# Lasso and Bayes: A Demostration using Real State Market data

*Final Project*

**Iker Caballero & Victor Jimenez**

**Bayesian Analysis** - Prof. Xavier Puig & Prof. Jesus Corral

# Contents

# 1 Introduction

Since the start of the 2000s, learning methodologies and technologies have gained incredible relevance in all layers of modern societies, given their immense applications and their improving capabilities. In the era of data, however, it is important for scientists and professionals in different fields to use the "right" data to fulfill their purposes and avoid what is known as "overfitting", which consists in the models capturing noisy patterns that lead to a worse generalization of the results.

Hence, a common approach is to use regularization methods, which are techniques that allow to avoid this problem. One of the most famous regularization approaches for the regression problem is the Lasso regression, introduced by Tibshirani (1996) [1]. The Lasso allows to regularize through eliminating variables in the model following an $L_1$ penalization (the absolute value) weighted through a penalty parameter (normally denoted by $\lambda$), which weights the relevance of the penalization imposed to the coefficients. The challenge in practice is to choose an adequate penalty term, and this is usually done through cross-validation.

Nevertheless, Tibshirani himself noted that, given the form of the penalty term in the Lasso regression, the estimates could be regarded as posterior mode estimates if the parameters have independent and identical Laplace priors. Therefore, various researchers have proposed different priors and generalizations in order to consider different ways of approaching the problem. A Lasso from a Bayesian point of view could allow to potentially choose the penalty parameter, include the estimation uncertainty for the error variances and other parameters and obtain more stable Lasso estimates than using cross-validation.

Hence, our aim for this project is to apply the Bayesian analysis for the Lasso real data in order to illustrate how this method works with different priors and models, to compare the results obtained regarding prediction and compare the common approach with the Bayesian approach. The motivation of this project is purely theoretical, as the Lasso can be applied to various different context and our goal is to use Bayesian analysis for different aspects related to the Lasso. Despite this fact, because it is widely used in economics and finance application, we decided to illustrate the different aspects of the model choosing a data set about the real state market in Boston, Massachussets, USA.

This project will be organized as follows: first, we briefly develop some important theory about the Lasso and its relation with Bayesian analysis. Then, we present the data set we will use for the whole analysis and do a preliminary analysis. We proceed to the Bayesian analysis using different models to discuss and interpret the results obtained. We finally highlight the most important conclusions of this project.

# 2 About the Dataset

The real state market data that we will use in our project is no other but the Boston Housing dataset, which is a benchmark database which has been widely used in the statistical learning literature. The dataset contains information collected by the U.S Census Service concerning housing in the area of Boston, Massachussets, USA. It was obtained from the StatLib archive, and has been used extensively throughout the literature to benchmark algorithms.

This dataset contains 506 observations and 14 variables, so this is a small dataset. The variables are sociodemographic and economic, and these are the following:

- CRIM - per capita crime rate by town
- ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS - proportion of non-retail business acres per town.
- CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
- NOX - nitric oxides concentration (parts per 10 million)
- RM - average number of rooms per dwelling
- AGE - proportion of owner-occupied units built prior to 1940

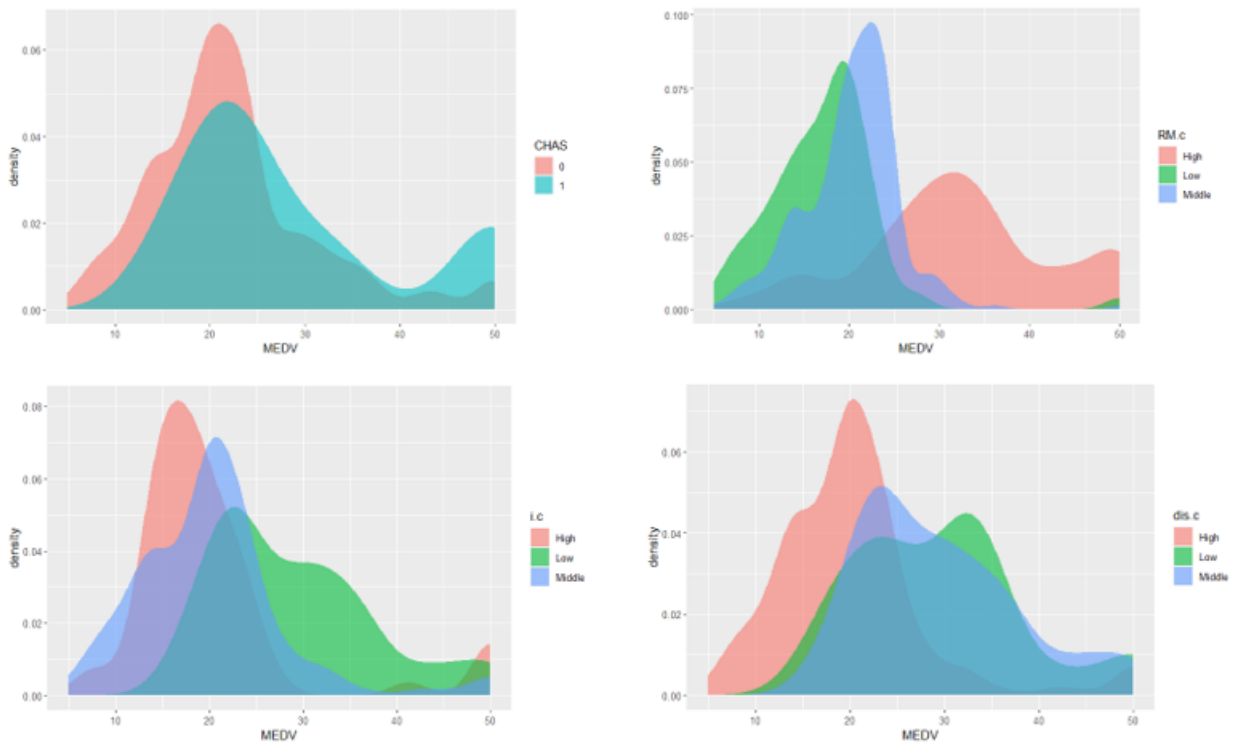


Figure 1: Distribution of MEDV segmented by different levels of CHAS (upper left), RM (upper right), INDUS (bottom left) and DIST (bottom right)

- DIS - weighted distances to five Boston employment centres
- RAD - index of accessibility to radial highways
- TAX - full-value property-tax rate per $10,000
- PTRATIO - pupil-teacher ratio by town
- B - $1000(Bk - 0.63)^2$ where $Bk$ is the proportion of blacks by town
- LSTAT - percentage lower status of the population
- MEDV - median value of owner-occupied homes in $1000's

Given our focus on the theoretical and the practical aspects of the lasso in the Bayesian framework, we are not specifically interested in a complete exploratory analysis of the data. However, we just check data quality and give some examples that motivate the usage of regularization methods in this dataset. In Figure 1 we have plotted the median value of houses depending on different levels of other variables, such as CHAS, INDUS, RM or DIST. Note that some of the variables have been converted to categorical for plotting (dividing them into three categories corresponding to their quartiles).

The different diagrams clearly show that the distribution of MEDV does not seem to be the same independently of the other variables. More importantly, it seems that there are variables where the distribution seems to be the same (like CHAS) and others where the distributions are clearly different (like RM). Hence, using MEDV as the dependent variable in a regression model, one could say that some variables do not seem to contribute significantly to the values of MEDV, so that applying regularization methods can lead to less noise in our model.

This motivates the usage of the lasso regression, and it is a perfect context for us to apply the Bayesian framework regarding the Lasso. However, we first discuss the lasso regularization methodology and then we link it with Bayes.

As usual with this procedure, we normalized the explanatory variables and centered the target one (MEDV in our case), as the lasso procedure is not invariant to the scales. Moreover, we look for extreme values, but we do not detect any, so we continue with the transformed data to develop our project.

# 3 Frequentist Lasso

In order to perform inference on a parameter or find a model for a certain learning task, one must validate and assess the model and also decide with which data to feed it, and find ways to quantify these decisions so that an optimal configuration is found. For that reason, a supervised learning model is usually defined as a minimization problem consisting of two terms:

$$\min_{f \in \mathcal{F}} E(\lambda) := \frac{1}{n} \sum_{i=1}^{n} L\left(y_i, f\left(x_i\right)\right) + \lambda \Omega(f) \quad for \quad \lambda > 0 \tag{1}$$

The first term quantifies the error of the prediction, by means of a loss function $L$ that compares predictions $f\left(x^i\right)$ and original target values $y^i$ using some distance or similarity metric. The second term quantifies the complexity of the model by means of a functional $\Omega(f)$ called regularizer, which is vitally important for the problem to reach an optimal value that avoids complete interpolation of the data, as this situation would yield to overfitting of the training dataset that would not generalize to new data points. The parameter $\lambda$ is the penalty parameter, and controls the trade-off between the minimization of the error and the complexity, and is associated with the bias-variance trade-off in the sense that lowering regularization yields lower variance but higher bias and vice-versa. This parameter is usually selected using validation techniques such as cross-validation (GCV, k-fold CV...), and our Bayesian approach will enable us to define a probability distribution for the parameter that will allow us to set aside arbitrariness in its choice.

Several regularization functionals and techniques have been proposed over the years, and in this project we will focus on the Lasso regularizer (least absolute shrinkage and selection operator), which is a particular approach aimed at linear regression models and allows for a sparse solution by minimizing the $L_1$-norm of the parameter vector. Considering a normal linear regression model like

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \tag{2}$$

where $\mathbf{y}$ is an $n \times 1$ vector of response, $\mathbf{X}$ is the $n \times p$ matrix of standardized regressors, which includes a column of 1s for the constant parameter, $\beta$ is the vector of coefficients $p \times 1$ and $\epsilon$ is the $n \times 1$ vector of independent and identically distributed normal errors with mean 0 and unknown variance $\sigma^2$, the Lasso estimates are obtained by solving the minimization problem

$$\min_{\beta} (\tilde{\mathbf{y}} - \mathbf{X}\beta)^T (\tilde{\mathbf{y}} - \mathbf{X}\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \tag{3}$$

for some $\lambda > 0$, where $\tilde{\mathbf{y}} = \mathbf{y} - \bar{\mathbf{y}}\mathbf{1_n}$ and $\mathbf{1_n}$ is a $n$-dimensional vector of 1's. The reason behind this choice is illustrated in Figure 2, where the optimization problem in a 2 dimensional parameter space is displayed for the $L_1$ (lasso) and $L_2$ (ridge) case:

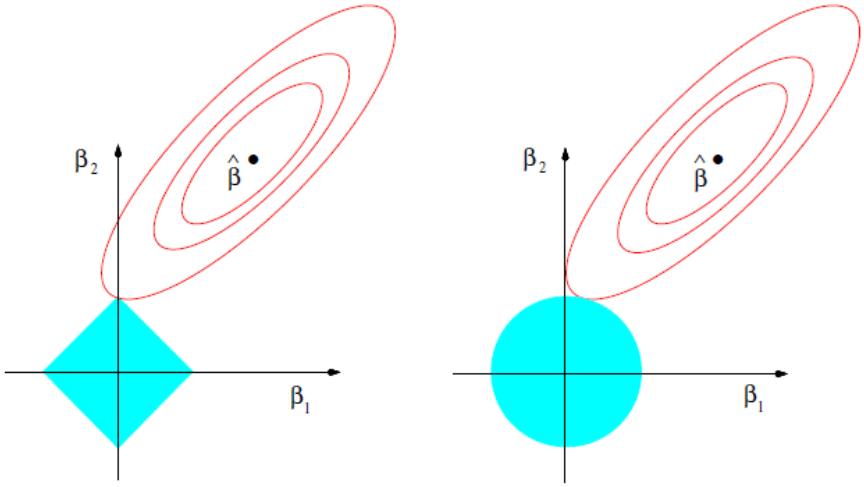


Figure 2: Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \leq t$ and $\beta_1^2 + \beta_2^2 \leq t^2$, respectively, while the red ellipses are the contours of the least squares error function. [3]

The regularization via $L_1$-norm simultaneously allows for a sparse solution while maintaining the convexity of the problem, and thus guaranteeing the uniqueness of the optimal. As we see, the contact between the feasibility region and the level curves of the loss function happens in a vertex that is aligned with the axis, thus setting this parameter $\beta_2$ to zero and allowing the solution to be sparse. For problems with high-dimension parameter space, sparsity in the solution will be required to obtain an interpretable model and the tuning of the penalty parameter $\lambda$ will determine how sparse the optimal solution is.

In our case, the dataset contains 13 explanatory variables, and we can represent the value of $\beta$ as a function of $\lambda$ as well as the predictive MSE obtained via cross-validation. This combined representation is very useful, as it allows us to identify how the model is tuned and how the most significant parameters are regularized to zero at the highest values of $\lambda$. Sometimes we will select a value of $\lambda$ that results in a slightly higher predictive error but that allows for a more interpretable sparse model.

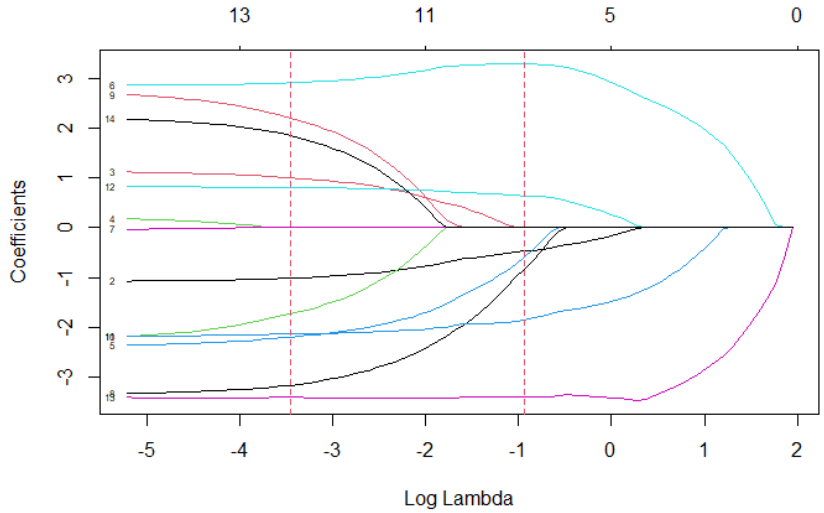


Figure 3: Beta paths representation. On top, the degrees of freedom of the model (i.e. the number of non-zero parameters).

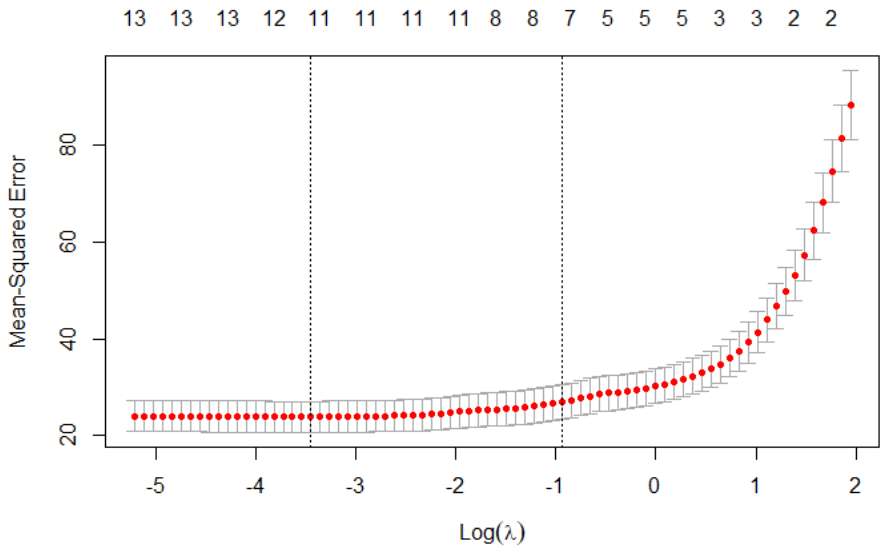


Figure 4: Predictive MSE as a function of $\lambda$. On top, the degrees of freedom of the model (i.e. the number of non-zero parameters).

As we can see, two values of $\lambda$ are highlighted via vertical lines in Figures 3 and 4. The smallest one corresponds to $\lambda_{min} = \min_\lambda MSE(\lambda)$; that is, the value with which fewer errors are to be expected in the predictions, and the second value corresponds to $\lambda_\sigma = \max_\lambda\{\lambda | MSE\left(\lambda_\sigma\right) = MSE\left(\lambda_{\min}\right) + \sigma\left(MSE\left(\lambda_{\min}\right)\right)\}$; which is the maximum value of $\lambda$ for which the mean error lays on the range of errors found for $\lambda_{min}$ (more specifically, at a standard deviation of the errors). This second parameter is significant in lasso regression, since the shrinkage of the vector is not the main factor determining the suitability of the model, but the number of coefficients that are made zero. In this sense, providing a more sparse vector that behaves almost as well as the optimal might be adequate in certain cases.

In this case, with $\lambda_{min}$ we obtain a model with 11 non-zero parameters, whereas with $\lambda_\sigma$ we obtain a model with 7 non-zero slopes. Furthermore, by looking at the beta paths plot, it is intuitive to consider as a sparser alternative the values $\lambda_3$ and $\lambda_5$, which are the minimum value of the penalty (therefore of MSE) that considers only three and five non-zero parameters, respectively. This sparser solutions are expected to contain the most relevant sources of correlation and thus provide a good prediction when new observations are centered in covariate distribution. Using a 30% of the dataset as a validation subset, we obtained the following metrics for the different selections:

| | $\lambda_{min}$ | $\lambda_{SE}$ | $\lambda_5$ | $\lambda_3$ |
|---|---|---|---|---|
| MAE | 3.236 | 3.235 | 3.432 | 3.563 |
| MSE | 20.680 | 23.044 | 24.068 | 25.344 |

Table 1: Test set metrics for the frequentist lasso model

As we can see, increasing complexity does not improve prediction metrics significantly. If we were able to obtain more observations or simulate them, the most sparse model would be sufficient to obtain an acceptable prediction of our covariate.

In this project we want to translate this idea to the Bayesian framework, so that the lack of accuracy in of the model is represented properly and contained in a distribution function associated with our estimated coefficients. In that direction, Tibshirani (1996) noted that Lasso estimates could be interpreted as posterior mode estimates whenever the parameters come from double-exponential or independent and identical Laplace priors. If we generalize both the ridge regression and the lasso and view them as Bayes estimates. Considering

$$\tilde{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left( y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\} \tag{4}$$

we can think of $|\beta_j|^q$ as the log-prior density for $\beta_j$. The value $q = 0$ corresponds to variable subset selection, as the penalty simply counts the number of nonzero parameters; $q = 1$ corresponds to the lasso, while $q = 2$ to ridge regression. The prior corresponding to the $q = 1$ case is an independent double exponential (or Laplace) distribution for each input, with density

$$f(\beta) = \frac{1}{2\tau} \exp\left( -\frac{|\beta|}{\tau} \right) \quad where \quad \tau = \frac{1}{\lambda} \tag{5}$$

Even though this interpretation seems reasonable, it remains unclear exactly how the Bayesian framework allows for sparsity. In the next section, we will elucidate how sparsity arises and in which way it requires a different interpretation.

# 4 Bayesian Sparsity

In general, sparsity is the assumption that, in situations where many of the covariates can be thought to possibly correlate with the outcome variable of interest, only a few of these have a meaningful correlation, and thus the rest can be discarded. This allows for a significant reduction of the size of the data that has to be considered for further predictions thus decreasing the computational cost and, more importantly, improving the interpretability of the model. In the previous section, the frequentist Lasso regularizer was presented and its natural expression as a prior for the regression parameters was derived. However, before dealing with the different Bayesian models that allow us to implement sparsity, we must discuss the difference between the frequentist and Bayesian approaches to sparsity here considered.

On the one hand, frequentist sparsity (e.g. the $\sum_{j=1}^{p} |\beta_j|^q$ term) is thought as a decision-making process that explicitly selects a small subset of covariates, which in the linear regression case is equivalent to identifying which slopes are zero and which are non-zero. This selection process is completely data-driven, as when the maximum likelihood estimator of the slope $\beta_j$ falls below the scale $\lambda$ (or $\lambda_j$, if the penalty term is specified for each covariate), it is regularized towards zero while keeping the estimates above $\lambda$ with negligible regularization. On the other hand, Bayesian sparsity avoids the decision altogether and instead manifests the significance of the covariates in the posterior distribution, which concentrates around a neighborhood in the parameter space where the insignificant slopes are zero.

This behavior can be induced via the choice of a specific prior function, but a problem associated with this new framework arises: including a penalty term in a frequentist-based regularization affects each point in the parameter space differently, whereas assigning a prior distribution for the estimates affects the whole parameter space at once. For that reason, using independent Laplace priors for each of the slopes does not yield the best possible posterior regularization, as the induced behavior for values above and below $\lambda$ (or $\lambda_j$) occur at the same time, and a heavy tail drags posterior probability far below or above this value, thus leaking probability mass towards undesired values. This is particularly problematic for our purposes, as relevant slopes could be reduced towards smaller values and be identified as non-relevant, while small slopes could still retain some concentration of mass and thus prevent an appropriate identification of the irrelevant parameters.

For these reasons, a more holistic approach to regularization must be considered, with a prior distribution that imposes a global scale while also giving the slopes the flexibility to surpass them. Some of the models used in the literature are the Horseshoe hierarchical model or the Finnish Horseshoe, which is a modification of this first. In any case, it is important to note that the choice of a penalty parameter in the frequentist approach has been substituted by the choice of suitable hyperpriors. We delve deeper into each one of the models (Laplace, Horseshoe, Finnish Horseshoe, etc.) and present results using real data in the following section.

# 5 Bayesian Analysis

Now that both the lasso and Bayesian sparsity have been discussed, in this section we aim to develop the details of each one of the models previously mentioned both theoretically and practically (through the dataset). For a better understanding, we first make use of a simple model to see how the Bayesian framework of Lasso works and then we construct and discuss other models that are more common in the literature and have nicer properties. The Bayesian models have been implemented in Stan, some of them with an additional covariate as an intercept for stability.

## 5.1 Strongly Regularizing Prior

Our first model does not use any Laplace or similar distributions, but instead we fix normal strongly regularizing priors for $\beta$ and $\sigma$. Hence, the model will have the following form:

$$\begin{aligned} y|\beta, X, \sigma &\sim N_n(X\beta, \sigma^2 I_{n\times n}) \\ \beta &\sim N(0,3) \\ \sigma &\sim N(0,2), \quad \sigma^2 > 0 \end{aligned} \tag{6}$$

where we fix the standard deviation of each of the priors such that they are strongly regularizing (a larger variance would allow betas to deviate more). The estimation of this model is done through a Stan code and using an MCMC algorithm. All chains have converged for all the parameters. An example of the chain behaviour (using the traces) is seen in Figure 5.

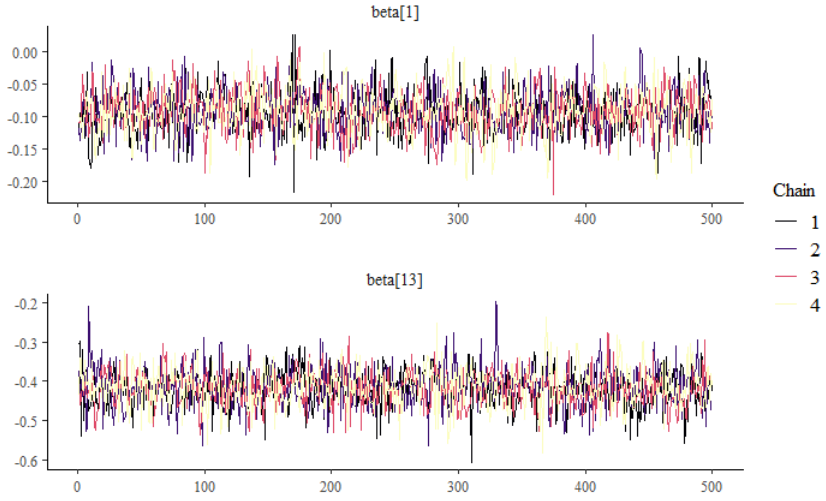


Figure 5: Conditional Monte Carlo chains for the parameter of CRIM and the parameter for B

The results of applying such a model are represented in Figure 6, where the different posterior distributions are plotted together with some other elements interesting for interpretation and comparison. The results show how some variables are concentrated around values far away from zero (not including the number), which means that the point estimators of these variables is expected to be different from zero, which indicates that these variables are relevant for predicting MEDV (the median value of houses). However, some of the variables are concentrated around 0 or contain it, such as INDUS or AGE, while the others are distributed near zero but are not concentrated around it or do not include it. If we use the frequentist estimates as a guide, we can draw vertical threshold lines indicating the values for which the estimates with $\lambda_5$ and $\lambda_3$ were different than zero.

The latter is drawn in red because some estimates below that threshold are clearly centered away from zero, and for that reason cannot be compared with the frequentist case. However, the $\lambda_5$ model indeed resembles the frequentist Lasso. It is significant, however, that even if there are also five components with a significantly high value, three of them are not coincident with the frequentist case: 'DIS', 'RAD' and 'NOX' have significantly high non-zero values. For the other two, however, the centers of the distribution coincide with the values given in the frequentist case.

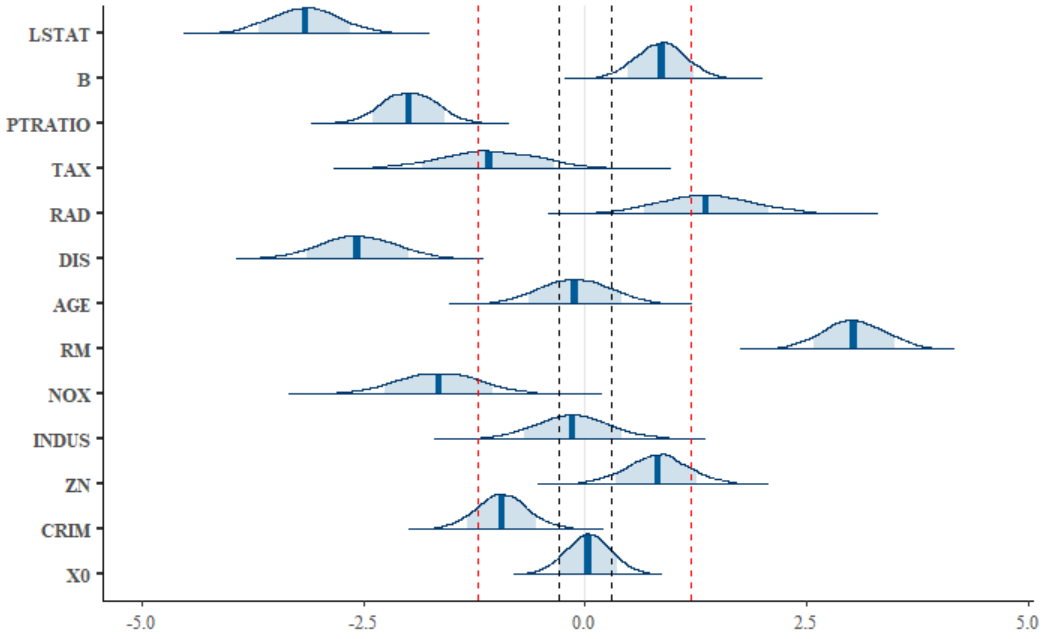


Figure 6: Posterior distributions for the different parameters for the strongly regularizing prior. The dashed lines the values for which the estimates with $\lambda_5$ (black) and $\lambda_3$ (red) were different than zero

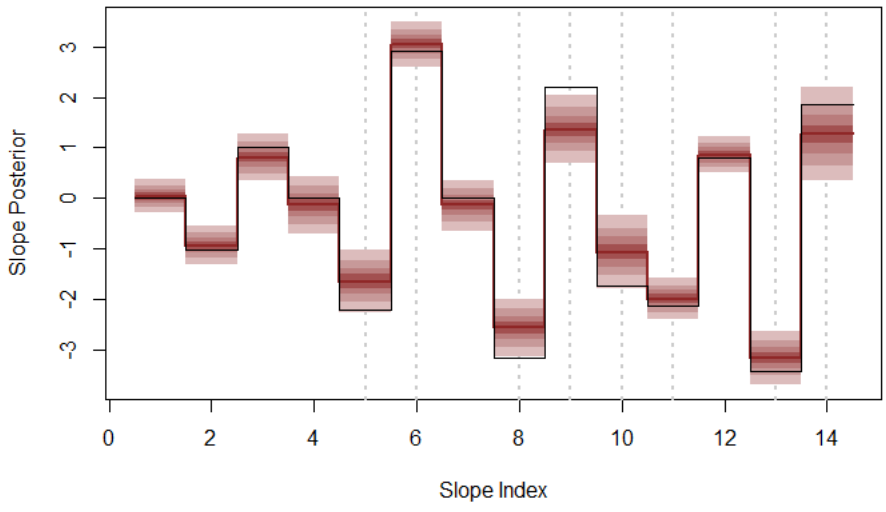


Figure 7: Strongly regularizing prior slopes (shaded regions) vs frequentist lasso slopes (black line)

Figure 6 also allows to obtain estimations for the credibility intervals of the different parameters (from the posterior distribution), which are represented with the blue shaded region inside each one of the posterior distri-

butions. We can see how some credibility intervals do include the value zero inside them. A clearer comparison is given using the diagram in Figure 7, where we represent the different values of the slopes estimated using the frequentist approach with the slopes induced by the posterior distributions obtained for this model. The diagram also represents visually the divergences between estimates that we were highlighting.

Henceforth, we have achieved some sparsity in the Bayesian sense, even though there is room for improvement. The center of the posteriors of $\beta_4$, $\beta_7$, $\beta_8$, $\beta_9$ and $\beta_{13}$ do not coincide with the frequentist lasso estimates. For that, in the next model we will consider the penalty term as in Tibshirani's interpretation of the estimates.

## 5.2 Laplace Prior

The second model we will use, then, is the naive translation of the frequentist lasso approach to the Bayesian framework, which deals with a Laplace prior for the different betas. This model will have the following form:

$$\begin{aligned} y|\beta, X, \sigma &\sim N_n(X\beta, \sigma^2 I_{n\times n}) \\ \beta &\sim \mathcal{L}(0, 41.6) \\ \sigma &\sim N(0, 2), \quad \sigma^2 > 0 \end{aligned} \tag{7}$$

In this case, the prior for the betas is a double exponential with mean zero and tau parameter 41.6, as this is the approximate value of the reciprocal of the optimal $\lambda$ parameter

$$\tau = \frac{1}{\lambda_{min}} \quad \Longrightarrow \quad \tau = \frac{1}{0.02404} \approx 41.6 \tag{8}$$

Because $\lambda$ has been selected rather than modelled with a distribution, in this model one could select a value of $\lambda$ through different methods, allowing to try different versions of the model and validating them so that one can choose an appropiate value in different context.

In this new case, we replicate the previous procedure to evaluate and discuss the results obtained with the model. Hence, the estimation of this model is done in the same way as before, and the chain convergence is as well satisfactory, which can be also visually seen in Figure 8.

We represent the posterior distributions of the same variables we used before in Figure 9. In here, there are some variations regarding the previous model: some of the posterior distributions seem to be wider and further from zero than before, such as those for 'TAX' or 'RAD', for example. Moreover, some distributions have shifted their center further from zero, so that the zero does not seem to be included in the distribution for some posteriors that included it before (like 'RAD'). However, we should take these results just as an illustration of choosing the double exponential, as in the previous model we chose a strongly regularizing prior and this, of course, delivered more sparsity (even though it is not based on the Bayesian interpretation of the lasso regression, which is the theoretical justification for this model).

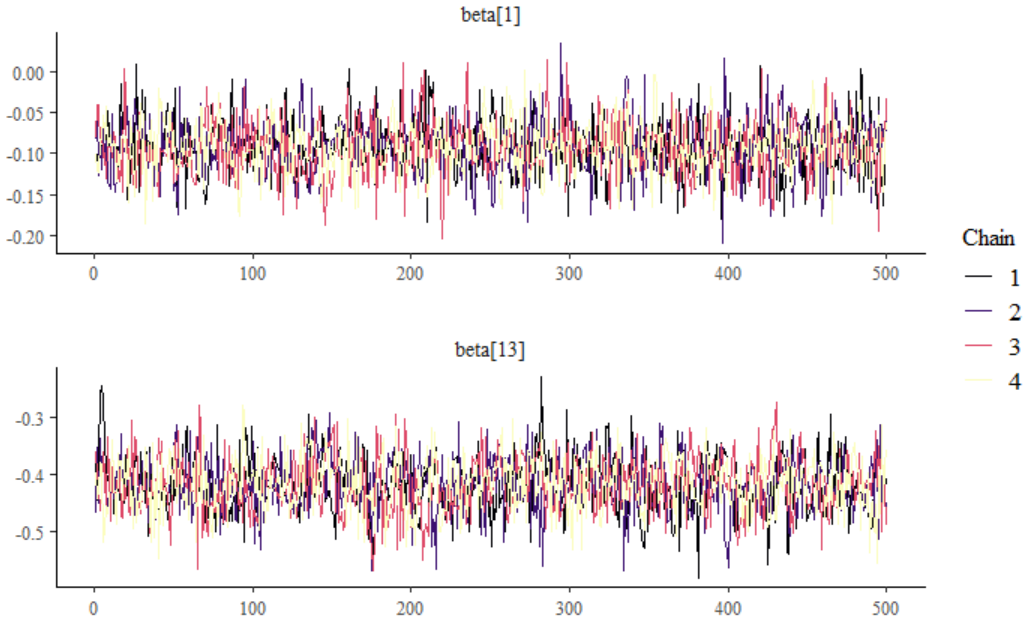


Figure 8: Conditional Monte Carlo chains for the parameter of CRIM and the parameter for B

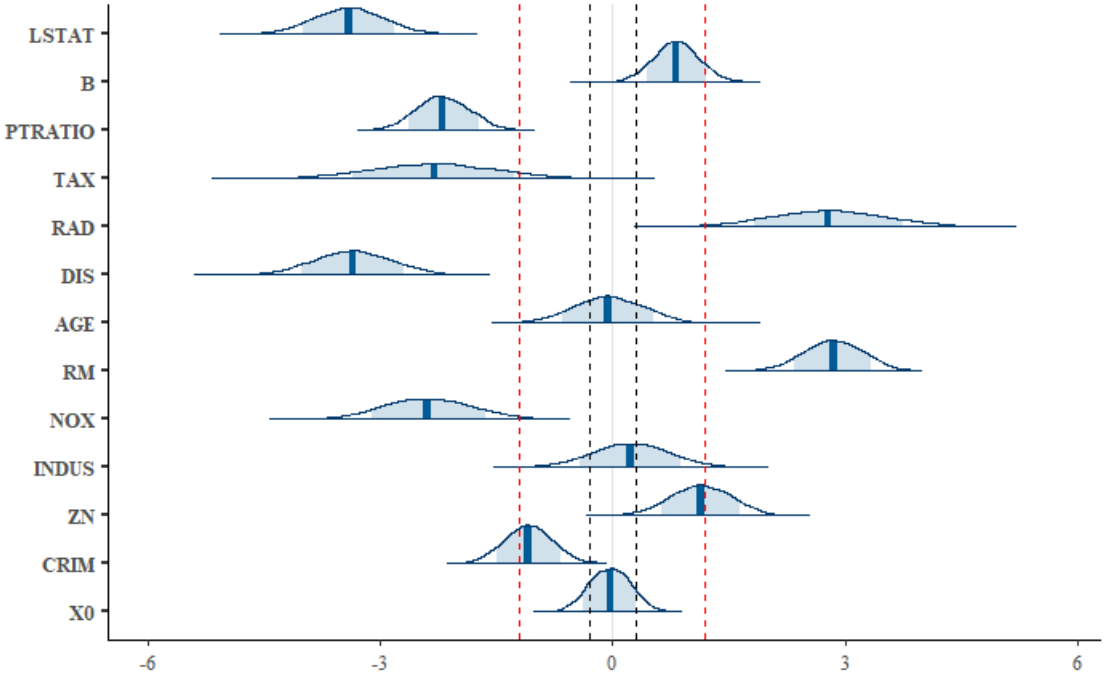


Figure 9: Posterior distributions for the different parameters for the strongly regularizing prior. The dashed lines the values for which the estimates with $\lambda_5$ (black) and $\lambda_3$ (red) were different than zero

Figure 10 shows the comparison between the frequentist estimates for the parameters and the results obtained using this model. Interpreting the elements of the diagram in the same way as before, one can see that there is a better fit of this model to the estimates obtained through the frequentist approach: results that differed from the frequentist estimation such as for the parameter corresponding the fifth index or the ninth index, now seem to

include the value of these estimates and also seem to be more centered towards them, indicating a better fit than the previous model.

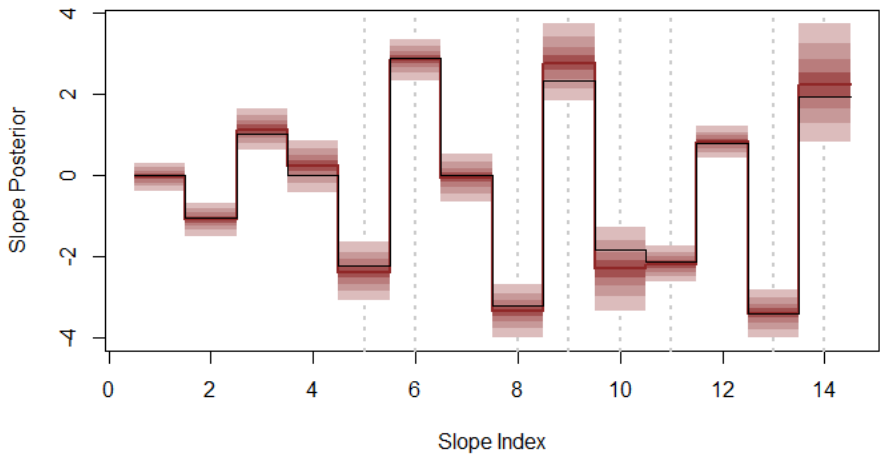


Figure 10: Laplace prior slopes (shaded regions) vs frequentist lasso slopes (black line)

As we discussed earlier, this model uses independent Laplace distributions for the betas, which makes the regularization parameter affect all quantities at the same time (affects the whole parameter space at once), rather than affecting each one separately (like in the frequentist approach). Hence, we can consider the Horseshoe model and its modification, the Finnish Horseshoe model, in order to overcome this problem and probably obtained a better fit.

## 5.3 Horseshoe Prior

The Horseshoe hierarchical model [4] accomplishes this flexibility by setting a scale for each component as the product of a local scale $\lambda_j$ equivalent to the penalty parameter and a global scale $\tau$:

$$\begin{aligned} y|\beta, X, \sigma &\sim N_n(X\beta, \sigma^2 I_{n\times n}) \\ \beta_j &\sim \mathcal{N}(0, \tau \cdot \lambda_j) \\ \lambda_j &\sim \text{Half} -\mathcal{C}(0,1) \\ \tau &\sim \text{Half} -\mathcal{C}(0, \tau_0)\,. \end{aligned} \tag{9}$$

The choice of the Half-Cauchy distribution is justified for being a heavy-tailed distribution, meaning that it assigns substantial probability mass to both small and large values. When used as a prior for the $\lambda_j$, it allows for substantial variability in the shrinkage of the different $\beta_j$. Specifically, when $\lambda_j$ is close to zero, it effectively shrinks the corresponding $\beta_j$ towards zero, implying that the corresponding covariate has little to no effect on the outcome. On the other hand, when $\lambda_j$ takes on a large value, it allows the corresponding $\beta_j$ to be far from zero, implying that the corresponding covariate has a significant effect on the outcome.

Its choice for the parameter $\tau$ can be justified in a similar way. In this case, when $\tau$ is close to zero, it shrinks all of the $\beta_j$ towards zero, implying that all covariates have little to no effect on the outcome. On the other hand, when $\tau$ takes on a large value, it allows the $\tau$ to be far from zero, implying that the covariates can have significant effects on the outcome. This enables the model to adapt to the data and identify the overall level of sparsity that is appropriate for each context.

In this case, the tuning of the parameter $\tau_0$ is of utmost importance. Thankfully, a proper interpretation can be given to this parameter, so that the selection is far from arbitrary: $\tau_0$ will determine the irrelevance of all slopes at the same time (i.e. global scale). For that reason, the specification of $\tau_0$ can be associated with the measurement process, so that the contribution of a slope is considered negligible when it is indistinguishable from the inherent variability of the observations. For a linear regression model:

$$y_i \sim N\left(\mathbf{x}_i^T \boldsymbol{\beta} + \alpha, \sigma\right) \tag{10}$$

a reasonable assumption for the hyperprior would be $\tau_0 = \frac{\sigma}{\sqrt{N}}$, where the denominator accounts for the dependence of the variability effects on the size of the data. Another proposed threshold parameter is the following, which includes an additional term associated with the number of expected relevant slopes $m_0$:

$$\tau_0 = \frac{m_0}{M - m_0} \frac{\sigma}{\sqrt{N}} \tag{11}$$

This extra term balances the contribution of several tails together when we consider the contribution from many slopes at once. In our case, we have selected $m_0 = 5$ to test the degree of sparsity obtained with the model, and we obtained the following results:

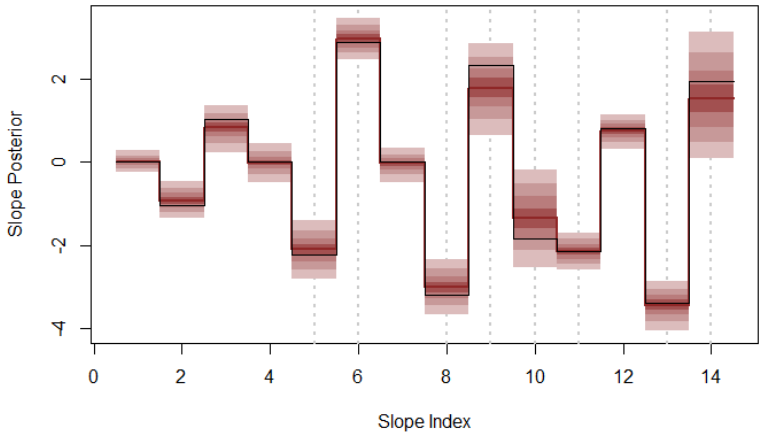


Figure 11: Horseshoe prior slopes vs Lasso slopes.

As we see, most of the posteriors are centered in the frequentist lasso estimates, which is an indication of a good fit. Nevertheless, we still are not able to obtain accurate point estimates for the posteriors of $\beta_8$, $\beta_9$ and $\beta_{13}$.

## 5.4 Finnish Horseshoe Prior

An immediate problem that arises using this prior is that those slopes transcending the global scale $\tau$ are left unregularized, which could yield to extremely large estimate posterior values and thus compromise the validity of the analysis. For that, a modified version, called Finnish horseshoe [5], is proposed, adding an additional layer associated with a new scale $c$, which modifies the local scale as follows:

$$\tilde{\lambda}_j = \frac{c\lambda_j}{\sqrt{c^2 + \tau^2 \lambda_j^2}} \tag{12}$$

$$c^2 \sim \text{Inv} - \mathcal{G}\left(\frac{v}{2}, \frac{v}{2} s^2\right) \tag{13}$$

where $v$ and $s$ must be set appropriately.

The results obtained with this model are the following:

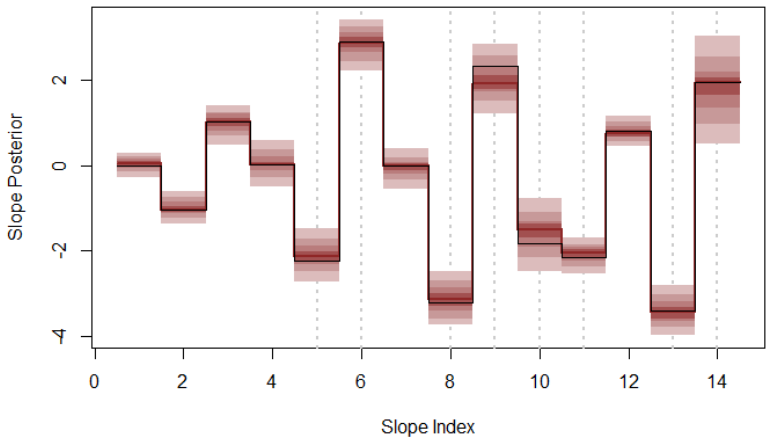


Figure 12: Finnish horseshoe prior slopes vs Lasso slopes.

This clearly is the best model so far, as the posterior distributions of the parameters are centered near the frequentist Lasso estimates. This solution, then, is not only able to obtain predictions with a comparable degree of accuracy but also incorporates uncertainty into the model. For that reason, we will use this last model to perform predictions and provide metrics on the validation set.

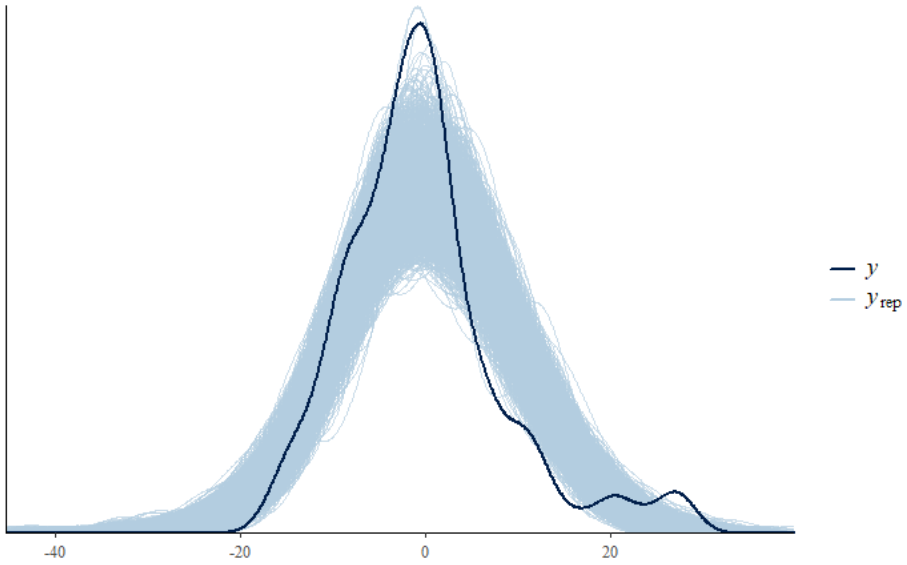


Figure 13: Posterior predictive density overlay. The p-value is set to 0.488.

As we can see, the real distribution of the target variable appears as a possible draw of the posterior predictive simulation, especially for values smaller than 10. The Bayesian p-value, which corresponds to the probability, given the data, that a future observation is more extreme than the data, is computed to be $0.488 \approx 0.5$, which suggests that the model is a good fit for the data. We obtain a MAE of 3.231 and a MSE of 20.689, which are comparable to the frequentist metrics.

All in all, both the distributions of the predicted MEDV values using frequentist Lasso with $\lambda_{min}$ and the mean

of the posterior predictive distributions simulated from the Finnish horseshoe model are the same, which indicates that the Bayesian model, using point estimates for the posterior distributions, is equivalent to the frequentist model.

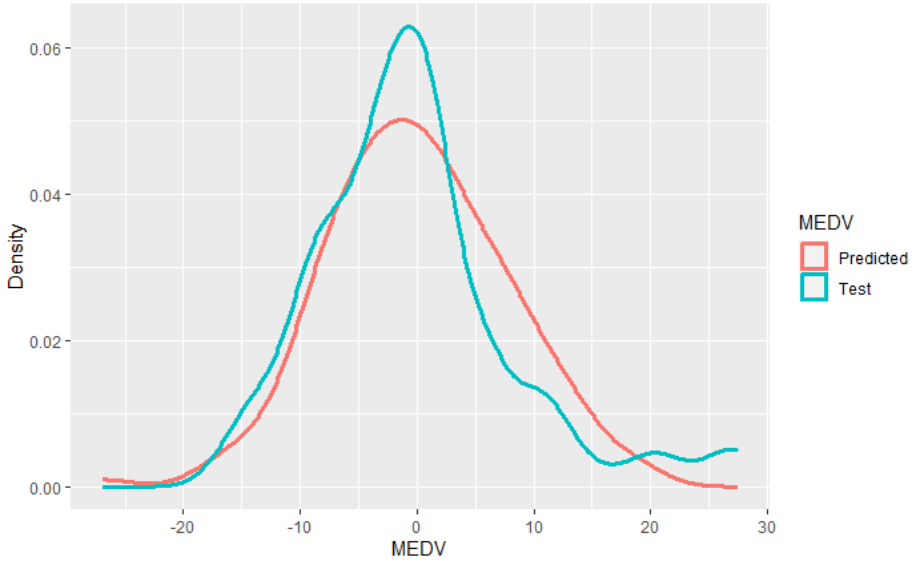


Figure 14: Model vs true values of MEDV.

For that same reason, the prediction error metrics obtained using both approaches are equivalent:

| | **Frequentist Lasso with $\lambda_{min}$** | **Bayesian Finnish Horseshoe** |
|---|---|---|
| MAE | 3.236 | 3.231 |
| MSE | 20.680 | 20.689 |

Table 2: Predictive capability metrics for the frequentist lasso and the Bayesian Finnish Horseshoe model

Nevertheless, as we showed earlier, only the Bayesian model is able to account for the uncertainty in the predictions that make true values deviate from the normal distribution.

# 6 Conclusions

In this project, we have discussed the lasso procedure in the usual way and introduced the idea of Bayesian sparsity and how, from the frequentist lasso, one can construct models that allow to obtain sparse results using the Bayesian modelling framework.

Regarding the interpretation of the results, we see that attributes LSTAT, RM, RAD and DIS are the ones that carry most of the weight, and their sign indicates the nature of the correlation. On the one hand, slopes estimated for LSTAT and DIS indicate that the higher percentage of the lower-status population and the higher the distance to employment centers, the lower the median value of the house. On the other hand, the higher amount of rooms per dwelling and the more accessible radial highways (i.e. better connectivity), the higher house prices are. Besides, our model also indicates that the proportion of non-retail businesses (INDUS) and the proportion of old buildings in the zone do not affect significantly the variation of house prices.

Throughout the Bayesian analysis done, we have seen how different priors for the $\beta$ parameters and for other parameters affect the results and allow to obtain more similar results to the frequentist lasso. In the first model, we ignored Tibshirani's interpretation of the Laplace prior and used a strongly regularizing normal prior, yielding sparsity but not too accurate results (when it comes to comparing it with the frequentist estimates). Then, we

improved the model by using a naive translation of the lasso using the Laplace prior, and obtained more similar results.

However, this model carries problems such as the penalization parameter $\lambda$ affecting the whole parameters' space at once, when the frequentist lasso divides the effect for each parameter space. Hence, we used two other models proposed in the literature, such as the Horseshoe model and the Finnish Horseshoe model, a modification of this previous model. The Horseshoe model provided a way of selecting or incorporating the selection of the penalization parameter $\lambda$ through considering a hierarchical model which defines a distribution for it, and also another one for the threshold $\tau$, which also affects the distribution of the covariates' parameters.

Slightly modifying the definition of this last model, we obtain the Finnish Horseshoe, which allows to obtain a great fit compared to the frequentist lasso (better than the previous model) and, because of that, we exemplified the predictive methodology we would use with a Bayesian lasso model. For this last model, we could obtain very similar and even better (depending on the measure) predictive capability, demonstrating the power of the Bayesian lasso modelling.

All in all, this whole project has illustrated the theory and the practice of Bayesian modelling for the lasso regularization procedure, allowing for alternative ways to interpret the sparsity of the resultant parameter vector and providing alternatives to cross-validation and other techniques to select a $\lambda$ parameter (in this case, modelling it through a hyperprior distribution in a hierarchical model).

# References

[1] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological), [online] 58(1), pp.267–288. Available at: https://www.jstor.org/stable/2346178.

[2] Park, T. and Casella, G. (2008). The Bayesian Lasso. Journal of the American Statistical Association, 103(482), pp.681–686. doi:https://doi.org/10.1198/016214508000000337.

[3] T. Hastie, R. Tibshirani, J. Friedman. The Elements of Statistical Learning. Springer, Second Edition (2008).

[4] C. Carvalho, N. Polson and J. G. Scott (2008). The Horseshoe Estimator for Sparse Signals. Discussion Paper 2008-31. Duke University Department of Statistical Science.

[5] Juho Piironen and Aki Vehtari (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. Electron. J. Statist., 11(2):5018-5051.