Received: 14 December 2018 | Revised: 16 February 2019 | Accepted: 6 March 2019

DOI: 10.1002/wics.1463

ADVANCED REVIEW

WIREs COMPUTATIONAL STATISTICS WILEY

Bayesian regularization: From Tikhonov to horseshoe

Nicholas G. Polson[1] | **Vadim Sokolov**[2]

[1]The University of Chicago, Booth School of Business, Chicago, Illinois

[2]Department of Systems Engineering and Operations Research, George Mason University, Fairfax, Virginia

Correspondence
Nicholas G. Polson, The University of Chicago, Booth School of Business, Chicago, IL.
Email: ngp@chicagobooth.edu

Bayesian regularization is a central tool in modern-day statistical and machine learning methods. Many applications involve high-dimensional sparse signal recovery problems. The goal of our paper is to provide a review of the literature on penalty-based regularization approaches, from Tikhonov (Ridge, Lasso) to horseshoe regularization.

This article is categorized under:

Statistical and Graphical Methods of Data Analysis > Robust Methods

Statistical Models > Linear Models

Statistical Models > Bayesian Models

KEYWORDS

Bayesian regression, horseshoe, lasso, regularization

1 | INTRODUCTION

Regularization is a machine learning technique that allows for an optimal trade-off between model complexity (bias) and out-of-sample performance (variance). To fix ideas, consider regularization in the context of a linear model, where an output y is generated by

$$y = x^T\beta + \epsilon, \;\; \epsilon \sim p(\epsilon). \tag{1}$$

Assuming normally distributed errors, $p(\epsilon) = N\left(0, \sigma_\epsilon^2\right)$, the corresponding regularized maximum likelihood optimization problem is finding the solution to

$$\text{minimize}_\beta \quad \|y - X\beta\|_2^2 \qquad \text{subject to} \sum_{i=1}^{p} \phi(\beta_i) \le s. \tag{2}$$

Here, y is the vector of observed outputs, X is a design matrix, and β are the model parameters. Each β_i has a regularization penalty $\phi(\beta_i)$ and s is a hyper-parameter that controls the bias-variance trade-off.

Regularization can be viewed as constraint on the model space. The techniques were originally applied to solve ill-posed problems where a slight change in the initial data could significantly alter the solution. Regularization techniques were then proposed for parameter reconstruction in a physical system modeled by a linear operator implied by a set of observations. It had long been believed that ill-conditioned problems offered little practical value, until Tikhonov published his seminal paper (Tikhonov, 1943) on regularization. Tihonov (1963) proposed methods for solving regularized problems of the form

$$\text{minimize}_\beta \|y - X\beta\|_p^p + \lambda \left\| \left(\beta - \beta^{(0)}\right) \right\|_q^q.$$

Here, λ is the weight on the regularization penalty and the ℓ_q-norm is defined by $\|\beta\|_q^q = \sum_i \beta_i^q$. This optimization problem is a Lagrangian form of the constrained problem given in Equation (2) with $\phi(\beta_i) = \left(\beta_i - \beta_i^{(0)}\right)^q$.

The subsequent developments were proposed in Ivanov (1962) and numerical algorithms were then developed by Bakushinskii (1967). All of these methods required developing approximations by well-posed problems, parameterized by the regularization parameter. Most of the early work in Soviet literature focused on proving convergence of the solutions of well-posed problems to the ill-posed problems. Numerical schemes were proposed much later. For a detailed overview of earlier convergence and numerical results, see Tikhonov and Arsenin (1977) and Ivanov, Vasin, and Tanana (2013).

In the context of linear models in statistics Hoerl and Kennard (1970) derived statistical properties of regularized estimators in case when penalty has ℓ_2 norm and $p = q = 2$. This estimator was called the Ridge regression.

Later, sparsity became a primary driving force behind new regularization methods Candès and Wakin (2008). When the penalty term has ℓ_1 norm ($p = 2$, $q = 1$), the solution to regularized problem is sparse, for example, has many zeros (Alliney, 1992; Aster, Borchers, & Thurber, 2018; Donoho, 1992; Donoho & Johnstone, 1995). Use of ℓ_0 (Polson & Sun, in press) pseudo-norm, which counts the number of nonzero entries in a vector, leads to a NP hard optimization problem. ℓ_1 penalty can be viewed as a convex approximation of ℓ_0 penalty which still has the required property of recovering sparse vectors of parameters. An algorithm for estimating ℓ_1 regularized linear statistical model was proposed by Alliney and Ruzinsky (1994). Williams (1995) used Bayesian approach that assigns Laplace prior for parameters of nonlinear neural network models. Tibshirani (1996) derived statistical properties of ℓ_1 regularization based estimators for linear models and coined the term lasso for this problem. For brief historical accounts on the use of the ℓ_1 penalty in statistics and signal processing, see Tibshirani (1996); Miller (2002), and the total variational denoising literature Claerbout and Muir (1973); Taylor, Banks, and McCoy (1979).

2 | BAYESIAN REGULARIZATION: FROM TIKHONOV TO HORSESHOE

Mathematically, one can to think of defining a regularized solution by constraining the topology of a search space to a ball. From a Bayesian perspective instead assigns a prior distribution to each of the model's parameters. From a historical perspective, James-Stein (a.k.a L^2-regularization) Stein (1964) provided a global shrinkage rule for improving statistical estimation. There are no local parameters to learn about sparsity, which led to horseshoe regularization.

2.1 | Bayes risk

A simple sparsity example illustrates the issue with L^2-regularization and the James-Stein estimator. Consider the sparse r-spike problem and focus solely on rules with the same shrinkage weight (albeit benefiting from pooling of information). Let the true parameter value be $\theta_p = \left(\sqrt{d/p}, \ldots, \sqrt{d/p}, 0, \ldots, 0\right)$. James-Stein is equivalent to the model

$$y_i = \theta_i + \epsilon_i \text{ and } \theta_i \sim N\left(0, \tau^2\right)$$

This dominates the plain MLE but loses admissibility because a "plug-in" estimate of global shrinkage $\hat{\tau}$ is used. Original "closed-form" analysis is particularly relevant here (Tiao & Tan, 1965). They point out that the mode of $p(\tau^2|y)$ is zero exactly when the shrinkage weight turns negative (their condition 6.6). From a risk perspective $E\|\hat{\theta}^{JS} - \theta\| \leq p, \forall\theta$ showing the inadmissibility of the MLE. At origin the risk is 2, but

$$\frac{p\|\theta\|^2}{p + \|\theta\|^2} \leq R\left(\hat{\theta}^{JS}, \theta_p\right) \leq 2 + \frac{p\|\theta\|^2}{d + \|\theta\|^2}.$$

This implies that $R\left(\hat{\theta}^{JS}, \theta_p\right) \geq (p/2)$. Hence, simple thresholding rule beats James-Stein this with a risk given by $\sqrt{\log p}$. This simple example, shows that the choice of penalty should not be taken for granted as different estimators will have different risk profiles.

2.2 | Bayesian regularization duality

From a Bayesian perspective regularization is performed by defining a prior distribution over the model parameters. A Bayesian linear regression model is defined as

TABLE 1 Prior distributions and corresponding penalty functions (negative log-probability)

	Ridge	Lasso	Cauchy	Horseshoe
Prior $p(\beta_i\|\tau)$	$\frac{1}{\sqrt{2\pi}\tau}\exp\left(-\frac{\beta_i^2}{2\tau^2}\right)$	$\frac{1}{2\tau}\exp\left(-\frac{\|\beta\|}{\tau}\right)$	$\frac{\tau}{\pi\tau^2+\pi\beta_i^2}$	$\leq\pi\sqrt{\frac{\pi}{2}}\log\left(1+\frac{2\tau^2}{\beta_i^2}\right)$
Penalty $\varphi_\tau(\beta_i)$	$\frac{1}{2\tau^2}\beta_i^2$	$\frac{\|\beta_i\|}{\tau}$	$\log\left(\tau^2+\beta_i^2\right)$	$-\log\log\left(1+\frac{2\tau^2}{\beta_i^2}\right)$

$$y=x^T\beta+\epsilon,\ \epsilon\sim N\left(0,\sigma_\epsilon^2\right),\ \beta\sim p(\beta\,|\,\tau), \tag{3}$$

the log of the posterior distribution is then given by

$$-\log p(\beta\,|\,X,y)=(1/2)\sigma_\epsilon^2\sum_i\left(y_i-x_i^T\beta\right)^2+\log p(\beta\,|\,\tau).$$

A regularized maximum a posteriori probability (MAP) estimator can be found by minimizing the negative log-posterior

$$\hat{\beta}_{\text{MAP}}=\underset{\beta\in R^p}{\text{argmin}}\|y-X\beta\|_2^2+\phi_\tau(\beta), \tag{4}$$

where $\phi_\tau(\beta)\propto\log p(\beta\mid\tau)$. The penalty term is interpreted as the log of the prior distribution, and is parameterized by the hyper-parameters τ. The resulting maximum a posteriori probability (MAP) is equivalent to the classical approach of constraining a search space by adding a penalty.

Table 1 provides penalty functions and their corresponding prior distributions, including lasso, ridge, Cauchy and horseshoe.

Figure 1 compares the geometry of a unit ball which is used as a constraint in traditional approach and the corresponding prior distribution as used in Bayesian approach, we show ridge, lasso, Cauchy, and horseshoe penalties.

A typical approach in Bayesian analysis is to define normal scale mixture priors which are constructed as a hierarchical model of the form

$$\beta_i\,|\,\lambda_i,\tau_i\sim N\left(0,\tau_i^2\lambda_i^2\right),\ p\left(\lambda^2,\sigma^2,\tau\right)=p\left(\lambda^2\right)p\left(\sigma^2\right)p(\tau) \tag{5}$$

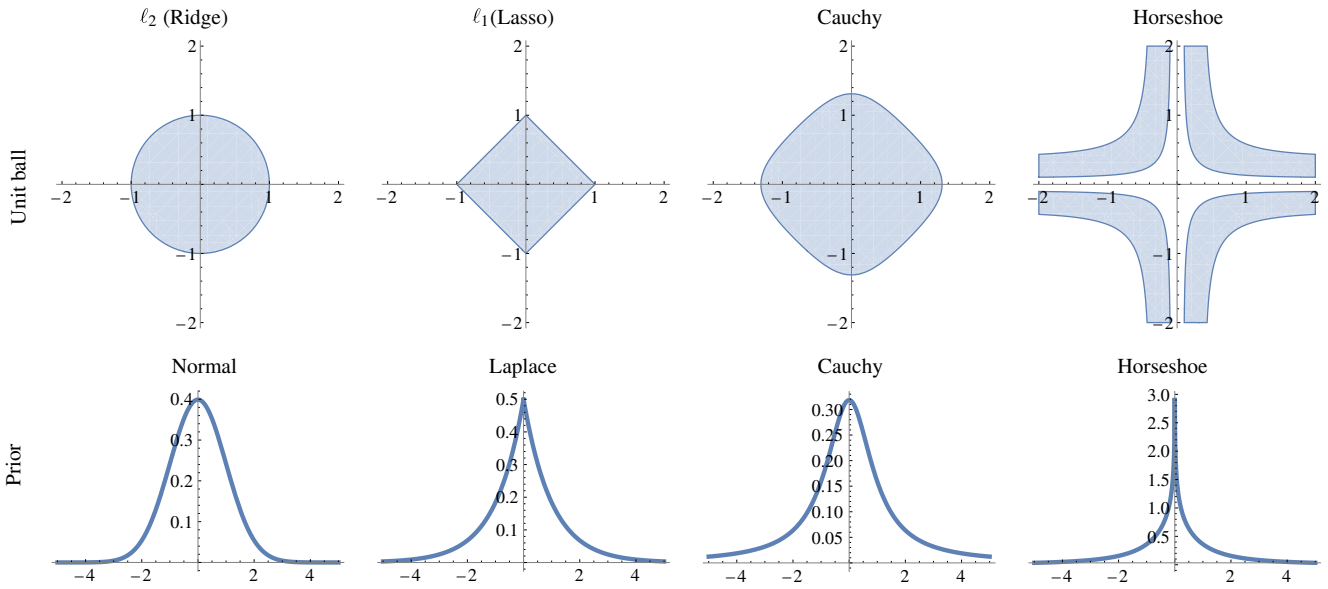

FIGURE 1 Comparison of geometry of a unit ball induced by Normal, Laplace, Cauchy and Horseshoe priors

While classical approach requires solving an optimization problem, the Bayesian approach requires calculating integrals. While in conjugate models, for example, when both likelihood and priors are normal (Ridge), we can calculate those integrals analytically, it is not possible in general case. An efficient numerical techniques for calculating samples from posterior distributions are required. George and McCulloch (1993) proposed a Gibbs sample for finding posterior of the following problem

$$\beta_i \mid \gamma_i \sim (1-\gamma_i)N\left(0,\tau_i^2\right)+\gamma_i N\left(0,c_i^2\tau_i^2\right),\ \ p(\gamma_i=1)=p_i,$$

where τ_i is chosen to be small, so that for $\gamma_i = 0$, the estimated β_i is close to zero and c_i is large so that when $\gamma_i = 1$ the estimated β_i does not get shrunk. Then variable selection is performed by calculating the posterior distribution over γ.

$$p(\gamma \mid X,y) \propto p(y \mid X,\gamma)p(\gamma).$$

Carlin and Polson (1991) proposed Gibbs sampling MCMC for the class of scale mixtures of Normals, taking the form

$$\epsilon_j \mid \sigma,\lambda_j \sim N\left(0,\lambda_j\sigma^2\right),\ \ \lambda_j \sim p\left(\lambda_j\right)$$

We now turn to lasso and horseshoe as special cases.

2.3 | Lasso

From a Bayesian perspective, lasso (Tibshirani, 1996) is equivalent to specifying double exponential (Laplace) prior distribution Carlin and Polson (1991) for each parameter β_i with σ^2 fixed

$$p(\beta_i \mid b) = (b/2)\exp(-|\beta_i|/b).$$

Bayes rule then calculates the posterior as a product of Normal likelihood and the Laplace prior to yield

$$\log p(\beta \mid X,y,b) \propto \|y-X\beta\|_2^2 + \frac{2\sigma^2}{b}\|\beta\|_1.$$

For $b > 0$, the posterior mode is equivalent to the LASSO estimate with $\lambda = 2\sigma^2/b$. Large variance b of the prior is equivalent to the small penalty weight λ in the Lasso objective function.

The Laplace prior used in Lasso can be represented as scale mixture of Normal distribution (Andrews & Mallows, 1974; Carlin & Polson, 1991)

$$\begin{aligned}\beta_i \mid \sigma^2,\tau &\sim N\left(0,\tau^2\sigma^2\right)\\ \tau^2 \mid \alpha &\sim \exp\left(\alpha^2/2\right)\\ \sigma^2 &\sim \pi\left(\sigma^2\right).\end{aligned}$$

There is an equivalence with the lasso penalty obtained by integrating out τ

$$p\left(\beta_i \mid \sigma^2,\alpha\right) = \int_0^\infty \frac{1}{\sqrt{2\pi\tau}}\exp\left(-\frac{\beta_i^2}{2\sigma^2\tau}\right)\frac{\alpha^2}{2}\exp\left(-\frac{\alpha^2\tau}{2}\right)d\tau = \frac{\alpha}{2\sigma}\exp(-\alpha/\sigma|\beta_i|).$$

Thus it is a Laplace distribution with location 0 and scale α/σ.

Carlin and Polson (1991); Carlin, Polson, and Stoffer (1992); Park and Casella (2008) used representation of Laplace prior is a scale Normal mixture to develop a Gibbs sampler that iteratively samples from $\beta|a$, y and $b|\beta$, y to estimate joint distribution over $\left(\hat{\beta},\hat{b}\right)$. Thus, we so not need to apply cross-validation to find optimal value of b, the Bayesian algorithm does it "automatically". Given data $D = (X, y)$, where X is the $n \times p$ matrix of standardized regressors and y is the n-vector of

outputs. Implement a Gibbs sampler for this model when Laplace prior is used for model coefficients β_i. Use scale mixture normal representation.

$$\begin{aligned}
&\beta \mid \sigma^2, \tau_1, \ldots, \tau_p \sim N\left(0, \sigma^2 D_\tau\right) \\
&D_\tau = \operatorname{diag}\left(\tau_1^2, \ldots, \tau_p^2\right) \\
&\tau_i^2 \mid \lambda \sim \exp\left(\lambda^2/2\right) \\
&\sigma^2 \sim 1/\sigma^2.
\end{aligned}$$

Then the complete conditional required for Gibbs sampling are given by

$$\begin{aligned}
&\beta \mid D, D_\tau \sim N\left(A^{-1}X^T y, \sigma^2 A^{-1}\right),\ A = X^T X + D_\tau^{-1} \\
&\sigma^2 \mid \beta, D, D_\tau \sim \text{InverseGamma}\left((n-1)/2 + p/2, (y - X\beta)^T (y - X\beta)/2 + \beta^T D_\tau^{-1}\beta/2\right) \\
&1/\tau_j^2 \mid \beta_j, \lambda \sim \text{InverseGaussian}\left(\sqrt{\frac{\lambda^2\sigma^2}{\beta_j^2}}, \lambda^2\right)
\end{aligned}$$

The formulas above assume that X is standardized, for example, observations for each feature are scaled to be of mean 0 and standard deviation one, and y is centered $y = y - \bar{y}$.

You can use empirical priors and initialize the parameters as follows

$$\begin{aligned}
&\beta = \left(X^T X + I\right)^{-1} X^T y \\
&r = y - X\beta \\
&\sigma^2 = r^T r/n \\
&\tau^{-2} = 1/(\beta \odot \beta) \\
&\lambda = p\sqrt{\sigma^2}/\sum |\beta|.
\end{aligned}$$

Here n is number of rows (observations) and p is number of columns (inputs) in matrix X.

2.4 | Ridge

When the prior is Normal $\beta_i \sim N\left(0, \sigma_\beta^2\right)$, the posterior mode is equivalent to the ridge estimate Hoerl and Kennard (1970) . The relation between variance of the prior and the penalty weight in ridge regression is inverse proportional $\lambda \propto 1/\sigma_\beta^2$.

Thus, Lasso and Ridge regressions are both maximum a posteriori (MAP) estimates for Laplace and Normal priors.

Given design matrix X and observed output values $y = (y_1, \ldots, y_n)$, and assuming $\epsilon \sim N(0, \sigma^2)$, the MLE is given by the solution to the following optimization problem

$$\text{minimize}_\beta \quad \left\|X\beta^T - y\right\|_2^2$$

and the solution is given by:

$$\beta = \left(X^T X\right)^{-1} X^T y.$$

However, when matrix X is close to being rank-deficient, the $X^T X$ will be ill-conditioned. This means that the problem of estimating β will also be ill-conditioned. For a linear model, we can quantify the sensitivity to perturbation in y by

$$\frac{\|\Delta\beta\|}{\|\beta\|} \leq \frac{\kappa(X^T X)}{\cos\theta}\frac{\|\delta X^T y\|}{\|X^T y\|},$$

here θ is the angle between $X^T y$ and the range of $X^T X$ and $\kappa(X^T X)$ is the condition number which is the ratio of the largest to smallest eigenvalues of $X^T X$.

A trivial example is shown when y is nearly orthogonal to x

$$x = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \; y^{(1)} = \begin{bmatrix} \epsilon \\ 1 \end{bmatrix}$$

The solution to the problem is $\beta^{(1)} = \epsilon$; but the solution for

$$x = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \; y^{(2)} = \begin{bmatrix} -\epsilon \\ 1 \end{bmatrix}$$

is $\hat{\beta}^{(2)} = -\epsilon$. Note that $\|y^{(1)} - y^{(2)}\| / \|y^{(1)}\| \approx 2\epsilon$ is small, but $|\hat{\beta}^{(1)} - \hat{\beta}^{(2)}| / |\hat{\beta}^{(1)}| = 2$, is huge.

Another case of interest is when a least squares problem is ill-conditioned is when the observations are close to be linearly dependent. It happens, for example, when input variables are correlated. Consider an example

$$X = \begin{pmatrix} 1 & 1 \\ 1 & \epsilon + 1 \end{pmatrix}, \; y = \begin{pmatrix} 2 \\ \delta + 2 \end{pmatrix}$$

The MLE estimate is given by

$$\beta = \left\{ 2 - \frac{\delta}{\epsilon}, \frac{\delta}{\epsilon} \right\}$$

For $\delta = 0$, we have $\hat{\beta}^{(1)} = (2, 0)$ but for $\delta = \epsilon$, we have $\hat{\beta}^{(2)} = (1, 1)$ with both ϵ and δ being arbitrarily small. We can analytically calculate the condition number

$$\kappa(X^T X) = \frac{\epsilon^2 + (\epsilon + 2)\sqrt{\epsilon^2 + 4} + 2\epsilon + 4}{\epsilon^2 - (\epsilon + 2)\sqrt{\epsilon^2 + 4} + 2\epsilon + 4}$$

It goes to infinity as ϵ goes to zero. Since condition number is the ratio of eigenvalues

$$\kappa(X^T X) = \frac{\lambda_{\max}(X^T X)}{\lambda_{\min}(X^T X)}$$

and in our case $\lambda_{\min}(X^T X)$ is close to zero, we can improve the condition number by shifting the spectrum $\lambda(A + \alpha I) = \lambda(A) + \alpha$, thus

$$\kappa(X^T X + \alpha I) = \frac{\lambda_{\max}(X^T X) + \alpha}{\lambda_{\min}(X^T X) + \alpha}.$$

Figure 2 compares the condition number of the original $X^T X$ matrix and the one with spectrum shifted by one $X^T X + I$.

Thus, the spectrum shift allows to address the issue of numerical instability when $X^T X$ is ill-conditioned, which is always a case whenever p is large. The solution is then given by

$$\hat{\beta} = (X^T X + \lambda I)^{-1} X^T y.$$

The corresponding objective function that leads to this regularized solution is

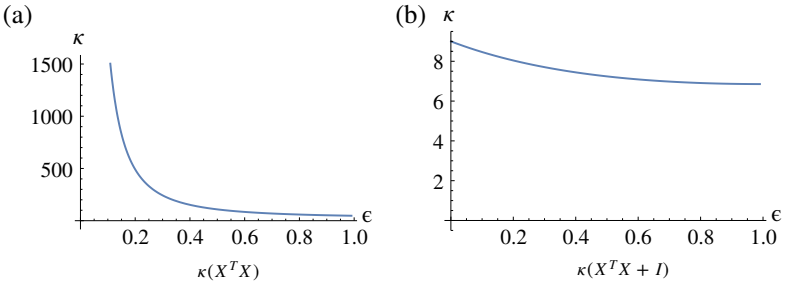

FIGURE 2 Condition number of original problem (left) and the regularized one (right)

$$\underset{\beta}{\text{minimize}} \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2. \tag{6}$$

An alternative formulation is

$$\underset{\beta}{\text{minimize}} \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 \quad \text{subject to} \quad \|\beta\|_2^2 \le s. \tag{7}$$

We can think of the constrain is of a budget on the size of β. In statistics the problem of solving (6) is called ridge regression.

2.5 | Spike-and-slab prior

Under spike-and-slab, prior for each β_i is defined as a mixture of a point mass at zero, and a Gaussian distribution centered at zero

$$\beta_i \,|\, \theta, \sigma^2 \sim (1-\theta)\delta_0 + \theta N\left(0, \sigma^2\right) . \tag{8}$$

Here $\theta \in (0, 1)$ determines the overall sparsity in β and σ^2 accommodates nonzero signals. This family is termed as the Bernoulli-Gaussian mixture model in the signal processing community.

A useful reparametrization, the parameters β is given by two independent random variable vectors $\gamma = (\gamma_1, \ \dots \ , \gamma_p)$ and $\alpha = (\alpha_1, \ \dots \ , \alpha_p)$ such that $\beta_i = \gamma_i\alpha_i$, with probabilistic structure

$$\begin{aligned} \gamma_i \,|\, \theta &\sim \text{Bernoulli}(\theta) \ ; \\ \alpha_i \,|\, \sigma^2 &\sim N\left(0, \sigma^2\right) . \end{aligned} \tag{9}$$

Since γ_i and α_i are independent, the joint prior density becomes

$$p\left(\gamma_i, \alpha_i \,|\, \theta, \sigma^2\right) = \theta^{\gamma_i}(1-\theta)^{1-\gamma_i} \frac{1}{\sqrt{2\pi}\sigma_\beta} \exp\left\{ -\frac{\alpha_i^2}{2\sigma^2} \right\} , \quad \text{for } 1 \le i \le p \ .$$

The indicator $\gamma_i \in \{0, 1\}$ can be viewed as a dummy variable to indicate whether β_i is included in the model. Under this re-parameterization, the posterior is given by

$$\begin{aligned} &-\log p\left(\gamma, \alpha \,|\, \theta, \sigma^2, \sigma_e^2, y\right) \propto -\log p\left(\gamma, \alpha \,|\, \theta, \sigma^2\right) p\left(y \,|\, \gamma, \alpha, \theta, \sigma_e^2\right) \\ &\propto \frac{1}{2\sigma_e^2}\left\|y - X_\gamma \alpha_\gamma\right\|_2^2 + \frac{1}{2\sigma^2}\|\alpha\|_2^2 + \log\left(\frac{1-\theta}{\theta}\right)\sum_{i=1}^{p}\gamma_i. \end{aligned}$$

By construction, the γ.

$\in \{0, 1\}^p$ will directly perform variable selection. Note, that the problem of minimizing the negative log-posterior is a mixed integer program with each γ_1 being constraint to take values 0 or 1. This optimization problem is NP-hard, for example,

we cannot solve it efficiently for any meaningful value of p. Efficient algorithms for MAP estimation for high-dimensional linear models were proposed in Moran, Rockova, and George (2018); Ročková and George (2018). A sampling algorithm was proposed in Atchade and Bhattacharyya (2018) For a recent review of sampling algorithms for spike–and-slab, see Rockova and McAlinn (2017).

3 | HORSESHOE

In a global-local class of priors, τ does not depend on index i, therefore, we have

$$\beta_i \,|\, \lambda_i, \tau \sim N\left(0, \tau^2 \lambda_i^2\right).$$

Global hyper-parameter τ shrinks all parameters towards zero, while the prior for the local parameter λ_i has a tail that decays slower than an exponential rate, and thus allows β_i not to be shrunk. A particular representative of global–local shrinkage prior is horseshoe, which assumes half-Cauchy distribution over λ_i and τ

$$\lambda_i \sim C^+(0,1), \;\; \tau \sim C^+(0,1).$$

Being constant at the origin, the half-Cauchy prior has nice risk properties near the origin (Polson & Scott, 2009). Polson and Scott (2010) warn against using empirical-Bayes or cross-validation approaches to estimate τ, due to the fact that MLE estimate of τ is always in danger of collapsing to the degenerate $\hat{\tau} = 0$ (Tiao & Tan, 1965).

A feature of the horseshoe prior is that it possesses both tail-robustness and sparse-robustness properties (Bhadra, Datta, Polson, Willard, et al., 2017); meaning that an infinite spike at the origin and very heavy tail that still ensures integrability. The horseshoe prior can also be specified as

$$\beta_i \,|\, \lambda_i, \tau \sim N\left(0, \lambda_i^2\right), \;\; \lambda_i \,|\, \tau \sim C^+(0,\tau), \;\; \tau \sim C^+(0,1)$$

The log-prior of the horseshoe cannot be calculated analytically, but a tight lower bound (Carvalho, Polson, & Scott, 2010) can be used instead

$$\phi_{HS}(\beta_i|\tau) = -\log p_{HS}(\beta_i|\tau) \geq -\log\log\left(1 + \frac{2\tau^2}{\beta_i^2}\right). \tag{10}$$

The motivation for the horseshoe penalty arises from the analysis of the prior mass and influence on the posterior in **both** the tail and behavior at the origin. The latter provides the key determinate of the sparsity properties of the estimator (Figure 3).

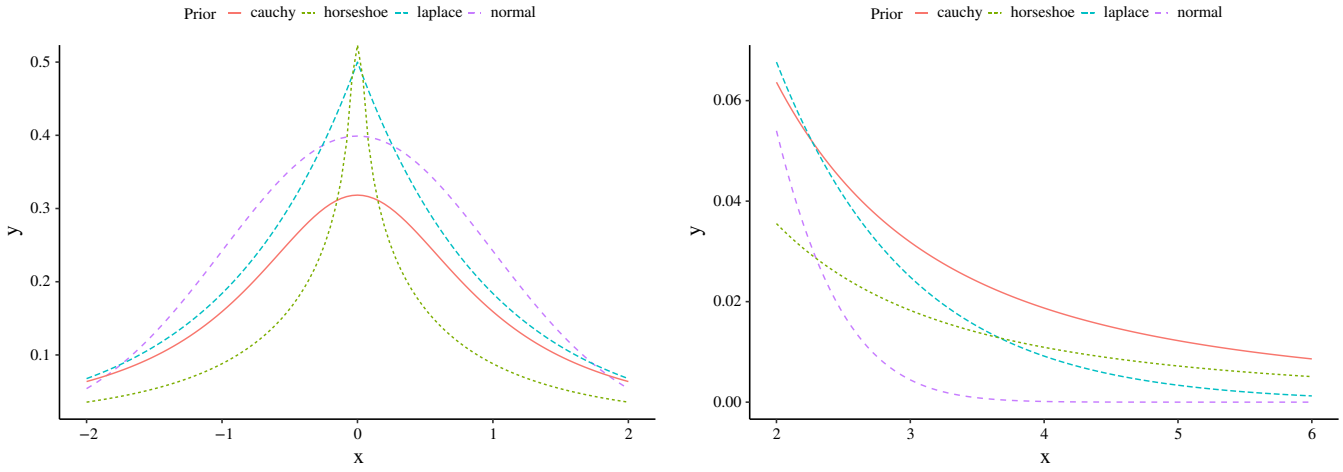

FIGURE 3 Comparison of Laplace (LASSO), Normal (Ridge), Cauchy and Horseshoe priors

When Metropolis-Hasting MCMC is applied to horseshoe regression, it suffers from sampling issues. The funnel shape geometry of the horseshoe prior is makes it challenging for MCMC to efficiently explore the parameter space. Piironen, Vehtari, et al. (2017) proposed to replace Cauchy prior with half-t prior with small degrees of freedom and showed improved convergence behavior for NUTS sampler Hoffman and Gelman (2014). Makalic and Schmidt (2016) proposed using a scale mixture representation of half-Cauchy which leads to conjugate hierarchy and allows a Gibbs sample to be used. Johndrow, Orenstein, and Bhattacharya (2017) proposed two MCMC algorithms to calculate posteriors for horseshoe priors. The first algorithm addresses computational cost problem in high dimensions by approximating matrix–matrix multiplication operations. For further details on computational issues and packages for horseshoe sampling, see Bhadra, Datta, Polson, and Willard (2017). An issue of high dimensionality was also addressed by Bhattacharya, Chakraborty, and Mallick (2016).

One approach is to replace the thick-tailed half-Cauchy prior over λ_j with half-t priors using small degrees of freedom. This leads to the sparsity-sampling efficiency trade-off problem. Larger degrees of freedom for a half-t distribution will lead to more efficient sampling algorithms, but will be less sparsity inducing. For cases with large degrees of freedom, tails of half-t are slimmer and we are required to choose large τ to accommodate large signals. However, priors with a large τ are not able to shrink coefficients towards zero as much.

4 | EMPIRICAL RESULTS

We use the half-Cauchy priors and a slice sampler for Bayesian linear regression models proposed in Hahn, He, and Lopes (in press) and implemented in the **`bayesml`** package. The sampler does not rely on latent variables and it is automated, so that it can work with any prior that can be evaluated up to a normalizing constant. The **`bayeslm`** package uses an elliptical slice sampler and can efficiently handle high-dimensional problems. Besides horseshoe priors, it also supports and spike-and-slab priors.

We then apply the slice sample to a synthetic data set. This data is generated by setting $\beta = (2, 2.5, 3, 0, 0, 0, 0, 0, 0, 0)$, generating matrix $X \in R^{100 \times 10}$ of 100 samples, with each uniformly distributed in $[-1, 1]$. We also set $y = X\beta + e$ with $e_i \sim N\left(0, \kappa^2 \|\beta\|_2^2\right)$ with $\kappa = 1$.

Figure 4 shows the MAP estimates using different prior assumptions as well as ordinary least squares (OLS) estimated coefficients. We can see that horseshoe was the only approach to correctly identify all zero-valued coefficients. Nonzero coefficients were recovered with a similar level of accuracy by all four methods, but we can see the shrinkage effect of the lasso estimator.

5 | CONCLUSION

There are several major advantages to using the Bayesian approach compared to the classical regularization method:

- It allows for a more flexible set of models that closely match the data generating process, and assumptions appear explicitly in the model.

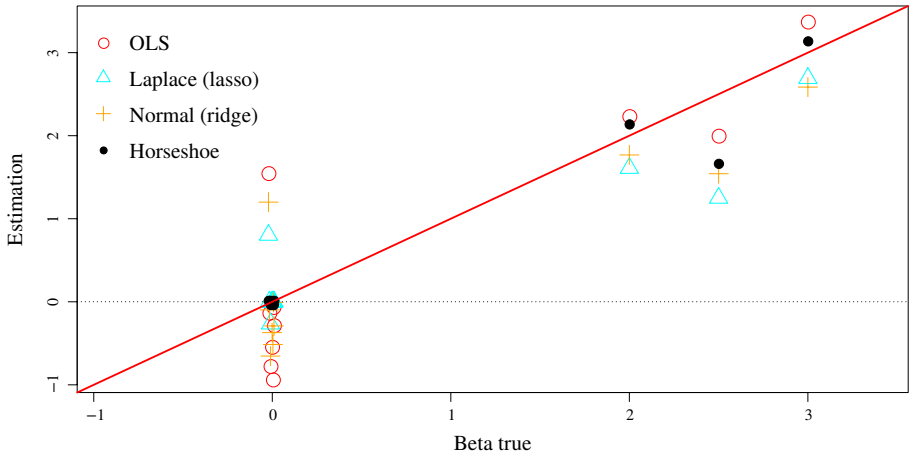

FIGURE 4 Posterior mode for each of each of the 10 betas estimated using Laplace, normal and horseshoe Bayesian models as well as OLS estimates

- Bayesian sampling algorithms are flexible enough and existing libraries can easily handle a wide range of model formulations without the need to design custom algorithms and implementations
- Bayesian estimates are optimal on the bias-variance scale. The parameters of the prior distribution (penalty function parameters) can be estimated using the training data set (X, y) (Kitagawa & Gersch, 1985) rather using brute-force search.
- Bayesian estimation procedures result in distributions over parameters and enable improved analysis of uncertainty in estimates and predictions.
- Ability to incorporate prior information based on expert opinion or previously observed data.

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

RELATED WIREs ARTICLES

Inverse problems: From regularization to Bayesian inference
Variable selection in linear models

REFERENCES

Alliney, S. (1992). Digital filters as absolute norm regularizers. *IEEE Transactions on Signal Processing*, *40*(6), 1548–1562.

Alliney, S., & Ruzinsky, S. (1994). An algorithm for the minimization of mixed l/sub 1/and l/sub 2/norms with application to bayesian estimation. *IEEE Transactions on Signal Processing*, *42*(3), 618–627.

Andrews, D. F., & Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *36*, 99–102.

Aster, R. C., Borchers, B., & Thurber, C. H. (2018). *Parameter estimation and inverse problems*. Amsterdam: Elsevier.

Atchade, Y., & Bhattacharyya, A. (2018). Regularization and computation with high-dimensional spike-and-slab posterior distributions. *arXiv preprint arXiv*, *1803*, 10282.

Bakushinskii, A. B. (1967). A general method of constructing regularizing algorithms for a linear incorrect equation in hilbert space. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, *7*(3), 672–677 [English translation: *U.S.S.R. Comput. Math. Math. Phys.*, 7(3) (1967), pp. 279—287].

Bhadra, A., Datta, J., Polson, N. G., & Willard, B. (2017). The horseshoe+ estimator of ultra-sparse signals. *Bayesian Analysis*, *12*(4), 1105–1131.

Bhadra, A., Datta, J., Polson, N. G., & Willard, B. T. (2017). Lasso meets horseshoe. *arXiv preprint arXiv*, *1706*, 10179.

Bhattacharya, A., Chakraborty, A., & Mallick, B. K. (2016). Fast sampling with gaussian scale mixture priors in high-dimensional regression. *Biometrika*, *103*(4).

Candès, E. J., & Wakin, M. B. (2008). An introduction to compressive sampling [a sensing/sampling paradigm that goes against the common knowledge in data acquisition]. *IEEE Signal Processing Magazine*, *25*(2), 21–30.

Carlin, B. P., & Polson, N. G. (1991). Inference for nonconjugate bayesian models using the gibbs sampler. *Canadian Journal of Statistics*, *19*(4), 399–405.

Carlin, B. P., Polson, N. G., & Stoffer, D. S. (1992). A Monte Carlo approach to nonnormal and nonlinear state-space modeling. *Journal of the American Statistical Association*, *87*(418), 493–500.

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, *97*(2), 465–480.

Claerbout, J. F., & Muir, F. (1973). Robust modeling with erratic data. *Geophysics*, *38*(5), 826–844.

Donoho, D. L. (1992). Superresolution via sparsity constraints. *SIAM Journal on Mathematical Analysis*, *23*(5), 1309–1331.

Donoho, D. L., & Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, *90*(432), 1200–1224.

George, E. I., & McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, *88*(423), 881–889.

Hahn, P. R., He, J., & Lopes, H. F. (in press). Efficient sampling for gaussian linear regression with arbitrary priors. *Journal of Computational and Graphical Statistics*.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Hoffman, M. D., & Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623.

Ivanov, V. K. (1962). On linear problems which are not well-posed. In *Doklady akademii nauk* (Vol. 145, pp. 270–272). Moscow: Russian Academy of Sciences.

Ivanov, V. K., Vasin, V. V., & Tanana, V. P. (2013). *Theory of linear ill-posed problems and its applications* (Vol. 36). Berlin: Walter de Gruyter.

Johndrow, J. E., Orenstein, P., & Bhattacharya, A. (2017). Scalable mcmc for bayes shrinkage priors. *arXiv preprint arXiv*, *1705*, 00841.

Kitagawa, G., & Gersch, W. (1985). A smoothness priors time-varying ar coefficient modeling of nonstationary covariance time series. *IEEE Transactions on Automatic Control*, *30*(1), 48–56.

Makalic, E., & Schmidt, D. F. (2016). A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, *23*(1), 179–182.

Miller, A. (2002). *Subset selection in regression*. Boca Raton: Chapman and Hall/CRC.

Moran, G. E., Rockova, V., & George, E. I. (2018). Variance prior forms for high-dimensional bayesian variable selection. *arXiv preprint arXiv*, *1801*, 03019.

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

Piironen, J., & Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, *11*(2), 5018–5051.

Polson, N. G., & Scott, J. G. (2009). *Alternative global–local shrinkage rules using hypergeometric-beta mixtures.* Technical report 14.

Polson, N. G., & Scott, J. G. (2010). Shrink globally, act locally: Sparse bayesian regularization and prediction. *Bayesian Statistics*, *9*, 501–538.

Polson, N. G., & Sun, L. (in press). Bayesian l 0-regularized least squares. *Applied Stochastic Models in Business and Industry.*

Ročková, V., & George, E. I. (2018). The spike-and-slab lasso. *Journal of the American Statistical Association*, *113*(521), 431–444.

Rockova, V., & McAlinn, K. (2017). Dynamic variable selection with spike-and-slab process priors. *arXiv preprint arXiv*, *1708*, 00085.

Stein, C. (1964). Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean. *Annals of the Institute of Statistical Mathematics*, *16*(1), 155–160.

Taylor, H. L., Banks, S. C., & McCoy, J. F. (1979). Deconvolution with the ℓ_1 norm. *Geophysics*, *44*(1), 39–52.

Tiao, G. C., & Tan, W. (1965). Bayesian analysis of random-effect models in the analysis of variance. i. Posterior distribution of variance-components. *Biometrika*, *52*(1/2), 37–53.

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

Tihonov, A. N. (1963). Solution of incorrectly formulated problems and the regularization method. *Soviet Math.*, *4*, 1035–1038.

Tikhonov, A., & Arsenin, V. Y. (1977). *Methods for solving ill-posed problems*. Hoboken: John Wiley and Sons, Inc.

Tikhonov, A. N. (1943). On the stability of inverse problems. In *Doklady akademii nauk SSSR* (Vol. 39, pp. 195–198). Moscow.

Williams, P. M. (1995). Bayesian regularization and pruning using a Laplace prior. *Neural Computation*, *7*(1), 117–143.

How to cite this article: Polson NG, Sokolov V. Bayesian regularization: From Tikhonov to horseshoe. *WIREs Comput Stat*. 2019;11:e1463. https://doi.org/10.1002/wics.1463