Enforcing sparsity with horseshoe prior

Timothé Boulet **Ali Ramlaoui** **Théo Saulus**
Master MVA
{timothe.boulet, ali.ramlaoui, theo.saulus}@student-cs.fr

Abstract

This report presents the article Carvalho et al. [2009], where the authors present their use of the horseshoe prior for sparse supervised learning. This prior is able to handle unknown sparsity and large outlying signals, from both theoretical and experimental standpoint. In this report, we will present the strengths and weaknesses of the paper and re-implement the experiments. Additionally, we propose an original contribution to challenge the Horseshoe prior with a logistic regression task to complete the paper's discussion, and extend the comparison with more recent priors. Our code is available publicly at `https://github.com/theosaulus/BML-horseshoe-prior`.

1 Introduction

Consider a multivariate linear model, with multiple input variables.

$$y_i = \beta^T x_i + \epsilon_i, \quad i = 1, \dots, n,$$

where y_i is the response variable, x_i is the input vector, β is the vector of coefficients and ϵ_i is the error term. The goal is to estimate the coefficients β from the data. Using a Bayesian approach, we can express the posterior density of β as:

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

The problem here lies in the choice of the prior $p(\beta)$. This choice will have to incorporate any relevant information we have about the coefficients or the nature of the problem. In the case where we have no prior information, non-informative priors can typically be used. However, in the case of sparse models, where we expect a lot of the coefficients to be zero, we can use shrinkage priors to enforce sparsity. A class of shrinkage priors is the global-local shrinkage priors, which are designed to shrink small coefficients to zero while allowing large coefficients to remain large, as is usually the case with such models. Globqal-local shrinkage priors are defined as follows [Bhadra et al., 2016]:

$$\beta_i \mid \lambda_i, \tau \sim \mathcal{N}(0, \lambda_i^2), \quad \lambda_i \mid \tau \sim p(\lambda_i, \tau), \quad \tau \sim p(\tau), \quad \lambda_i > 0, \tau > 0,$$

where $p(\lambda_i \mid \tau)$ has a tail that decays exponentially. The intuition behind this prior is that the λ_i's are the local shrinkage parameters allowing to individually enforce coefficients to be equal to 0, while τ is the global shrinkage parameter, which will define an overall level of sparsity in the problem. The horseshoe prior [Carvalho et al., 2009] is a member of this family of priors. Moreover, the LASSO formalism of linear regression [Tibshirani, 1996] can be seen as a special case of the global-local shrinkage priors, where the prior on the local shrinkage parameters is a double exponential distribution. These models are more formally defined in section 2. We propose to re-implement the experiments of the paper and to challenge the horseshoe prior in different settings to extend the paper's discussion in section 3. Moreover, we compare the initial prior to a newly proposed priors adapted to handle sparsity in different settings, and analyze those differences.

Preprint. Under review.

2 Priors definition

2.1 Horseshoe and other shrinkage priors

Global-local shrinkage priors. The *horseshoe prior* [Carvalho et al., 2009] is a global-local shrinkage prior that has been introduced to handle the problem of unknown sparsity.

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

The horseshoe prior has the advantage of being robust to unknown sparsity, but also robust to handling large outlying signals. This is due to the fact that the tails of the prior on the local shrinkage parameters λ_i are heavy, which allows for large coefficients to remain large, while the global shrinkage parameter τ will enforce small coefficients to be zero. The prior profile of the horseshoe is therefore an infinite spike at the origin and heavy tails, which is a desirable property for sparse models as seen in Figure 1. Note that we opted for a fully Bayesian framework for τ, as advised by Carvalho et al. [2010] in a later paper, which allows to respect the prescriptions of Carvalho et al. [2009] to avoid using plug-in.

Using the global-local shrinkage prior formalism, we can also compare the horseshoe prior to other shrinkage priors such as the LASSO shrinkage prior, a.k.a *Laplace prior*, defined as follows:

$$\beta_i \mid \lambda_i, \tau \sim \mathcal{N}(0, \lambda_i^2 \tau^2), \quad \lambda_i^2 \mid \tau \sim \text{Exp}\left(\frac{1}{2}\right), \quad \tau \sim \mathcal{C}^+(0,1),$$

and the *Student-t shrinkage prior* from the relevance vector machine model (RVM) Tipping [1999]:

$$\beta_i \mid \lambda_i, \tau \sim \mathcal{N}(0, \lambda_i^2 \tau^2), \quad \lambda_i^2 \mid \tau \sim \text{InvGamma}(a, b), \quad \tau \sim \mathcal{C}^+(0,1).$$

More recent priors. Evolutions of the LASSO have been proposed after the article was published, one of them being the *hyperlasso prior* [Griffin and Brown, 2011], with a heavier tail than LASSO:

$$\beta_i \mid \lambda_i^2 \sim \mathcal{N}(0, \lambda_i^2), \quad \lambda_i^2 \mid \tau_i \sim \text{Exp}(\tau_i), \quad \tau_i \mid \nu, \phi^2 \sim \text{Gamma}\left(\nu, \frac{1}{\phi^2}\right), \quad \phi \sim \mathcal{C}^+(0,1)$$

Note that we use here the formulation of Van Erp et al. [2019], who use a prior on ϕ instead of a plug-in method.

While the horseshoe prior has been shown to be robust to unknown sparsity, it struggles in situations where coefficients are weakly identified (for example two covariates that are highly correlated, a small sample size) as shown by Piironen and Vehtari [2017]. Indeed, because of the heavy tails, the horseshoe prior has trouble shrinking coefficients by a small amount. This problem can be observed in logistic regression, where the horseshoe prior can lead to poor performance as seen in section 3.4. The *regularized horseshoe prior* [Piironen and Vehtari, 2017] was introduced to address this issue:

$$\beta_i \mid \lambda_i, \tau, c \sim \mathcal{N}(0, \tilde{\lambda}_i^2 \tau^2), \quad \tilde{\lambda}_i^2 = \frac{c^2 \lambda_i^2}{c^2 + \tau^2 \lambda_i^2}, \quad \lambda_i \sim C^+(0,1), \quad \tau \sim \mathcal{C}^+(0,1), \tag{1}$$

The idea here is to shift the infinite value of the prior of the shrinkage coefficients κ_i at 0 towards values than can be larger than 0When $c \ll \tau^2 \lambda_i^2$, we retrieve the original horseshoe prior. On the other hand, when $c \gg \tau^2 \lambda_i^2$, the prior regularizes large coefficients. The authors propose to impose an inverse-gamma prior on c^2, which has a heavy right tail avoiding accumulation near 0.

Spike-and-slab prior and Bayesian model-averaging The *spike-and-slab prior* is a popular prior for sparse models, introduced by Mitchell and Beauchamp [1988]. It is defined as follows:

$$\beta_i \mid \lambda_i, c \sim \mathcal{N}(0, c^2 \lambda_i^2) \quad \lambda_i \sim \text{Ber}(\pi), \quad i = 1, \ldots, D,$$

where λ_i is a binary variable that indicates whether the coefficient β_i is zero or not (note that this prior is non continuous), and π is the prior probability of $\lambda_i = 1$. This allows to get a distribution where the prior on the non-zero coefficients is a normal distribution with a large variance c, and the prior on the zero coefficients is a spike.

This prior have a shrinkage profile similar to the horseshoe, which can be related to the fact that both have a global sparsity parameter, which controls how much of the coefficients are set to zero, respectively π and τ (in both cases, a lower value indicates more shrinkage). This similarity is illustrated with the shrinkage profile of both priors, in Figure 1, where we observe that the density of both prior is concentrated on full shrinkage or little shrinkage, with little in between.

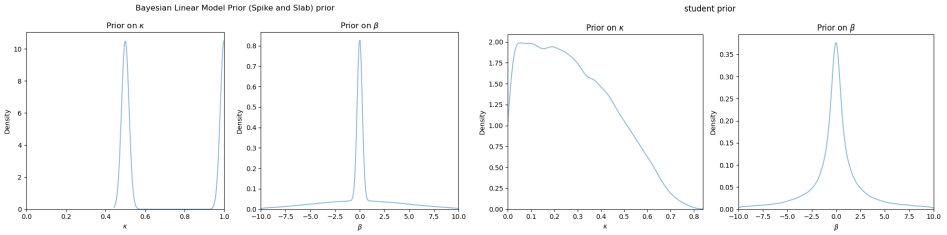
2.2 Shrinkage and robustness

Shrinkage analysis. As proposed by [Carvalho et al., 2009], an interesting way to compare these priors for sparse regression is to look at the distribution of the *shrinkage coefficient*'s prior $\kappa_i = \frac{1}{1+\lambda_i^2}$, which is found in the expectancy of the coefficients β_i:

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

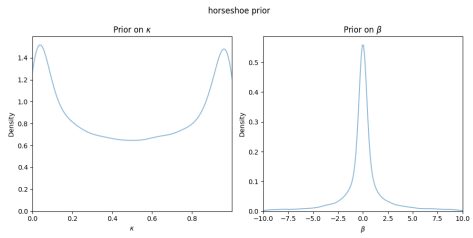
The shrinkage coefficient κ_i is a measure of the amount of shrinkage applied to the coefficient β_i, and it typically ranges between 0 and 1. When the posterior distribution of the κ_i is high around 1, then it means that the model is capable of shrinking coefficients to 0 easily, and if it is high around 0, then the model is able to predict large coefficients correctly.

We reproduce bellow the shrinkage plots of the article, and add new ones not present in the paper (in particular the spike and slab, which we found is missing). Although our plots look similar to the ones of the authors, we did not manage to obtain as smooth curves, probably due to numerical instabilities.

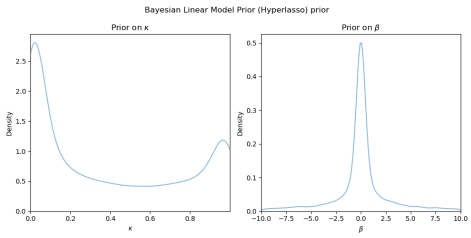


(a) Spike and Slab Prior

(b) Student-t Prior



(c) Horseshoe Prior



(d) Hyperlasso Prior

Figure 1: Distribution of the shrinkage coefficients κ_i for different priors and the priors on β. The density is here unnormalized.

Robustness analysis. The authors also develop a robustness analysis by studying (experimentally and analytically) the posterior mean of β, to show that the horseshoe maintains a "bounded influence" on extreme data points, i.e. values far from zero are not shrinked, contrary to LASSO for example. We found this analysis to be a strength of the paper, however we do not have enough space to further discuss and make significant contributions to this part.

3 Experiments

We implemented in Python all previously stated models using the PyMC package [Wiecki et al., 2024], and NumPyro [Bingham et al., 2019], a package based on Jax [Bradbury et al., 2018], for MCMC chains parallelization. For each experiment, we use the NUTS sampler to estimate the posterior and sample 10.000 times and report metrics averaged on multiple datasets for statistical significance. In all experiments, we used a constant $c = 5$ for the spike and slab prior.

3.1 Estimate the mean of a multivarate Normal

The goal of the first experiment is to estimate the coefficients of a sparse vector β using X as the identity. This means that there will be as many data points as features, and the goal is to check the

average errors between the predicted values and the true values of β under a sparse context. The sparsity level varies between 60% and 90%, and we run 100 estimations for every prior used for both a noise level $\sigma^2 \in \{1, 9\}$. We report the average errors over all the experiments in Table 1. Our implementation of hyperlasso diverges in terms of loss, so we ignore it from this experiment.

We observe that both priors based on the horseshoe are more robust to the noise introduced in the observations in terms of ℓ_1 loss. Our assumption is that this is due the better sparsity handling with most zero coefficients correctly predicted, therefore decreasing the ℓ_1 significantly while the ℓ_2 loss is harder to minimize because of the correct prediction of non-zero coefficients.

	LASSO		Student-t		Horseshoe		Regularised Horseshoe		Spike-and-slab	
	$\sigma^2 = 1$	$\sigma^2 = 9$	$\sigma^2 = 1$	$\sigma^2 = 9$	$\sigma^2 = 1$	$\sigma^2 = 9$	$\sigma^2 = 1$	$\sigma^2 = 9$	$\sigma^2 = 1$	$\sigma^2 = 9$
L2 Loss	1.82	5.74	**1.52**	6.14	1.55	5.11	3.53	**4.83**	1.75	5.44
L1 Loss	0.73	1.64	0.68	1.59	**0.52**	0.91	0.58	**0.86**	0.54	1.38

Table 1: Average errors in estimating sparse vector β with varying noise levels. The experiments are run with 100 estimations for every prior and every noise level, and the average ℓ_1 and ℓ_2 losses are reported between the true β and the estimated vector.

3.2 Padded linear regression

The second experiment is a linear regression with padded coefficients β. The idea is to take a (fixed) vector $\tilde{\beta} = [2, 2, 2, 2, 2, 2, 2, 2, 5, 50]$, and to take $\beta = [\tilde{\beta} \cdot \mathbf{0}]$ of size p, with a varying value of p. The matrices X are generated from multivariate normal with a 20% correlation between every features. The idea behind this experiment is to check the ability of the priors to shrink the coefficients to zero, and to check the ability of the priors to predict the large coefficients correctly in a normal regression context where the features are slightly correlated to each other.

In the paper, this experiment suffers from two drawbacks that we address here: first, the authors only compare the horseshoe with LASSO. Second, they use a plug-in method with cross-validation to choose the τ coefficient, after advocating strongly in favour of fully Bayesian techniques. We tackle both issues by comparing all previously introduced methods for completeness, and we use a HalfCauchy prior on τ, as previously explained in Section 2.1. Table 2 shows the results of this experiment.

Experiment		Prior					
n	p	LASSO	Student-t	Horseshoe	Reg-horseshoe	Hyperlasso	Spike-and-slab
24	20	3.06	2.76	1.99	**1.45**	2.08	3.93
60	50	1.41	0.90	0.73	**0.66**	0.91	1.37
120	100	0.53	0.36	0.38	**0.34**	0.59	0.84
240	200	0.38	0.24	0.17	**0.15**	0.52	0.53
480	400	0.20	0.09	**0.08**	**0.08**	0.42	0.53
500	250	0.22	0.11	0.10	**0.09**	0.24	0.29

Table 2: Averaged Mean Squared Error between the true β vector and the estimated one is reported for every prior and different number of data points and dimensions on 10 datasets.

We observe that Student-t prior has similar MSE compared to horseshoe, indicating the absence of outliers. Its averaged Mean Absolute Error (not reported here) is slightly worse, which shows that Student-t is less precise overall. From the posterior distribution profiles, the regularized horseshoe adapts better to small non-zero values: we observe in Fig. 3 that densities are slightly shifted for $\beta_i = 2$ compared to $\beta_i = 0$, which does not happen as much with other priors except spike and slab.

3.3 Linear regression with varying levels of sparsity

This experiment follows a similar setting as the previous one, with varying sparsity levels and randomly generated values for non-zero β_i, drawn from a Student-t distribution. We take $n = 50$ data points and $p = 50$ features and change the level of sparsity. The matrix X is generated the same way as in the previous experiment. The results of this experiment are used to compare the priors in a more general context of linear regression and are available in Table 3.

We notice that for high sparsity levels, the horseshoe prior performs the best, but as the sparsity level decreases, other priors such as the regularized horseshoe perform better because they are able to better predict the actual values of the coefficients due to the regularisation applied on the predictions of non-zero coefficients.

Sparsity Level	LASSO		Student-t		Horseshoe		Reg-horseshoe		Hyperlasso		Spike-and-slab	
	ℓ_1	ℓ_2	ℓ_1	ℓ_2	ℓ_1	ℓ_2	ℓ_1	ℓ_2	ℓ_1	ℓ_2	ℓ_1	ℓ_2
50%	1.50	0.93	1.43	0.88	1.42	0.76	1.15	0.76	**1.07**	**0.72**	1.10	**0.72**
60%	1.45	0.90	**0.83**	0.64	1.02	0.67	1.19	0.74	0.94	0.65	0.91	**0.63**
70%	0.91	0.70	0.74	0.60	0.80	0.52	1.07	0.67	0.70	0.54	**0.62**	**0.50**
90%	2.82	0.77	0.30	0.34	**0.18**	**0.16**	0.23	0.20	0.41	0.39	0.36	0.36

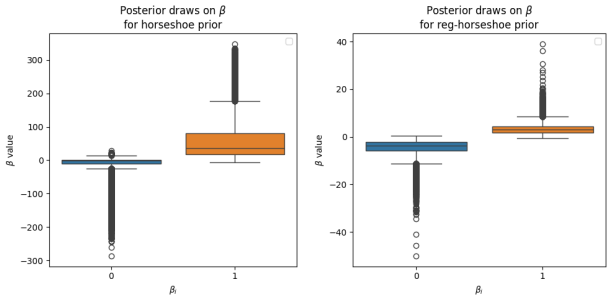
Table 3: Linear regression results with varying levels of sparsity and different priors. Metrics are averaged on 10 different datasets.

3.4 New experiment: classification

As suggested by [Piironen and Vehtari, 2017], the horseshoe prior does not control the magnitude of the coefficients, specifically in classification tasks where the goal is identify a separating hyperplane and therefore where no control of the mean of the coefficients is available. We propose this experiment to check the values of the predicted coefficients using the horseshoe prior and then comparing it with the regularized horseshoe.

The dataset we create is composed of 2 different classes in the target variable and 20 different features where only the two first features are relevant for the classes predictions. The first class is obtained with observations generated from a gaussian centered at $(1, -1)$, with a standard deviation of 0.5 for both relevant features respectively while for the second class, the means are swapped. We then train two regressors with the original horseshoe prior and then with the regularized horseshoe. Figure 2 shows boxplots of the posterior draws of both relevant features for the prediction.

We notice that while the F1-score of both priors is the same, the horseshoe prior predicts a posterior for the coefficients with a heavier tail than the regularized horseshoe.



(a) Horseshoe Prior (b) Regularized Horseshoe Prior

Figure 2: 10.000 Posterior draws of the magnitude of the predicted β's for the two relevant features for the classification task. The horseshoe prior draws are on the left plot and the regularized horseshoe prior draws are on the right plot.

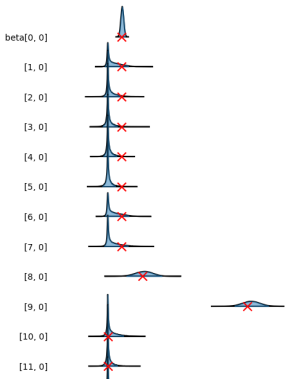


Figure 3: Posterior on β, for $n = 240, p = 200$ and regularized horseshoe prior (first 11 components), regression task. The red cross is the true value.

4 Conclusion

We acknowledge the clear contribution of the paper, which is to apply the horseshoe prior to supervised learning tasks, while proposing explanations for its shrinkage power. We tried to address what appeared as weak points to us, namely using an actually fully Bayesian framework instead of plug-in during the experiments, providing more comprehensive comparisons with other priors, and doing an experiment on a task designed to challenge the prior. More specifically, we notice that in many experiments, the Student-t prior approximately matches the performances of the horseshoe, which was not obvious when reading the article. Thus, although the horseshoe (or its regularized version) is a very efficient and flexible prior, it may not perform significantly different from Student-t, which should have been mentioned in the first place.

References

Anindya Bhadra, Jyotishka Datta, Nicholas G Polson, and Brandon Willard. Default bayesian analysis with global-local shrinkage priors. *Biometrika*, 103(4):955–969, 2016.

Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul A. Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep universal probabilistic programming. *J. Mach. Learn. Res.*, 20:28:1–28:6, 2019. URL `http://jmlr.org/papers/v20/18-403.html`.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL `http://github.com/google/jax`.

Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. Handling sparsity via the horseshoe. In David van Dyk and Max Welling, editors, *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 73–80, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR. URL `https://proceedings.mlr.press/v5/carvalho09a.html`.

Carlos M Carvalho, Nicholas G Polson, and James G Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

Jim E Griffin and Philip J Brown. Bayesian hyper-lassos with non-convex penalization. *Australian & New Zealand Journal of Statistics*, 53(4):423–442, 2011.

Toby J Mitchell and John J Beauchamp. Bayesian variable selection in linear regression. *Journal of the american statistical association*, 83(404):1023–1032, 1988.

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

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Michael Tipping. The relevance vector machine. *Advances in neural information processing systems*, 12, 1999.

Sara Van Erp, Daniel L Oberski, and Joris Mulder. Shrinkage priors for bayesian penalized regression. *Journal of Mathematical Psychology*, 89:31–50, 2019.

Thomas Wiecki, John Salvatier, Ricardo Vieira, Maxim Kochurov, Anand Patil, Michael Osthege, Brandon T. Willard, Bill Engels, Colin Carroll, Osvaldo A Martin, Adrian Seyboldt, Austin Rochford, Luciano Paz, rpgoldman, Kyle Meyer, Peadar Coyle, Oriol Abril-Pla, Marco Edward Gorelli, Ravin Kumar, Junpeng Lao, Virgile Andreani, Taku Yoshioka, George Ho, Thomas Kluyver, Kyle Beauchamp, Alexandre Andorra, Demetri Pananos, Eelke Spaak, and Benjamin Edwards. pymc-devs/pymc: v5.10.4, February 2024. URL `https://doi.org/10.5281/zenodo.10656993`.