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

Juho Piironen, Aki Vehtari

Helsinki Institute for Information Technology, HIIT
Department of Computer Science, Aalto University

juho.piironen@aalto.fi

AISTATS, 2017

Introduction

- *Sparse estimation:* large number of parameters $\boldsymbol{\theta} = (\theta_1, \dots, \theta_D)$, assume only a few are nonzero
  - Regression/classification with many candidate predictors
  - Example dataset: Leukemia classification $D = 7129$, $n = 72$
- Non-Bayesian approaches: LASSO, elastic net etc.
- Bayesian approach: sparsifying prior + integrate over uncertainty

Introduction

- *Sparse estimation:* large number of parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_D)$, assume only a few are nonzero
  - Regression/classification with many candidate predictors
  - Example dataset: Leukemia classification $D = 7129$, $n = 72$
- Non-Bayesian approaches: LASSO, elastic net etc.
- Bayesian approach: sparsifying prior + integrate over uncertainty
- Horseshoe prior
  - Continuous shrinkage prior
  - Computationally convenient alternative to the spike-and-slab, with similar or better performance

Introduction

- *Sparse estimation:* large number of parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_D)$, assume only a few are nonzero
  - Regression/classification with many candidate predictors
  - Example dataset: Leukemia classification $D = 7129$, $n = 72$
- Non-Bayesian approaches: LASSO, elastic net etc.
- Bayesian approach: sparsifying prior + integrate over uncertainty

- Horseshoe prior
  - Continuous shrinkage prior
  - Computationally convenient alternative to the spike-and-slab, with similar or better performance
- However:
  - Previously not clear how to encode prior assumptions about the sparsity to the model (trivial in spike-and-slab)

    $\Rightarrow$ This talk

Horseshoe prior

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

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

Horseshoe prior

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

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

- 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 many inputs $\mathbf{x} = (x_1, \ldots, x_D)$

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

- The horseshoe prior:

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

- *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

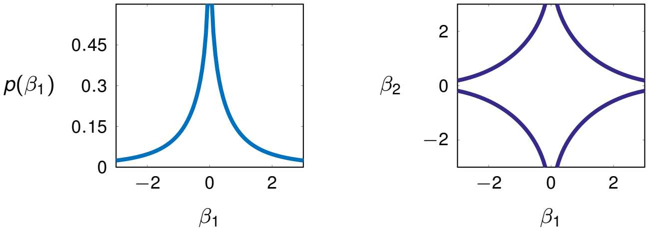
- Linear regression model with many inputs $\mathbf{x} = (x_1, \ldots, x_D)$

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

- The horseshoe prior:

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

- *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

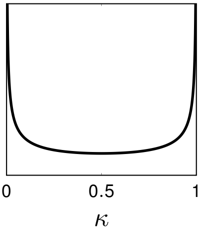
- Given the hyperparameters, the posterior mean satisfies approximately

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

where κ_j is the *shrinkage factor*

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

$n\sigma^{-2}\tau^2 = 1.0$



We expect both

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

Horseshoe prior

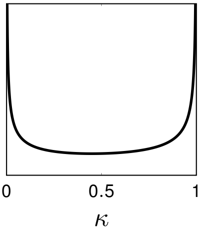
- Given the hyperparameters, the posterior mean satisfies approximately

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

where κ_j is the *shrinkage factor*

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

$n\sigma^{-2}\tau^2 = 0.9$



We expect both

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

Horseshoe prior

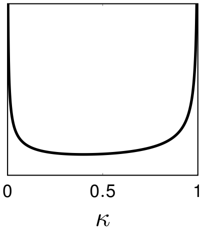
- Given the hyperparameters, the posterior mean satisfies approximately

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

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

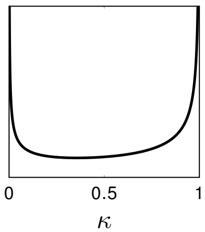
- Given the hyperparameters, the posterior mean satisfies approximately

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

where κ_j is the *shrinkage factor*

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

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



We expect both

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

Horseshoe prior

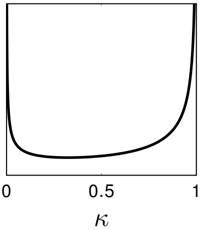
- Given the hyperparameters, the posterior mean satisfies approximately

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

where κ_j is the *shrinkage factor*

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

$n\sigma^{-2}\tau^2 = 0.6$



We expect both

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

Horseshoe prior

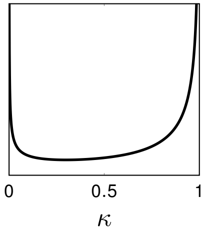
- Given the hyperparameters, the posterior mean satisfies approximately

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

where κ_j is the *shrinkage factor*

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

$n\sigma^{-2}\tau^2 = 0.5$



We expect both

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

Horseshoe prior

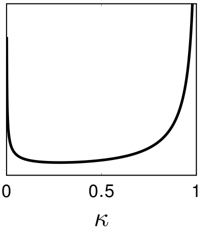
- Given the hyperparameters, the posterior mean satisfies approximately

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

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

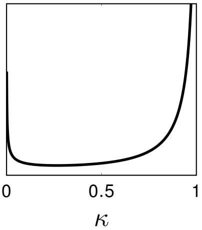
- Given the hyperparameters, the posterior mean satisfies approximately

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

  where κ_j is the *shrinkage factor*

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

$n\sigma^{-2}\tau^2 = 0.3$



We expect both

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

Horseshoe prior

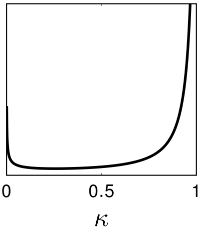
- Given the hyperparameters, the posterior mean satisfies approximately

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

where κ_j is the *shrinkage factor*

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

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



We expect both

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

Horseshoe prior

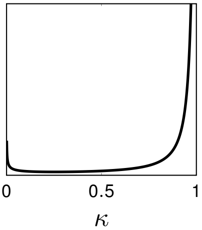
- Given the hyperparameters, the posterior mean satisfies approximately

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

where κ_j is the *shrinkage factor*

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

$n\sigma^{-2}\tau^2 = 0.1$



We expect both

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

Horseshoe prior

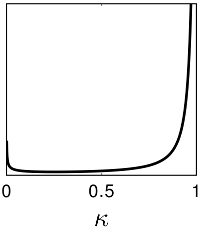
- Given the hyperparameters, the posterior mean satisfies approximately

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

where κ_j is the *shrinkage factor*

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

$n\sigma^{-2}\tau^2 = 0.1$



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

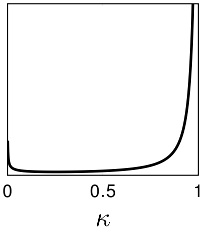
- Given the hyperparameters, the posterior mean satisfies approximately

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

  where κ_j is the *shrinkage factor*

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

$n\sigma^{-2}\tau^2 = 0.1$



We expect both

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

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

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

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

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

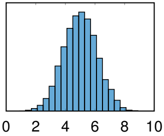
Illustration $p(\tau)$ vs. $p(m_{\text{eff}})$

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

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

$\tau = \tau_0$

$D = 10$



$D = 1000$

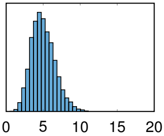


Illustration $p(\tau)$ vs. $p(m_{\text{eff}})$

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

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

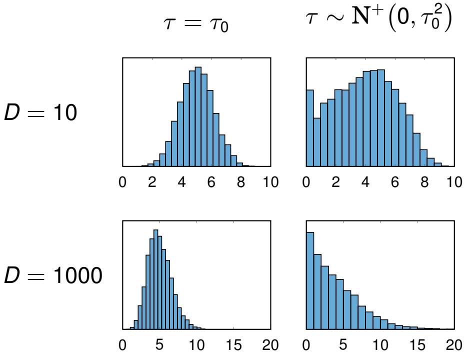


Illustration $p(\tau)$ vs. $p(m_{\text{eff}})$

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 $p(\tau)$:

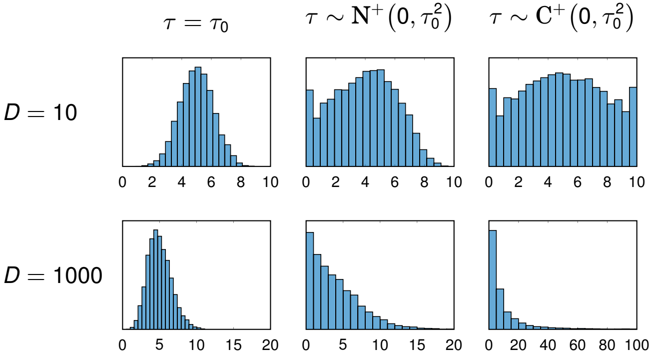
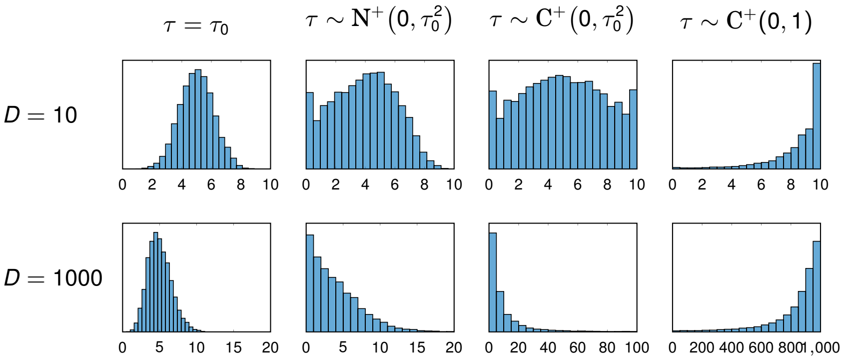


Illustration $p(\tau)$ vs. $p(m_{\text{eff}})$

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 (reminder):

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

- 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

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
Corn (4 targets)	Regression	700	80

- Models implemented and posterior inference using Stan[1].

[1] http://mc-stan.org/

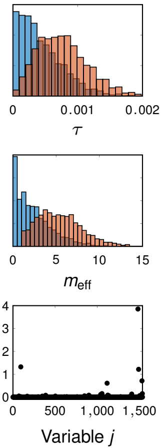
Effect of $p(\tau)$ on parameter estimates

Ovarian cancer data ($n = 54, D = 1536$).

Choose τ_0 according to a prior guess $p_0 = 3$.

Effect of $p(\tau)$ on parameter estimates

$\tau \sim \mathrm{N}^{+}(0, \tau_0^2)$

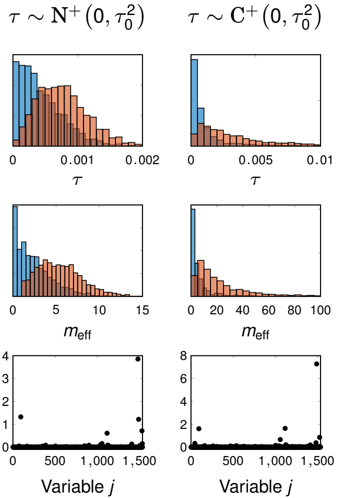


Ovarian cancer data ($n = 54$, $D = 1536$).

Choose τ_0 according to a prior guess $p_0 = 3$.

Prior and posterior samples for τ and m_{eff}, and absolute posterior mean coefficients $|\bar{\beta}_j|$.

Effect of $p(\tau)$ on parameter estimates

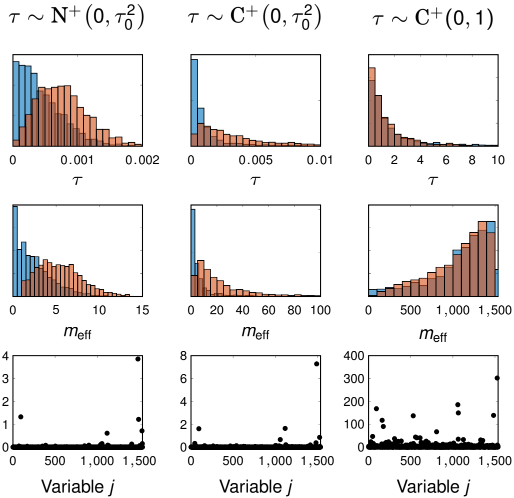


Ovarian cancer data ($n = 54$, $D = 1536$).

Choose τ_0 according to a prior guess $p_0 = 3$.

Prior and posterior samples for τ and m_{eff}, and absolute posterior mean coefficients $|\bar{\beta}_j|$.

Effect of $p(\tau)$ on parameter estimates



Ovarian cancer data ($n = 54$, $D = 1536$).

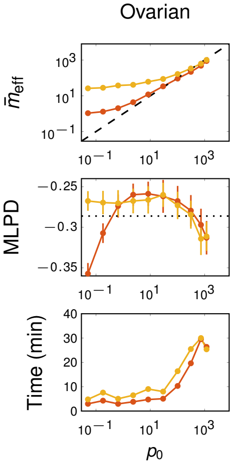
Choose τ_0 according to a prior guess $p_0 = 3$.

Prior and posterior samples for τ and m_{eff}, and absolute posterior mean coefficients $|\bar{\beta}_j|$.

Effect of $p(\tau)$ on prediction accuracy (1/2)

$\tau \sim \mathrm{N}^+(0, \tau_0^2)$ (red) and $\tau \sim \mathrm{C}^+(0, \tau_0^2)$ (yellow),
for various p_0 transformed into τ_0 (largest p_0 corresponds to $\tau_0 = 1$).

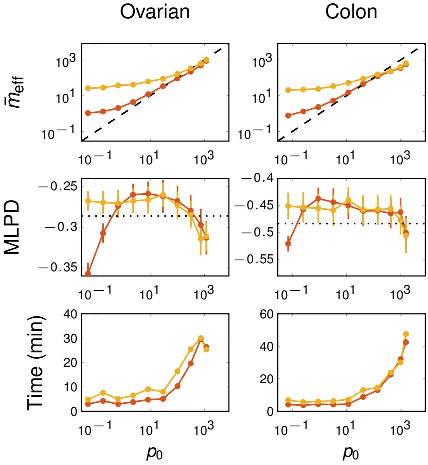
Effect of $p(\tau)$ on prediction accuracy (1/2)



$\tau \sim \mathrm{N}^+(0, \tau_0^2)$ (red) and $\tau \sim \mathrm{C}^+(0, \tau_0^2)$ (yellow),
for various p_0 transformed into τ_0 (largest p_0 corresponds to $\tau_0 = 1$).

Posterior mean $\bar{m}_{\rm eff}$, mean log predictive density (MLPD) on test data (dashed line denotes LASSO), and computation time.

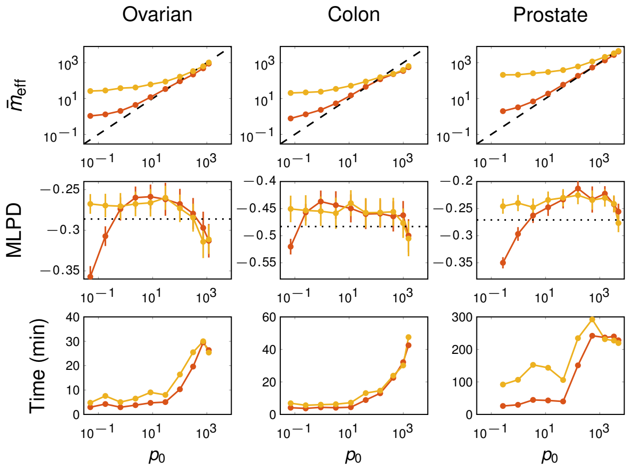
Effect of $p(\tau)$ on prediction accuracy (1/2)



$\tau \sim \mathrm{N}^+(0, \tau_0^2)$ (red) and $\tau \sim \mathrm{C}^+(0, \tau_0^2)$ (yellow),
for various p_0 transformed into τ_0 (largest p_0 corresponds to $\tau_0 = 1$).

Posterior mean $\bar{m}_{\mathrm{eff}}$, mean log predictive density (MLPD) on test data (dashed line denotes LASSO), and computation time.

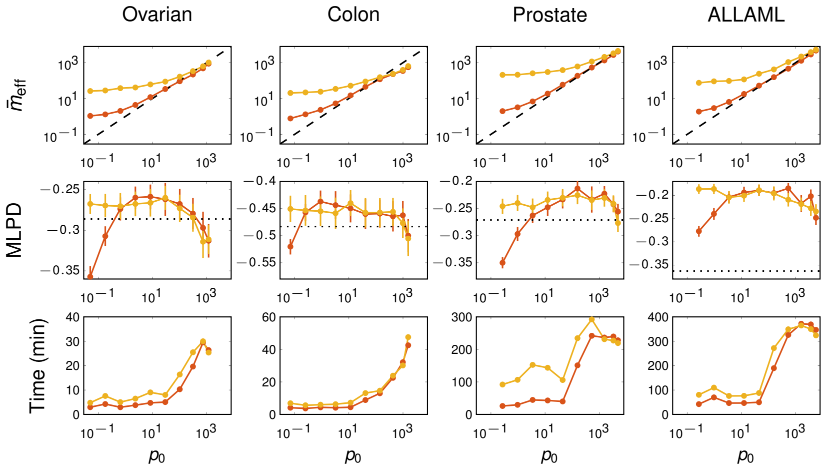
Effect of $p(\tau)$ on prediction accuracy (1/2)



$\tau \sim \mathrm{N}^+(0, \tau_0^2)$ (red) and $\tau \sim \mathrm{C}^+(0, \tau_0^2)$ (yellow),
for various p_0 transformed into τ_0 (largest p_0 corresponds to $\tau_0 = 1$).

Posterior mean $\bar{m}_{\text{eff}}$, mean log predictive density (MLPD) on test data (dashed line denotes LASSO), and computation time.

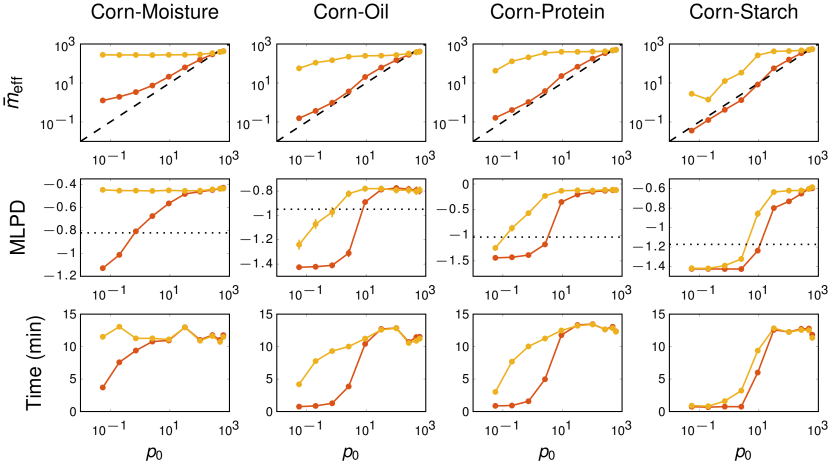
Effect of $p(\tau)$ on prediction accuracy (1/2)



$\tau \sim \mathrm{N}^+(0, \tau_0^2)$ (red) and $\tau \sim \mathrm{C}^+(0, \tau_0^2)$ (yellow),
for various p_0 transformed into τ_0 (largest p_0 corresponds to $\tau_0 = 1$).

Posterior mean $\bar{m}_{\text{eff}}$, mean log predictive density (MLPD) on test data (dashed line denotes LASSO), and computation time.

Effect of $p(\tau)$ on prediction accuracy (2/2)



$\tau \sim \mathrm{N}^+(0, \tau_0^2)$ (red) and $\tau \sim \mathrm{C}^+(0, \tau_0^2)$ (yellow),
for various p_0 transformed into τ_0 (largest p_0 corresponds to $\tau_0 = 1$).

Posterior mean $\bar{m}_{\text{eff}}$, mean log predictive density (MLPD) on test data (dashed line denotes LASSO), and computation time.

Summary

- The global shrinkage parameter τ effectively determines the level of sparsity
- The prior for $p(\tau)$ can have a significant effect on the inference results
  - "Uninformative" $\tau \sim \mathrm{C}^+(0, 1)$ often poor choice
- Our framework allows the user to calibrate the prior for τ based on the prior beliefs about the sparsity
- The concept of effective number of nonzero regression coefficients m_{eff} could be applied also to other shrinkage priors

Implementation

- Horseshoe prior is implemented at least in R-packages *rstanarm* and *brms*
  - Both allow prior specification for the global parameter τ
- Demo about the model fitting and the subsequent projective variable selection using our R-package *projpred*:

  `https://users.aalto.fi/~jtpiiron/projpred/quickstart.html`