Regularized horseshoe

Aki Vehtari

Aalto University, Finland
aki.vehtari@aalto.fi
@avehtari

Joint work with Juho Piironen

Outline

- Large p, small n regression
- Prior on weights vs. prior on shrinkage
- Horseshoe prior
- Regularized horseshoe prior

Large p, small n regression

- Linear or generalized linear regression
  - number of covariates p
  - number of observations n
- Large p, small n common e.g.
  - in modern medical/bioinformatics studies (e.g. microarrays, GWAS)
  - brain imaging
  - in our examples p is around 1e2–1e5, and usually $n < 100$

Large p, small n regression

- If noiseless observations we can fit uniquely identified regression in $n-1$ dimensions

Large p, small n regression

- If noiseless observations we can fit uniquely identified regression in $n - 1$ dimensions
- If noisy observations, more complicated

Large p, small n regression

- If noiseless observations we can fit uniquely identified regression in $n - 1$ dimensions
- If noisy observations, more complicated
- If correlating covariates, more complicated

Large p, small n regression

- Priors!

Large p, small n regression

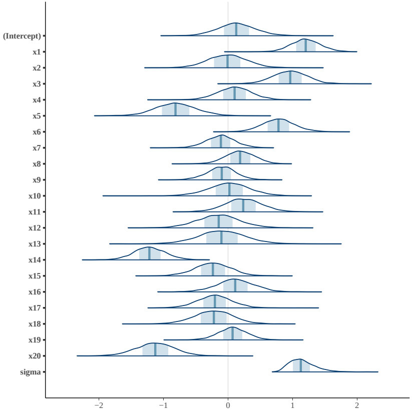
- Priors!
- Non-sparse priors assume most covariates are relevant, but may have strong correlations

  $\rightarrow$ factor models

Large p, small n regression

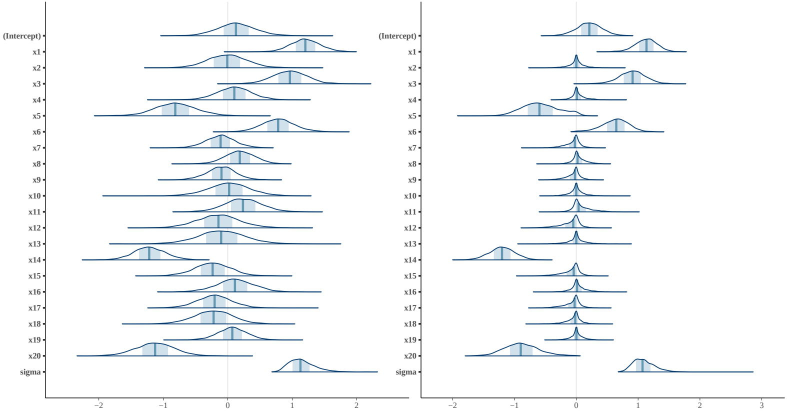
- Priors!
- Non-sparse priors assume most covariates are relevant, but may have strong correlations
  - $\rightarrow$ factor models
- Sparse priors assume only small number of covariates effectively non-zero $m_{\mathrm{eff}} \ll n$

Example



Gaussian prior

Example



Gaussian prior | Horseshoe prior

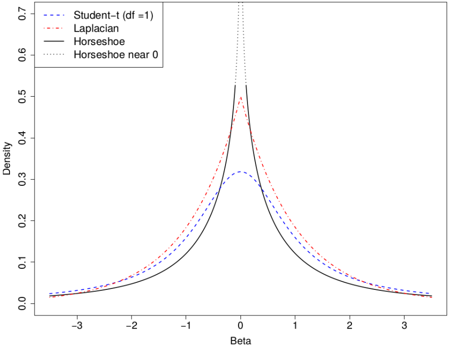
rstanarm + bayesplot

- Gaussian vs. Horseshoe predictive performance using cross-validation (loo package, more in Friday Model selection tutorial)

```
> compare(loog,loohs)
 elpd_diff        se
       7.9       2.8
```

Large p, small n regression

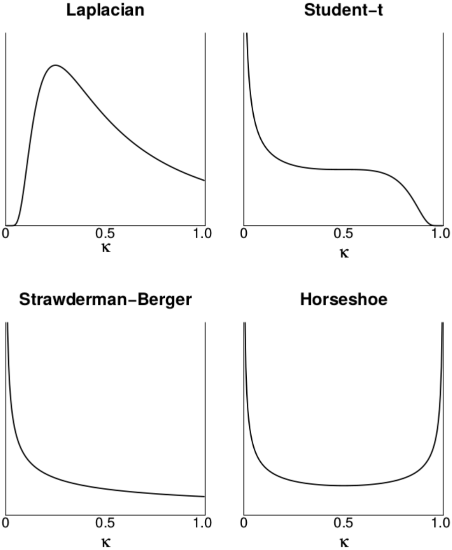
- Sparse priors assume only small number of covariates effectively non-zero $m_{\mathrm{eff}} \ll p$
  - Laplace prior ("Bayesian lasso")
    - computationally convenient (continuous and log-concave), but not really sparse
  - spike-and-slab (with point-mass at zero)
    - prior on number of non-zero covariates, discrete
  - Horseshoe and hierarchical shrinkage priors
    - prior on amount of shrinkage, continuous



Carvalho et al (2009)

Prior on shrinkage

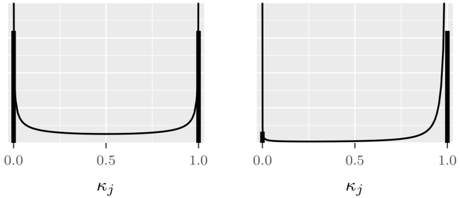
- Slope of the prior at specific value determines the amount of shrinkage



Carvalho et al (2009)

Spike-and-slab vs horseshoe prior

- Spike and slab prior (with point-mass at zero) has mix of continuous prior and probability mass at zero
  - parameter space is mixture of continuous and discrete
- Hierarchical shrinkage and horseshoe priors are continuous



Piironen and Vehtari (2017a)

Horseshoe prior

- Linear regression model with covariates $\mathbf{x} = (x_1, \ldots, x_D)$

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

Horseshoe prior

- Linear regression model with covariates $\mathbf{x} = (x_1, \ldots, x_D)$

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

- The horseshoe prior:

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

Horseshoe prior

- Linear regression model with covariates $\mathbf{x} = (x_1, \ldots, x_D)$

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

- The horseshoe prior:

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

- *The global parameter* τ shrinks all β_j towards zero
- *The local parameters* λ_j allow some β_j to escape the

  shrinkage

Horseshoe prior

- Linear regression model with covariates $\mathbf{x} = (x_1, \ldots, x_D)$

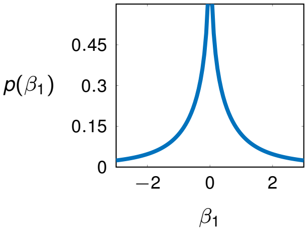
$$y_i = \boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}\left(0, \sigma^2\right), \quad i = 1, \ldots, n,$$

- The horseshoe prior:

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

- *The global parameter* τ shrinks all β_j towards zero
- *The local parameters* λ_j allow some β_j to escape the

shrinkage



Horseshoe prior

- Given the hyperparameters, the posterior mean satisfies approximately

$$\bar{\beta}_j = (1 - \kappa_j)\beta_j^{\mathrm{ML}}, \qquad \kappa_j = \frac{1}{1 + n\sigma^{-2}\tau^2\lambda_j^2},$$

where κ_j is the *shrinkage factor*

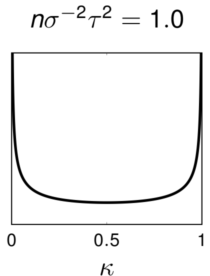
Horseshoe prior

- Given the hyperparameters, the posterior mean satisfies approximately

$$\bar{\beta}_j = (1-\kappa_j)\beta_j^{\mathrm{ML}}, \qquad \kappa_j = \frac{1}{1+n\sigma^{-2}\tau^2\lambda_j^2},$$

where κ_j is the *shrinkage factor*

- With $\lambda_j \sim \mathrm{C}^+(0,1)$, the prior for κ_j looks like:



We expect both

- relevant ($\bar{\beta}_j \approx \beta_j^{\mathrm{ML}}$) features
- irrelevant ($\bar{\beta}_j \approx 0$) features

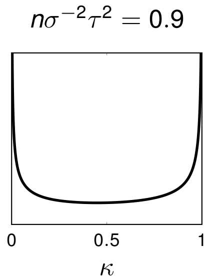
Horseshoe prior

- Given the hyperparameters, the posterior mean satisfies approximately

$$\bar{\beta}_j = (1 - \kappa_j)\beta_j^{\text{ML}}, \qquad \kappa_j = \frac{1}{1 + n\sigma^{-2}\tau^2\lambda_j^2},$$

where κ_j is the *shrinkage factor*

- With $\lambda_j \sim \mathrm{C}^+(0, 1)$, the prior for κ_j looks like:



We expect both

- relevant ($\bar{\beta}_j \approx \beta_j^{\text{ML}}$) features
- irrelevant ($\bar{\beta}_j \approx 0$) features

Horseshoe prior

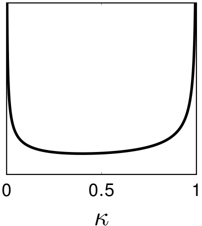
- Given the hyperparameters, the posterior mean satisfies approximately

$$\bar{\beta}_j = (1 - \kappa_j)\beta_j^{\mathrm{ML}}, \qquad \kappa_j = \frac{1}{1 + n\sigma^{-2}\tau^2\lambda_j^2},$$

where κ_j is the *shrinkage factor*

- With $\lambda_j \sim \mathrm{C}^+(0, 1)$, the prior for κ_j looks like:

$n\sigma^{-2}\tau^2 = 0.8$



We expect both

- relevant ($\bar{\beta}_j \approx \beta_j^{\mathrm{ML}}$) features
- irrelevant ($\bar{\beta}_j \approx 0$) features

Horseshoe prior

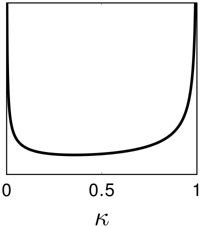
- Given the hyperparameters, the posterior mean satisfies approximately

$$\bar{\beta}_j = (1 - \kappa_j)\beta_j^{\mathrm{ML}}, \qquad \kappa_j = \frac{1}{1 + n\sigma^{-2}\tau^2\lambda_j^2},$$

  where κ_j is the *shrinkage factor*

- With $\lambda_j \sim \mathrm{C}^+(0, 1)$, the prior for κ_j looks like:

$n\sigma^{-2}\tau^2 = 0.7$



We expect both

- relevant ($\bar{\beta}_j \approx \beta_j^{\mathrm{ML}}$) features
- irrelevant ($\bar{\beta}_j \approx 0$) features

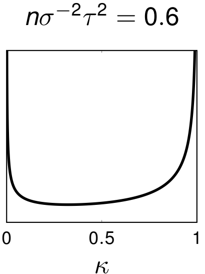
Horseshoe prior

- Given the hyperparameters, the posterior mean satisfies approximately

$$\bar{\beta}_j = (1 - \kappa_j)\beta_j^{\mathrm{ML}}, \qquad \kappa_j = \frac{1}{1 + n\sigma^{-2}\tau^2\lambda_j^2},$$

where κ_j is the *shrinkage factor*

- With $\lambda_j \sim \mathrm{C}^+(0, 1)$, the prior for κ_j looks like:



We expect both

- relevant ($\bar{\beta}_j \approx \beta_j^{\mathrm{ML}}$) features
- irrelevant ($\bar{\beta}_j \approx 0$) features

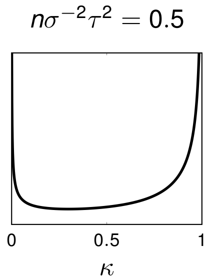
Horseshoe prior

- Given the hyperparameters, the posterior mean satisfies approximately

$$\bar{\beta}_j = (1-\kappa_j)\beta_j^{\mathrm{ML}}, \qquad \kappa_j = \frac{1}{1+n\sigma^{-2}\tau^2\lambda_j^2},$$

where κ_j is the *shrinkage factor*

- With $\lambda_j \sim \mathrm{C}^+(0,1)$, the prior for κ_j looks like:



We expect both

- relevant ($\bar{\beta}_j \approx \beta_j^{\mathrm{ML}}$) features
- irrelevant ($\bar{\beta}_j \approx 0$) features

Horseshoe prior

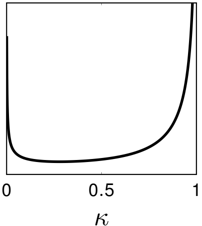
- Given the hyperparameters, the posterior mean satisfies approximately

$$\bar{\beta}_j = (1 - \kappa_j)\beta_j^{\mathrm{ML}}, \qquad \kappa_j = \frac{1}{1 + n\sigma^{-2}\tau^2\lambda_j^2},$$

where κ_j is the *shrinkage factor*

- With $\lambda_j \sim \mathrm{C}^+(0, 1)$, the prior for κ_j looks like:

$n\sigma^{-2}\tau^2 = 0.4$



We expect both

- relevant ($\bar{\beta}_j \approx \beta_j^{\mathrm{ML}}$) features
- irrelevant ($\bar{\beta}_j \approx 0$) features

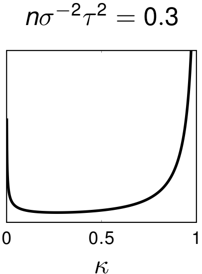
Horseshoe prior

- Given the hyperparameters, the posterior mean satisfies approximately

$$\bar{\beta}_j = (1 - \kappa_j)\beta_j^{\mathrm{ML}}, \qquad \kappa_j = \frac{1}{1 + n\sigma^{-2}\tau^2\lambda_j^2},$$

where κ_j is the *shrinkage factor*

- With $\lambda_j \sim \mathrm{C}^+(0, 1)$, the prior for κ_j looks like:



We expect both

- relevant ($\bar{\beta}_j \approx \beta_j^{\mathrm{ML}}$) features
- irrelevant ($\bar{\beta}_j \approx 0$) features

Horseshoe prior

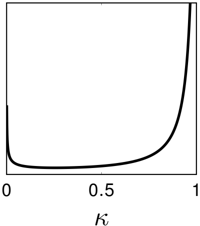
- Given the hyperparameters, the posterior mean satisfies approximately

$$\bar{\beta}_j = (1 - \kappa_j)\beta_j^{\mathrm{ML}}, \qquad \kappa_j = \frac{1}{1 + n\sigma^{-2}\tau^2\lambda_j^2},$$

where κ_j is the *shrinkage factor*

- With $\lambda_j \sim \mathrm{C}^+(0, 1)$, the prior for κ_j looks like:

$n\sigma^{-2}\tau^2 = 0.2$



We expect both

- relevant ($\bar{\beta}_j \approx \beta_j^{\mathrm{ML}}$) features
- irrelevant ($\bar{\beta}_j \approx 0$) features

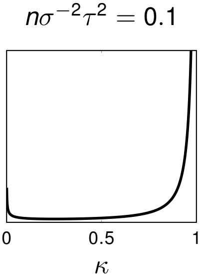
Horseshoe prior

- Given the hyperparameters, the posterior mean satisfies approximately

$$\bar{\beta}_j = (1 - \kappa_j)\beta_j^{\mathrm{ML}}, \qquad \kappa_j = \frac{1}{1 + n\sigma^{-2}\tau^2\lambda_j^2},$$

where κ_j is the *shrinkage factor*

- With $\lambda_j \sim \mathrm{C}^+(0, 1)$, the prior for κ_j looks like:



We expect both

- relevant ($\bar{\beta}_j \approx \beta_j^{\mathrm{ML}}$) features
- irrelevant ($\bar{\beta}_j \approx 0$) features

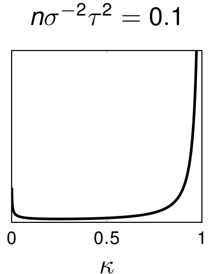
Horseshoe prior

- Given the hyperparameters, the posterior mean satisfies approximately

$$\bar{\beta}_j = (1-\kappa_j)\beta_j^{\mathrm{ML}}, \qquad \kappa_j = \frac{1}{1+n\sigma^{-2}\tau^2\lambda_j^2},$$

where κ_j is the *shrinkage factor*

- With $\lambda_j \sim \mathrm{C}^+(0,1)$, the prior for κ_j looks like:



We expect both

- relevant ($\bar{\beta}_j \approx \beta_j^{\mathrm{ML}}$) features
- irrelevant ($\bar{\beta}_j \approx 0$) features

Small $\tau \Rightarrow$ more coefficients ≈ 0

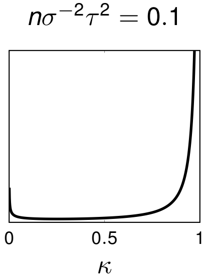
Horseshoe prior

- Given the hyperparameters, the posterior mean satisfies approximately

$$\bar{\beta}_j = (1 - \kappa_j)\beta_j^{\mathrm{ML}}, \qquad \kappa_j = \frac{1}{1 + n\sigma^{-2}\tau^2\lambda_j^2},$$

where κ_j is the *shrinkage factor*

- With $\lambda_j \sim \mathrm{C}^+(0, 1)$, the prior for κ_j looks like:



We expect both

- relevant ($\bar{\beta}_j \approx \beta_j^{\mathrm{ML}}$) features
- irrelevant ($\bar{\beta}_j \approx 0$) features

Small $\tau \Rightarrow$ more coefficients ≈ 0
How to specify prior for τ?

The global shrinkage parameter τ

- *Effective number of nonzero coefficients*

$$m_{\text{eff}} = \sum_{j=1}^{D} (1 - \kappa_j)$$

The global shrinkage parameter τ

- *Effective number of nonzero coefficients*

$$m_{\text{eff}} = \sum_{j=1}^{D} (1 - \kappa_j)$$

- The prior mean can be shown to be

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

The global shrinkage parameter τ

- *Effective number of nonzero coefficients*

$$m_{\text{eff}} = \sum_{j=1}^{D} (1 - \kappa_j)$$

- The prior mean can be shown to be

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

- Setting $\mathrm{E}[m_{\text{eff}} \,|\, \tau, \sigma] = p_0$ (prior guess for the number of nonzero coefficients) yields for τ

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

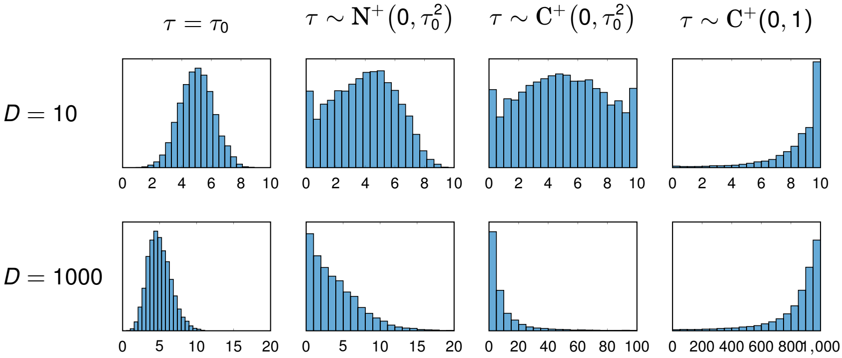
$\Rightarrow$ Prior guess for τ based on our beliefs about the sparsity

Illustration $p(\tau)$ vs. $p(m_{\text{eff}})$

Let $n = 100, \quad \sigma = 1, \quad p_0 = 5, \quad \tau_0 = \frac{p_0}{D-p_0}\frac{\sigma}{\sqrt{n}}, \quad D =$ dimensionality

Let $n = 100$, $\sigma = 1$, $p_0 = 5$, $\tau_0 = \frac{p_0}{D-p_0}\frac{\sigma}{\sqrt{n}}$, $D =$ dimensionality

$p(m_{\text{eff}})$ with different choices of $p(\tau)$:



Non-Gaussian observation models

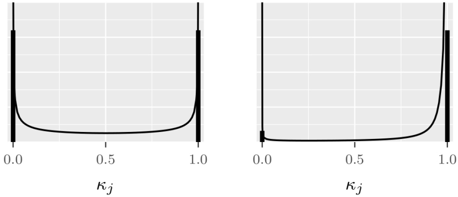
- The reference value:

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

- The framework can be applied also to non-Gaussian observation models by deriving appropriate plug-in values for σ
  - Gaussian approximation to the likelihood
  - E.g. $\sigma = 2$ for logistic regression

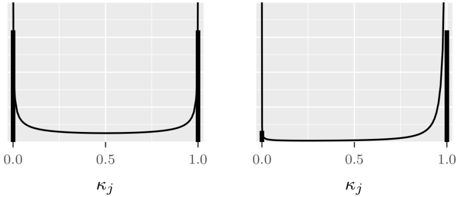
Regularized horseshoe

- HS allows some coefficients to be completely unregularized
  - allows complete separation in logistic model with $n \ll p$

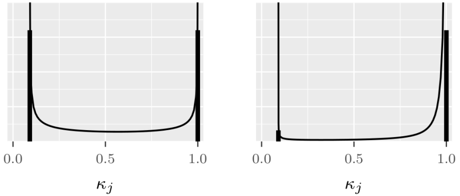


Regularized horseshoe

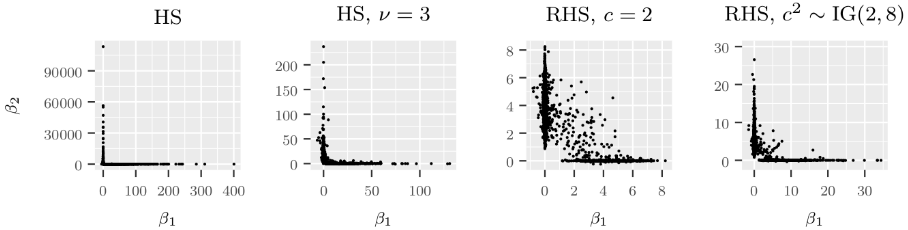
- HS allows some coefficients to be completely unregularized
  - allows complete separation in logistic model with $n \ll p$



- Regularized horseshoe adds additional wide slab
  - maintains division to relevant and non-relevant variables



- Regularized horseshoe helps regularize relevant variables



Regularized horseshoe in rstanarm

- Easy in rstanarm (thanks to Ben Goodrich)

  ```
  p0 <- 5
  tau0 <- p0/(D-p0) * sigmaguess/sqrt(n)
  fit <- stan_glm(y ~ x, gaussian(), hs(global_scale=tau0, slab_scale=2.5,
  slab_df=4))
  ```

- Note: rstanarm does not condition on σ, and thus need to scale tau0 with a guess of expected value of σ
  - luckily the result is not sensitive to the exact value

Regularized horseshoe in rstanarm

- Easy in rstanarm (thanks to Ben Goodrich)

  p0 <- 5
  tau0 <- p0/(D-p0) * sigmaguess/sqrt(n)

  fit <- stan_glm(y ~ x, gaussian(), hs(global_scale=tau0, slab_scale=2.5, slab_df=4))

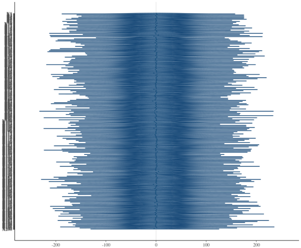
- Note: rstanarm does not condition on σ, and thus need to scale tau0 with a guess of expected value of σ
  - luckily the result is not sensitive to the exact value
- Note 2: hs() prior is called "hierarchical shrinkage" prior, as it is extension of Horseshoe (Horseshoe has local_df=1)
  - luckily the result is not sensitive to the exact value

Regularized horseshoe in rstanarm

- Simulated regression example
  $n = 100$, $p = 200$, true $p_0 = 7$
- Gaussian vs. “Bayesian LASSO” vs. Reg. Horseshoe

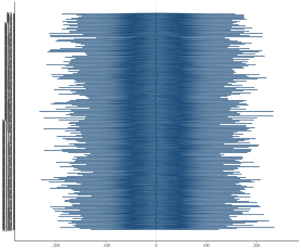
Regularized horseshoe in rstanarm

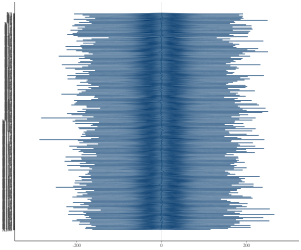
- Simulated regression example
  $n = 100$, $p = 200$, true $p_0 = 7$
- Gaussian vs. "Bayesian LASSO" vs. Reg. Horseshoe



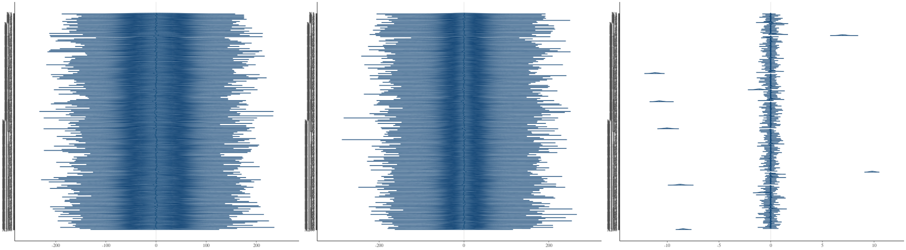
Regularized horseshoe in rstanarm

- Simulated regression example
  $n = 100$, $p = 200$, true $p_0 = 7$
- Gaussian vs. "Bayesian LASSO" vs. Reg. Horseshoe



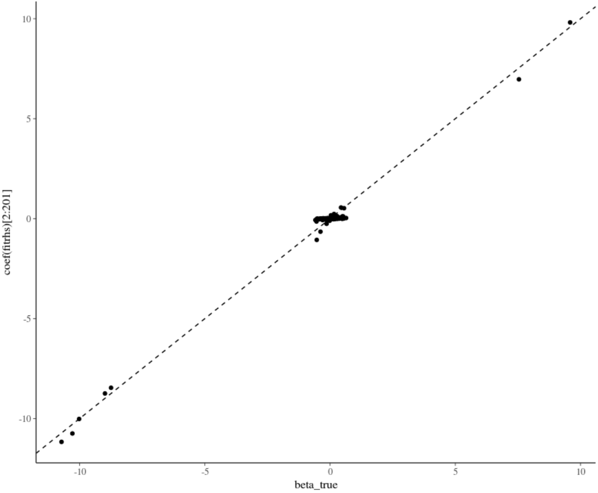


- Simulated regression example
  $n = 100$, $p = 200$, true $p_0 = 7$
- Gaussian vs. "Bayesian LASSO" vs. Reg. Horseshoe



Regularized horseshoe in rstanarm

- Simulated regression example
  $n = 100$, $p = 200$, true $p_0 = 7$

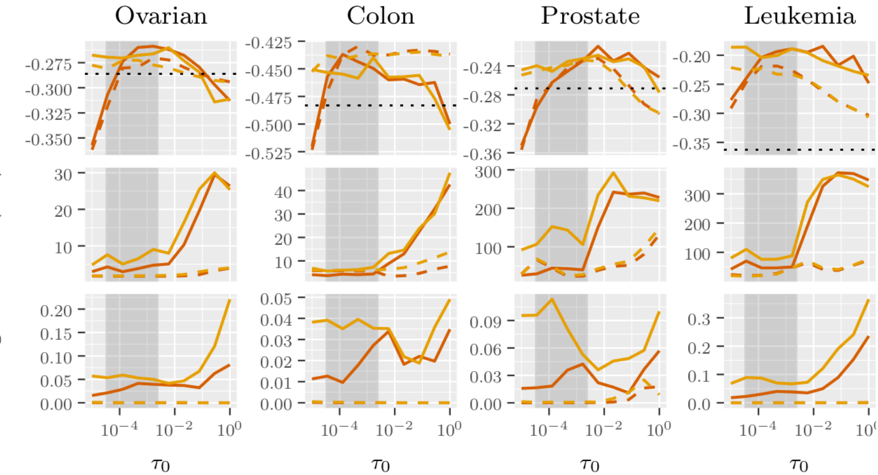


Experiments

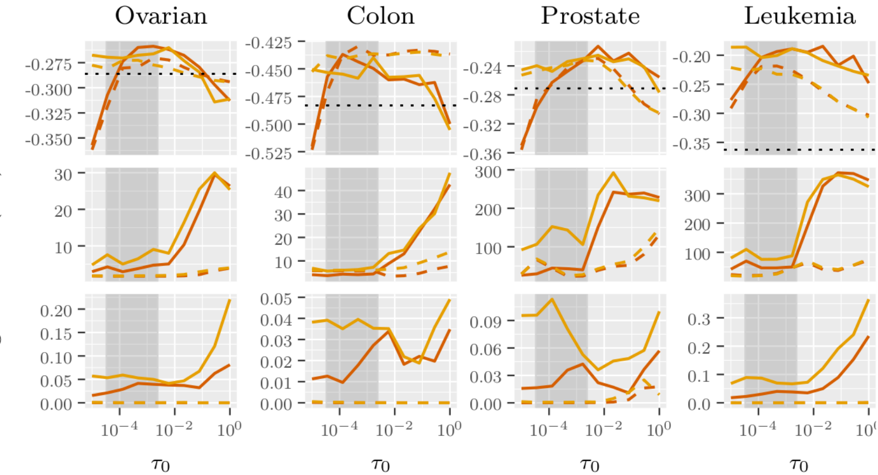
Table: Summary of the real world datasets, D denotes the number of predictors and n the dataset size.

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

Horseshoe vs regularized horseshoe

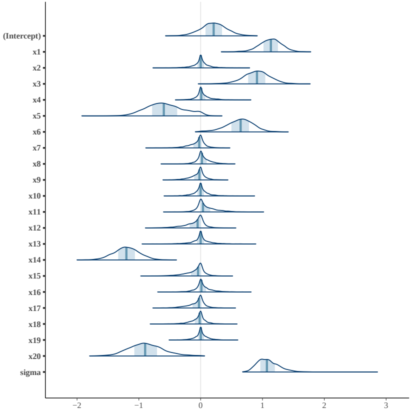


Horseshoe vs regularized horseshoe



- Regularized horseshoe helps to reduce the number of divergences, too

Example



- Even if Horseshoe shrinks a lot, coefficient posterior has unecrtainty and it’s not exactly zero
- Tomorrow in Model selection tutorial
  - how to select most relevant variables
  - how to do the inference after the selection while taking into account the uncertainties in the full model

Summary of regularized horseshoe prior

- Sparse as horseshoe, but more robust inference and computation
- Better performance than LASSO and Bayesian LASSO

References (with code examples for Stan included)

- Juho Piironen and Aki Vehtari (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. In Electronic Journal of Statistics, 11(2):5018-5051. https://projecteuclid.org/euclid.ejs/1513306866
- Juho Piironen and Aki Vehtari (2017). On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, PMLR 54:905-913. http://proceedings.mlr.press/v54/piironen17a.html
- Juho Piironen, and Aki Vehtari (2018). Iterative supervised principal components. Proceedings of the 21th International Conference on Artificial Intelligence and Statistics, accepted for publication. https://arxiv.org/abs/1710.06229
- See also model selection tutorial with some notebooks using regularized horseshoe https://github.com/avehtari/modelselection_tutorial