On the Hyperprior Choice for the Global Shrinkage Parameter in the Horseshoe Prior

Aalto University
School of Science

Juho Piironen **Aki Vehtari**

Helsinki Institute for Information Technology, HIIT

Department of Computer Science, Aalto University, Finland

Introduction

- The horseshoe prior [1] has proven to be a noteworthy choice for sparse Bayesian estimation, being a computationally convenient alternative to the spike-and-slab prior.
- The level of sparsity is determined by the global shrinkage hyperparameter.
- However, we demonstrate that the results can be sensitive to the hyperprior choice for this parameter.
- We show how one can specify this hyperprior based on the prior beliefs about the number of nonzero parameters in the model.
- We show that one can improve the parameter estimation and predictive accuracy by transforming even a crude prior guess about the sparsity into the model using our framework.

Horseshoe Prior

- Consider the standard linear regression model

$$y_i = \boldsymbol{\beta}^{\mathsf{T}} \mathbf{x}_i + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}\left(0, \sigma^2\right), \quad i = 1, \ldots, n,$$

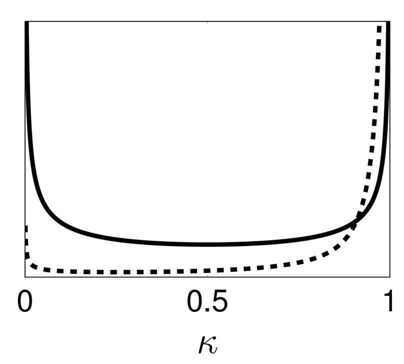
  where $\mathbf{x}$ is the D-dimensional vector of predictors, $\boldsymbol{\beta}$ denotes the corresponding coefficients and σ^2 is the noise variance.
- The horseshoe prior for the regression coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_D)$ is given by

$$\begin{aligned} \beta_j \,|\, \lambda_j, \tau &\sim \mathrm{N}\left(0, \lambda_j^2 \tau^2\right), \\ \lambda_j &\sim \mathrm{C}^+(0, 1), \qquad j = 1, \ldots, D. \end{aligned} \tag{1}$$

- Given the hyperparameters λ_j and τ, and assuming uncorrelated predictors (with zero mean and unit variance), the posterior mean satisfies approximately

$$\bar{\beta}_j = (1 - \kappa_j)\hat{\beta}_j, \qquad \kappa_j = \frac{1}{1 + n\sigma^{-2}\tau^2\lambda_j^2}. \tag{2}$$

- Here $\hat{\beta}$ is the maximum likelihood solution and κ_j the shrinkage factor.
- The prior for each shrinkage factor κ_j looks like a horseshoe:



Density for the shrinkage factor κ_j (Eq. (2)) for the horseshoe prior (Eq. (1)) when $n\sigma^{-2}\tau^2 = 1$ (solid) and when $n\sigma^{-2}\tau^2 = 0.1$ (dashed).

- Intuition: we expect both relevant ($\bar{\beta}_j \approx \hat{\beta}_j$) and irrelevant ($\bar{\beta}_j \approx 0$) variables; which one is favored, depends on the global shrinkage τ.

The Global Hyperparameter

- We define the effective number of nonzero coefficients as

$$m_{\mathrm{eff}} = \sum_{j=1}^{D} (1 - \kappa_j). \tag{3}$$

- The prior mean and variance for m_{eff} can be derived analytically

$$\mathrm{E}\left[m_{\mathrm{eff}} \,|\, \tau, \sigma\right] = \frac{\tau\sigma^{-1}\sqrt{n}}{1 + \tau\sigma^{-1}\sqrt{n}} D,$$

$$\mathrm{Var}\left[m_{\mathrm{eff}} \,|\, \tau, \sigma\right] = \frac{\tau\sigma^{-1}\sqrt{n}}{2(1 + \tau\sigma^{-1}\sqrt{n})^2} D.$$

- Thus, if our prior guess for the number of relevant variables is p_0, it is reasonable to choose the prior so that $\mathrm{E}\left[m_{\mathrm{eff}} \,|\, \tau, \sigma\right] = p_0$, which yields for τ

$$\tau_0 = \frac{p_0}{D - p_0} \frac{\sigma}{\sqrt{n}}. \tag{4}$$

- This equation captures the relationship between the global shrinkage parameter and the prior assumptions about the sparsity, and indicates where $p(\tau)$ should have most of its mass.
- It is insightful to visualize the prior imposed on m_{eff} for different prior choices for τ by drawing samples for m_{eff} (see the figure below).
- Theoretical result: assuming a true $\boldsymbol{\beta}_*$ exists, if our prior guess is $p_0 = p_*$ (the true number of relevant variables), then τ_0 is asymptotically the optimal choice in terms of posterior contraction rates and mean squared error in comparison to the true $\boldsymbol{\beta}^*$ [2, 3].
- The result (4) can be generalized also to non-Gaussian observation models by deriving appropriate plug-in values for σ, for instance $\sigma = 2$ for the logistic regression [3].

Conclusions

- We have shown how to specify the hyperprior for the global shrinkage parameter in the horseshoe prior based on our prior beliefs about the number nonzero parameters in the model.
- Setting up the prior for τ based on the prior beliefs regarding the sparsity improves the results even when the prior knowledge is rough.
- The presented framework could also be generalized to other shrinkage priors than the horseshoe.

Implementation

- The horseshoe prior is implemented in the R-package *rstanarm* (`https://github.com/stan-dev/rstanarm`).
- A demo about model fitting and the subsequent projective variable selection using our R-package *projpred* (`https://github.com/stan-dev/projpred`) can be found in the vignette `https://users.aalto.fi/~jtpiiron/projpred/quickstart.html`.

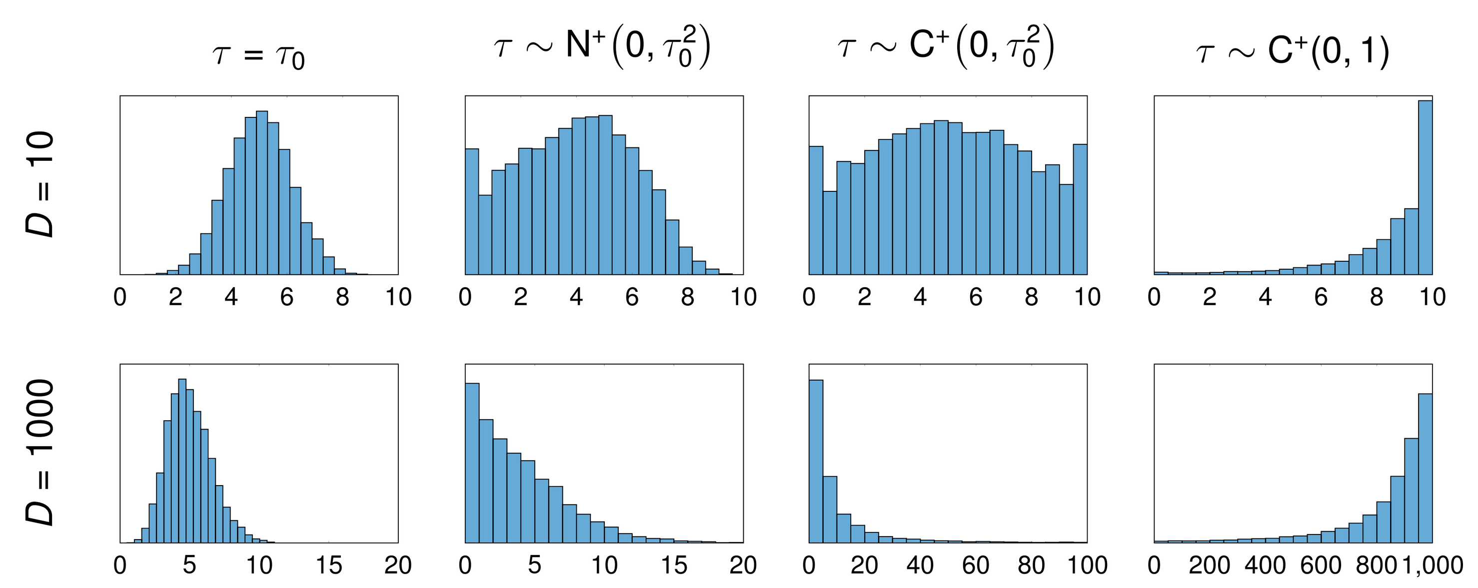
References

[1] Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. Handling sparsity via the horseshoe. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*. JMLR W&CP, 2009. 1

[2] S. L. van der Pas, B. J. K. Kleijn, and A. W. van der Vaart. The horseshoe estimator: posterior concentration around nearly black vectors. *Electronic Journal of Statistics*, 8(2):2585–2618, 2014. 1

[3] Juho Piironen and Aki Vehtari. On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. JMLR W&CP, 2017. 1

Illustration of the Prior Choice



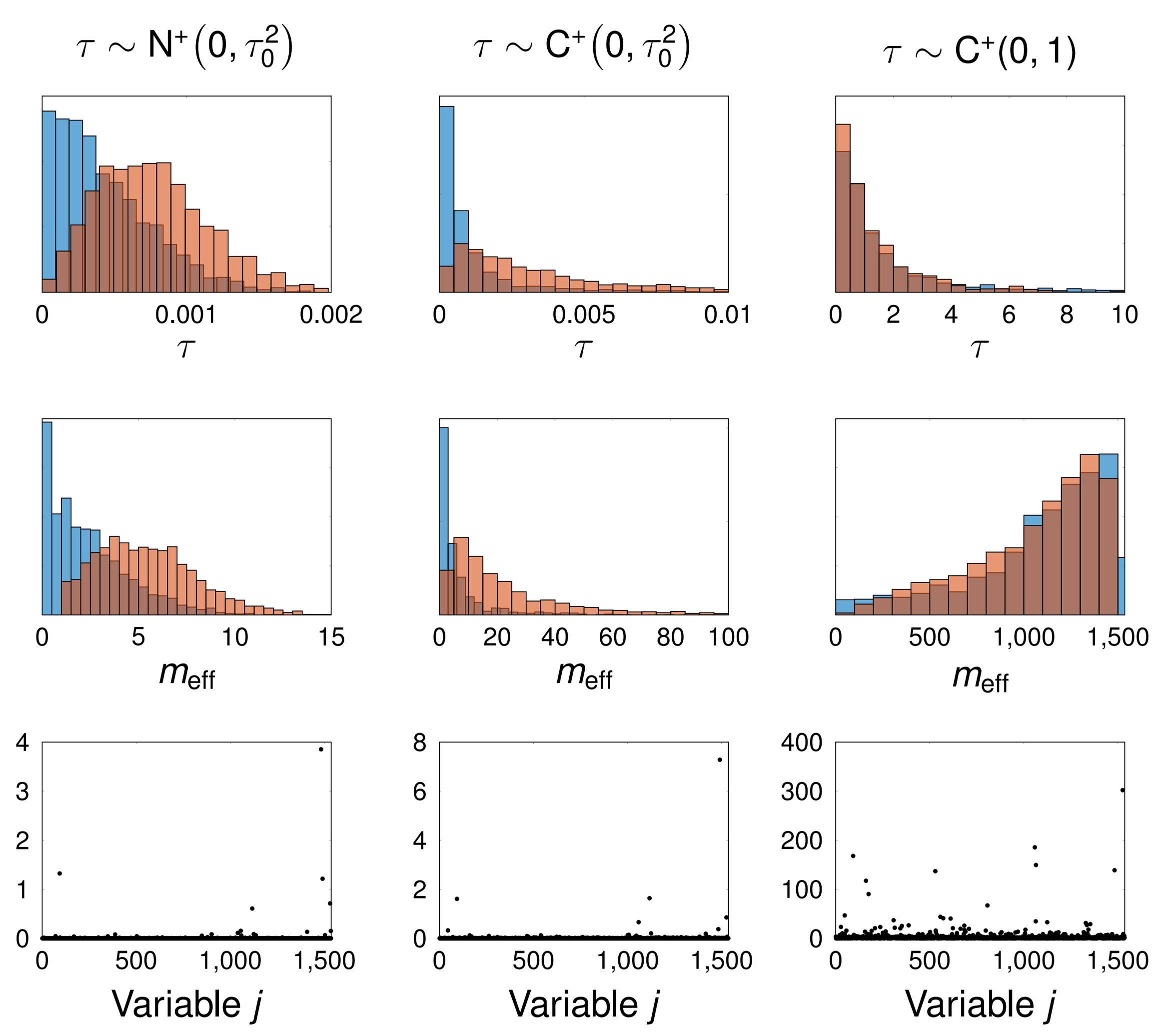
Various priors for m_{eff}: Histograms of prior draws for m_{eff} (Eq. (3)) imposed by different prior choices for τ. D denotes the total number of variables, and τ_0 is computed from formula (4) assuming $n = 100$ observations with $\sigma = 1$ and $p_0 = 5$ as the prior guess for the number of relevant variables. Notice how the "uninformative" $\tau \sim \mathrm{C}^+(0, 1)$ results in a rather dubious prior for m_{eff}.

Datasets

Dataset	n	D
Ovarian	54	1536
Colon	62	2000
Prostate	102	5966
ALLAML	72	7129

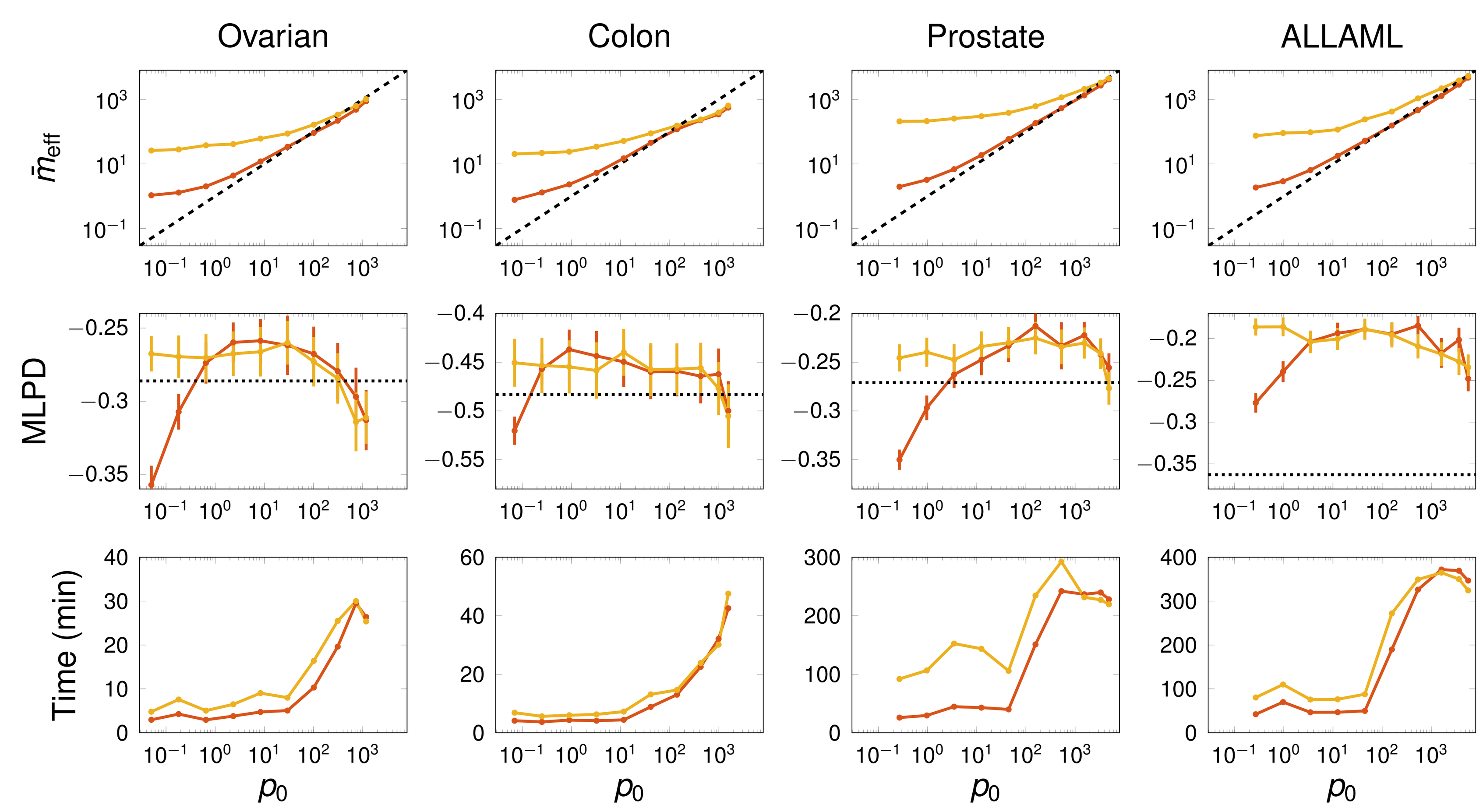
Summary of the microarray cancer datasets (binary classification) used for the real world illustrations.

Effect on Parameter Estimates



Ovarian cancer dataset: Histograms of prior and posterior samples for τ (top row) and m_{eff} (middle row), and absolute values of the posterior means for the regression coefficients $|\bar{\beta}_j|$ (bottom row) imposed by three different prior choices for τ. τ_0 corresponds to a prior guess $p_0 = 3$ relevant variables (Eq. (4)).

Effect on Predictive Accuracy and Computational Efficiency



Microarray classification datasets: Posterior mean for m_{eff}, mean log predictive density (MLPD) on test data ($\pm$ one standard error), and computation time for two priors for the global hyperparameter: $\tau \sim \mathrm{N}^+\left(0, \tau_0^2\right)$ (red), and $\tau \sim \mathrm{C}^+\left(0, \tau_0^2\right)$ (yellow), where τ_0 is computed from (4) varying p_0 (horizontal axis). For each curve, the largest p_0 corresponds to $\tau_0 = 1$. For comparison, the dotted line in the middle row plots denotes the MLPD for LASSO.

Contact: juho.piironen@aalto.fi, web page: https://users.aalto.fi/jtpiiron/