A brief introduction to Bayesian LASSO

Shenyu Zhou

2024-12-09

1. Introduction

Consider the linear regression model:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i,$$

with data (X, Y), where $X \in \mathbb{R}^{n \times p}$ and $Y \in \mathbb{R}^n$, and the assumption for the error term that $\epsilon_i \sim N(0, \sigma^2)$. The parameters $\beta = (\beta_1, ..., \beta_p)$ and β_0 are typically estimated by minimizing the Residual Sum of Squares (RSS):

$$\text{RSS}(\beta) = \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2,$$

which leads to the ordinary least squares (OLS) estimates.

However, OLS may not be robust when p is large or when collinearity exists. In order to counter this, shrinkage methods are introduced. LASSO (Least Absolute Shrinkage and Selection Operator) is one such technique that replaces the minimization of RSS by a penalized version:

$$\hat{\beta}_{\text{lasso}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

The penalty term $\lambda \sum_{j=1}^{p} |\beta_j|$ forces some of the coefficients to become exactly zero, thereby performing variable selection and leading to a sparse model. The hyperparameter $\lambda > 0$ determines the amount of shrinkage. From a Bayesian perspective, parameters are treated as random variables with prior distributions. Following the Bayes' theorem, the likelihood of the observed data with these priors produces a posterior distribution for the parameters:

$$p(\beta | X, Y) \propto \underbrace{p(Y | X, \beta)}_{\text{likelihood}} \cdot \underbrace{p(\beta)}_{\text{prior}}$$

- The likelihood reflects how plausible the observed data Y are given parameters β and data X.
- The prior expresses pre-established beliefs about β before gathering the data. Different priors will lead to different forms of regularization.

Going back to the standard linear regression, the likelihood for the data is:

$$p(Y|X,\beta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \beta_0 - X_{i,\cdot}\beta\right)^2\right),$$

- where $X_{i,\cdot}$ denotes the i-th row of X.

It can be shown that choosing a Laplace prior (or double-exponential prior) for each β_j:

$$\beta_j \sim \text{Laplace}(0, b)$$

with density:

$$p(\beta_j) = \frac{1}{2b}\exp\left(-\frac{|\beta_j|}{b}\right)$$

yields a posterior that the mode (the maximum a posterior, or MAP, estimate) is the lasso solution.

To continue, the joint prior on β is:

$$p(\beta) = \prod_{j=1}^{p}\frac{1}{2b}\exp\left(-\frac{|\beta_j|}{b}\right) = \left(\frac{1}{2b}\right)^p \exp\left(-\frac{1}{b}\sum_{j=1}^{p}|\beta_j|\right).$$

The choice of β_0 does not affect the argument regarding β.

Substituting the above likelihood and prior distribution to the Bayes' theorem, we have:

$$p(\beta|X,Y) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - X_{i,\cdot}\beta)^2\right)\cdot\left(\frac{1}{2b}\right)^p \exp\left(-\frac{1}{b}\sum_{j=1}^{p}|\beta_j|\right) \tag{1}$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \left(\frac{1}{2b}\right)^p \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - X_{i,\cdot}\beta)^2 - \frac{1}{b}\sum_{j=1}^{p}|\beta_j|\right). \tag{2}$$

By ignoring multiplicative constants that do not depend on β, the posterior simplifies to:

$$p(\beta|X,Y) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - X_{i,\cdot}\beta)^2 - \frac{1}{b}\sum_{j=1}^{p}|\beta_j|\right).$$

Again, the MAP estimate $\hat{\beta}$ is defined as the parameter value that maximizes the posterior distribution. Since the exponential is monotonic, we can minimize the negative log-posterior for simplicity and canceling out the negative signs. The negative log-posterior is:

$$-\log p(\beta|X,Y) \propto \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - X_{i,\cdot}\beta)^2 + \frac{1}{b}\sum_{j=1}^{p}|\beta_j|,$$

which leads the MAP estimate to become:

$$\hat{\beta}_{\text{MAP}} = \arg|_{\beta} \left\{ \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - X_{i,\cdot}\beta)^2 + \frac{1}{b} \sum_{j=1}^{p} |\beta_j| \right\}.$$

Next, recall that the lasso estimate $\hat{\beta}_{\text{lasso}}$ is defined as the optimization problem that:

$$\hat{\beta}_{\text{lasso}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

Placing the MAP objective function and the lasso objective function side by side, we can see that they differ only by a constant scalar factor. If we multiply the MAP objective function by $2\sigma^2$, we obtain:

$$2\sigma^2 \left[\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - X_{i,\cdot}\beta)^2 + \frac{1}{b} \sum_{j=1}^{p} |\beta_j| \right] = \sum_{i=1}^{n} (y_i - \beta_0 - X_{i,\cdot}\beta)^2 + \frac{2\sigma^2}{b} \sum_{j=1}^{p} |\beta_j|.$$

Then, it becomes obvious that by equating $\lambda = \frac{2\sigma^2}{b}$:

$$\hat{\beta}_{\text{MAP}} = \arg|_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - X_{i,\cdot}\beta)^2 + \frac{b}{2\sigma^2} \sum_{j=1}^{p} |\beta_j| \right\} = \hat{\beta}_{\text{lasso}}.$$

Thus, we have shown that the Bayesian MAP estimator under a Laplace prior for regression coefficients corresponds exactly to the lasso estimator. In Appendix A., you can find a more intuitive explanation of the Bayesian way of thinking and the reason why Laplace distribution can be used to zero-out coefficients.

2. Scale Mixture Representation of the Laplace Prior

As established in the introduction, placing an independent Laplace prior on each regression coefficient β_j leads to a posterior whose mode is equivalent to the lasso solution.

This prior is not conjugate to the Gaussian likelihood, which complicates direct posterior inference. However, Park and Casella (2008) offered a key insight: the Laplace distribution can be expressed as a scale mixture of Gaussian and exponential distributions. This reparameterization introduces an auxiliary variable, significantly simplifying posterior sampling and enabling full exploitation of Bayesian inference. Specifically, for parameters $\sigma^2 > 0$ and $\lambda > 0$, if $\beta_j | \lambda_j \sim N(0, \sigma^2 \lambda_j)$ and $\lambda_j \sim \text{Exp}\left(\frac{\lambda^2}{2}\right)$, then the marginal distribution of β_j is Laplace$(0, b)$ where $b = \frac{\sigma^2}{\lambda}$.

The corresponding densities for the aforementioned hierarchical model are:

$$p(\beta_j | \lambda_j, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2\lambda_j}} \exp\left(-\frac{\beta_j^2}{2\sigma^2\lambda_j} \right),$$

and

$$p(\lambda_j) = \frac{\lambda^2}{2} \exp\left(-\frac{\lambda^2}{2} \lambda_j \right), \quad \lambda_j > 0.$$

To find the marginal distribution of β_j, integrate out λ_j:

$$p(\beta_j|\sigma^2) = \int_0^\infty p(\beta_j|\lambda_j, \sigma^2)p(\lambda_j)\, d\lambda_j.$$

Substituting the pdfs, we obtain:

$$p(\beta_j|\sigma^2) = \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2\lambda_j}} \exp\left(-\frac{\beta_j^2}{2\sigma^2\lambda_j}\right) \cdot \frac{\lambda^2}{2} \exp\left(-\frac{\lambda^2}{2}\lambda_j\right)\, d\lambda_j.$$

This integral is known in the literature as a standard result in scale mixtures of normal distributions. Performing the integration yields:

$$p(\beta_j|\sigma^2) = \frac{\lambda}{2} \exp\left(-\frac{\lambda|\beta_j|}{\sigma^2}\right).$$

Define $b = \frac{\sigma^2}{\lambda}$. Then we can rewrite the above as:

$$p(\beta_j|\sigma^2) = \frac{1}{2b} \exp\left(-\frac{|\beta_j|}{b}\right).$$

This is exactly the Laplace distribution with location 0 and scale b:

$$\beta_j|\sigma^2 \sim \text{Laplace}(0, b).$$

Since this derivation holds for any given $\sigma^2 > 0$ and $\lambda > 0$, we conclude that:

$$\beta_j \sim \text{Laplace}(0, b) \iff \beta_j|\lambda_j \sim N(0, \sigma^2\lambda_j) \quad \text{and} \quad \lambda_j \sim \text{Exp}\left(\frac{\lambda^2}{2}\right),$$

with $b = \frac{\sigma^2}{\lambda}$.

3. Full Bayesian Hierarchy

We now consider the full Bayesian Lasso model. Following from the above scale mixture representation of Laplace as the prior on each β_j, we also assign an inverse-gamma prior to σ^2 to complete the Bayesian specification:

$$\sigma^2 \sim IG(a_0, b_0), \quad a_0, b_0 > 0.$$

The joint posterior distribution is given by:

$$p(\beta, \lambda, \sigma^2|y) \propto p(y|\beta, \sigma^2) \prod_{j=1}^{p} p(\beta_j|\lambda_j, \sigma^2)p(\lambda_j)p(\sigma^2).$$

Substituting the known formulas:

- $y|\beta, \sigma^2 \sim N(X\beta, \sigma^2 I_n)$,
- $\beta_j|\lambda_j, \sigma^2 \sim N(0, \sigma^2\lambda_j)$,

- $\lambda_j \sim \text{Exp}\left(\frac{\lambda^2}{2}\right)$,
- $\sigma^2 \sim IG(a_0, b_0)$.

This encompasses the full hierarchical Bayesian model.

4. Posterior Sampling

A key advantage of the hierarchical representation is that each full conditional distribution is either conjugate or has a known form, enabling convenient MCMC Gibbs sampling.

1. **Full Conditional for β:**

Conditional on λ and σ^2, the posterior for β is Gaussian. Let $D_\lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$. Then:

$$p(\beta|\lambda, \sigma^2, y) \sim N\left(\left(X^\top X + D_\lambda^{-1}\right)^{-1} X^\top y,\, \sigma^2\left(X^\top X + D_\lambda^{-1}\right)^{-1}\right)$$

2. **Full Conditional for λ_j:**

For each j, given β_j and σ^2, the full conditional distribution of λ_j is Inverse-Gaussian.

$$\lambda_j|\beta_j, \sigma^2 \sim \text{Inverse-Gaussian}(\mu_j, \gamma),$$

where the parameters are:

$$\mu_j = \frac{\sqrt{\sigma^2}}{\lambda|\beta_j|}, \quad \gamma = \lambda^2.$$

The density is given by:

$$p(z) = \sqrt{\frac{\gamma}{2\pi z^3}} \exp\left(-\frac{\gamma(z-\mu)^2}{2\mu^2 z}\right), \quad z > 0.$$

3. **Full Conditional for σ^2:**

Given β and λ, the full conditional for σ^2 is Inverse-Gamma:

$$\sigma^2|\beta, \lambda, y \sim IG\left(a_0 + \frac{n+p}{2},\, b_0 + \frac{\|y - X\beta\|^2}{2} + \frac{1}{2}\sum_{j=1}^{p} \frac{\beta_j^2}{\lambda_j}\right).$$

This inverse-gamma distribution arises from the conjugacy of the normal likelihood and the scaled normal prior for σ^2.

4.1 Gibbs Sampler Algorithm

1. **Initialization**:

Initialize $\beta^{(0)}, \lambda^{(0)}, \sigma^{2(0)}$.

2. **Update β:**

Given $\lambda^{(t-1)}$ and $\sigma^{2(t-1)}$:

$$\beta^{(t)} \sim N\left(\left(X^\top X + D^{-1}_{\lambda^{(t-1)}}\right)^{-1} X^\top y,\, \sigma^{2(t-1)}\left(X^\top X + D^{-1}_{\lambda^{(t-1)}}\right)^{-1}\right).$$

3. **Update λ_j for each $j = 1, \ldots, p$:**

Given $\beta_j^{(t)}$ and $\sigma^{2(t-1)}$:

$$\lambda_j^{(t)} \sim \text{Inverse-Gaussian}\left(\frac{\sqrt{\sigma^{2(t-1)}}}{\lambda|\beta_j^{(t)}|},\, \lambda^2\right).$$

4. **Update σ^2:**

Given $\beta^{(t)}$ and $\lambda^{(t)}$:

$$\sigma^{2(t)} \sim IG\left(a_0 + \frac{n+p}{2},\, b_0 + \frac{\|y - X\beta^{(t)}\|^2}{2} + \sum_{j=1}^{p} \frac{\beta_j^{(t)^2}}{\lambda_j^{(t)}}\right).$$

5. **Iterate:**

Repeat steps 2-4 for $t = 1, \ldots, T$ iterations. After suitable burn-in (e.g., decided by traceplot), use the remaining samples to approximate the posterior distributions of β and σ^2.

4.2 Posterior Inference:

From the collected MCMC Gibbs samples $\beta^{(t)}, \sigma^{2(t)}$, we can compute posterior summaries, including posterior mean ($\hat{\beta}_m = \frac{1}{M}\sum_{t=1}^{M} \beta^{(t)}$ and credible intervals.

5. Simulated Example:

Next, we will run a simple simulation to compare the traditional (frequentist) Lasso, Lasso with bootstrapping for empirical uncertainty assessment, and the Bayesian Lasso in terms of point estimation accuracy, uncertainty qualification, and variable selection effectiveness.

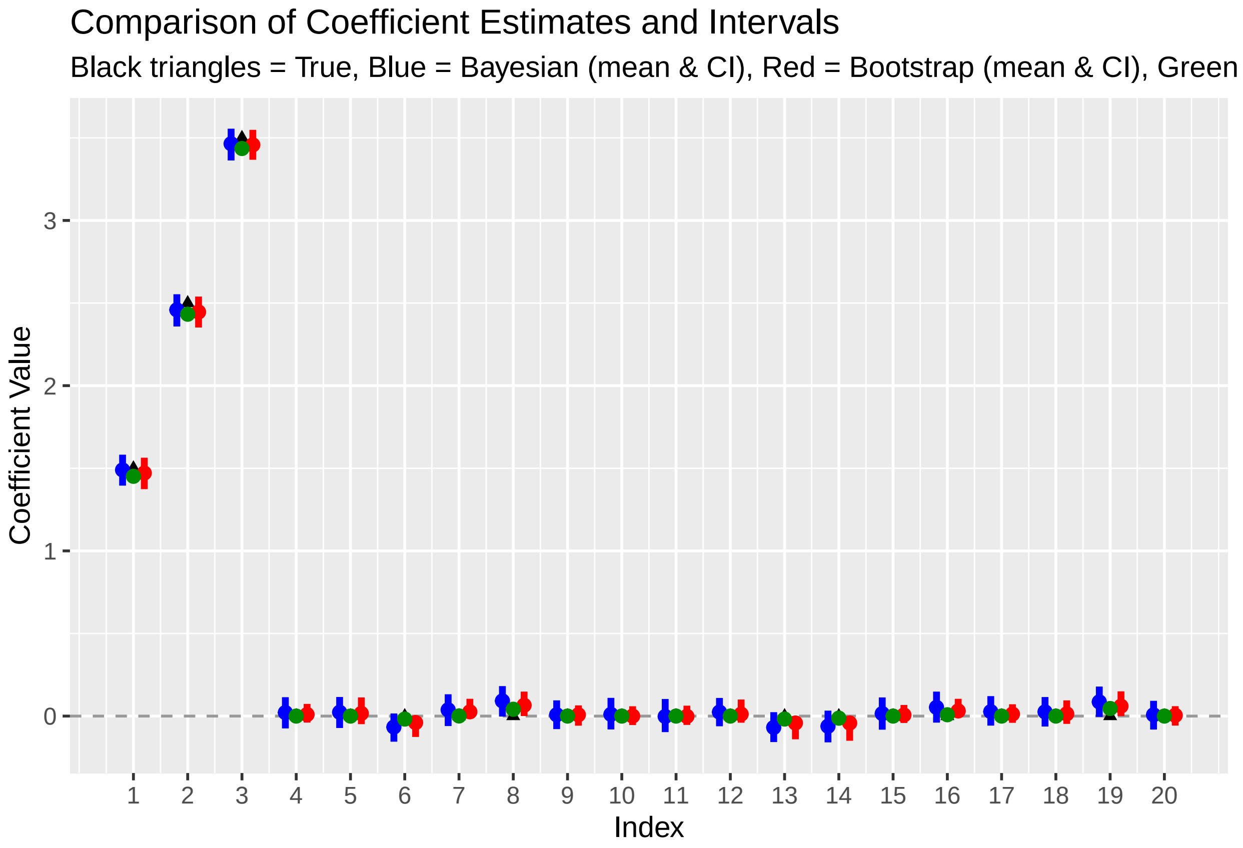
The set-up is as follows:

- Sample size: $n = 500$
- Number of predictors: $p = 20$

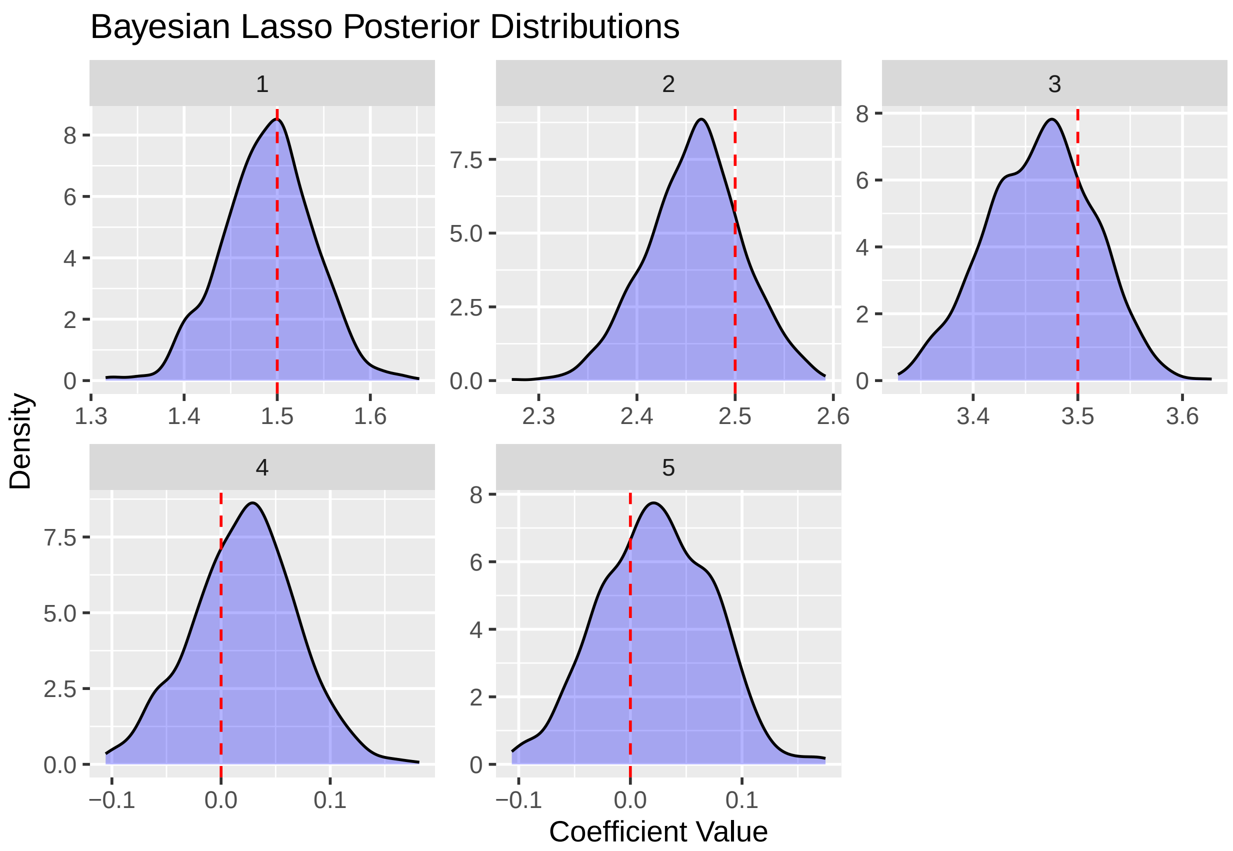
- True coefficients: $\beta^* = (1.5, 2.5, 3.5, 0, \dots, 0)$, with only three nonzero true values to simulate a sparse scenario.
- $X_{ij} \sim_{i.i.d} N(0,1)$
- $y = X\beta^* + \epsilon, \epsilon \sim_{i.i.d} N(0,1)$

The intention is to test the similarity in terms of performance between frequentist Lasso and Bayesian Lasso; and the differences between Lasso with Bootstrapping that yield empirical intervals and posterior credible intervals.

```
## Frequentist Lasso Estimates:

##  [1]  1.4513956897  2.4324151397  3.4355699619  0.0000000000  0.0000000000
##  [6] -0.0181063290  0.0008068596  0.0413408229  0.0000000000  0.0000000000
## [11]  0.0000000000  0.0000000000 -0.0179382157 -0.0125672986  0.0000000000
## [16]  0.0080060524  0.0000000000  0.0000000000  0.0458186691  0.0000000000

##
## Bootstrapped Lasso Means and 95% Intervals:

##              [,1]     [,2]     [,3]         [,4]        [,5]        [,6]
## mean     1.471133 2.446005 3.457501  0.009607898  0.01613904 -0.03935249
## ci_lower 1.373514 2.352196 3.367407 -0.039184316 -0.04886426 -0.12620823
## ci_upper 1.563989 2.539126 3.548582  0.073839278  0.11288932  0.00000000
##                 [,7]       [,8]         [,9]        [,10]        [,11]
## mean      0.02589403 0.06547572  0.007932695  2.642478e-05 -0.001745804
## ci_lower -0.01290876 0.00000000 -0.058153373 -5.425888e-02 -0.053470641
## ci_upper  0.10468195 0.14804731  0.064584540  5.947082e-02  0.062938355
##                [,12]       [,13]         [,14]       [,15]      [,16]
## mean      0.01161351 -0.04186976 -0.0421790768  0.00667215 0.03133238
## ci_lower -0.03995028 -0.14040612 -0.1492642160 -0.04159837 0.00000000
## ci_upper  0.10004003  0.00000000  0.0008926861  0.06717149 0.10452442
##                [,17]       [,18]      [,19]        [,20]
## mean      0.01217462  0.01348161 0.06011742  0.004675554
## ci_lower -0.04063421 -0.04715406 0.00000000 -0.057687255
## ci_upper  0.07150023  0.09498035 0.14959496  0.059895024

##
## Bayesian Lasso Posterior Means and 95% Credible Intervals:

##              [,1]     [,2]     [,3]        [,4]        [,5]        [,6]
## mean     1.489961 2.458935 3.464444  0.02139147  0.02305214 -0.06632575
## ci_lower 1.395016 2.358254 3.363268 -0.07458665 -0.07218371 -0.15487936
## ci_upper 1.582228 2.553421 3.555510  0.11428027  0.11560833  0.01576948
##                 [,7]         [,8]         [,9]       [,10]        [,11]
## mean      0.03774316  0.091826876  0.008084577  0.01041076 -0.002268714
## ci_lower -0.06027940 -0.002965315 -0.078762204 -0.08087050 -0.097045694
## ci_upper  0.13211690  0.181392104  0.095022368  0.11024738  0.103488751
##                [,12]       [,13]       [,14]       [,15]       [,16]
## mean      0.02418686 -0.06898059 -0.06154494  0.01520955  0.05280079
## ci_lower -0.06159162 -0.15712832 -0.15902972 -0.08227585 -0.03957666
## ci_upper  0.10982833  0.02381479  0.03293029  0.11291128  0.14802858
##                 [,17]       [,18]        [,19]        [,20]
## mean       0.02704109  0.02560291  0.086079654  0.007602938
## ci_lower -0.05760811 -0.06307989 -0.004802462 -0.081495506
## ci_upper  0.12069999  0.11518782  0.178723196  0.092350040

##
## MSE of Estimates:

## Lasso: 0.000788102

## Lasso (Bootstrap mean): 0.001056956

## Bayesian Lasso (Posterior mean): 0.001973984

##
## Variable Selection (qualitative):

## Lasso sets coefficients exactly to zero or not, Bayesian gives a probability.

## Number of exact zeros (Lasso): 10

## Bayesian CI containing zero (count): 17 out of 20 coefficients
```


Comparison of Coefficient Estimates and Intervals

Black triangles = True, Blue = Bayesian (mean & CI), Red = Bootstrap (mean & CI), Green



The frequentist Lasso estimates for the three nonzero coefficients (approximately 1.45, 2.43, and 3.44) are quite close to their true values (1.5, 2.5, and 3.5). The Bayesian Lasso's posterior mean estimates for these coefficients (around 1.49, 2.46, 3.46) are similarly close but slightly more shrunk towards zero. Both methods capture the true nonzero coefficients reasonably well.

The bootstrapped intervals are relatively tight and include values near the frequentist estimates. They offer an approximation of variability but are based on resampling and do not directly represent a "probability" that a coefficient is in a given range. Thus, interpreting these intervals are challenging, even though the resulting intervals are pretty similar.

On the other hand, the Bayesian Lasso provides credible intervals that have a direct probabilistic interpretation. For the true coefficients, the intervals do not include zero, confirming a high posterior probability that these coefficients are nonzero. For most of the irrelevant predictors, the Bayesian Lasso's intervals often include zero, reflecting uncertainty and the possibility that these coefficients may be effectively negligible.

The Lasso sets several coefficients exactly to zero, enforcing sparsity. In this run, it sets 10 out of 20 coefficients to zero. This is a binary selection that can be easy to interpret, but does not convey how close a "zero" coefficient might be to being included.

The Bayesian Lasso rarely produces exact zeros but shows posterior distributions heavily centered near zero for irrelevant predictors. Most of these variables have credible intervals that include zero, suggesting that the Bayesian model does not support these predictors as significant. This approach offers more insight into the model specifics by providing how plausible a coefficient differs from zero.

The provided MSE values show that in this scenario, the frequentist Lasso's point estimates achieve a slightly lower MSE than the Bayesian Lasso's posterior means. This can happen because the Bayesian method's shrinkage and uncertainty modeling may lead to more conservative estimates, sometimes increasing average squared error slightly. But, overall the two methods achieved very similar results.

In essence, the Bayesian approach offers more “interpretability” through its unique posterior distribution, while at the cost of greater computational demand. In terms of performance, the frequentist Lasso and Bayesian Lasso methods do not exhibit significant differences, and this alignment is expected to persist or even strengthen as the sample size increases, consistent with the Bernstein–von Mises theorem.

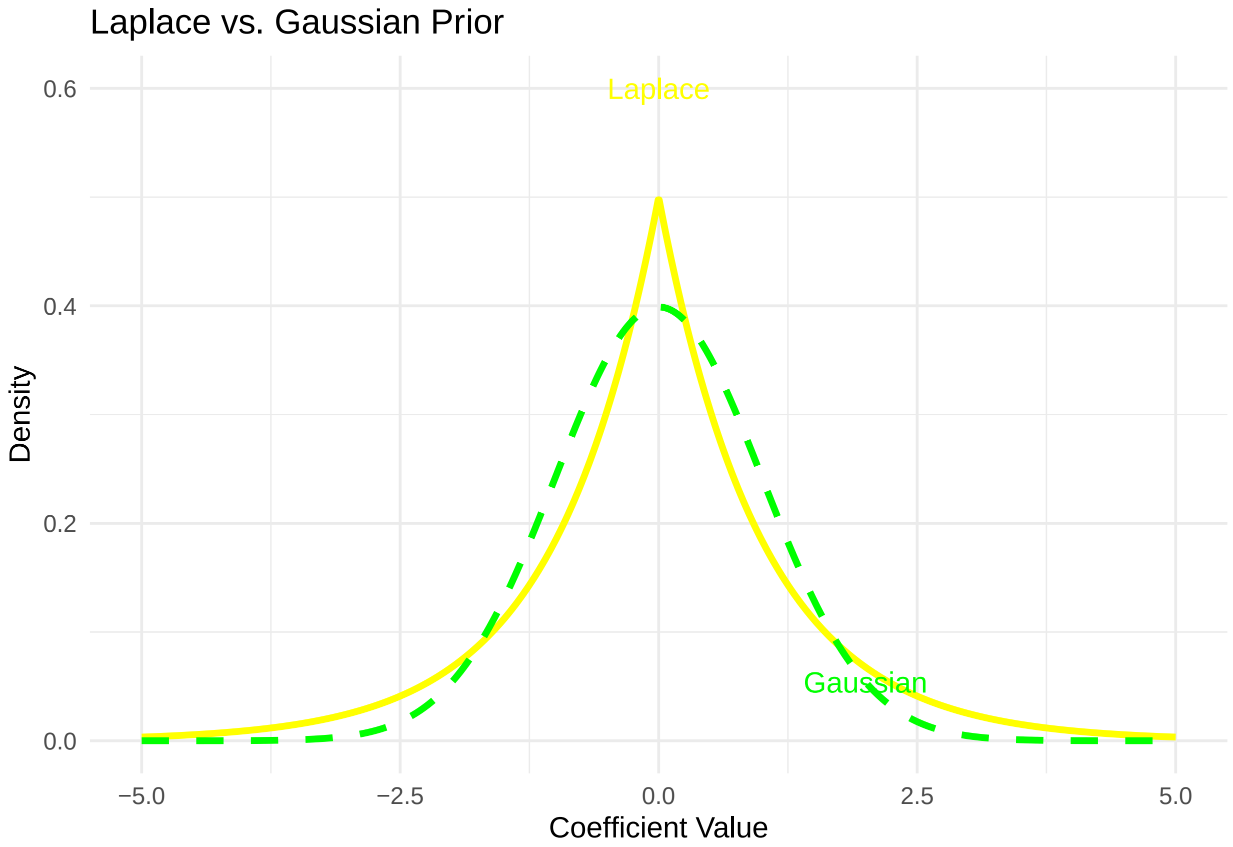
6. Discussion and Limitation:

The Bayesian Lasso provides a natural probabilistic interpretation of Lasso shrinkage by using Laplace priors. Unlike the frequentist Lasso, which offers only point estimates, the Bayesian Lasso delivers full posterior distributions, making credible intervals and probabilistic statements about coefficient values possible. This richer inference power can be especially valuable when assessing uncertainty about variable inclusion and magnitude.

This report focuses primarily on the theoretical foundation and a specific simulation scenario. It does not thoroughly explore the sensitivity of results to hyperparameter choices, investigate performance across a wide range of simulation seettings, or delve deeply into model diagnostics such as MCMC convergence checks. Moreover, the comparisons are illustrative rather than exhaustive, providing a conceptual understanding rather than a comprehensive empirical assessment of Bayesian Lasso under various complex modeling conditions.

Appendix A.

A Bayesian viewpoint starts by establishing beliefs about parameter values before seeing data. The Laplace prior places a very high density near zero, expressing a strong initial belief that coefficients should be small or zero. To deviate far from zero, the data must exhibit extreme evidence. This “sharp peak” at zero effectively shrinks weak effects towards zero, yielding a sparse solution where only truly necessary parameters become large.



Reference:

Chen, Y. (2021). STA521: Predictive Modelling and Statistical Learning, Lecture 9: Bayesian Regression I. Duke University. Retrieved from https://www2.stat.duke.edu/courses/Fall21/sta521.001/post/week05-1/main.pdf

Kyung, Minjung & Gill, Jeff & Ghosh, Malay & Casella, George. (2010). Penalized Regression, Standard Errors, and Bayesian Lassos. Bayesian Analysis. 5. 369-412. 10.1214/10-BA607.

Park, T., & Casella, G. (2008). The Bayesian Lasso. Journal of the American Statistical Association, 103(482), 681–686.

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