# A Survey of Hahn and Cavarlho's Posterior Model Selection Technique

Michael Price
Rice University

December 4, 2020

**Abstract**

Many different Bayesian variable selection techniques exist for linear regression analysis. Although the posterior output of these methods can provide great insight, for simplicity, a single model output is often preferred. Hahn and Cavarlho (2015) introduced a posterior summary technique that transforms a posterior distribution into sequence of sparse predictors and a method to pick one of these sparse predictors. Our goal in this analysis is to assess the validity of this technique. We will accomplish this by first obtaining multiple posterior distributions from different Bayesian variable selection methods and simulated datasets. We will then compare the accuracy and predictive power of the models chosen by Hahn and Cavarlho's method to the model output from other frequentist techniques. We will restrict our analysis on highly multicollinear datasets, as they pose a greater challenge to a data analyst.

## Introduction

In linear regression analysis, we are often interested in the following model:

$$y = \boldsymbol{X}\boldsymbol{\beta} + \epsilon$$

where $\boldsymbol{X}$ is a $n\ x\ p$ design matrix, $\boldsymbol{\beta}$ is a $p\ x\ 1$ matrix of real numbered coefficients, $y$ is a $n\ x\ 1$ matrix of response variables, and $\epsilon \sim N(0, \sigma^2 I_n)$, with $I_n$ being a $n\ x\ n$ identity matrix. However, the data analyst is often given predictors that have no significant relationship with the response variable, or put more mathematically, $\boldsymbol{\beta}_j = 0$, for $j \in J \subset \{1, 2, ..., p\}$

Many frequentist and Bayesian methods have been developed to not only find the irrelevant predictors, but also provide accurate estimates of the coefficients where $\boldsymbol{\beta}_k \neq 0$.

Popular frequentist methods include Forward Stepwise Selection, LASSO (Tibshirani 1996) and Elastic Net (Zou and Hastie 2004). These methods are notable for their high accuracy and computational efficiency, and are often preferred to Bayesian methods because they output one model, while the posterior distribution theoretically provides an infinite number of models to choose from. While metrics like posterior mean and median may provide "one" model, the estimates for $\boldsymbol{\beta}_k$ almost never equal 0, therefore inducing almost no sparsity in the model choice.

Hahn and Cavarlho (2015) introduced a technique which not only outputs one model from a posterior distribution, but one which exhibits sparsity. In addition, the shrinkage and selection procedures are "decoupled", unlike most other methods, where shrinkage and selection are performed simultaneously. The goal of this analysis is to assess the validity of Hahn and Cavarlho's method by applyitng it to multiple Bayesian variable selection methods and simulated datasets. We will assess the outputted model through three metrics, which measure the accuracy of the estimates for $\hat{\boldsymbol{\beta}}$ and the predictive power the chosen model. In addition, we will perform a similar analysis on the frequentist methods LASSO and Elastic Net Regression in order to assess how Hahn and Cavarhlo's procedure performs relative to popular frequentist counterparts. We will restrict our analysis to datasets that exhibit high multicollinarity, as they prove especially challenging to a data analyst.

We will begin with an extensive summary and discussion of Hahn and Cavarlho's method. Next, we will provide a brief explaination of the Bayesian variable selection techniques that will be used in this analysis, namely Bayesian LASSO (Park and Casella 2008), the "Horseshoe Prior" (Carvalho, Polson, and Scott 2010), Stochastic Search Variable Selection (George and McCulloch 1993) and the "spike and slab" method provided by Ishwaran and Rao (2005). We will then explain the setup for our empirical experiment, and follow up with an in depth discussion on the results of our experiment. We will conclude with a survey of the strengths and weaknesses of Hahn and Cavarlho's method, and possible avenues for future work.

# A Survey of Hahn and Cavarlho's Procedure

## The DSS Loss Function

Hahn and Cavarlho's method solves a Bayesian decision problem, where the goal is to minimize the expected loss with respect to the posterior distribution.

$$\mathbb{E}(L(\hat{\gamma}, \gamma)) = \int L(\hat{\gamma}, \gamma) p(\gamma|y)$$

Hahn and Cavarlho choose the following loss function

$$L(\hat{y}, \gamma) = \lambda ||\gamma||_0 + \tfrac{1}{n} ||X\gamma - \hat{y}||_2^2$$

where $||\gamma||_0 = \sum_{i=1}^{p} \mathbb{I}(\gamma_j \neq 0)$ and $\lambda > 0$. This resembles many loss functions used in different shrinkage and variable selection techniques, namely LASSO and Adaptive LASSO (Zou 2006). Integrating over $\hat{y}$ conditional on $\boldsymbol{\beta}$ and $\sigma^2$ yields

$$L(\boldsymbol{\beta}, \sigma^2, \gamma) \equiv \lambda ||\gamma||_0 + \tfrac{1}{n} ||\boldsymbol{X}\gamma - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \sigma^2$$

Integrating over the posterior $p(\boldsymbol{\beta}, \sigma^2|y)$ finally yields

$$L(\gamma) \equiv \lambda ||\gamma||_0 + \bar{\sigma^2} + \text{trace}(\boldsymbol{X^T X} \Sigma_{\boldsymbol{\beta}}) + \tfrac{1}{n} ||\boldsymbol{X}\gamma - \boldsymbol{X}\bar{\boldsymbol{\beta}}||_2^2$$

where $\bar{\sigma^2} = \mathbb{E}(\sigma^2)$, $\bar{\boldsymbol{\beta}} = \mathbb{E}(\boldsymbol{\beta})$ and $\Sigma_{\boldsymbol{\beta}} = \text{cov}(\boldsymbol{\beta})$, all with respect to the posterior. As $\bar{\sigma^2}$ and $\text{trace}(\boldsymbol{X^T X} \Sigma_{\boldsymbol{\beta}})$ are both constants, the loss function we are explicitly interested in is

$$L(\gamma) = \lambda ||\gamma||_0 + \tfrac{1}{n} ||\boldsymbol{X}\gamma - \boldsymbol{X}\bar{\boldsymbol{\beta}}||_2^2$$

in which the authors call the "decoupled shrink and selection" (DSS) loss function. The full derivation is provided in the Appendix under "DSS Loss Function Derivation". This loss function exhibits many desirable properties, namely that a solution to the minimization problem:

$$argmin_{\gamma} \;\; \lambda ||\gamma||_0 + \tfrac{1}{n} ||\boldsymbol{X}\gamma - \boldsymbol{X}\bar{\boldsymbol{\beta}}||_2^2$$

exists if $\bar{\boldsymbol{\beta}}$ exists as well. Therefore, this selection procedure can be applied for any prior distribution as long as it yields a proper posterior distribution, thereby expanding the versatility of this method.

## Estimating $\boldsymbol{\beta}_\lambda$

Unfortunately, no efficient algorithms exist that finds the exact $\gamma$ which minimizes $L(\gamma)$. One possible approach is to replace $||\gamma||_0$ with $||\gamma||_1$, which would make $L(\gamma)$ identical to the LASSO minimization function, with $Y = X\bar{\boldsymbol{\beta}}$. However, this may cause more shrinkage than desired, given how the prior can provide additional regularization. To avoid this, Hahn and Cavarlho advise to minimize the following function instead,

$$L'(\gamma) = \sum_{i=1}^{p} \frac{\lambda}{|\bar{\omega_j}|}\gamma_i + \frac{1}{n}||\boldsymbol{X}\gamma - \boldsymbol{X}\bar{\boldsymbol{\beta}}||_2^2$$

where $w_j = \bar{\boldsymbol{\beta}}_j$. This function is identical to the loss function from the Adaptive LASSO, which can be minimized via the `lars` procedure (Efron et al. 2004), only requiring a simple re-scaling of the design matrix (Zou 2006). The procedure is outlined below:

1. For each column $\boldsymbol{x}_j$ in $\boldsymbol{X}$, set $\boldsymbol{x}_j^* = \boldsymbol{x}_j * |\bar{\boldsymbol{\beta}}_j|$.
2. Input $\boldsymbol{X^*}$ and $y = \boldsymbol{X}\bar{\boldsymbol{\beta}}$ into the `lars` algorithm, which outputs a sequence of different $\boldsymbol{\beta}^*$ for a variety of $\lambda$ values.
3. For each $\boldsymbol{\beta}^*$, set $\boldsymbol{\beta_{j\lambda}} = \boldsymbol{\beta_j^*} * |\bar{\boldsymbol{\beta}}_j|$.

The final step of the procedure involves choosing the most suitable $\boldsymbol{\beta_\lambda}$.

## Choosing the Proper $\boldsymbol{\beta_\lambda}$

Hahn and Cavarlho define "variation-explained" for $\boldsymbol{\beta}_\lambda$ as

$$\boldsymbol{\rho_\lambda} = \frac{\frac{1}{n}||X\boldsymbol{\beta}||_2^2}{\frac{1}{n}||X\boldsymbol{\beta}||_2^2 + \sigma_2 + \frac{1}{n}||X\boldsymbol{\beta} - X\boldsymbol{\beta_\lambda}||_2^2}$$

where $\boldsymbol{\beta}$ is a draw from the posterior distribution. Thus, we can obtain a sample of $\boldsymbol{\rho_\lambda}$'s for each $\boldsymbol{\beta_\lambda}$. $\boldsymbol{\rho_\lambda}$ bars a resemblance to the $R^2$ statistic used in frequentist regression analysis, due to each $\boldsymbol{\rho_{\lambda j}}$ having a value in between 0 and 1 and the metric describing some form of "variation explained by the model".

Finally, with a sample of $\boldsymbol{\rho_\lambda}$ we can obtain 90% credible interval estimates for $\boldsymbol{\rho_\lambda}$. With these intervals, Hahn and Cavarlho recommend choosing the sparsest $\boldsymbol{\beta_\lambda}$ whose credible interval for $\boldsymbol{\rho_\lambda}$ contains $\mathbb{E}(\boldsymbol{\rho_0})$, which will the model whose coefficients are $\bar{\boldsymbol{\beta}}$. In other words, we desire the smallest model that is not significantly different from the model provided by the posterior mean.

## Summary Plots

In addition to providing a posterior selection technique, Hahn and Cavalrho also detail two different summary plots that provide additional insight into the model selection process. The first one they recommend is to plot the credible intervals for each of the $\boldsymbol{\rho_\lambda}$ considered, where the models are ordered from most to least sparse. An example is plotted below.

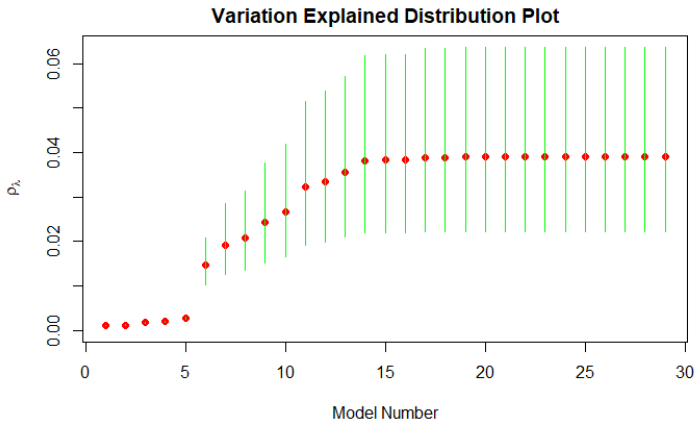



Figure 1: $\boldsymbol{\rho_\lambda}$ Distribution Plot Example

We can see that the distribution of the $\boldsymbol{\rho_\lambda}$'s start to look similar around Model 14. Therefore, the predictors added to model past Model 15 do not help explain the variation in the response variable, thus rendering them insignificant.

In addition, Hahn and Cavarlho introduce another metric, "excess error", defined as:

$$\boldsymbol{\psi_\lambda} = \sqrt{\tfrac{1}{n}||X\boldsymbol{\beta} - X\boldsymbol{\beta_\lambda}||_2^2 + \sigma_2} - \sigma$$

where $\boldsymbol{\beta}$ is a draw from the posterior distribution. Similarly to $\boldsymbol{\rho_\lambda}$, credible intervals for $\boldsymbol{\psi_\lambda}$

can be easily derived, and plotted for each $\boldsymbol{\beta_\lambda}$. Like $\boldsymbol{\rho_\lambda}$, the models are ordered from most to least sparse.

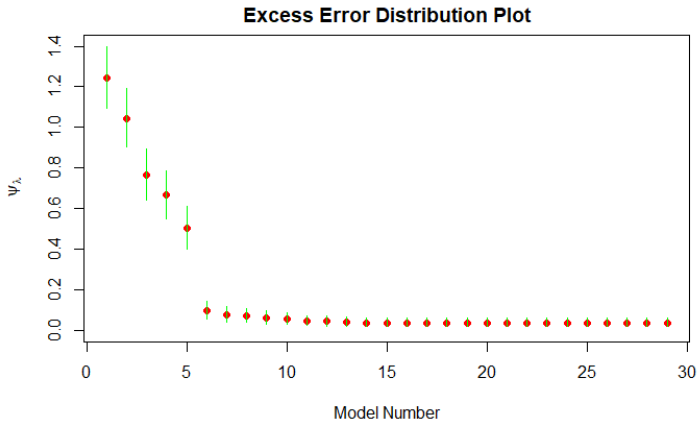



Figure 2: $\boldsymbol{\psi_\lambda}$ Distribution Plot Example

Again, given how around Model 14 the excess error fails to noticeable decrease, indicating that predictors added after Model 14 are insignificant.

# Common Bayesian Variable Selection Techniques

For each of the Bayesian variable selection methods presented in this section, the likelihood remains $y \sim N(\boldsymbol{X\beta}, \sigma^2 I_n)$. However, each of the priors have high density on values around 0 in order the shrink the insignificant predictors' coefficients toward 0. We present two categories, shrinkage priors, which mitigate spurious correlations, and spike and slab priors, which more aggressively push the coefficients toward 0. For each of these models, the prior for the variance will be $\tau = \sigma^{-2} \sim Gamma(\alpha, \beta)$

## Shrinkage Priors

### Bayesian LASSO

The Bayesian LASSO (Park and Casella 2008) induces to following prior on each $\boldsymbol{\beta_j}$, where each $\boldsymbol{\beta_j}$ is i.i.d:

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

which is a double exponential distribution with parameters $(0, \frac{\lambda}{\sigma})$. Conditioning on $\sigma^2$ guarantees a unimodal posterior, which are more computationally efficent to sample from. In addition, $\log p(\boldsymbol{\beta}|\sigma^2)$ is proportional up to a constant to $-\frac{\lambda}{\sigma}\sum_{j=1}^{p}|\boldsymbol{\beta_j}|$, which resembles the penalization term for the LASSO minimization function. In addition, for $\lambda$, the authors recommend putting a disperse hyperprior on $\lambda$, namely $\lambda^2 \sim Gamma(a_\lambda, b_\lambda)$.

### Horseshoe Prior

Carvalho, Polson, and Scott (2010) introduced the following prior for $\boldsymbol{\beta_j}$, where the $\boldsymbol{\beta_j}$'s are independent.

$$\boldsymbol{\beta_j} \overset{\text{ind}}{\sim} N(0, c^2\lambda_j^2)$$
$$\lambda_j \overset{\text{i.i.d}}{\sim} C^+(0,1)$$

where $C^+$ is a half-Cauchy distribution. The horseshoe name comes from the fact that if $y_j \sim N(\theta_j, 1)$ and $\theta_j \sim N(0, \lambda_j^2)$, then $\mathbb{E}(\theta_j|y_j) = (1-\kappa_j)y_j$, where $\kappa_j = \frac{1}{1+\lambda_j^2}$. This implies that $\kappa_j \sim Beta(\frac{1}{2}, \frac{1}{2})$, where the highest density regions are near 0 and 1, giving the p.d.f. a distinct horseshoe shape. In addition, this implies that the posterior $\boldsymbol{\beta_j}$ are either near 0 or far away, entailing more aggressive shrinking than the Bayesian LASSO.

## Spike and Slab Priors

### Stochastic Search Variable Selection

SSVS, proposed by George and McCulloch in 1993, induces a normal mixture prior for $\boldsymbol{\beta_j}$, namely,

$$\boldsymbol{\beta_j}|\pi_j, \gamma_j, c_j^2 \overset{\text{ind}}{\sim} (1-\pi_j)N(0, \gamma_j^2) + \pi_j N(0, c_j^2\gamma_j^2)$$

In theory, $\gamma_j^2$ takes on a small value, while $c_j^2$ takes on a sufficiently large number. However, SSVS proves to be a relatively non-robust method, and results can vary greatly for

different choices of $\gamma_j^2$ and $c_j^2$. The authors recommend trying numerous different values for $\gamma_j^2$ and $c_j^2$, and to extract information from each of the different posteriors. $\pi_j$ is traditionally chosen to be $\frac{1}{2}$ for each of the coefficients.

### Ishwaran and Rao's Continuous Bimodal Prior

In response to the shortcomings of SSVS, Ishwaran and Rao (2005) proposed the following prior for $\boldsymbol{\beta_j}$, where $\tau_k$ represents a precision:

$$\begin{aligned}
\boldsymbol{\beta_j} &\overset{\text{ind}}{\sim} N(0, \gamma_k \tau_k^2) \\
\gamma_k &\overset{\text{i.i.d}}{\sim} (1-w)\delta_{v_0} + w\delta_1 \\
\tau_k &\overset{\text{i.i.d}}{\sim} Gamma(a_\beta, b_\beta) \\
w &\sim Uniform(0,1)
\end{aligned}$$

where $v_0$ is a positive number very close to 0, and $\delta_x$ are point masses located at $x$. In this setup, $\gamma_k \tau_k^2$ has a continuous bimodal distribution unlike in SSVS, where a bimodoal prior has to be manually set (Ishwaran and Rao 2005).

# Experiment Setup

## Generating Data with High Multicollinearity

To ensure that the each regressor has identical mean and variance, as well as exhibit significant multicollineairty, $\boldsymbol{X}$ will be generated from a multivariate normal distribution, with mean 0 and covariance $\Sigma_X$ where $\Sigma_X$ is a correlation matrix. In addition, with mean 0, we will not need to include an intercept term into any of our models. Joe (2006) provides a method that generates a random correlation matrix where the multicollinearity can be adjusted easily. Essentially, the algorithm takes in some positive real number $\phi$, where the partial correlations are first drawn from a $Beta(\phi, \phi)$ distribution, scaled to in between $-1$ and 1. As $\phi \to 0$, more probability mass is concentrated around $-1$ and 1, thus leading

to much more correlation among the variables in $\boldsymbol{X}$. Finally, the partial correlations are adjusted to the standard Pearson correlation coefficients. This process can be achieved automatically via the `rcorrmatrix` function in the `clusterGeneration` package in R. The correlation matrix of the first generated dataset, with $\phi = 0.05$ is plotted below:

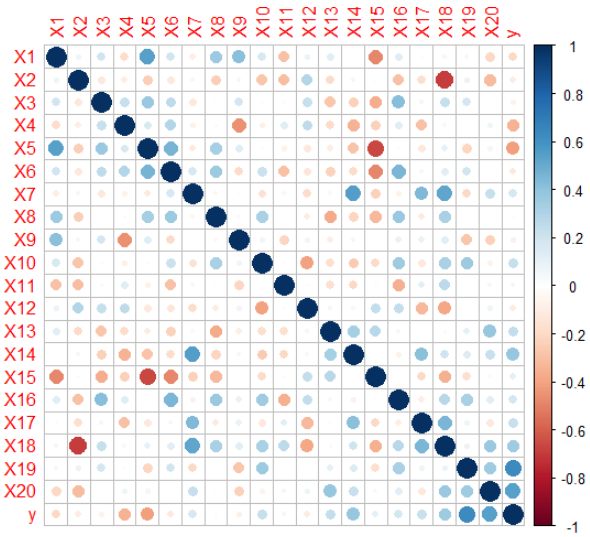



Figure 3: Example Correlation Matrix Generated by Joe's Method

Next, the ground truth $\boldsymbol{\beta_j}$'s were generated, where the significant predictors' coefficients were drawn from a uniform distribution with some user defined minimum and maximum, and the insignificant predictors' coefficents remained 0. The significant predictors were chosen randomly, with the number of significant predictors inputted by the user. Finally, $y$ was generated by setting $y = \boldsymbol{X\beta} + \epsilon$, where $\epsilon \sim N(0, \sigma^2 I_n)$, with $\sigma^2 > 0$. In total, three different datasets were generated, each of them detailed below. $p$ represented the total number of predictors in the dataset, and $p_{rel}$ is the number of predictors whose $\boldsymbol{\beta_j} \neq 0$.

| Dataset # | $n$ | $p$ | $p_{rel}$ | $\phi$ | $\sigma$ |
|---|---|---|---|---|---|
| 1 | 1000 | 20 | 5 | 0.05 | 3 |
| 2 | 100 | 50 | 5 | 0.2 | 5 |
| 3 | 100 | 50 | 25 | 0.2 | 2 |

Table 1: Dataset Generation Hyperparameters

## Model Fitting and Posterior Selection

Each dataset was fitted to the four Bayesian variable selection models discussed earlier. Each individual Bayesian method, dataset pair was inputted into RStan to obtain 10000 posterior draws. The prior hyperparameters were chosen to be weakly informative, which are generally the priors Bayesian statisticians recommend. The hyperparameters used are listed in the Appendix section titled "Prior Hyperparameters".

The posterior draws were then inputted into Hahn and Cavarlho's algorithm, where the sequence of sparse models and the distributions for $\boldsymbol{\rho_\lambda}$ and $\boldsymbol{\psi_\lambda}$ were obtained. Finally, using Hahn and Cavarlho's selection criterion, a final model was selected from the posterior distribution.

Next, both LASSO and Elastic Net Regression were applied to each of the three datasets. The optimal values for $\lambda$ and $\alpha$ were found using the `cv.glmnet` function in the `glmnet` package, which employs a 10-fold cross validation.

## Model Evaluation

Finally, all outputted models were evaluated based on three criteria. First, the false positive rate was calculated, meaning the proportion of $\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}$ whose true value was 0, but was incorrectly estimated to be nonzero. Next, the coefficient error was calculated, simply defined as:

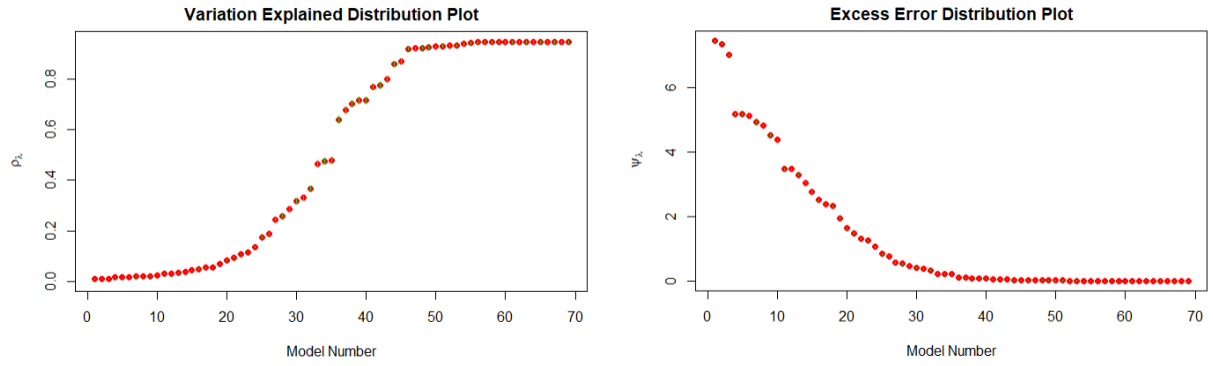
$$\text{Coefficient Error } (\hat{\boldsymbol{\beta}}) = \frac{1}{p}\sum_{j=1}^{p} |\boldsymbol{\beta_j} - \hat{\boldsymbol{\beta}}_{\boldsymbol{j}}|$$

where $\hat{\boldsymbol{\beta}}$ is the outputted model. Because the scale of each predictor is approximately 1, a weighted average is not needed. Lastly, we will calculate the mean absolute error for the model evaluated on some test set to evaluate its predictive power.

# Experiment Results

## Posterior Sampling

As mentioned above, all Bayesian models were implemented via Stan. Overall, posterior sampling was successful more most models, given that the effective sample sizes were all between 9000 and 11000. In addition, divergent transitions were minimal, indicating an accurate estimation of the posterior. However, the one notable exception was the Horseshoe Prior for Dataset 3. Although the effective sample sizes were reasonably high, again around 9000 and 11000, almost every sample resulted in a divergent transition, resulting in extremely long computation time and an inaccurate estimation of the posterior. This poor performance is also reflected in Hahn and Cavarlho's diagnostic plots.




(a) Variation Explained

(b) Excess Error

Figure 4: Variation Explained and Excess Error Plots for Horseshoe Prior, Dataset 3

Thus, given the minimal length of the credible intervals, the model chosen by Hahn and Cavarlho's procedure will not exhibit much sparsity, leading a highly inaccurate model choice. These small credible intervals are most likely due to the draws for $\boldsymbol{\beta}$ and $\sigma^2$ being extremely similar, indicating minimal exploration of the posterior distribution.

## Comparison of all Models

Summary bar charts for each of the three metrics are provided below, with the model type on the x-axis and the error on the y axis, grouped by dataset. The exact metric outputs can be found in the appendix under the section "Exact Experiment Metric Outputs".

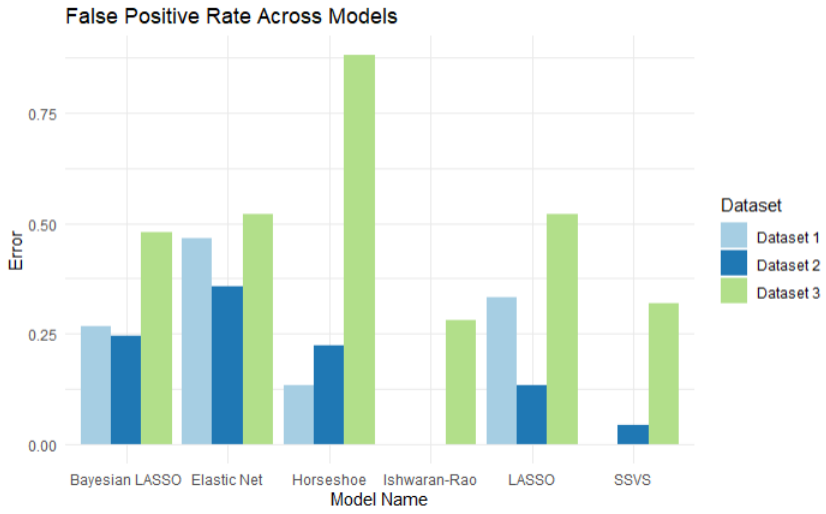



Figure 5: False Positive Error Metric

Interestingly, Hahn and Cavalrho's method combined with Ishwaran and Rao's prior did not misidentify any insignificant predictors in Datasets 1 and 2, and had the lowest false positive rate for Dataset 3, with SSVS not following far behind. In addition, LASSO actually performed worse than both the Bayesian spike and slab models, with false positive rates around 0.33 0.13 and 0.52 respectively. Thus, in terms of feature selection, Hahn and Cavarlho's method combined with spike and slab methods seem superior.

The shrinkage priors' false positive rate showed to be higher than the spike and slab models. However, this makes sense, given that they tend to be less aggressive shrinking coefficients to 0. Still, besides the poor performing Horseshoe Prior for Dataset 3, these methods had lower false positive rates than Elastic Net, which also tries to perform regularization and feature selection.

Finally, all methods did not perform spectacularly well on dataset 3, indicating that all

these methods slightly struggle when a large quantity of predictors are significant.


Figure 6: Coefficent Error Metric

In terms of coefficient error, all models performed similarly across all three datasets, with an exception being Elastic Net on Dataset 2. Therefore, even through the shrinkage priors and elastic net had higher false positive rates, the coefficient error was similar to the spike and slab methods and LASSO. Therefore, for the shrinkage priors and Elastic Net, the insignificant predictor's coefficients are reasonably close to zero. Finally, even though Ishwaran-Rao and SSVS had the lowest coefficient error across the three datasets, the difference was minimal, most certainly caused by the over-shrinkage of the non-zero coefficients. Thus, the estimates from the spike and slab models should by no means be considered completely accurate.

For the final metric, the mean absolute error of the model applied to a testing set. As seen in the plot, no model significantly outperformed the others. Thus, no matter the prior, Hahn and Cavarlho's method leads to models with similar predictive power to popular frequentist methods.

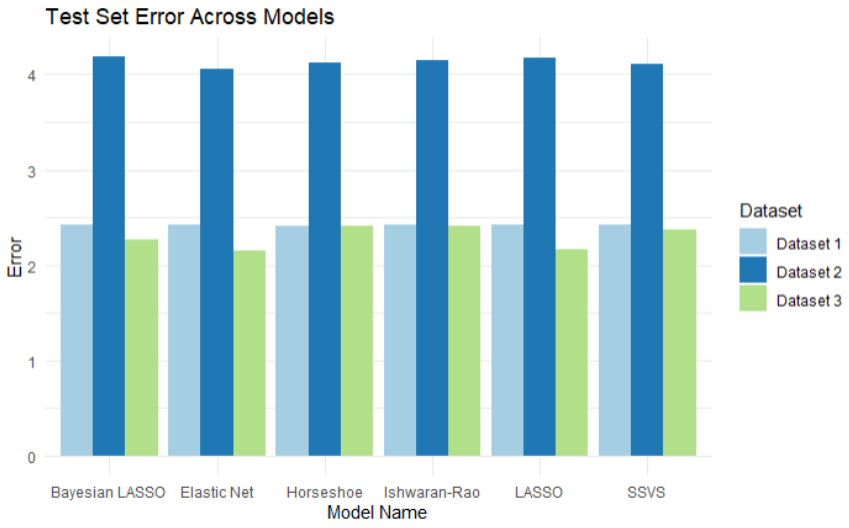
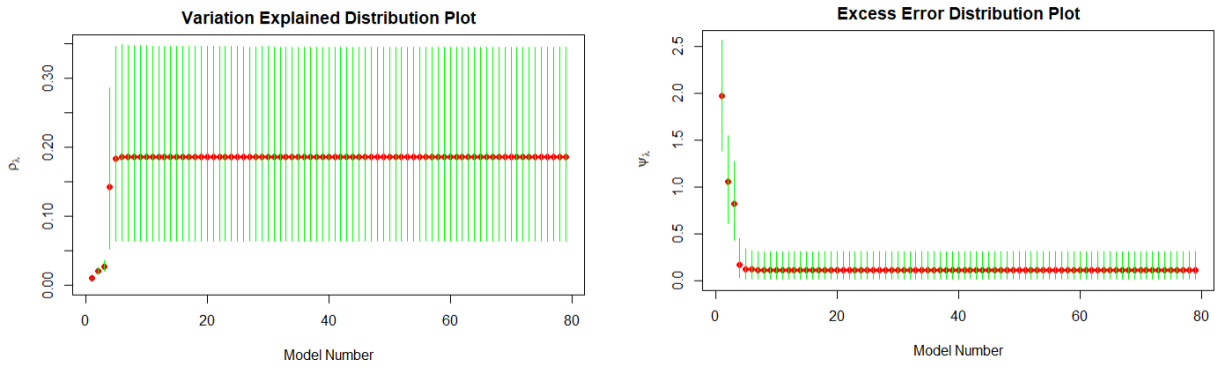



Figure 7: Test Error Metric

Therefore, given that Hahn and Cavalrho's procedure for posterior model selection, especially in combination with spike and slab priors, consistently performed as well or better than the frequentist methods, the technique proves worthwhile, especially when faced with data that exhibits high multicollinearity. However, because sampling from a posterior can be somewhat computationally expensive, it is worth noting that the frequentist methods still excel in terms of their efficiency.




(a) Variation Explained

(b) Excess Error

Figure 8: Variation Explained and Excess Error Plots for Ishwaran-Rao, Dataset 2

# Conclusion

Sparsity is desirable in linear regression models in order to mitigate spurious correlations. Despite numerous Bayesian method existing to identify these insignificant predictors, frequentist methods are often preferred due to them outputting one model, in contrast to the infinite models that the posterior distribution provides.

Hahn and Cavarlho (2015) introduced a method that not only selects one model from the posterior, but also one that exhibits sparsity. In addition, they claim that this procedure is highly interpretable, given that the metrics used to assess the posterior can easily be plotted. In this experiment, we investigated the validity of this method by applying Hahn and Cavarlho's techniques to numerous different Bayesian variable shrinkage and selection priors, and comparing the outputs to popular frequentist methods. In addition, we restricted our analysis to simulated datasets that exhibit high multicollinearity, as these prove especially challenging for data analysts.

Overall, Hahn and Cavarhlo's method shows extreme promise, especially when used with spike and slab priors. In terms of average coefficient error and prediction accuracy, the models chosen by Hahn and Cavarhlo's technique displayed similar results among the priors to LASSO and Elastic Net. However, it truly stands out in filtering out the insignificant predictors. For example, when using the Ishwaran-Rao prior, the method successfully set all the useless regressors' coefficients to zero in two of the datasets. The other priors also produced promising results, with Bayesian LASSO, the Horseshoe and SSVS producing similar or even better results compared to the frequentist counterparts. Thus, Hahn and Cavarlho's method shows extreme promise as a variable selection technique.

In addition, the "variation explained" and "excess error" metrics can easily be plotted, providing a visually appealing way to summarize the posterior selection process. The plots resemble "elbow" methods commonly used in machine learning, like PCA and K-Means Clustering, where the chosen model is the one where the change in "variation explained" starts to diminish.

However, there are future paths to consider when evaluating this method. First, most notably, the Gaussian error assumption was perfectly realized in the simulated datasets, which is rare in real world applications. Therefore, further analysis on datasets that do not perfectly satisfy the Gaussian assumption should be pursed. In addition, because Hahn and Cavarlho's approach of "decoupling" shrinkage and selection shows extreme promise, it is worth investigating if this method can be applied to other models beside linear regression.

# Appendix

## DSS Loss Function Derivation

$$
\begin{aligned}
\mathbb{E}(L(\hat{y},\gamma)) &= \lambda||\gamma||_0 + \mathbb{E}(\tfrac{1}{n}(\boldsymbol{X}\gamma-\hat{y})^T(\boldsymbol{X}\gamma-\hat{y})) \\
&= \lambda||\gamma||_0 + \tfrac{1}{n}[\gamma^T\boldsymbol{X}^T\boldsymbol{X}\gamma - 2\gamma^T\boldsymbol{X}^T\mathbb{E}(\hat{y}) + \mathbb{E}(\hat{y}^T\hat{y})] \\
&= \lambda||\gamma||_0 + \tfrac{1}{n}[\gamma^T\boldsymbol{X}^T\boldsymbol{X}\gamma - 2\gamma^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + n\sigma^2 + (\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{X}\boldsymbol{\beta})] \\
&= \lambda||\gamma||_0 + \tfrac{1}{n}||\boldsymbol{X}\gamma - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \sigma^2 \\
\mathbb{E}(L(\boldsymbol{\beta},\sigma^2,\gamma)) &= \lambda||\gamma||_0 + \mathbb{E}(\sigma^2) + \tfrac{1}{n}[\gamma^T\boldsymbol{X}^T\boldsymbol{X}\gamma - 2\gamma^T\boldsymbol{X}^T\boldsymbol{X}\mathbb{E}(\boldsymbol{\beta}) + \mathbb{E}((\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{X}\boldsymbol{\beta}))] \\
&= \lambda||\gamma||_0 + \bar{\sigma^2} + \tfrac{1}{n}[\gamma^T\boldsymbol{X}^T\boldsymbol{X}\gamma - 2\gamma^T\boldsymbol{X}^T\boldsymbol{X}\bar{\boldsymbol{\beta}} + \bar{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\bar{\boldsymbol{\beta}} + trace(\boldsymbol{X}^T\boldsymbol{X}\Sigma_{\boldsymbol{\beta}})] \\
&= \lambda||\gamma||_0 + \bar{\sigma^2} + \tfrac{1}{n}||\boldsymbol{X}\gamma - \boldsymbol{X}\bar{\boldsymbol{\beta}}||_2^2 + trace(\boldsymbol{X}^T\boldsymbol{X}\Sigma_{\boldsymbol{\beta}}) \\
&\stackrel{c}{=} \lambda||\gamma||_0 + \tfrac{1}{n}||\boldsymbol{X}\gamma - \boldsymbol{X}\bar{\boldsymbol{\beta}}||_2^2 \\
&= L(\gamma)
\end{aligned}
$$

## Prior Hyperparameters

All 3 datasets used the same weakly informative priors. All notation here is identical to the notation used in section "Baeysian Variable Selection Techniques".

| $\alpha$ | $\beta$ | $a_\lambda$ | $b_\lambda$ |
|---|---|---|---|
| 1 | 1 | 0.1 | 0.1 |

Table 2: Bayesian LASSO Prior Hyperparameters

| $\alpha$ | $\beta$ | $c_j$ |
|---|---|---|
| 1 | 1 | 10 |

Table 3: Horseshoe Prior Hyperparameters

| $\alpha$ | $\beta$ | $\pi_j$ | $c_j$ | $\gamma_j$ |
|---|---|---|---|---|
| 1 | 1 | 1/2 | 1000 | 0.01 |

Table 4: SSVS Prior Hyperparameters

| $\alpha$ | $\beta$ | $a_\beta$ | $b_\beta$ | $v_0$ |
|---|---|---|---|---|
| 1 | 1 | 2 | 200 | 0.000001 |

Table 5: Ishwaran-Rao Prior Hyperparameters

## Exact Experiment Metric Outputs

| | Dataset 1 | Dataset 2 | Dataset 3 |
|---|---|---|---|
| Bayesian LASSO | 0.26667 | 0.24444 | 0.36048 |
| Horseshoe Prior | 0.13333 | 0.22222 | 0.88 |
| SSVS | 0 | 0.04444 | 0.32 |
| Elastic Net | 0.46667 | 0.35556 | 0.52 |
| LASSO | 0.33333 | 0.13333 | 0.52 |

Table 6: False Positive Metric

| | Dataset 1 | Dataset 2 | Dataset 3 |
|---|---|---|---|
| Bayesian LASSO | 0.031242 | 0.132806 | 0.360487 |
| Horseshoe Prior | 0.017847 | 0.117854 | 0.36215 |
| SSVS | 0.018203 | 0.090882 | 0.331264 |
| Elastic Net | 0.051178 | 0.23388 | 0.326043 |
| LASSO | 0.033508 | 0.11728 | 0.329287 |

Table 7: Average Coefficient Error Metric

| | Dataset 1 | Dataset 2 | Dataset 3 |
|---|---|---|---|
| Bayesian LASSO | 2.420402 | 4,184336 | 2.267713 |
| Horseshoe Prior | 2.413127 | 4.117588 | 2.416304 |
| SSVS | 2.421959 | 4.110726 | 2.37274 |
| Elastic Net | 2.42825 | 4.153906 | 2.155725 |
| LASSO | 2.421658 | 4.16949 | 2.161671 |

Table 8: Average Absolute Test Error Metric

# References


Carvalho, C., Polson, N., & Scott, J. (2010). The horseshoe estimator for sparse signals. Biometrika, 97(2), 465-480. Retrieved December 6, 2020, from http://www.jstor.org/stable/25734098

Efron, B., Hastie, T. Johnstone, I. Tibshirani, R. (2004). Least angle regression. Ann. Statist. 32 no. 2, 407–499. doi:10.1214/009053604000000067. https://projecteuclid.org/euclid.aos/1083178935

George, E., & McCulloch, R. (1993). Variable Selection Via Gibbs Sampling. Journal of the American Statistical Association, 88(423), 881-889. doi:10.2307/2290777

Hahn, P.R. & Carvalho C.M. (2015) Decoupling Shrinkage and Selection in Bayesian Linear Models: A Posterior Summary Perspective, Journal of the American Statistical Association, 110:509, 435-448, DOI: 10.1080/01621459.2014.993077

Ishwaran, H. & Rao, J. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. Ann. Statist. 33 , no. 2, 730–773. doi:10.1214/009053604000001147. https://projecteuclid.org/euclid.aos/1117114335

Joe, H. (2006). Generating random correlation matrices based on partial correlations. Journal of Multivariate Analysis, 97(10), 2177-2189.

Park, T. & Casella G. (2008). The Bayesian Lasso, Journal of the American Statistical Association, 103:482, 681-686, DOI: 10.1198/016214508000000337

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. Royal. Statist. Soc B., Vol. 58, No. 1, pages 267-288).

Zou, H. (2006) The Adaptive Lasso and Its Oracle Properties, Journal of the American Statistical Association, 101:476, 1418-1429, DOI: 10.1198/016214506000000735

Zou, H., & Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 67(2), 301-320. Retrieved December 6, 2020, from http://www.jstor.org/stable/3647580