Bayesian Statistics — Assignment 2

On the Bayesian Lasso

Caio Peixoto

July 7, 2024

Contents

1 LASSO regression and the Laplace prior

In[1] this work we will study the usual linear regression model, given by

$$\boldsymbol{y} \mid \mu, \boldsymbol{\beta}, \mathbf{X}, \sigma^2 \sim \mathcal{N}(\mu \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n), \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{n \times p}, \boldsymbol{\beta} \in \mathbb{R}^p, \mu \in \mathbb{R}$, $\mathbf{1}_n$ denotes the vector in $\mathbb{R}^n$ with all coordinates set to 1 and $\mathbf{I}_n$ denotes the n-dimensional identity matrix. The goal is to conduct inference on $\mu, \boldsymbol{\beta}$ and σ^2, and we will follow the Bayesian approach in [PC08].

In [Tib96], the author presented an estimator for $\boldsymbol{\beta}$ and μ, called the LASSO[2], with the following definition:

$$(\hat{\boldsymbol{\beta}}_{\text{LASSO}}, \hat{\mu}_{\text{LASSO}}) = \underset{\|\boldsymbol{\beta}\|_1 \leq t, \mu \in \mathbb{R}}{\arg\min} \|\boldsymbol{y} - \mu \mathbf{1}_n - \mathbf{X}\boldsymbol{\beta}\|_2^2, \tag{2}$$

where $t > 0$ is a tuning parameter. His goal was to address two of the main problems of the usual OLS estimator, which does not include the L^1 norm constraint: *high variance*, which lowers prediction accuracy, and *lack of interpretability*, particularly when the number of predictors is high. The effect produced by the L^1 constraint is the shrinkage of the overall coefficient magnitude and, more importantly, the collapse of the less important coefficients to exactly 0. This effect introduces bias in the estimation, but the gain in variance is enough to improve predictive performance. Furthermore, it provides a principled way to perform variable selection, by discarding the variables whose coefficients were collapsed to 0.

One might wonder what is special about the L^1 norm, since this collapsing behavior is not observed for the Ridge estimator, which uses L^2 constraint instead of L^1. It turns out the reason is geometric. The L^1 ball in $\mathbb{R}^p$ has sharp edges and vertices, in contrast with the smooth surface of the L^2 ball. When solving problem (2), the solution will often be located on an edge of the L^1 ball of radius t. Along these edges, one or more coefficients are exactly zero, which produces the collapsing effect.

Although the methodology in [Tib96] is classical, in Section 5 the authors point out that problem (2) is equivalent to MAP estimation when employing a Laplace prior. To see this, note that the constrained optimization problem is equivalent to the following regularized version:

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^p, \mu \in \mathbb{R}}{\arg\min} \|\boldsymbol{y} - \mu \mathbf{1}_n - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1,$$

[1] Corresponds to items a) and b).

[2] *Least absolute shrinkage selection operator.*

where λ depends on t. Much like the squared Euclidean norm in the objective corresponds to the normal likelihood, the L^1 norm can be seen as the log of the kernel of the Laplace distribution. Hence, the LASSO is the MAP estimator when the prior for $\boldsymbol{\beta}$ is chosen as

$$\pi(\boldsymbol{\beta}) = \left(\frac{\lambda}{2}\right)^n \exp\left\{-\lambda\|\boldsymbol{\beta}\|_1\right\}.$$

Expanding upon this observation, in [PC08] the authors discuss a Bayesian formulation for LASSO regression by employing the following joint prior on $\boldsymbol{\beta}$ and σ:

$$\begin{aligned} \pi(\sigma^2) &\propto \frac{1}{\sigma^2} \cdot \mathbb{I}\{\sigma^2 > 0\}, \\ \pi(\boldsymbol{\beta} \mid \sigma^2) &= \left(\frac{\lambda}{2\sqrt{\sigma^2}}\right)^n \exp\left\{-\frac{\lambda}{\sqrt{\sigma^2}}\|\boldsymbol{\beta}\|_1\right\}, \end{aligned} \tag{3}$$

where λ is a hyperparameter which can be given a hyperprior or be estimated using marginal maximum likelihood methods. The choice for a Laplace prior for $\boldsymbol{\beta}$ is necessary to recover the LASSO estimator as the MAP. With a normal prior, for instance, the MAP would become the Ridge regression estimator. This conditional structure in the prior is justified in [PC08] by a computational argument: without it, the posterior may exhibit multimodality, which can pose difficulties for MCMC sampling schemes such as the Gibbs sampler. From a statistical perspective, this prior encapsulates the idea that the larger σ^2 is, the less informative the prior on $\boldsymbol{\beta}$ should be. This makes *some* sense, since, for more spread out y values, we would like the prior not to be as restrictive and not to "get in the way" of the likelihood. On another note, it would also be reasonable not to assume $\boldsymbol{\beta}$ and σ^2 depend on each other in any way a *priori*. Anyway, since we are following the methodology of [PC08], our prior of choice will be (3).

The main motivation for this Bayesian formulation is the already mentioned fact that the LASSO estimator is recovered as the MAP. However, Bayesian estimation goes far beyond maximizing the posterior, and one might wonder if other forms of inference based on the prior (3) exhibit shrinkage-like properties. This, among other things, is what we will investigate in the following sections. Apart from the usual advantages of having a full probability distribution describing our belief about the parameters, it is not clear if this Bayesian approach does what we might desire it to do, e.g., produce credibility intervals which include 0 for the irrelevant covariates and exclude it for the relevant ones. In fact, there are statisticians which assert that the Bayesian LASSO simply *does not work*, as can be seen, for instance, in [SS21]. Simpson's argument is that the prior (3) either shrinks *almost all* parameters to 0, or does not produce *any shrinkage at all*.

2 The Gibbs sampler

In[3] [PC08], the authors propose a Gibbs sampler to draw samples from the posterior. This sampler uses the following decomposition of the Laplace distribution: if $s \sim \text{Exp}(a^2/2)$[4] and $z \mid s \sim \mathcal{N}(0, s)$, then, marginally, $z \sim \mathcal{L}(a)$, that is,

$$p(z) = \frac{a}{2}\exp\{-a|z|\}.$$

Using this decomposition, one may rewrite the model given by (1) and (3) with auxiliary variables $\tau_1^2, \ldots, \tau_p^2$:

$$\begin{aligned} \boldsymbol{y} \mid \mu, \boldsymbol{\beta}, \mathbf{X}, \sigma^2 &\sim \mathcal{N}(\mu\mathbf{1}_n + \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n), \\ \boldsymbol{\beta} \mid \sigma^2, \tau_1^2, \ldots, \tau_p^2 &\sim \mathcal{N}(\mathbf{0}_p, \sigma^2\mathbf{D}_\tau), \\ \mathbf{D}_\tau &= \text{diag}(\tau_1^2, \ldots, \tau_p^2), \\ \tau_i^2 &\overset{\text{iid}}{\sim} \text{Exp}(\lambda^2/2), \end{aligned}$$

and σ^2 is distributed independently of $\tau_{1:p}^2$ according to a distribution π with positive support, while μ is given an independent, flat, prior. Park and Casella chose to use $\pi \propto \frac{1}{\sigma^2}\mathbb{I}\{\sigma^2 > 0\}$, but, as they mentioned

[3] Corresponds to items c) and d).

[4] This is the *rate* parameter

in [PC08], any inverse-gamma prior could be used without affecting conjugacy. This is a way to introduce subjective knowledge about σ^2 to the prior. It is necessary to introduce the auxiliary variables $\tau_1^2, \ldots, \tau_p^2$ because the L^1 norm which appears inside the exponential in the Laplace distribution makes the full conditional of $\boldsymbol{\beta}$ hard to sample from. With this new hierarchical structure, the full conditional of $\boldsymbol{\beta}$ is multivariate normal, while that of σ^2 and $1/\tau_i^2$ are inverse gamma and inverse Gaussian, respectively. Of course, this is only needed if one is interested in fully exploring the posterior distribution. If all one desires is to compute the MAP, then it is best to go back to the prior in (3) and use any available package to obtain the LASSO estimator.

The intercept parameter, μ, is only related to the location of $\boldsymbol{y}$, and, thus, does not contain any information on which covariates are relevant and which are not. If the goal of the data analyst is only to analyze the effect of each covariate on the outcome, there is no reason to conduct inference on (and, hence, sample from) the intercept μ. It can be easily marginalized from the joint distribution of data and parameters, and this was the rout taken in [PC08]. The authors mention that, should one desire to sample from μ, its full conditional is normal with mean $\bar{\boldsymbol{y}}$ and variance σ^2/n.

3 But can it actually shrink?

In[5][6] this section, we will investigate the shrinking properties of the Bayesian LASSO using the artificial data scenarios in presented in the Section 7.2 of [Tib96]. All the scenarios consist of generating n samples from the model in (1), where the lines of $\mathbf{X}$ are independent samples from a multivariate normal distribution. We considered two sample sizes: $n = 20$ and $n = 100$. For the prior distribution, we employed the prior in (3), with an additional Gamma(δ, r) prior[7] on λ^2 (not λ), as suggested in Section 3.2 of [PC08]. We specified a weakly-informative prior by choosing $\delta = 1.0$ and $r = 0.1$. The sampling scheme we chose was NUTS-HMC, utilized through `Stan` [Sta24] and `CmdStanPy`. For each dataset, we sampled from 5 chains, using 1000 warm-up samples and 10000 samples from (hopefully) the posterior. On a last note, we also standardized the columns of $\mathbf{X}$ to have mean 0 and variance 1. This is a good practice also employed in [PC08], as it puts all covariates in the same scale. If one of the betas was much smaller than the other ones simply because of the larger scale of the corresponding covariate, then it might be incorrectly shrunk to 0 by the LASSO regression.

Scenario 1

For this scenario, the parameters were

$$\begin{aligned}\boldsymbol{\beta} &= (3, 1.5, 0, 0, 2, 0, 0, 0)\\ \sigma &= 3.\end{aligned}$$

The correlation between x_i and x_j was set to $\rho^{|i-j|}$, with $\rho = 0.5$. All $(x_i)_{1\leq i\leq 8}$ had mean 0.

Effective sample size and $\hat{R}$ statistics are in Table 1. We can see that, for both values of n, the $\hat{R}$ values are very close to one, indicating that we are indeed sampling from the posterior. The effective sample sizes are also quite high, which allows us to trust the credibility intervals computed from these samples. In order to further assess chain mixing, we show trace plots of the parameters in Figure 1. From them, we can see that the chains are generally well mixed for both dataset sizes. As a final diagnostic, we also conducted a posterior predictive check, whose results are in Figure 2. To do this, for each sample from the posterior, we sampled $\boldsymbol{y}$ according to (1) and plotted a histogram of the coordinates. The actual data is included for comparison. We can see that there is little resemblance between actual and simulated data for $n = 20$, but the situation is better for $n = 100$. This can be due to the large variance ($\sigma^2 = 9$) in the data generating process, which makes it hard to extract information from datasets with few samples.

To assess shrinkage, we computed 95% credibility intervals using the 2.5% and 97.5% quantiles of the 50000 posterior samples. We interpret that a parameter should be set to 0 when the corresponding 95% credibility interval contains 0. These are shown in Figure 3. For comparison, we also plotted the

[5] Corresponds to item e).

[6] All code for the experiments in Sections 3 and 4 is openly available in Github

[7] Here, δ and r are the *shape and rate* parameters, respectively. This is the *opposite* of what is done in [PC08], however, I realized this mistake far too late to fix it. Fortunately, the only place where this might be annoying is in the next section.

$n = 20$			$n = 100$		
	N_Eff	R_hat		N_Eff	R_hat
lp___	15200.000000	1.000000	lp___	19470.000000	1.000000
mu	47660.000000	1.000000	mu	54950.000000	1.000000
beta[1]	31305.000000	1.000000	beta[1]	41583.000000	1.000000
beta[2]	30485.000000	1.000000	beta[2]	38976.000000	1.000000
beta[3]	33833.000000	1.000000	beta[3]	32296.000000	1.000000
beta[4]	39068.000000	1.000000	beta[4]	35541.000000	0.999900
beta[5]	36640.000000	0.999900	beta[5]	36941.000000	1.000000
beta[6]	35091.000000	1.000000	beta[6]	34982.000000	1.000000
beta[7]	33731.000000	1.000000	beta[7]	31183.000000	1.000000
beta[8]	35182.000000	1.000000	beta[8]	33972.000000	1.000000
sigma	32499.237400	1.000100	sigma	45289.799000	1.000000
lambda	36077.141000	1.000000	lambda	44630.841000	1.000000

Table 1: Effective sample sizes and $\hat{R}$ statistics for Scenario 1, reported with four significant digits.

true $\boldsymbol{\beta}_i$ values, as well as the estimate produced by the classical LASSO when λ is chosen through cross validation.

We can see that, for the smaller sample size, the Bayesian LASSO correctly produced credibility intervals which are centered around 0 for the $\boldsymbol{\beta}_i$'s which were actually zero. However, it had difficulty producing credibility intervals which clearly put the non-zero parameters away from zero, getting $\boldsymbol{\beta}_5$ completely wrong, for example. On a positive note for Bayesians, the classical LASSO was guilty of the same mistakes, since its estimates always fell within the $(25\%, 75\%)$ credibility interval of the Bayesian LASSO. For the larger sample size, however, the Bayesian LASSO showed a significant improvement. It managed to correctly separate the null parameters from the non-null ones, while the classical LASSO failed to set any parameter to 0.

Scenario 2

For this scenario, the parameters were

$$\begin{aligned}\boldsymbol{\beta} &= (0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85)\\ \sigma &= 3.\end{aligned}$$

The distribution of $\mathbf{X}$ was the same as in Scenario 1. Effective sample sizes and $\hat{R}$ statistics are in Table 2, while trace plots, posterior predictive checks and 95% credibility intervals are in Figures 4, 5 and 6, respectively. Once again, effective sample sizes and $\hat{R}$ statistics are very healthy, as well as the trace plots. The posterior predictive distribution for $\boldsymbol{y}$ were again misaligned with the data for $n = 20$, but were much more faithful for $n = 100$. This scenario differs from Scenario 1 in that *no true coefficients were* 0. This creates an awkward situation, in which, for $n = 20$, both the classical LASSO and the Bayesian LASSO shrunk all coefficients to 0, although it should be noted that all but two Bayesian LASSO credibility intervals correctly included the true $\boldsymbol{\beta}_i$ value. When $n = 100$, however, the situation changes. Four Bayesian LASSO credibility intervals correctly did not shrink the parameter to zero and included the true parameter value. The other four wrongly included 0, but just barely. Once again, for both sample sizes the classical LASSO estimates tended to fall within the $(25\%, 75\%)$ credibility interval of the Bayesian LASSO.

Scenario 3

For this scenario, the parameters were

$$\begin{aligned}\boldsymbol{\beta} &= (5, 0, 0, 0, 0, 0, 0, 0)\\ \sigma &= 2.\end{aligned}$$

	$n = 20$			$n = 100$	
	N_Eff	R_hat		N_Eff	R_hat
lp___	13250.000000	1.000000	lp___	22100.000000	1.000000
mu	46830.000000	1.000000	mu	53480.000000	0.999900
beta[1]	29538.000000	1.000000	beta[1]	42594.000000	1.000000
beta[2]	30893.000000	1.000000	beta[2]	38128.000000	1.000000
beta[3]	32954.000000	1.000000	beta[3]	38743.000000	1.000000
beta[4]	30811.000000	1.000000	beta[4]	33994.000000	1.000000
beta[5]	32159.000000	1.000000	beta[5]	44015.000000	1.000000
beta[6]	33505.000000	1.000000	beta[6]	35119.000000	1.000000
beta[7]	28825.000000	1.000000	beta[7]	35445.000000	1.000000
beta[8]	29206.000000	1.000000	beta[8]	39788.000000	1.000000
sigma	34513.738000	1.000000	sigma	46622.085200	1.000100
lambda	32343.862000	1.000000	lambda	52132.501200	0.999900

Table 2: Effective sample sizes and $\hat{R}$ statistics for Scenario 2, reported with four significant digits.

The distribution of $\mathbf{X}$ was the same as in Scenarios 1 and 2. Effective sample sizes and $\hat{R}$ statistics are in Table 3, while trace plots, posterior predictive checks and 95% credibility intervals are in Figures 7, 8 and 9, respectively. The comments about effective sample size, $\hat{R}$ statistics, trace plots and posterior predictive checks made previously also apply here.

	$n = 20$			$n = 100$	
	N_Eff	R_hat		N_Eff	R_hat
lp___	15020.000000	1.001000	lp___	16390.000000	1.000000
mu	45560.000000	1.000000	mu	49930.000000	0.999900
beta[1]	29434.000000	1.000000	beta[1]	45798.000000	1.000000
beta[2]	33079.000000	1.000000	beta[2]	39123.000000	1.000000
beta[3]	39256.000000	1.000000	beta[3]	44796.000000	1.000000
beta[4]	40002.000000	1.000000	beta[4]	40713.000000	1.000000
beta[5]	40608.000000	1.000000	beta[5]	46209.000000	1.000000
beta[6]	35598.000000	1.000000	beta[6]	41420.000000	1.000000
beta[7]	41152.000000	1.000000	beta[7]	43684.000000	1.000000
beta[8]	34751.000000	1.000000	beta[8]	42667.000000	0.999900
sigma	28577.832000	1.000000	sigma	46355.200300	1.000100
lambda	32480.464000	1.000000	lambda	46537.422700	1.000000

Table 3: Effective sample sizes and $\hat{R}$ statistics for Scenario 3, reported with four significant digits.

This scenario is much more friendly for sparse regression methods, since the variance of $\boldsymbol{y}$ is smaller and almost all coefficients are exactly 0, except for $\boldsymbol{\beta}_1$, which is considerably large, given the scale of $\mathbf{X}$. This is apparent in the credibility intervals, which managed to always include the true parameter value, as well as shrink only the parameters which were actually 0. The increase in sample size only decreased the size of the intervals. The classical LASSO also had a good performance, staying very close to the median of the Bayesian LASSO.

Scenario 4

For this scenario, the parameters were

$$\boldsymbol{\beta} = (0, 0, \ldots, 0, 2, 2, \ldots, 2, 0, 0, \ldots, 0, 2, 2, \ldots, 2)$$
$$\sigma = 15,$$

where each block in $\boldsymbol{\beta}$ has size 10, meaning the total number of covariates is $p = 40$. The correlation between different predictors was set to 0.5. Since this model is much larger than the other three, the

sample sizes we analyzed were $n = 100$ and $n = 500$. Effective sample sizes and $\hat{R}$ statistics are in Table 4, while posterior predictive checks and 95% credibility intervals are in Figures 10 and 11, respectively. Given the number of covariates, we did not analyze trace plots. The comments about effective sample size, $\hat{R}$ statistics and posterior predictive checks made previously also apply here. While all diagnostics

	$n = 100$			$n = 500$	
	N_Eff	R_hat		N_Eff	R_hat
lp___	14800.000000	1.000000	lp___	17720.000000	1.000000
mu	46810.000000	1.000000	mu	39490.000000	1.000000
beta[1]	28299.000000	1.000000	beta[1]	35021.000000	1.000000
beta[2]	36454.000000	1.000000	beta[2]	34017.000000	1.000000
beta[3]	26406.000000	1.000000	beta[3]	34293.000000	1.000000
beta[4]	31612.000000	1.000000	beta[4]	32462.000000	1.000000
beta[5]	32506.000000	1.000000	beta[5]	30194.000000	1.000000
beta[6]	33365.000000	1.000000	beta[6]	29062.000000	1.000000
beta[7]	31769.000000	1.000000	beta[7]	27408.000000	1.000000
beta[8]	32778.000000	1.000000	beta[8]	27918.000000	1.000000
beta[9]	32183.000000	1.000000	beta[9]	26745.000000	1.000000
beta[10]	28522.000000	1.000000	beta[10]	24627.000000	1.000000
beta[11]	35286.000000	1.000000	beta[11]	37007.000000	1.000000
beta[12]	30690.000000	1.000000	beta[12]	37694.000000	1.000000
beta[13]	36391.000000	1.000000	beta[13]	40944.000000	1.000000
beta[14]	32888.000000	1.000000	beta[14]	33475.000000	1.000000
beta[15]	33140.000000	1.000000	beta[15]	32878.000000	1.000000
beta[16]	30367.000000	1.000000	beta[16]	37431.000000	1.000000
beta[17]	29747.000000	0.999900	beta[17]	38631.000000	1.000000
beta[18]	40266.000000	1.000000	beta[18]	36876.000000	1.000000
beta[19]	31150.000000	1.000000	beta[19]	39417.000000	1.000000
beta[20]	36288.000000	0.999900	beta[20]	36417.000000	1.000000
beta[21]	41201.000000	1.000000	beta[21]	26103.000000	1.000000
beta[22]	30010.000000	1.000000	beta[22]	37153.000000	1.000000
beta[23]	31646.000000	1.000000	beta[23]	28222.000000	1.000000
beta[24]	29263.000000	1.000000	beta[24]	27295.000000	1.000000
beta[25]	37481.000000	1.000000	beta[25]	28212.000000	1.000000
beta[26]	36954.000000	1.000000	beta[26]	33869.000000	1.000000
beta[27]	35958.000000	1.000000	beta[27]	31385.000000	1.000000
beta[28]	30299.000000	1.000000	beta[28]	24276.000000	1.000000
beta[29]	32467.000000	1.000000	beta[29]	36815.000000	1.000000
beta[30]	37781.000000	1.000000	beta[30]	26786.000000	1.000000
beta[31]	34519.000000	1.000000	beta[31]	36241.000000	1.000000
beta[32]	30874.000000	1.000000	beta[32]	36215.000000	1.000000
beta[33]	32701.000000	1.000000	beta[33]	32350.000000	1.000000
beta[34]	33755.000000	1.000000	beta[34]	28491.000000	1.000000
beta[35]	32160.000000	1.000000	beta[35]	38596.000000	1.000000
beta[36]	29151.000000	1.000000	beta[36]	26368.000000	1.000000
beta[37]	35907.000000	1.000000	beta[37]	28086.000000	1.000000
beta[38]	37879.000000	1.000000	beta[38]	37653.000000	1.000000
beta[39]	30782.000000	1.000000	beta[39]	38766.000000	1.000000
beta[40]	28028.000000	1.000000	beta[40]	37611.000000	1.000000
sigma	36537.714000	1.000000	sigma	39951.620600	1.000100
lambda	35204.500000	1.000000	lambda	42298.545900	1.000100

Table 4: Effective sample sizes and $\hat{R}$ statistics for Scenario 4, reported with four significant digits.

employed indicate our sampling scheme is performing as intended, the Bayesian LASSO faced great

difficulty in this scenario, collapsing all but two covariates to 0 in the $n = 100$ regime. In this sample size, the classical LASSO showed a better performance in terms of shrinkage, managing to correctly not collapse 14 non-zero parameters, in spite of failing to collapse 2 actually null ones. When n was increased to 500, the credibility intervals for the parameters equal to 2 generally shifted towards the positive side, but this was still not enough to correctly decide which parameters are actually 0. These results show that the Bayesian LASSO may face difficulties when p is large and the proportion of non-null parameters is high.

4 Choosing λ

In[8] [PC08], two approaches are presented to "choose" the LASSO parameter λ. The first one is based on an EM algorithm to choose the λ value which maximizes the marginal likelihood of the data. The second one is arguably more suitable for a Bayesian analysis[9] and consists of simply incorporating λ as an additional pararameter. As was seen in the previous section, we chose to use the second option for our experiments. To maintain conjugacy, one should specify a Gamma prior on λ^2 [PC08]. However, the choice of hyperparameters for this prior in [PC08] was based on a desired relationship between the prior mean and the MLE. This makes little sense in the Bayesian context, as the prior should not be chosen based on information gained from the data values.

To analyze what effects the choice of the λ hyperprior hyperparameters has on the final estimate, we applied the Bayesian Lasso to the `diabetes` dataset of [Efr+04] using three combinations of hyperparameters. The hyperprior was $\lambda^2 \sim \text{Gamma}(\delta, r)$. The computational tools used to fit the models were the same ones described in Section 3.

- $\delta = 1, r = 1.78$: These are the hyperparameters suggested in [PC08]. Summary statistics are presented in Table 5 and credibility intervals can be visualized in Figure 12. Interestingly enough, our 95%, equal-tailed credibility interval for λ was $(10.06, 11.498)$, completely different from the one reported in [PC08] of $(0.139, 0.486)$. Since our $\hat{R}$ values are all close to 1 and the effective sample sizes surpass 25k for all variables, we have no reason to doubt our results.

	2.5\%	50\%	97.5\%	N_Eff	R_hat
lp___	-2850.000000	-2843.000000	-2839.000000	20490.000000	1.000000
mu	148.300000	152.100000	155.600000	64750.000000	1.000000
beta[1]	-3.628000	-0.091650	3.040000	62712.000000	1.000000
beta[2]	-13.730000	-9.503000	-5.623000	58193.000000	1.000000
beta[3]	20.230000	24.950000	29.180000	58450.000000	1.000000
beta[4]	9.721000	14.260000	18.440000	58566.000000	1.000000
beta[5]	-15.180000	-5.186000	1.582000	23732.000000	1.000000
beta[6]	-8.897000	-1.252000	4.845000	30117.000000	1.000000
beta[7]	-15.730000	-8.760000	-1.482000	26419.000000	1.000000
beta[8]	-3.072000	2.991000	11.270000	31179.000000	1.000000
beta[9]	17.840000	23.810000	29.430000	37253.000000	1.000000
beta[10]	-1.067000	2.696000	6.847000	56814.000000	1.000000
sigma	36.946300	38.781400	40.606400	74059.453500	0.999900
lambda	10.059600	10.798400	11.498100	70911.294400	1.000000

Table 5: Summary statistics for the hyperparameters $\delta = 1$ and $r = 1.78$.

- $\delta = 1.0, r = 0.1$: These are the hyperparameter values we used in Section 3. Summary statistics are presented in Table 6 and credibility intervals can be visualized in Figure 13. The 95% credibility interval we obtain for λ using these values is $(39.54, 45.54)$. This is to be expected since the prior is way more diffuse and assigns more probability for larger values of λ.

[8] Corresponds to item f).

[9] And is considerably easier to implement using `Stan` 😅.

	2.5\%	50\%	97.5\%	N_Eff	R_hat
lp___	-2282.000000	-2275.000000	-2271.000000	18320.000000	1.000000
mu	148.100000	152.100000	155.800000	48840.000000	1.000000
beta[1]	-2.120000	0.044360	2.117000	38068.000000	1.000000
beta[2]	-8.704000	-4.295000	-0.559700	34044.000000	1.000000
beta[3]	19.480000	24.270000	28.660000	35405.000000	1.000000
beta[4]	6.507000	11.250000	15.480000	33648.000000	1.000000
beta[5]	-5.095000	-0.849800	1.059000	21870.000000	1.000000
beta[6]	-4.232000	-0.699300	1.133000	27689.000000	1.000000
beta[7]	-12.620000	-7.433000	-2.497000	25298.000000	1.000000
beta[8]	-1.333000	0.866200	5.211000	21181.000000	1.000000
beta[9]	16.220000	21.450000	26.290000	33718.000000	1.000000
beta[10]	-0.778000	1.429000	5.136000	31419.000000	1.000000
sigma	38.602000	40.575000	42.532000	41530.225000	1.000000
lambda	39.537000	42.611000	45.542000	44663.473000	1.000000

Table 6: Summary statistics for the hyperparameters $\delta = 1$ and $r = 0.1$.

- $\delta = 1.0, r = 10.0$: These hyperparameter values correspond to an informative prior which puts λ^2 close to zero, meaning the shrinking effect is smaller. Summary statistics are presented in Table 7 and credibility intervals can be visualized in Figure 14. The 95% credibility interval we obtain for λ using these values is $(4.31, 4.63)$.

	2.5\%	50\%	97.5\%	N_Eff	R_hat
lp___	-3216.000000	-3210.000000	-3206.000000	21540.000000	1.000000
mu	148.400000	152.100000	155.600000	58280.000000	1.000000
beta[1]	-4.074000	-0.203500	3.227000	54814.000000	1.000000
beta[2]	-14.770000	-10.560000	-6.703000	55482.000000	0.999900
beta[3]	20.400000	25.000000	29.200000	51029.000000	1.000000
beta[4]	10.390000	14.880000	19.010000	53606.000000	1.000000
beta[5]	-27.930000	-10.440000	1.294000	20926.000000	1.000000
beta[6]	-9.521000	1.161000	14.490000	23264.000000	1.000000
beta[7]	-15.700000	-6.624000	1.620000	25770.000000	1.000000
beta[8]	-3.622000	4.713000	13.680000	38928.000000	1.000000
beta[9]	18.560000	25.730000	33.190000	26433.000000	1.000000
beta[10]	-1.140000	3.042000	7.202000	53886.000000	1.000000
sigma	36.592800	38.420700	40.227500	55799.672500	1.000000
lambda	4.311800	4.632000	4.928400	57694.368400	0.999900

Table 7: Summary statistics for the hyperparameters $\delta = 1$ and $r = 10$.

It is worth noting that the credibility intervals did not change considerably while varying the hyperparameter values. In all three cases, only the parameters $\boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\beta}_4$ and $\boldsymbol{\beta}_9$ were not collapsed to 0, while the median of the other parameters was similar across hyperparameter values. This indicates that the model is robust to the choice of prior, which is a desirable feature.

References

[Efr+04] Bradley Efron et al. "Least Angle Regression". In: *The Annals of Statistics* 32.2 (2004), pp. 407–451. ISSN: 00905364. URL: `http://www.jstor.org/stable/3448465` (visited on 07/07/2024).

[PC08] Trevor Park and George Casella. "The Bayesian Lasso". In: *Journal of the American Statistical Association* 103.482 (2008), pp. 681–686. DOI: 10.1198/016214508000000337. eprint: https://doi.org/10.1198/016214508000000337. URL: https://doi.org/10.1198/016214508000000337.

[SS21] Dan Simpson and Dan Simpson. *The King Must Die (Repost)*. Dec. 8, 2021. URL: https://dansblog.netlify.app/2021-12-08-the-king-must-die-repost.

[Sta24] Stan Development Team. *Stan Modeling Language Users Guide and Reference Manual, Version 2.35*. 2024. URL: http://mc-stan.org/.

[Tib96] Robert Tibshirani. "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1 (1996), pp. 267–288. ISSN: 00359246. URL: http://www.jstor.org/stable/2346178 (visited on 07/05/2024).

List of Figures

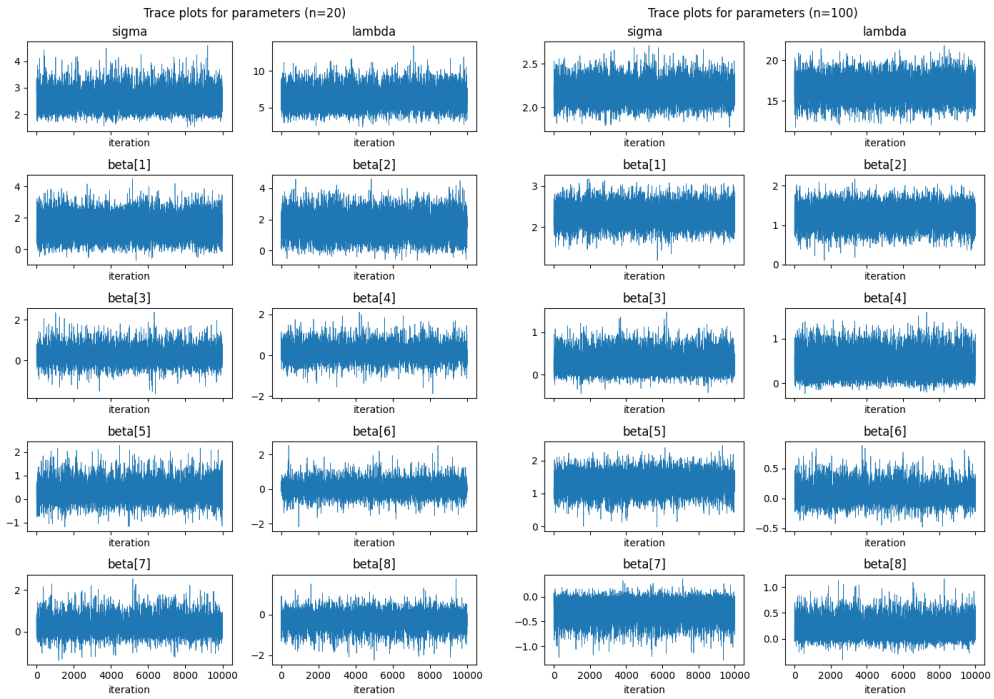

Figure 1: Trace plots for the parameters in Scenario 1. Each plot is from one of the 5 chains, chosen randomly and uniformly.

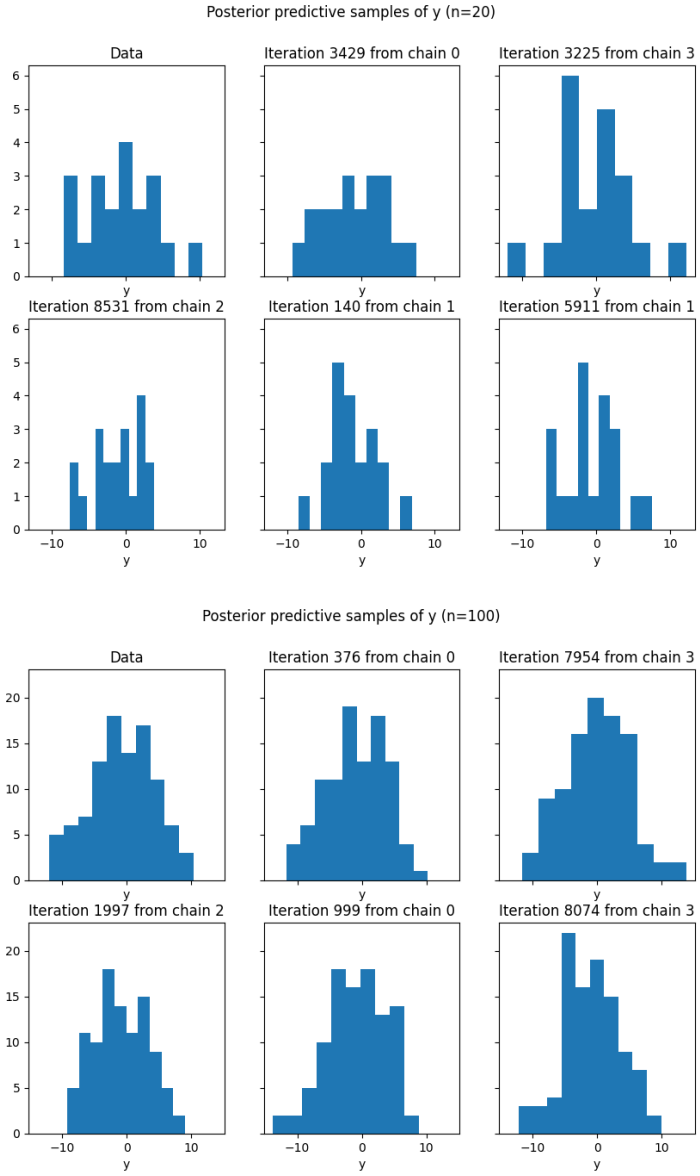

Figure 2: Posterior predictive checks for Scenario 1. The parameters corresponding to these 5 (for each n) simulated values from $\boldsymbol{y}$ were randomly and uniformly chosen across all samples from the posterior.

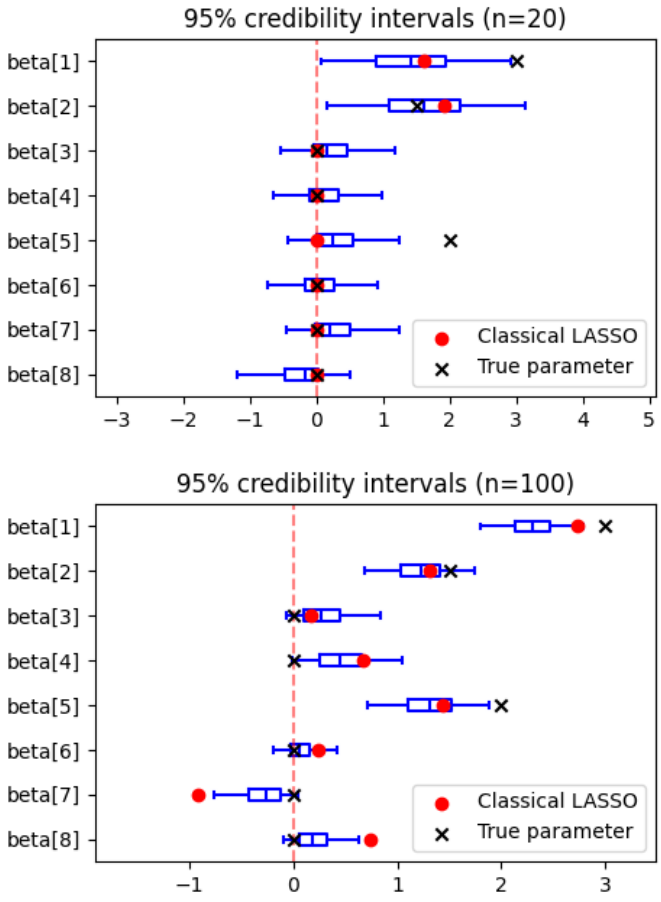

Figure 3: 95% credibility intervals for Scenario 1. The red dots are the classical LASSO's predictions when fit using cross validation, and the black crosses are the true parameter values. The vertical red line corresponds to $x = 0$.

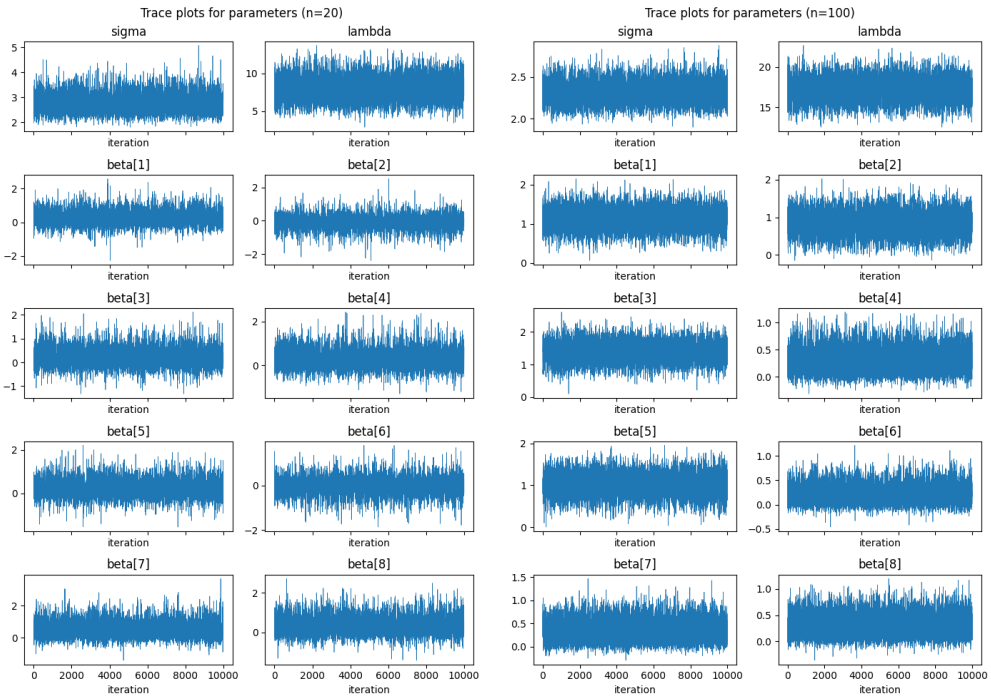

Figure 4: Trace plots for the parameters in Scenario 2. Each plot is from one of the 5 chains, chosen randomly and uniformly.

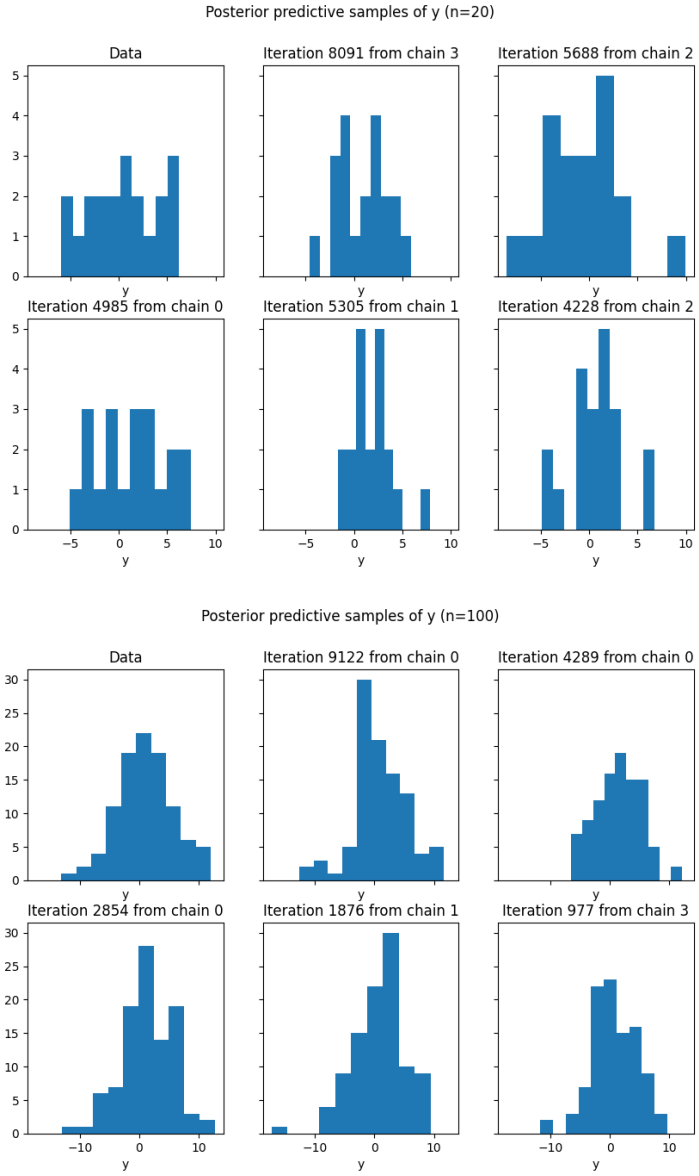

Figure 5: Posterior predictive checks for Scenario 2. The parameters corresponding to these 5 (for each n) simulated values from $\boldsymbol{y}$ were randomly and uniformly chosen across all samples from the posterior.

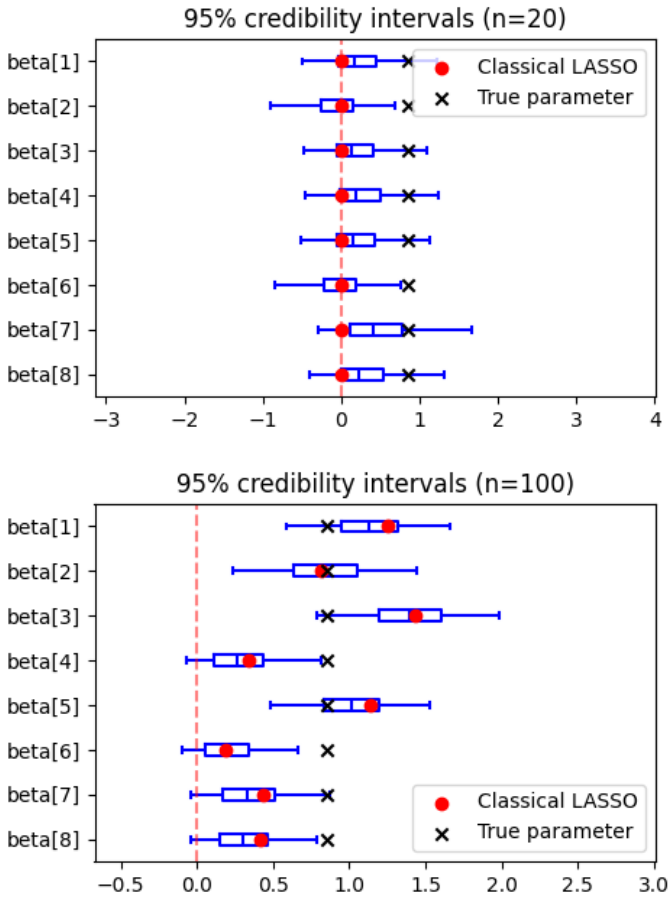

Figure 6: 95% credibility intervals for Scenario 2. The red dots are the classical LASSO's predictions when fit using cross validation, and the black crosses are the true parameter values. The vertical red line corresponds to $x = 0$.

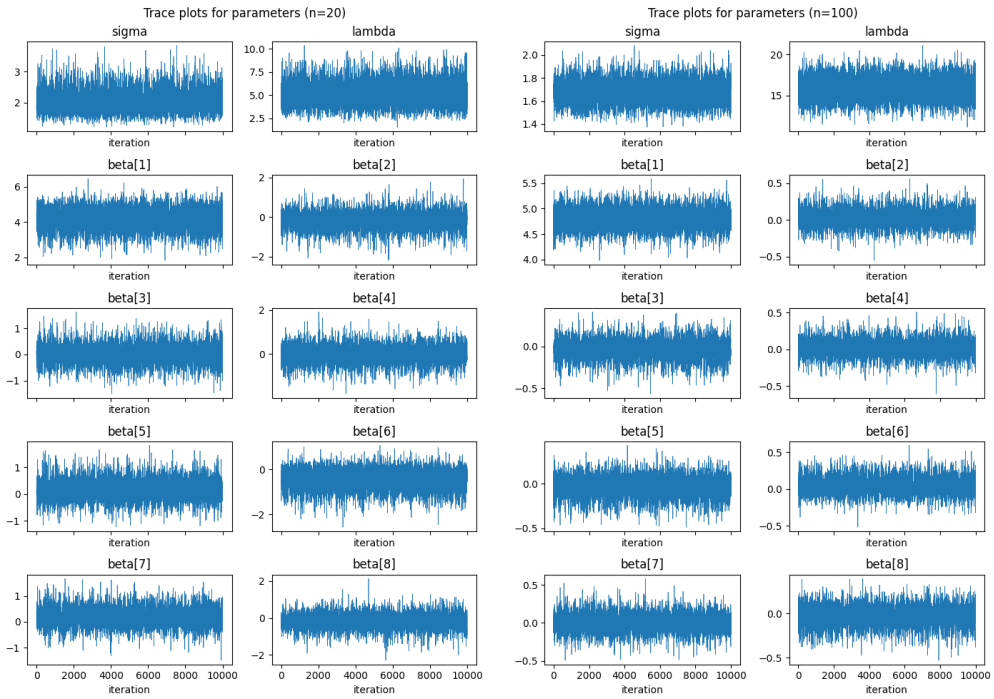

Figure 7: Trace plots for the parameters in Scenario 3. Each plot is from one of the 5 chains, chosen randomly and uniformly.

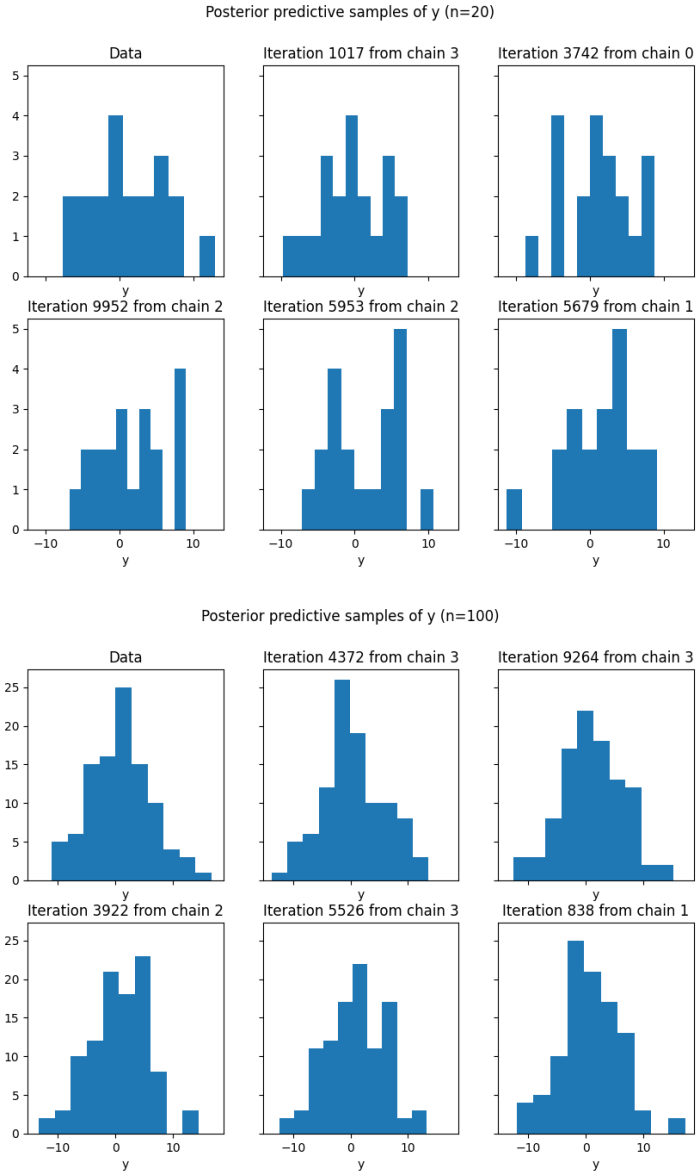

Figure 8: Posterior predictive checks for Scenario 3. The parameters corresponding to these 5 (for each n) simulated values from $\boldsymbol{y}$ were randomly and uniformly chosen across all samples from the posterior.

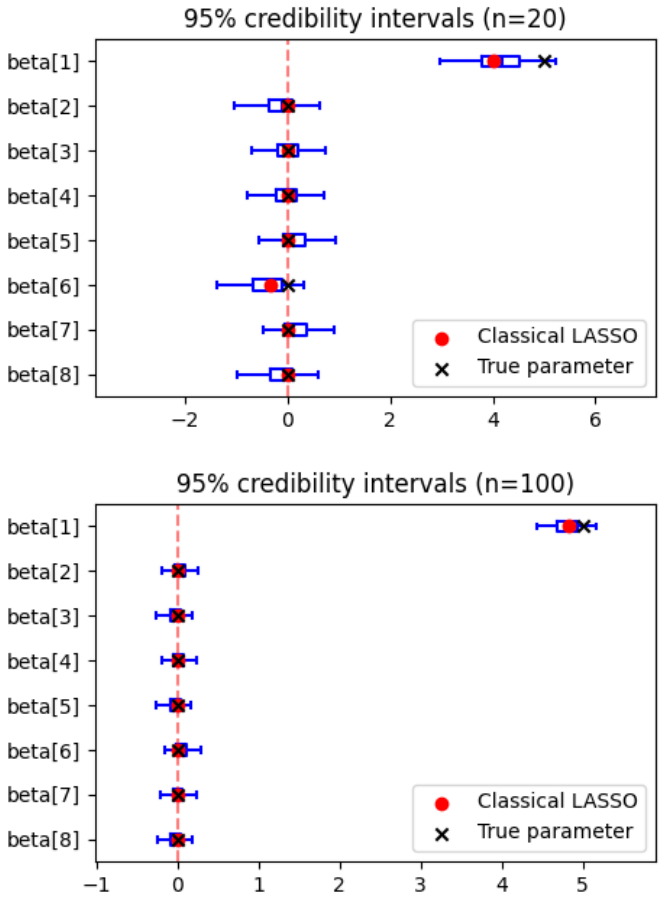

Figure 9: 95% credibility intervals for Scenario 3. The red dots are the classical LASSO's predictions when fit using cross validation, and the black crosses are the true parameter values. The vertical red line corresponds to $x = 0$.

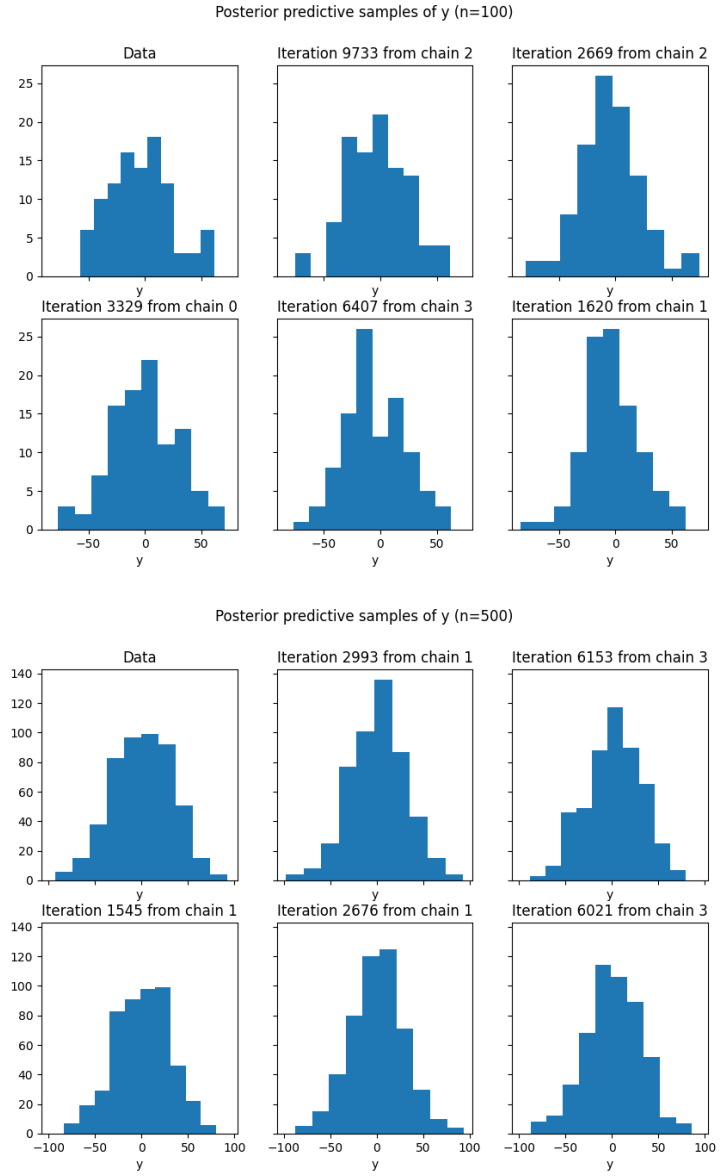

Figure 10: Posterior predictive checks for Scenario 4. The parameters corresponding to these 5 (for each n) simulated values from $\boldsymbol{y}$ were randomly and uniformly chosen across all samples from the posterior.

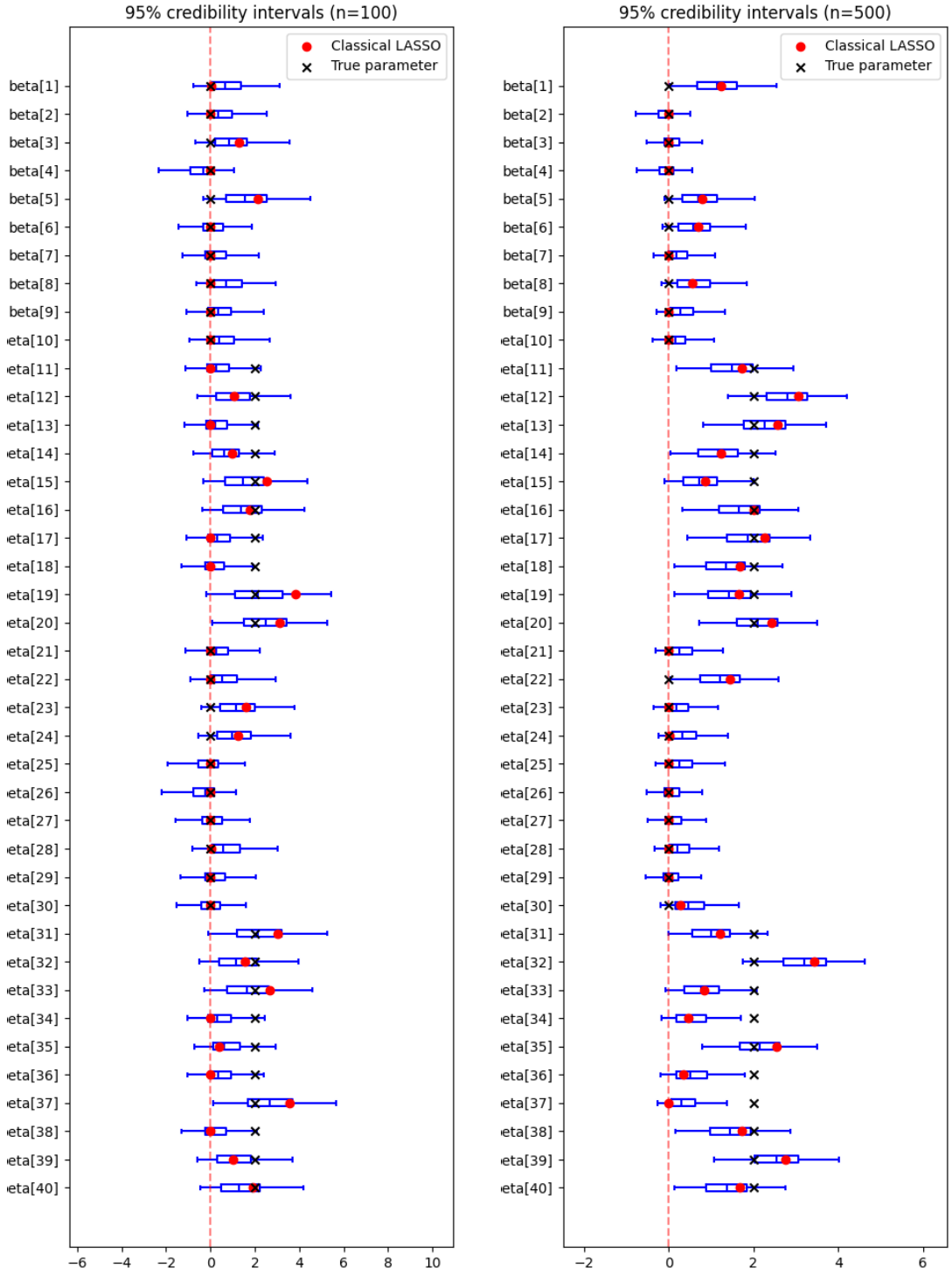

Figure 11: 95% credibility intervals for Scenario 4. The red dots are the classical LASSO's predictions when fit using cross validation, and the black crosses are the true parameter values. The vertical red line corresponds to $x = 0$.

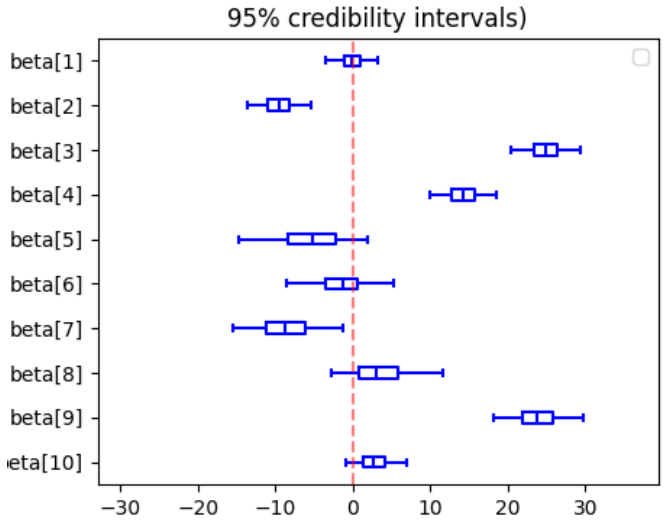

Figure 12: 95% credibility intervals for Diabetes data with $\delta = 1$ and $r = 1.78$.

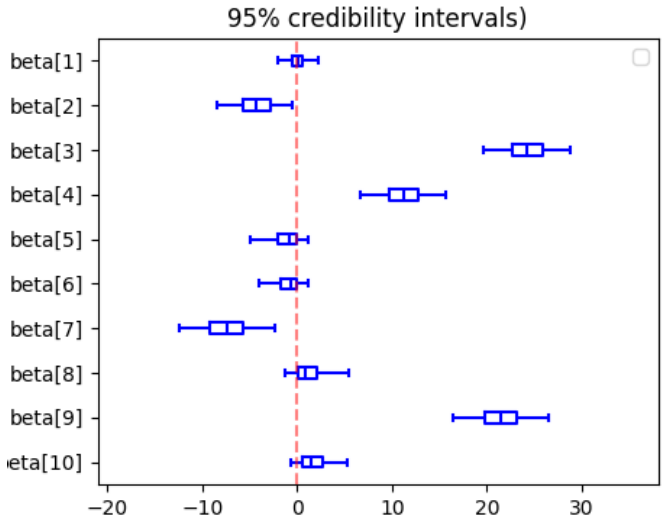

Figure 13: 95% credibility intervals for Diabetes data with $\delta = 1$ and $r = 0.1$.

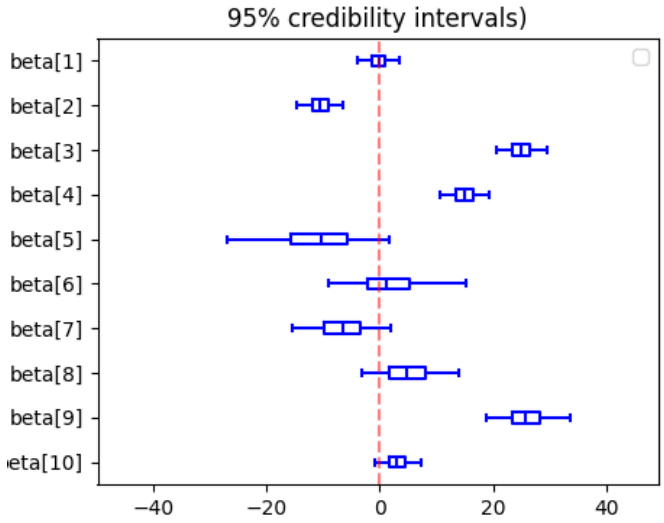

Figure 14: 95% credibility intervals for Diabetes data with $\delta = 1$ and $r = 10$.