The horseshoe prior and its limitations

Gaëtan Ecrepont
ENS Paris-Saclay
Paris, FRANCE
gaetan.ecrepont@polytechnique.edu

Guillaume Martin
ENS Paris-Saclay
Paris, FRANCE
guillaume.martin@polytechnique.edu

Abstract

The horseshoe prior is a popular choice for Bayesian variable selection in the context of sparse supervised learning. After presenting the horseshoe and its key properties, we argue that it is more robust than the popular LASSO method through an experiment. We then discuss a classic limitation of the horseshoe prior and highlight solutions that have been proposed in the literature.

Introduction

Supervised learning is the branch of machine learning that deals with the prediction of a target variable y given a set of predictors $\mathbf{x} = (x_1, \dots, x_p)$. The mapping from $\mathbf{x}$ to y is approximated by learning a vector of coefficients $\beta = (\beta_1, \dots, \beta_p)$ that minimizes a loss function. In the canonical case of linear regression, the loss function is the mean squared error (MSE) and the estimation is given by $\hat{y} = \mathbf{x}^T\beta$.

Supervised learning is a very general framework which encompasses many estimation problems including regression, classification and function estimation. Though it is a well-understood domain of machine learning, supervised learning becomes challenging when the number of predictors p becomes large, in particular if the number of observations n is comparatively small. In this case the estimation of β can be very far from the ground truth if we do not take precautions. Intuitively, we want to reduce the dimensionality p of the problem to make the estimation more robust. We do so by looking for *sparse* solutions, meaning that we want to set most coefficients of β to zero, which amounts to performing variable selection.

The most straightforward way to enforce sparsity to our estimated coefficients β is to add a sparsity-inducing penalty term to the loss function. The most used penalty is the L^1 norm of β, which is the basis of the LASSO (Least Absolute Shrinkage and Selection Operator) method. One can also consider using the L^0 norm of β, which counts the number of non-zero coefficients, but this is a combinatorial problem and is computationally intractable.

A more principled way to enforce sparsity is to endow β with a prior distribution $p(\beta)$ that attributes a high probability to zero or near-zeros values. This is the Bayesian approach to variable selection. Interestingly, the LASSO can be recovered as the Maximum A Posterior (MAP) estimator of β when its prior is the Laplace distribution. Likewise, L^0 regularization can be seen as the MAP estimator of β when its prior is a mixture of a point mass at zero and a Gaussian distribution. [4]

In fact, LASSO and L^0 regularization illustrate the two main approaches to Bayesian variable selection: either we specify a *shrinkage prior* that tends to bring coefficients β_i closer to 0, or we adopt the so-called *spike-and-slab prior* that allocates a positive probability mass to $\beta_i = 0$ combined with a continuous alternative. Intuitively, the point mass at zero makes the spike-and-slab much superior to the shrinkage approach. In practice however, shrinkage priors such as LASSO are much easier to implement and computationally cheaper, making them the default choice in many applications.

Preprint. Under review.

In an attempt to reconcile the computational tractability of shrinkage priors with the superior performance of spike-and-slab priors, Carvalho et al. introduced the horseshoe prior in 2009 [2]. The horseshoe prior belongs to the family of shrinkage priors but it is designed to mimick the properties of the spike-and-slab approach. It has since become a popular choice in Bayesian variable selection.

In this paper, we first present the horseshoe prior and discuss its key properties before comparing it to the LASSO. We then discuss a classic limitation of the horseshoe prior: the choice of a prior for τ. In particular, our contribution is to propose a heuristic to choose a prior for τ when we believe the number of non-zero parameters to be around $\hat{p_0}$.

1 The horseshoe prior

In this section, instead of the classic linear model $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2\mathbf{I})$, we will consider for the sake of interpretation the simplified model $\mathbf{y} \sim \mathcal{N}(\beta, \sigma^2\mathbf{I})$, where $\mathbf{y}$ is the vector of observations, β is the vector of coefficients and σ^2 is the variance of the noise.

We have n observations $\mathbf{y}_1, \dots, \mathbf{y}_n$ and p predictors $x_1, \dots, x_p$.

1.1 Definition

The horseshoe is a simple hierarchical prior defined as follows:

$$\beta_i \sim \mathcal{N}(0, \lambda_i^2\tau^2)$$
$$\lambda_i \sim \mathrm{C}^+(0, 1)$$

where $\mathrm{C}^+(0, 1)$ denotes the half-Cauchy distribution with location 0 and scale 1.

The parameter τ is a *global* shrinkage parameter that controls the amount of shrinkage applied to β overall, whereas the λ_i are *local* shrinkage parameters that control the amount of shrinkage applied to each individual coefficient β_i. Note that τ can either be set as a hyperparameter by the user or it can be endowed with a prior distribution. In the original paper, the authors advocate for a fully Bayesian approach and set $\tau \sim \mathrm{C}^+(0, 1)$.

We can write the marginal distribution of β_i with τ fixed as

$$p(\beta_i|\tau) = \int p(\beta_i|\lambda_i, \tau)p(\lambda_i)d\lambda_i = \int \mathcal{N}(\beta_i|0, \lambda_i^2\tau^2)p(\lambda_i)d\lambda_i. \tag{1}$$

As such, the horseshoe prior is a *scale mixture* of normal distributions.

Note that using the half-Cauchy distribution for the λ_i is a design choice that allows for heavy tails in the marginal distribution of β_i. These heavy tails ensure that large signals are detected. In fact, although there is not closed-form expression of the marginal distribution of β_i, it can be shown that $p(\beta_i|\tau) \propto \log(1 + \frac{2}{\beta_i^2})$ such that $p(\beta_i|\tau) \propto_{\beta_i \to \pm\infty} \frac{1}{\beta_i^2}$ i.e. β_i has Cauchy-like tails. In addition, we have $p(\beta_i|\tau) \xrightarrow[\beta_i \to 0]{} +\infty$ which means that the horseshoe prior has an infinite spike at zero and as such it can also provide severe shrinkage.

Interestingly, different priors on the λ_i can provide closed-form expressions for the marginal distribution of β_i given in (1). In particular, setting $\lambda_i^2 \sim \mathrm{Exp}(2)$ leads to a Laplace prior on β_i. As such, LASSO can also be seen as a scale mixture of normals, but with prior on λ_i that is not heavy-tailed. Likewise, setting $\lambda_i^2 \sim \mathrm{InvGamma}(\alpha, \beta)$ leads to Student-t prior on β_i. To illustrate this, we set $\tau = 1$ and $\sigma = 1$ and specify three different priors for λ_i: $\lambda_i \sim \mathrm{C}^+(0, 1)$, $\lambda_i \sim \mathrm{Exp}(2)$ and $\lambda_i \sim \mathrm{InvGamma}(1, 1)$. We then sample $M = 10^6$ values of β_i from the marginal distribution on β_i for each prior and plot the results in Figure 1. We use the `numpyro` python library for the sampling.

1.2 Key properties

The three key properties of the horseshoe prior are:

1. it has Cauchy-like tails, allowing for the detection of large signals *a posteriori* and without shrinking them

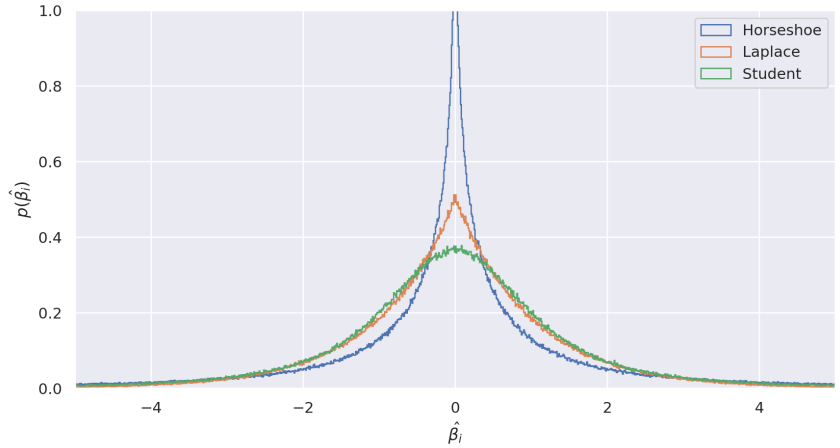

Figure 1: Marginal distribution of β_i for different priors on λ_i. We observe that the horseshoe prior (half-Cauchy prior on the λ_i) has the heaviest tails and an infinite spike at zero, making it an interesting candidate for a prior on β_i.

2. it has an infinite spike at zero, meaning that it can provide severe shrinkage to small signals
3. it is a global-local shrinkage prior, meaning that it can leverage τ to adapt to the *overall* sparsity of the problem while allowing individual coefficients to escape τ's shrinkage effect with the help of the λ_i

In a nutshell, the horseshoe prior is versatile because it can both detect strong signals and shrink weak signals, which is a desirable property in variable selection problems. One way to intuitively understand why the horseshoe is more versatile than other classic priors is to look at $\kappa_i = \frac{1}{1+\lambda_i^2}$(We place ourselves in the case $\tau = 1, \sigma = 1$ for this formula), which can be interpreted as the mass that the posterior mean for β_i places at zero once the data $\mathbf{y}$ has been observed. Indeed, by Bayes' rule for Gaussians we have $p(\beta_i|\lambda_i, y_i) = \mathcal{N}(\beta_i; (1-\kappa_i)y_i, 1-\kappa_i)$ such that

$$\mathbb{E}[\beta_i|\lambda_i, y_i] = (1-\kappa_i)y_i = (1-\kappa_i)y_i + \kappa_i \times 0, \tag{2}$$

which in turn implies

$$\mathbb{E}[\beta_i|y_i] = (1 - \mathbb{E}[\kappa_i|y_i])y_i. \tag{3}$$

Thus, studying the priors on κ_i induced by different priors on λ_i gives us an idea of how the model tries to distinguish noise from signal. We again sample $M = 10^6$ values of λ_i from the three priors and plot the resulting distribution on κ_i in Figure 2.

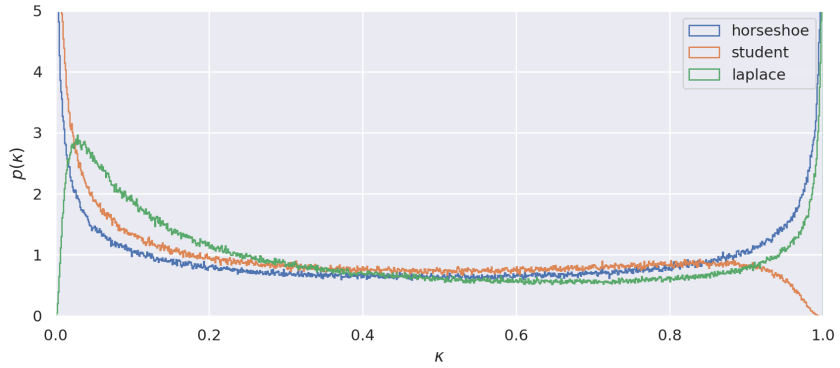

Figure 2: Distribution of κ_i for different priors on λ_i. We observe that the horseshoe prior (half-Cauchy prior on the λ_i) is the only prior that has infinite spikes both at zero and one, meaning that it can provide both no shrinkage ($\kappa_i \simeq 0$) and strong shrinkage ($\kappa_i \simeq 1$).

1.3 Comparison to LASSO

A crucial property for an estimator is robustness. In the case of the linear model $\mathbf{y} \sim \mathcal{N}(\beta, \sigma^2\mathbf{I})$, we would like an estimator of β which still performs well when $\mathbf{y}$ is very far from its prior mean i.e. when there are some $|y_i| \gg 1$.

A very simple setting to study robustness is to consider the one-dimensional single-sample observation model $y \sim \mathcal{N}(\beta, 1)$ (note that we fix $\sigma = 1$ without loss of generality) and then compute the posterior mean of β conditional on the unique sample y. To help us, we have the following representation theorem (Polson, 1991):

$$\mathbb{E}[\beta|y] = y + \nabla_y \ln p(y), \tag{4}$$

where $p(y) = \int_\beta^(y|\beta)p(\beta)d\beta$ is the marginal density for y. This representation hints that the tail behavior of $\nabla_y \ln p(y)$ is critical to the robustness of our estimator of β for large $|y|$. In particular, a robust estimator should verify the "bounded influence" property i.e. its score function $\nabla_y \ln p(y)$ should be bounded as $|y| \to \infty$. One can show that this is the case for both the horseshoe and the Laplace prior, but with a key difference:

$$\nabla_y \ln p_{\text{horseshoe}}(y) \xrightarrow[|y|\to\infty]{} 0, \tag{5}$$

whereas

$$\nabla_y \ln p_{\text{laplace}}(y) \xrightarrow[|y|\to\infty]{} \pm a, \tag{6}$$

where a is called the "nonrobustness parameter" and varies inversely with the global shrinkage parameter τ. This implies that in sparse supervised learning, LASSO will perform significantly worse than the horseshoe when the data $\mathbf{y}$ is far from its prior mean. And this effect will only grow worse as the sparsity of the problem increases.

This key difference in tail behavior is illustrated in Figure 3 where we plot the $\mathbb{E}[\beta|y]$ for the horseshoe and the Laplace prior as a function of y. Note that in our simplified regression setting the Laplace prior amounts to a soft-thresholding of the data i.e. $\hat{\beta}^{\text{LASSO}} = S_\Lambda(\hat{\beta}^{\text{OLS}})$ where S_Λ is the soft-thresholding operation with threshold Λ.

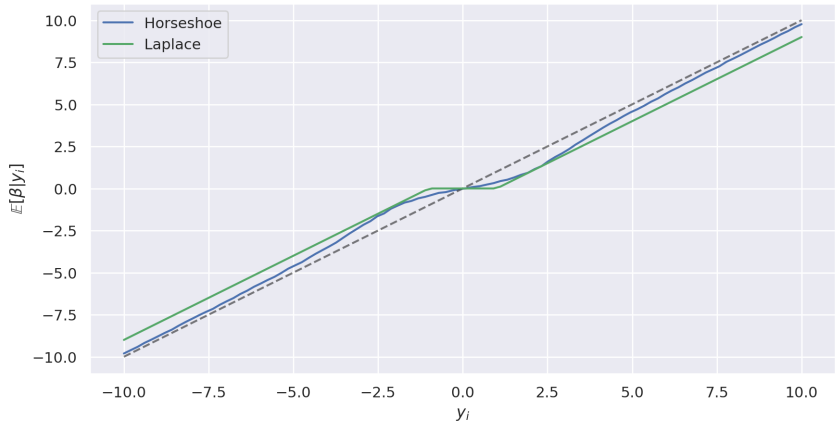

Figure 3: Posterior mean of β for the horseshoe and the Laplace prior as a function of y. We observe that both methods effectively shrink small signals by "bending" around zero. However the Laplace prior exhibits a "nonrobustness bias" for large values of $|y|$ whereas the horseshoe leaves large signals mostly unshrunk.

2 Choosing τ

The original paper advocates for a hierarchical Bayes approach for the choice of τ with a prior $\tau \sim \text{C}^+(0, 1)$ following [1]. This prior is chosen as it is quite uninformative and enables both small and large τ values thanks to its large tails. In this section, we will show intuitively that such a choice is questionable as pointed out in [3].

The κ_i coefficients have a high probability of being very close to either 0 (no shrinkage) or 1 (complete shrinkage) (see Figure 2). Thanks to this, we can effectively compute the effective number of non-zero parameters as

$$m_{eff} = \sum_{i=1}^{p} (1 - \kappa_i)$$

We will now study the prior induced on m_{eff} by τ to see that the choice of half Cauchy distribution of unit scale might not be appropriate.

The more general formula for κ_i (when we have n samples and τ, σ are not equal to 1) is:

$$\kappa_i = \frac{1}{1 + n\sigma^{-2}\tau^2\lambda_i^2}$$

such that:

$$\mathbb{E}(\beta \mid \tau, \sigma, \kappa_i, y_i^1, ..., y_i^n) = (1 - \kappa_i)\hat{\beta}^{MLE}$$

This formula for κ_i gives us:

$$\mathbb{E}(\kappa_i \mid \sigma^2, \tau) = \frac{1}{1 + \tau\sigma^{-1}\sqrt{n}}$$

which in turn yields, by linearity of the expectation:

$$\mathbb{E}(m_{eff} \mid \sigma^2, \tau) = \frac{\tau\sigma^{-1}\sqrt{n}}{1 + \tau\sigma^{-1}\sqrt{n}} p$$

Let's say we now have a given prior guess $\hat{p_0}$ for the number of effective parameters. In order to have $\mathbb{E}(m_{eff} \mid \sigma^2, \tau) = \hat{p_0}$, we should thus set τ close to:

$$\tau_0 = \frac{\sigma}{\sqrt{n}} \frac{\hat{p_0}}{p - \hat{p_0}}. \tag{7}$$

Thanks to (7), we can now design priors for τ which are more informed than the default $\mathrm{C}^+(0, 1)$. We can for instance propose $\delta_{\tau_0}, \mathcal{N}^+(0, \tau_0^2)$ or $\mathrm{C}^+(0, \tau_0^2)$. These priors in turn induce different priors on m_{eff}, which we present in Figure 4 and compare to the default prior.

To begin with, note that the default $\mathrm{C}^+(0, 1)$ prior implies a heavily skewed prior on m_{eff}, which does not encourage sparsity. On the contrary, all the other proposed priors are centered around $\hat{p_0}$ as expected and result in much more reasonable values of m_{eff} ; as such, these priors seem more relevant in the context of sparse estimation.

In particular, choosing $\tau \sim \mathrm{C}^+(0, \tau_0^2)$ as a prior seems like a good compromise: it allocates an important density mass around $\hat{p_0}$ while still allowing for larger values of m_{eff} thanks to the fat tails of the Cauchy distribution.

Thus, the default $\mathrm{C}^+(0, 1)$ prior on τ should be avoided in the context of sparse estimation. Instead, one should first estimate the number of non-zero parameters $\hat{p_0}$ in β, and based on that estimation compute τ_0 using (7), to finally use $\mathrm{C}^+(0, \tau_0^2)$ as prior on τ. In addition, if the user feels very confident in their estimation of the number of nonzero parameters then they can use priors on τ with thinner tail, like a normal distribution.

References

[1] C. M. Bishop. Prior distributions for variance parameters in hierarchical models. 2006.

[2] Carlos Carvalho, Nicholas Polson, and James Scott. Handling sparsity via the horseshoe. *Journal of Machine Learning Research*, 5:73–80, 04 2012.

[3] Aki Vehtari Juho Piironen. Sparsity information and regularization in the horseshoe and other shrinkage priors. *arXiv:1707.01694*, 2017.

[4] Nicholas G. Polson and Lei Sun. Bayesian l0-regularized least squares. *Applied Stochastic Models in Business and Industry*, 35(3):717–731, August 2018.

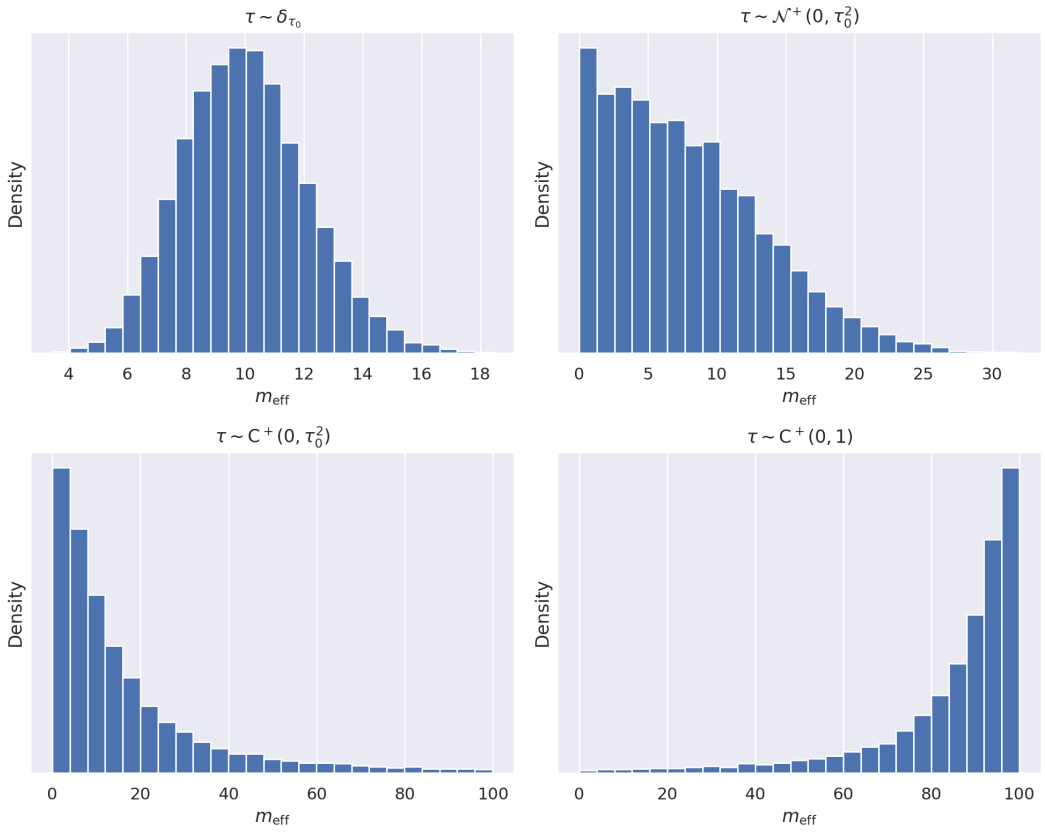

Figure 4: The distribution on the effective number of non-zero parameters m_{eff} for various priors on τ. We choose $\hat{p_0} = 10$ in this example and compute τ_0 according to the formula above, with $p = 100$, $n = 100$ and $\sigma = 1$