10
Bayesian Variable Selection and Sparsity

10.1 Introduction

Over the past three decades a great deal of attention has been paid to the problem of variable selection. Specifically, in considering a relatively long list of predictors such as shown in the linear regression example in Chapter 5, concern focuses on the trade-off between the bias that could occur if important variables are omitted from the model and the variance that could occur from overfitting the model with variables that do not play a very important role in the prediction of the outcome. Variable selection methods are designed to yield so-called *sparse* models that contain, more or less, the important predictors of the outcome.

This chapter concentrates on Bayesian methods for variable selection, although the two methods discussed here can be implemented in a frequentist framework and the results are often comparable. However, as pointed out by van Erp (2020), there are a number of important benefits in adopting a Bayesian framework for variable selection. First, as we will see, variable selection can be easily implemented through the priors placed on model parameters, and these are generically referred to as *shrinkage* priors or *sparsity-inducing* priors. Shrinkage priors can be specified to shrink small coefficients toward zero while allowing large coefficients to remain large. Sparsity is induced by specifying certain hyperparameters within the priors set on the model parameters. These hyperparameters are defined through their own hyperprior distributions. The hyperpriors can be manipulated to increase or decrease the amount of shrinkage in the estimated effects.

The second benefit of adopting a Bayesian perspective to variable selection is that the penalty term is estimated in the same step as the other model parameters. In other words, the penalty term is built into the model estimation process because it is incorporated directly into the model via a prior. In turn, that prior can be specified in a flexible manner through different settings, controlling for the degree of shrinkage as the researcher sees fit.

Finally, the third benefit of estimating Bayesian penalty terms is that many different forms of penalties can be implemented. There are frequentist-based penalty techniques, such as the ridge and lasso methods described, which have their

Bayesian counterparts. In addition, there are methods that are strictly Bayesian such as the spike-and-slab prior and the horseshoe prior (see van de Schoot et al., 2021, for more information on these priors.).

In this chapter, we focus on Bayesian variable selection methods in the context of linear models and consider four methods for variable selection: (1) the ridge prior (A. E. Hoerl & Kennard, 1970; Hsiang, 1975), (2) lasso prior, (Park & Casella, 2008; Tibshirani, 1996), (3) horseshoe prior (Carvalho, Polson, & Scott, 2010), and (4) regularized horseshoe prior (Piironen & Vehtari, 2017). The first two can also be implemented in a frequentist setting, but we will concentrate on their Bayesian counterparts. Although there are many more that could be considered (see, e.g., Hastie, Tibshirani, & Friedman, 2009), these methods are chosen to highlight the issues of variable selection and lead naturally into our discussion of Bayesian model averaging in Chapter 11. A representation of the different shrinkage prior distributions is given below in Figure 10.1, and a comparison of the performance of these priors will be given in Section 10.6 below.

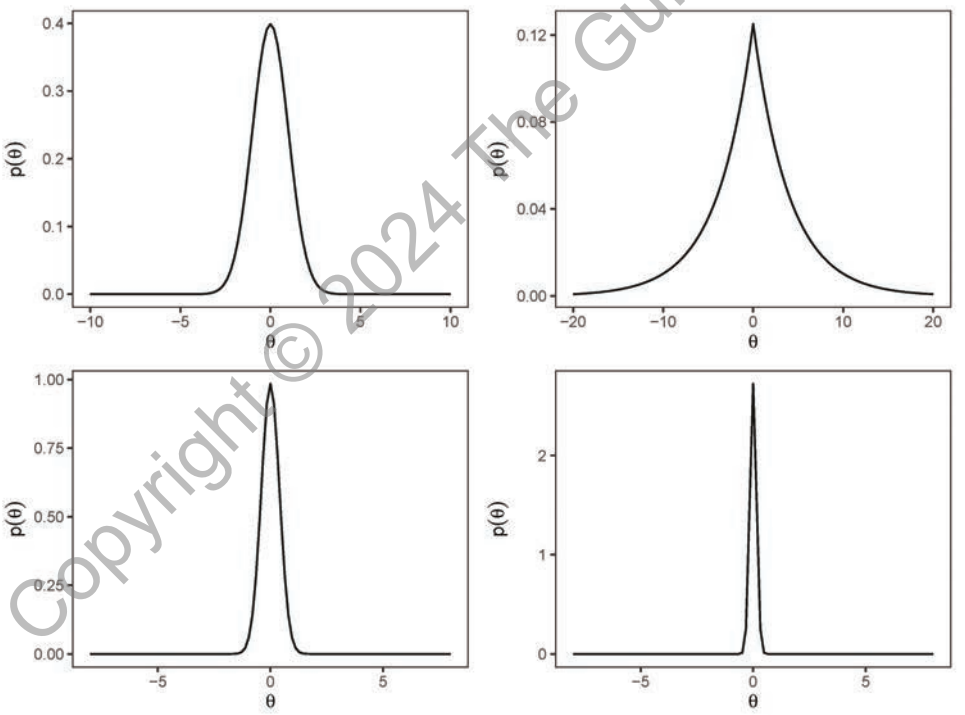

FIGURE 10.1. Four types of shrinkage priors. Top row left: Ridge prior $N(0,1)$; top row right: Laplace prior with location=0, scale=4; bottom row left: Horseshoe prior with $\lambda_p \sim C^+(0,1)$ and $\tau \sim C^+(0,1)$; bottom row right: Regularized horseshoe prior.

10.2 The Ridge Prior

As a regularization method, ridge regression (A. E. Hoerl & Kennard, 1970; R. W. Hoerl, 1985) aims to yield a parsimonious regularized regression model in the presence of highly correlated variables. The frequentist ridge estimator of β, denoted as β_{ridge} is obtained by solving the minimization

$$\boldsymbol{\beta}_{ridge} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left(\mathbf{y}'\mathbf{y} - \boldsymbol{\beta}'\boldsymbol{x}'\boldsymbol{x}\right) + \lambda \sum_{p=1}^{P} \beta_p^2 \tag{10.1}$$

where $\lambda \geq 0$ is a tuning parameter that controls the degree of regularization and the term $\lambda \sum_{p=1}^{P} \beta_p^2$) is referred to as an $L_2 - norm$. When $\lambda = 0$, we have ordinary least squares, and when $\lambda = \infty$, we obtain $\boldsymbol{\beta}_{ridge} = \mathbf{0}$. With ridge regression it can be seen that a large value of λ can lead to very heavy penalization.

Hsiang (1975) showed that if $\boldsymbol{\beta}$ has a mean of zero and covariance matrix $\boldsymbol{\Sigma} = (\sigma^2/\lambda)\mathbf{I}$, and if $\epsilon \sim N(0, \sigma_\epsilon^2\mathbf{I})$, then the posterior mean of $\boldsymbol{\beta}$ is $(\boldsymbol{x}'\boldsymbol{x} + \lambda\mathbf{I})^{-1}\boldsymbol{x}'\boldsymbol{y}$, which is an alternative specification of the ridge estimator. Hsiang (1975) also notes that if weakly informative or informative priors are placed on β_p, then the interpretation of the posterior mean of β as the ridge estimate is no longer valid.

In Bayesian ridge regression, the penalty term (λ) is captured through normally distributed independent priors placed on the regression slope parameters. These normal priors have mean hyperparameter values fixed at zero in order to control shrinkage toward zero. The variance hyperparameter is typically rescaled to be in standard deviation form and is set to define the degree of spread that the distribution exhibits. Note that we specify a $C^+(0,1)$ distribution for the residual standard deviation, but other priors could be specified as well. A representation of the ridge prior is given in the top left of Figure 10.1.

The specification of the ridge prior for the example that follows can be written as

$$\beta_j|\lambda, \sigma^2 \sim \mathcal{N}\left(0, \frac{\sigma^2}{\lambda}\right), \text{ for } j = 1, ..., p \tag{10.2}$$

where, for the following example, we assume $\sigma^2 = 1$ and we set $\lambda = 1$, inducing an $\mathcal{N}(0,1)$ prior on each regression coefficient. Note, however, that larger values of λ induce greater penalties insofar as the variance of the regression coefficients become smaller.

Example 10.1: Bayesian Ridge Regression

Before beginning, it is necessary when using any of the sparsity-inducing priors that the data be standardized beforehand. Standardizing the data beforehand provides a constant value that all parameters can be shrunk toward, namely, zero.

For our example of Bayesian ridge regression, we return to the Bayesian linear regression model in Chapter 5, using data from PISA 2018 to estimate a model of reading proficiency. For this example, we take a random sample of 100 observations to demonstrate the differences in the amount of shrinkage across the

methods. Preliminary analyses with the full sample reveal virtually no differences among the methods, as would be expected.[1]

In what follows, only the data and parameter blocks are provided insofar as the remaining code is the same as that in Example 5.1 and also across all other methods. For the ridge priors, we give an $\mathcal{N}(0, 1)$ prior to the regression coefficients and a $C^{+}(0,1)$ distribution to the standard deviation of the residuals. The likelihood follows the specification of the priors.

```
RidgeString = "
data {
    int<lower=0> n;
    vector [n] readscore;
    vector [n] Female;      vector [n] ESCS;
    vector [n] METASUM;     vector [n] PERFEED;
    vector [n] JOYREAD;     vector [n] MASTGOAL;
    vector [n] ADAPTIVITY;  vector [n] TEACHINT;
    vector [n] SCREADDIFF;  vector [n] SCREADCOMP;
}

parameters {
    real alpha;
    real beta1; real beta6;
    real beta2; real beta7;
    real beta3; real beta8;
    real beta4; real beta9;
    real beta5; real beta10;
    real<lower=0> sigma;
}

model {
    real mu[n];
      for (i in 1:n)
       mu[i] = alpha + beta1*Female[i] + beta2*ESCS[i] + beta3*METASUM[i]
          + beta4*PERFEED[i] + beta5*JOYREAD[i] + beta6*MASTGOAL[i]
          + beta7*ADAPTIVITY[i] + beta8*TEACHINT[i]
          + beta9*SCREADDIFF[i] + beta10*SCREADCOMP[i] ;
// Priors
      alpha ~ normal(0, 1);
      beta1 ~ normal(0, 1); beta6 ~ normal(0, 1);
      beta2 ~ normal(0, 1); beta7 ~ normal(0, 1);
      beta3 ~ normal(0, 1); beta8 ~ normal(0, 1);
      beta4 ~ normal(0, 1); beta9 ~ normal(0, 1);
      beta5 ~ normal(0, 1); beta10 ~ normal(0, 1);
      sigma ~ cauchy(0,1);
```


[1]We do not sample students within schools, thus this example should not be taken as a serious model of reading proficiency.

```
  // Likelihood
      readscore ~ normal(mu, sigma);

}
```

The important points to note about this code is that, first, the data should be standardized before estimation. Second, note that the specification of the $N(0, 1)$ priors induces the ridge shrinkage in the sense that regression coefficients that are close to zero will be shrunk toward the prior mean of zero, whereas large coefficients should be relatively unaffected by the prior. Again, as noted above, the extent of the shrinkage is determined by the value of λ.

10.3 The Lasso Prior

A drawback of ridge regression is that it does not improve parsimony in the sense that all of the variables still remain in the model after penalization (Zou & Hastie, 2005). A method that appears similar to ridge regression but is principally different in terms as yielding a parsimonious model is the *least absolute shrinkage and selection operator* or LASSO. The frequentist lasso involves solving the expression

$$\boldsymbol{\beta}_{lasso} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left(\mathbf{y}'\mathbf{y} - \boldsymbol{\beta}'\boldsymbol{x}'\boldsymbol{x}\right) + \lambda \sum_{p=1}^{P} |\beta_p| \tag{10.3}$$

The term $\lambda \sum_{p=1}^{P} |\beta_p|$ is referred to as an $L_1 - norm$ penalty, which allows less important coefficients to be set to zero, and thus the lasso provides for both shrinkage and variable selection.

Bayesian lasso penalization uses a different shrinkage prior as compared to the Bayesian ridge approach. Specifically, Tibshirani (1996) showed that $|\beta_p|$ is proportional to minus the log-density of the double exponential (Laplace) distribution. That is, the lasso estimate of the posterior mode of β_p can be obtained by using the prior

$$p(\beta_p) = \frac{1}{2\tau} \exp\left(-\frac{|\beta_p|}{\tau}\right) \tag{10.4}$$

where $\tau = 1/\lambda$.

The top right of Figure 10.1 shows the double exponential distribution. We see that the double exponential distribution is ideal because it peaks at zero, which shrinks small coefficients toward zero. However, the double exponential can be set to have thick tails (in both directions), allowing the larger coefficients to remain large. Given that the distribution is centered at zero to control shrinkage toward zero, the mean hyperparameter setting is fixed to zero. The scale, or dispersion, of the double exponential distribution is the hyperparameter that researchers can alter when implementing the shrinkage. This defines the amount of spread and

the thickness of the tails, which controls the degree of shrinkage in coefficients. Again, a $C^+(0,1)$ prior can be specified on the standard deviation of the residuals, if desired.

Although the ridge and lasso approaches are similarly implemented in the Bayesian framework, these techniques can produce different amounts of shrinkage depending on the hyperparameter settings. That is, the lasso approach can result in more shrinkage for the small estimates, but less shrinkage for the large estimates. This result is a function of the double exponential distribution implemented in the lasso approach. The double exponential distribution is more peaked around zero and it has heavier tails compared to the normal distribution used in the ridge approach. Regardless of the approach implemented, Bayesian penalization can be a useful tool when attempting to avoid overfitting a complex model to small samples. Indeed, the lasso is simultaneously a shrinkage and variable selection method. In addition, these approaches further highlight the modeling flexibility that Bayesian methods provide through the flexible implementation of priors. Next follows the specification for the lasso priors.

Example 10.2: Bayesian Lasso Regression

Below we show the Stan code for the lasso prior. Note in the parameter block the use of the double exponential(0,1) distribution to induce the lasso. Also, notice that we do not attempt to induce as much shrinkage in the intercept alpha.

```
modelString = "
data {
    int<lower=0> n;
    vector [n] readscore;
    vector [n] Female;      vector [n] ESCS;
    vector [n] METASUM;     vector [n] PERFEED;
    vector [n] JOYREAD;     vector [n] MASTGOAL;
    vector [n] ADAPTIVITY;  vector [n] TEACHINT;
    vector [n] SCREADDIFF;  vector [n] SCREADCOMP;
}
model {
    real mu[n];
      for (i in 1:n)
       mu[i] = alpha + beta1*Female[i] + beta2*ESCS[i] +
       beta3*METASUM[i]
        + beta4*PERFEED[i] + beta5*JOYREAD[i] + beta6*MASTGOAL[i]
        + beta7*ADAPTIVITY[i] + beta8*TEACHINT[i]
        + beta9*SCREADDIFF[i] + beta10*SCREADCOMP[i] ;

// Priors
alpha ~ normal(0, 1);
beta1 ~ double_exponential(0, 1); beta6 ~ double_exponential(0, 1);
```

```
beta2 ~ double_exponential(0, 1); beta7 ~ double_exponential(0, 1);
beta3 ~ double_exponential(0, 1); beta8 ~ double_exponential(0, 1);
beta4 ~ double_exponential(0, 1); beta9 ~ double_exponential(0, 1);
beta5 ~ double_exponential(0, 1); beta10 ~ double_exponential(0, 1);
sigma ~ cauchy(0,1);

// Likelihood
    readscore ~ normal(mu, sigma);

}
```

The lasso is not without limitations (see van Erp, Oberski, & Mulder, 2019). First, when the number of variables p are greater than the sample size n (which we might encounter in "big data" problems), the model selection algorithm will stop at n because the model will no longer be identified. Second, if there are groups of variables that are highly pairwise correlated, the lasso will select only one of the variables from that group rather arbitrarily. Third, when $n > p$, which is the motivating case in this chapter, and when variables are highly correlated, it has been shown that ridge regression will outperform the lasso with respect to predictive performance.

10.4 The Horseshoe Prior

An alternative to the lasso prior which has gained popularity in the Bayesian literature is the so-called *horseshoe prior*. The horseshoe prior belongs to a class of so-called *global-local* shrinkage priors.[2] Following the notation in Betancourt (2018a), the horseshoe prior can be specified as follows:

$$\beta_p \sim N(0, \tau\lambda_p) \tag{10.5a}$$
$$\lambda_p \sim C^+(0, 1) \tag{10.5b}$$
$$\tau \sim C^+(0, \tau_0) \tag{10.5c}$$

where τ_0 is a hyperparameter that controls the behavior of the global shrinkage prior τ. The intuition behind the horseshoe prior is that the global parameter τ shrinks all of the coefficients toward zero while the local parameter λ_p allows some large coefficients to bypass the shrinkage. The horseshoe prior can be seen in the bottom row left of Figure 10.1

Example 10.3: The Horseshoe Prior

[2]The horseshoe prior gets its name from the fact that under certain conditions, the probability distribution of the shrinkage parameter associated with horseshoe prior reduces to a Beta($\frac{1}{2}$, $\frac{1}{2}$) distribution, which has the shape of a horseshoe.

For this example, we specify λ_p as the local prior for each of the p regression coefficients and τ as the global prior in the Stan parameter block, where we set $\tau_0 = 1$. Note that in the Stan model block, the regression coefficients have mean zero and a scale mixture $\tau\lambda_p$.

```
Horseshoe = "
data {
  int<lower=1> n; // Number of data
  int<lower=1> p; // Number of covariates
  matrix[n,p] X;
  real readscore[n];
}
parameters {
  vector[p] beta;
  vector<lower=0>[p] lambda; // Local prior
  real<lower=0> tau; // Global prior
  real alpha;
  real<lower=0> sigma;
}
model {
  beta ~ normal(0, tau * lambda); // Scale mixture
  tau ~ cauchy(0, 1);
  lambda ~ cauchy(0, 1);
  alpha ~ normal(0, 1);
  sigma ~ cauchy(0, 1);

  readscore ~ normal(alpha + X * beta, sigma);
}
// For posterior predictive checking and loo cross-validation
generated quantities {
  vector[n] readscore_rep;
  vector[n] log_lik;
  for (i in 1:n) {
  readscore_rep[i] = normal_rng(alpha + X[i,:] * beta, sigma);
  log_lik[i] = normal_lpdf(readscore[i] | alpha + X[i,:]
    * beta, sigma);
}
  }
"
```

10.5 Regularized Horseshoe Prior

A limitation of the conventional horseshoe prior relates to the regularization of the large coefficients. Specifically, it is still the case that large coefficients can transcend the global scale set by τ_0, with the impact being that the posteriors of these large coefficients can become quite diffused, particularly in the case of weakly-identified coefficients (Betancourt, 2018a). To remedy this issue, Piironen and Vehtari (2017) proposed a *regularized* version of the horseshoe prior (also known as the *Finnish horseshoe prior*). Following the notation used in Betancourt (2018a),

$$\beta_p \sim N(0, \tau\tilde{\lambda}_p) \tag{10.6a}$$

$$\tilde{\lambda}_p = \frac{c\lambda_p}{\sqrt{c^2 + \tau^2\lambda_p^2}} \tag{10.6b}$$

$$\lambda_p \sim C^+(0, 1), \tag{10.6c}$$

$$c^2 \sim \text{inv-gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}s^2\right) \tag{10.6d}$$

$$\tau \sim C^+(0, \tau_0) \tag{10.6e}$$

where s^2 is the variance for each of the p predictor variables. As pointed out by Piironen and Vehtari (2017), those variables that have large variances would be considered more relevant a priori, and while it is possible to provide predictor specific values for s^2, generally we scale the variables ahead of time so that $s^2 = 1$. Finally, c^2 is the slab width which controls the size of the large regression coefficients.

To gain an intuition of the regularized horseshoe, first note that the form of Equation (10.6a) is quite similar to the horseshoe prior, however $\tilde{\lambda}_p$ places a control on the size of the coefficients by introducing a slab width c^2 in Equation (10.6b). Following Piironen and Vehtari (2017), notice that if $\tau^2\lambda_p^2 \ll c^2$, then this means that β_p is close to zero and $\tilde{\lambda}_p \to \lambda_p$, which is the original horseshoe in Section 10.4. However, if $\tau^2\lambda_p^2 \gg c^2$, then $\tilde{\lambda}_p \to c^2/\tau^2$ and the prior begins to approach the $\mathcal{N}(0, c^2)$, where, again, the choice of c^2 controls the size of the large coefficients. Because c^2 is a slab width that might not be well known, it follows that it should be given a prior distribution, and Piironen and Vehtari (2017) recommend the inverse-gamma distribution in Equation (10.6d), which induces a relatively non-informative Student's - t slab when coefficients are far from zero.

Example: 10.4: The Regularized Horseshoe Prior

In setting up Stan first recall that as with all of the methods for sparsity, the data are first standardized to have a mean of zero and standard deviation of one. Also, recall that Stan works with standard deviations and not variances or precisions. To start, for the regularized horseshoe we first need to indicate our belief regarding the number of large coefficients. This is required because the global scale parameter τ_0 inside the transformed parameter block is a function of

the number of large coefficients assumed by the researcher ahead of analyzing the data. In the `transformed data` block, this is indicated by the line `real p0=5;`.

```
PISA18sampleScale <- read.csv(file.choose(),header=T)

n <- nrow(PISA18sampleScale)
X <- PISA18sampleScale[,2:11]
readscore <- PISA18sampleScale[,1]
p <- ncol(X)

data.list <- list(n=n, p=p, X=X, readscore=readscore)

# Stan code adapted and modified from from Betacourt 2018 #

modelString = "
data {
  int <lower=1> n;          // number of observations
  int <lower=1> p;          // number of predictors
  real readscore[n];        // outcome
  matrix[n,p] X;            // inputs
}

transformed data {
  real p0 = 5;
}
```

Next, in the `parameters` block, we define the parameters of the regularized horseshoe given in Equations (10.6a) - (10.6e).

```
parameters {
  vector[p] beta;
  vector<lower=0>[p] lambda;
  real<lower=0> c2;
  real<lower=0> tau;
  real alpha;
  real<lower=0> sigma;

}
```

In the `transformed parameters` we specify `tau0` in line with Betancourt (2018a) and we write $\tilde{\lambda}$ as in Equation (10.6d).

```
transformed parameters {
  real tau0 = (p0 / (p - p0)) * (sigma / sqrt(1.0 * n));
  vector[p] lambda_tilde =
     sqrt(c2)  * lambda ./ sqrt(c2 + square(tau) * square(lambda));
}
```

We now put everything together in the model block.

```
model {
  beta ~ normal(0, tau * lambda_tilde);
  lambda ~ cauchy(0, 1);
  c2 ~ inv_gamma(2,8);
  tau ~ cauchy(0, tau0);

  alpha ~ normal(0, 2);
  sigma ~ cauchy(0, 1);

  readscore ~ normal(X * beta + alpha, sigma);

}
// For posterior predictive checking and loo cross-validation
generated quantities {
  vector[n] readscore_rep;
  vector[n] log_lik;
  for (i in 1:n) {
  readscore_rep[i] = normal_rng(alpha + X[i,:] * beta, sigma);
  log_lik[i] = normal_lpdf(readscore[i] | alpha + X[i,:]
 * beta, sigma);
}
  }
"
```


10.6 Comparison of Regularization Methods

It may be of interest to run a side-by-side comparison of the regularization methods in terms of their effects on parameter estimates and standard deviations and LOO cross-validation. The comparison is displayed below in Table 10.1. Bayesian linear regression with non-informative priors using the standardized data is given under the BLR column for comparison purposes.

TABLE 10.1. Comparison of posterior results based on different regularization methods

		Ridge	Lasso	Horseshoe[a]	Reg. horseshoe[b]
Variable	Parameter	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)
Intercept	alpha	−0.04 (0.12)	−0.05 (0.12)	−0.01 (0.10)	−0.01 (0.10)
FEMALE	beta1	0.08 (0.16)	0.08 (0.17)	0.02 (0.10)	0.02 (0.10)
ESCS	beta2	0.27 (0.09)	0.28 (0.09)	0.24 (0.10)	0.23 (0.10)
METASUM	beta3	0.37 (0.08)	0.36 (0.08)	0.33 (0.09)	0.33 (0.09)
PERFEED	beta4	−0.10 (0.10)	−0.10 (0.10)	−0.05 (0.08)	−0.05 (0.07)
JOYREAD	beta5	0.10 (0.09)	0.09 (0.10)	0.07 (0.08)	0.06 (0.08)
MASTGOAL	beta6	−0.20 (0.08)	−0.19 (0.09)	−0.14 (0.09)	−0.13 (0.09)
ADAPTIVITY	beta7	−0.05 (0.11)	−0.04 (0.11)	−0.02 (0.07)	−0.02 (0.06)
TEACHINT	beta8	0.03 (0.09)	0.02 (0.09)	0.00 (0.06)	0.00 (0.06)
SCREADDIFF	beta9	−0.13 (0.10)	−0.13 (0.10)	−0.11 (0.09)	−0.10 (0.09)
SCREADCOMP	beta10	0.18 (0.10)	0.17 (0.10)	0.11 (0.09)	0.13 (0.10)
Residual	sigma	0.83 (0.06)	0.83 (0.06)	0.83 (0.06)	0.83 (0.06)
LOO-IC[c]		258.1 (12.4)	258.0 (12.4)	256.9 (12.1)	257.1 (11.9)

[a] 262 divergent transitions generated after warmup.

[b] 29 divergent transitions generated after warmup.

[c] Value in parentheses are LOO-IC standard errors.

First, note that the horseshoe and regularized horseshoe methods generated a warning of *divergent transitions* after warmup. This message needs to be taken seriously and implies that the complexity of the model is such that the HMC/NUTS algorithm cannot pick up small changes in the curvature of the log posterior. As such, the estimates may be biased. A possible solution to this problem is to adjust the alpha_delta setting to beyond the default of 0.99 and max_treedepth setting to beyond the default value of 12, and of course to check the model and priors. For this example, we set adapt_delta=.9999 and max_treedepth=20 and still had divergent transitions. Generally speaking, however, if other diagnostics such as n_eff and Rhat look good, then one can proceed to interpret the results, albeit with caution. For more information on Stan program warnings, see `https://mc-stan.org/misc/warnings.html`.

With this caveat in mind, a visual inspection of the results in Table 10.1 indicates that the ridge and lasso priors provide results that are somewhat similar to Bayesian linear regression with non-informative priors that we found in Table 5.1 (when standardized). On the other hand, the original horseshoe prior and regularized horseshoe achieve slightly more shrinkage in the posterior estimates and standard deviations with the regularized horseshoe yielding the most shrinkage, and indeed shrinking some of the larger coefficients (e.g., *beta*2 and *beta*3), as expected. In terms of cross-validation, however, we find that the horseshoe prior yields the lowest value of the LOO-IC followed closely by the regularized horseshoe. A comparative analysis of this kind might be worthwhile in practice if the goal of the analysis is not only variable selection but also comparative predictive performance.

10.6.1 An Aside: The Spike-and-Slab Prior

In this chapter, we did not demonstrate the so-called *spike-and-slab* prior, which has been considered the "gold standard" for sparsity for quite some time (Mitchell & Beauchamp, 1988; E. I. George & McCulloch, 1993). However, in the interest of completeness, we should say a brief word about it.

The spike-and-slab prior gets its name because the prior distribution on the individual regression coefficients come from a two-component mixture of Gaussian distributions and can be written as

$$\beta_p \mid \lambda_p, c, \epsilon \sim \lambda_p \mathcal{N}(0, c^2) + (1 - \lambda_p)\mathcal{N}(0, \epsilon^2) \tag{10.7a}$$

$$\lambda_p \sim \text{Bernoulli}(\pi) \tag{10.7b}$$

where $\lambda_p \in \{0, 1\}$ is an indicator variable that determines whether the coefficient is close to zero, in which case it comes from the spike ($\lambda_p = 0$), or nonzero, in which case it comes from the slab ($\lambda_p = 1$). To create a spike, it is common to set $\epsilon = 0$. The slab width c and the inclusion probability π of the Bernoulli random variable is set by the user. Notice that with $\epsilon = 0$ the spike and slab prior can be rewritten as

$$\beta_p \mid \lambda_p, c \sim \lambda_p \mathcal{N}(0, c^2\lambda^2) \tag{10.8a}$$

$$\lambda_p \sim \text{Bernoulli}(\pi) \tag{10.8b}$$

The result of this setup is that λ is a discrete parameter that only takes on two values ($\lambda_p = 0, 1$).

It is necessary to note that Stan cannot incorporate discrete parameters. However, studies have shown the similarity in performance between the spike-and-slab prior and the horseshoe prior (see, e.g., Carvalho et al., 2010; Polson & Scott, 2011). Finally, the spike-and-slab prior is similar to the regularized horseshoe prior when the slab width $c < \infty$, thus providing some regularization on large coefficients.

10.7 Summary

This chapter considered the problem of Bayesian variable selection and sparsity. Many variable selection methods can be implemented in the frequentist and Bayesian framework, and some are explicitly Bayesian. However, both simulation studies and real data analyses seem to point to the original horseshoe prior or regularized horseshoe prior as the preferred methods for inducing sparsity, particularly with respect to out-of-sample predictive performance. As usual, in the case of large sample sizes, application of sparsity-inducing priors will likely lead to similar conclusions. Nevertheless, it may be prudent to examine results using different priors and choose the model that yields desirable shrinkage along with acceptable out-of-sample predictive performance.

In the end, however, a single model is selected for interpretation, and although the predictive performance of Bayesian shrinkage methods is often better than regression modeling without inducing sparsity, these methods do not account for the uncertainty that underlies the choice of a single model. An approach

to addressing the problem of model selection is simply not to choose a single model but to carefully average over the space of possible models that could have generated the data. The next chapter takes up the problem of model uncertainty.

Guilford Publications
370 Seventh Avenue
New York, NY 10001
212-431-9800
800-365-7006
www.guilford.com