a^2)^{1/2}}\Big]\exp\Big[-\frac{|b|^2(a-|b|\phi)^2}{|b|^2 2\phi^2 a}\Big]\,\mathrm{d}a \\ &= \frac{1}{2\phi^2}\exp(-\frac{|b|}{\phi})\frac{1}{\delta^{1/2}}\int_0^\infty a\frac{\delta^{1/2}}{(2\pi a^3)^{1/2}}\exp\Big[-\frac{\delta(a-\gamma)^2}{2\gamma^2 a}\Big]\,\mathrm{d}a \\ &= \frac{1}{2\phi^2}\frac{1}{\delta^{1/2}}\exp(-\frac{|b|}{\phi})\mathsf{E}\big[A\big] \end{aligned} \tag{63}$$

Where E$[A]$ is the expectation of A under the form of the inverse-Gaussian with mean γ and variance δ. Therefore

$$\begin{aligned} f_B(b) &= \frac{1}{2\phi^2}\exp(-\frac{|b|}{\phi})\mathsf{E}\Big[A\Big]\frac{1}{\delta^{1/2}} \\ &= \frac{1}{2\phi^2}\exp(-\frac{|b|}{\phi})\gamma = \frac{1}{2\phi^2}\exp(-\frac{|b|}{\phi})|b|\phi\frac{1}{|b|} \\ &= \frac{1}{2\phi}\exp(\frac{|b|}{\phi}) \end{aligned} \tag{64}$$

Which is identical to the Laplace$(0, \phi)$ formulation.

References

Abhijit, G. (2017). *Machine Learning with R*. Springer Singapore, Singapore.

Alhamzawi, R., Alhamzawi, A., and Ali, H. T. M. (2019). New gibbs sampling methods for bayesian regularized quantile regression. *Computers in Biology and Medicine*, 110:52 – 65.

Allenby, G., Rossi, P., and McCulloch, R. (2005). Hierarchical bayes models: A practitioners guide. *SSRN Electronic Journal*.

Auguie, B. (2017). *gridExtra: Miscellaneous Functions for ”Grid” Graphics*. R package version 2.3.

Betancourt, M. (2017). A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*.

Betancourt, M. and Girolami, M. (2015). Hamiltonian monte carlo for hierarchical models. *Current trends in Bayesian methodology with applications*, 79:30.

Brooks, S., Gelman, A., Galin, G. J., and Meng, X. (2011). *Handbook of Markov Chain Monte Carlo*. Chapman & Hall / CRC, New York.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, Articles*, 76(1):1–32.

Celeux, G., Anbari, M. E., Marin, J.-M., and Robert, C. P. (2012). Regularization in regression: comparing bayesian and frequentist methods in a poorly informative situation. *Bayesian Analysis*, 7(2):477–502.

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., and Liu, J. (2013). A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika*, 78(4):685–709.

Detmer, F. and Slawski, M. (2018). A note on coding and standardization of categorical variables in (sparse) group lasso regression. *arXiv preprint arXiv:1805.06915*.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marqué, J. R., Gruber, B., Lafourcade, B., Leitão, P. J., et al. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1):27–46.

Dowle, M. and Srinivasan, A. (2019). *data.table: Extension of ‘data.frame‘*. R package version 1.12.2.

Freedman, D. (2009). *Statistical models: theory and practice*. Cambridge University Press, Cambridge.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, Articles*, 33(1):1–22.

from Jed Wing, M. K. C., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., the R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., and Hunt., T. (2018). *caret: Classification and Regression Training*. R package version 6.0-80.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409.

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

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004). *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd edition.

Gelman, A. and Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472.

Grzenda, W. (2015). The advantages of bayesian methods over classical methods in the context of credible intervals. *Information Systems in Management*, 4.

Hans, C. (2009). Bayesian lasso regression. *Biometrika*, 96(4):835–845.

Harrison, D. and Rubinfeld, D. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5:81–102.

Herawati, N., Nisa, K., Setiawan, E., and Nusyirwan, T. (2018). Regularized multiple regression methods to deal with severe multicollinearity. *International Journal of Statistics and Applications*, 8(4):167–172.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hoff, P. D. (2009). *A First Course in Bayesian Statistical Methods*. Springer Publishing Company, Incorporated, 1st edition.

Hoffman, M. D., Matthew, D., and Gelman, A. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.

Hsiang, T. (1975). A bayesian view on ridge regression. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 24(4):267–268.

Kruschke, J. (2015). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Academic Press, 2nd edition.

Kyung, M., Gill, J., Ghosh, M., Casella, G., et al. (2010). Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, 5(2):369–411.

Li, Q., Lin, N., et al. (2010). The bayesian elastic net. *Bayesian analysis*, 5(1):151–170.

Liang, W., Dai, H., and He, S. (2019). Mean empirical likelihood. *Computational Statistics and Data Analysis*, 138:155 – 169.

Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2014). A significance test for the lasso. *Annals of statistics*, 42(2):413.

Mallick, H. (2015). *Some contributions to Bayesian regularization methods with applications to genetics and clinical trials*. PhD thesis, The University of Alabama at Birmingham.

Marquardt, D. W. and Snee, R. D. (1975). Ridge regression in practice. *The American Statistician*, 29(1):3–20.

Meen, G. (2009). Modelling local spatial poverty traps in england. *Housing Studies*, 24(1):127–147.

Moran, G. E., Ročková, V., George, E. I., et al. (2018). Variance prior forms for high-dimensional bayesian variable selection. *Bayesian Analysis*.

Newman, D., Hettich, S., Blake, C., and Merz, C. (1998). Uci repository of machine learning databases.

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

Pericchi, L. R. and Smith, A. F. M. (1992). Exact and approximate posterior moments for a normal location parameter. *Journal of the Royal Statistical Society. Series B (Methodological)*, 54(3):793–804.

Plummer, M. (2018). *rjags: Bayesian Graphical Models using MCMC*. R package version 4-8.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ribeiro Jr, P. J. and Diggle, P. J. (2018). *geoR: Analysis of Geostatistical Data*. R package version 1.7-5.2.1.

Rifkin, R. M. and Lippert, R. A. (2007). Notes on regularized least-squares. Technical report.

RStudio Team (2019). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA.

Schloerke, B., Crowley, J., Cook, D., Briatte, F., Marbach, M., Thoen, E., Elberg, A., and Larmarange, J. (2018). *GGally: Extension to 'ggplot2'*. R package version 1.4.0.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13.

Spiegelhalter, D. J. (2004). *Bayesian approaches to clinical trials and health-care evaluation*. John Wiley & Sons, Chichester.

Stan Development Team (2019a). RStan: the R interface to Stan. R package version 2.19.2.

Stan Development Team (2019b). *Stan Modeling Language Users Guide and Reference Manual*.

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

Titterington, D. M. (1997). *Introduction to Gelfand and Smith (1990) Sampling-Based Approaches to Calculating Marginal Densities*, pages 519–550. Springer New York, New York, NY.

Vats, D. and Knudson, C. (2018). Revisiting the gelman-rubin diagnostic. *arXiv preprint arXiv:1812.09384*.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P. (2019). Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of mcmc. *arXiv preprint arXiv:1903.08008*.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12):1–20.

Yue, S., Zhiguo, F., and Ka, F. C. Y. (2017). A descent method for least absolute deviation lasso problems. *Optimization Letters*.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.