

# Advanced Bayesian Learning

## Regularization and Variable Selection - Lecture 1

Mattias Villani

**Department of Statistics  
Stockholm University**



# Topic overview

## ■ Bayesian regularization priors

- ▶ Ridge prior
- ▶ Lasso prior
- ▶ Horseshoe prior
- ▶ Dynamic shrinkage priors

## ■ Bayesian variable selection

- ▶ Spike-and-slab variable selection regression
- ▶ Polya-Gamma augmentation for logistic regression
- ▶ Extensions

# Ridge regression (L2-regularized)

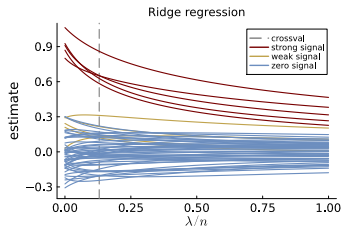
- Minimization of L2-penalized sum of squares

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^T \boldsymbol{\beta}$$

gives **Ridge regression**

$$\tilde{\boldsymbol{\beta}} = \left( \mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_p \right)^{-1} \mathbf{X}^T \mathbf{y}$$

- **Shrinkage** toward zero: as  $\lambda \rightarrow \infty$ ,  $\tilde{\boldsymbol{\beta}} \rightarrow 0$ .
- Prevents overfitting.
- Numerical stability. Can handle  $p \gg n$  case.
- Estimate  $\lambda$  by cross-validation.



# Ridge regression is an iid normal prior

- Ridge regression minimizes L2-penalized sum of squares

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^T \boldsymbol{\beta}$$

- Corresponds to the posterior mean under iid normal prior

$$\beta_j | \sigma^2 \stackrel{\text{iid}}{\sim} N\left(0, \frac{\sigma^2}{\lambda}\right)$$

- Note that

$$\log p(\boldsymbol{\beta} | \sigma^2, \mathbf{y}, \mathbf{X}) \propto -\frac{1}{2\sigma^2} \left[ (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^T \boldsymbol{\beta} \right]$$

so **penalty** = log **prior**.

- Gaussian has thin tails. No extreme values.
- Prior beliefs: all  $\beta_j$  are roughly of the same size.

# Recall: Linear regression - conjugate prior

## ■ Joint prior for $\beta$ and $\sigma^2$

$$\begin{aligned}\beta|\sigma^2 &\sim N(\mu_0, \sigma^2\Omega_0^{-1}) \\ \sigma^2 &\sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2)\end{aligned}$$

## ■ Posterior

$$\begin{aligned}\beta|\sigma^2, y &\sim N(\mu_n, \sigma^2\Omega_n^{-1}) \\ \sigma^2|y &\sim \text{Inv-}\chi^2(\nu_n, \sigma_n^2)\end{aligned}$$

$$\mu_n = \left(X^\top X + \Omega_0\right)^{-1} \left(X^\top X\hat{\beta} + \Omega_0\mu_0\right)$$

$$\Omega_n = X^\top X + \Omega_0$$

$$\nu_n = \nu_0 + n$$

$$\sigma_n^2 = \left(\nu_0\sigma_0^2 + y^\top y + \mu_0^\top \Omega_0 \mu_0 - \mu_n^\top \Omega_n \mu_n\right) / \nu_n$$

# Direct sampling L2-regularization parameter

- **Cross-validation** used to determine degree of smoothness,  $\lambda$ .
- Bayesian:  $\lambda$  is **unknown**  $\Rightarrow$  **put a prior** for  $\lambda$ !
- The joint posterior of  $\beta$ ,  $\sigma^2$  and  $\lambda$  is ( $\Omega_0(\lambda) = \lambda I$ )

$$\beta | \sigma^2, \lambda, y, X \sim N(\mu_n, \Omega_n^{-1})$$

$$\sigma^2 | \lambda, y, X \sim \text{Inv} - \chi^2(\nu_n, \sigma_n^2)$$

$$p(\lambda | y, X) \propto \sqrt{\frac{|\Omega_0(\lambda)|}{|\mathbf{X}^\top \mathbf{X} + \Omega_0(\lambda)|}} \left( \frac{\nu_n \sigma_n^2(\lambda)}{2} \right)^{-\nu_n/2} \cdot p(\lambda)$$

- This is the **conditional-marginal decomposition**

$$p(\beta, \sigma^2, \lambda | y, X) = p(\beta | \sigma^2, \lambda, y, X) p(\sigma^2 | \lambda, y, X) p(\lambda | y, X)$$

# Gibbs sampling for L2-regularized regression

- Prior:

$$\begin{aligned}\beta|\sigma^2, \lambda &\sim N\left(0, \frac{\sigma^2}{\lambda} I_p\right) \\ \sigma^2 &\sim \text{Inv} - \chi^2(\nu_0, \sigma_0^2) \\ \lambda^{-1} &\sim \text{Inv} - \chi^2(\omega_0, \psi_0^2) .\end{aligned}$$

- By Bayes' theorem

$$p(\lambda|\beta, \sigma^2, y) \propto p(y|\beta, \sigma^2, \lambda) p(\lambda|\beta, \sigma^2)$$

- $p(y|\beta, \sigma^2, \lambda)$  does not depend on  $\lambda$  once we condition on  $\beta$ :

$$p(\lambda|\beta, \sigma^2, y) \propto p(\lambda|\beta, \sigma^2)$$

- So using Bayes' theorem once more

$$p(\lambda|\beta, \sigma^2) \propto p(\beta|\sigma^2, \lambda) p(\lambda)$$

- In conditional posterior for  $\lambda$ , the  $\beta_1, \dots, \beta_p$  act like “data”.

# Gibbs sampling L2-regularized regression $\psi^2 = \frac{1}{\lambda}$

## Gibbs sampling linear regression - L2 regularization prior

The posterior for the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I_n), \quad (12.16)$$

with hierarchical L2 regularization prior

$$\begin{aligned}\boldsymbol{\beta} | \sigma^2, \psi^2 &\sim N(\mathbf{0}, \sigma^2 \psi^2 I_p) \\ \sigma^2 &\sim \text{Inv-}\chi^2(\nu_0, \tau_0^2) \\ \psi^2 &\sim \text{Inv-}\chi^2(\omega_0, \psi_0^2).\end{aligned}$$

can be sampled by a two-block Gibbs sampler:

$$\begin{aligned}\text{Block1 : } \boldsymbol{\beta} | \sigma^2, \psi^2, \mathbf{y} &\sim N(\hat{\boldsymbol{\beta}}_{L_2}, \sigma^2 (\mathbf{X}^\top \mathbf{X} + \psi^{-2} I_p)^{-1}) \\ \sigma^2 | \psi^2, \mathbf{y} &\sim \text{Inv-}\chi^2(\tau_n^2, \nu_n)\end{aligned}$$

$$\text{Block2 : } \psi^2 | \boldsymbol{\beta}, \sigma^2, \mathbf{y} \sim \text{Inv-}\chi^2(\omega_n, \psi_n^2),$$

where  $\hat{\boldsymbol{\beta}}_{L_2}$  is the ridge estimator

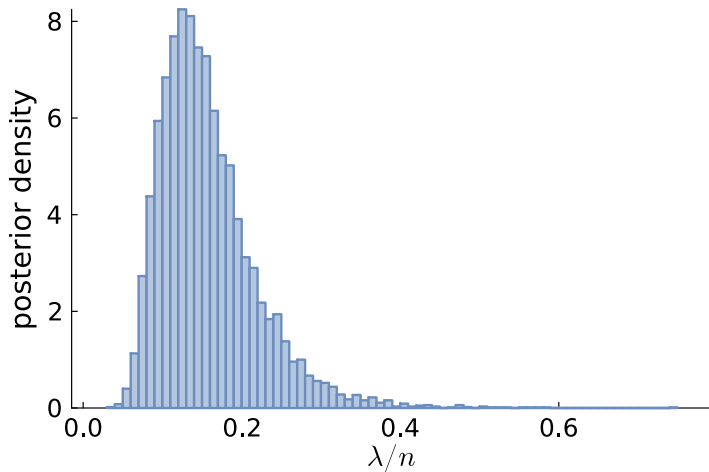
$$\hat{\boldsymbol{\beta}}_{L_2} = (\mathbf{X}^\top \mathbf{X} + \psi^{-2} I_p)^{-1} \mathbf{X}^\top \mathbf{y} = (\mathbf{X}^\top \mathbf{X} + \lambda I_p)^{-1} \mathbf{X}^\top \mathbf{y}.$$

The hyperparameters  $\nu_n$  and  $\tau_n^2$  are given in Figure 5.3.

Finally,  $\omega_n = \omega_0 + p$  and  $\psi_n^2 = (\sum_{i=1}^p (\beta_i / \sigma)^2 + \omega_0 \psi_0^2) / \omega_n$ .



## Marginal posterior of $\lambda$



# Regularization prior - Lasso

- **Lasso** is equivalent to posterior mode under Laplace prior

$$\beta_i | \sigma^2 \stackrel{\text{iid}}{\sim} \text{Laplace} \left( 0, \frac{\sigma^2}{2\lambda} \right)$$

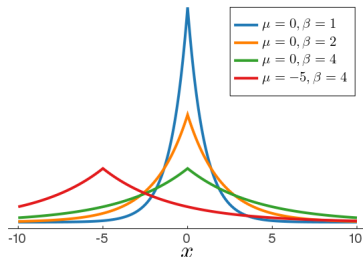
## Laplace distribution

$X \sim \text{Laplace}(\mu, \beta)$  for  $X \in \mathbb{R}$ .

$$p(x) = \frac{1}{2\beta} \exp \left( -\frac{|x - \mu|}{\beta} \right)$$

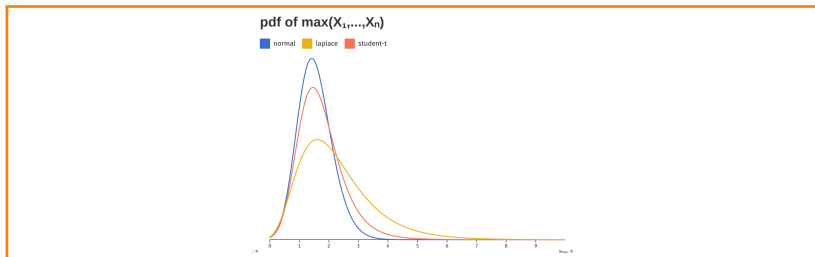
$$\mathbb{E}(X) = \mu$$

$$\mathbb{V}(X) = 2\beta^2$$

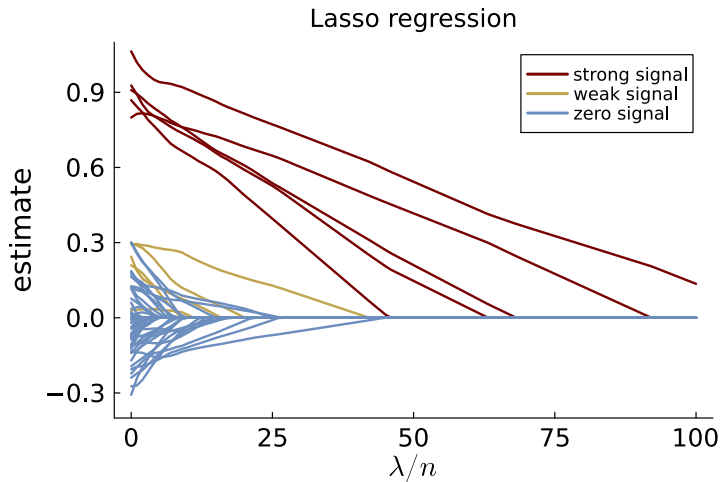


- Laplace distribution have heavy tails.
- **Laplace**: many  $\beta_i$  close to zero, but some  $\beta_i$  rather large.

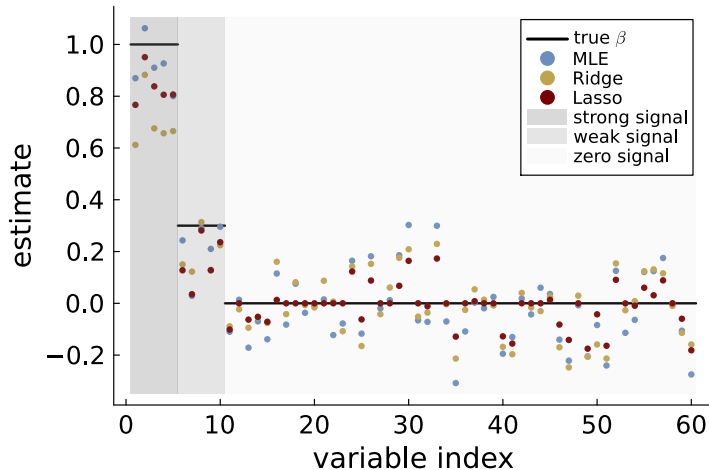
# A tale of tails



# Lasso/Laplace prior



# Ridge vs Lasso shrinkage



# Horseshoe prior

- Normal and Laplace - only one global shrinkage parameter  $\lambda$ .
- **Global-Local shrinkage**: global + local shrinkage for each  $\beta_j$ .
- **Horseshoe prior**:

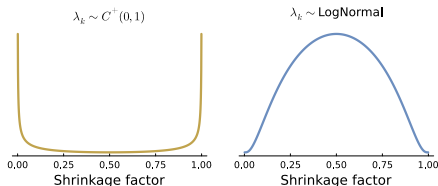
$$\beta_j | \lambda_j^2, \tau^2 \sim N(0, \tau^2 \lambda_j^2)$$

$$\lambda_j \stackrel{\text{iid}}{\sim} C^+(0, 1)$$

$$\tau \sim C^+(0, 1)$$

- **Shrinkage factor**  $c_j$  (orthogonal covariates)

$$\tilde{\beta}_j = (1 - c_j) \hat{\beta}_j, \quad c_j = \frac{1}{1 + (n/\sigma^2) \tau^2 \lambda_j^2}$$



# Gibbs sampling for regression with horseshoe prior

- $X \sim C^+(0, 1)$  can be generated by continuous mixture:

$$Y \sim \text{Inv-}\chi^2(1, 2)$$
$$X^2 | Y \sim \text{Inv-}\chi^2(1, 2/Y)$$

- **Horseshoe prior** in mixture formulation:

$$\beta | \lambda_1, \dots, \lambda_p, \tau^2, \sigma^2 \mathbf{\Lambda} \sim N(0, \sigma^2 \tau^2 \mathbf{\Lambda})$$
$$\lambda_j^2 | \nu_j \stackrel{\text{inde}}{\sim} \text{Inv-}\chi^2(1, 2/\nu_j)$$
$$\tau^2 | \xi \sim \text{Inv-}\chi^2(1, 2/\xi)$$
$$\nu_1, \dots, \nu_p, \xi \stackrel{\text{iid}}{\sim} \text{Inv-}\chi^2(1, 2)$$

where  $\mathbf{\Lambda} = \text{Diag}(\lambda_1^2, \dots, \lambda_p^2)$ .

# Gibbs sampling for regression with horseshoe prior

## ■ Gibbs sampler

$\beta, \sigma | \Lambda, \mathbf{y}, \mathbf{X} \sim \text{Linear regression with } \Omega_0^{-1} = \tau^2 \Lambda$

$$\nu_j | \lambda_j \stackrel{\text{iid}}{\sim} \text{Inv-}\chi^2(2, 1 + 1/\lambda_j^2)$$

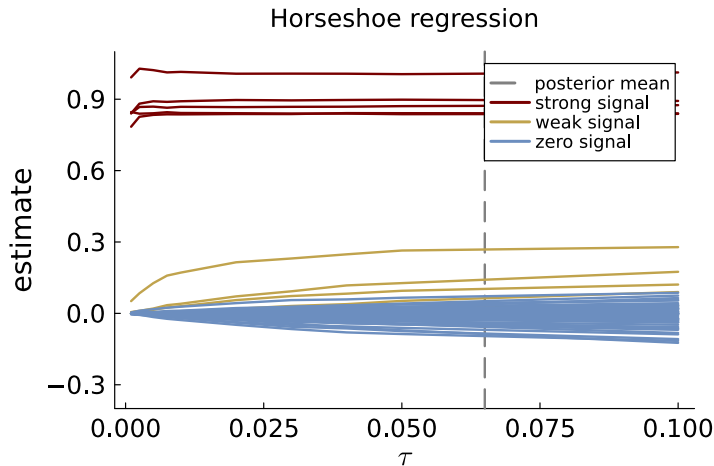
$$\lambda_j^2 | \nu_j, \tau, \beta, \sigma \sim \text{Inv-}\chi^2\left(2, \frac{1}{\nu_j} + \frac{1}{2} \left(\frac{\beta_j}{\sigma\tau}\right)^2\right)$$

$$\xi | \tau \stackrel{\text{iid}}{\sim} \text{Inv-}\chi^2(2, 1 + 1/\tau^2),$$

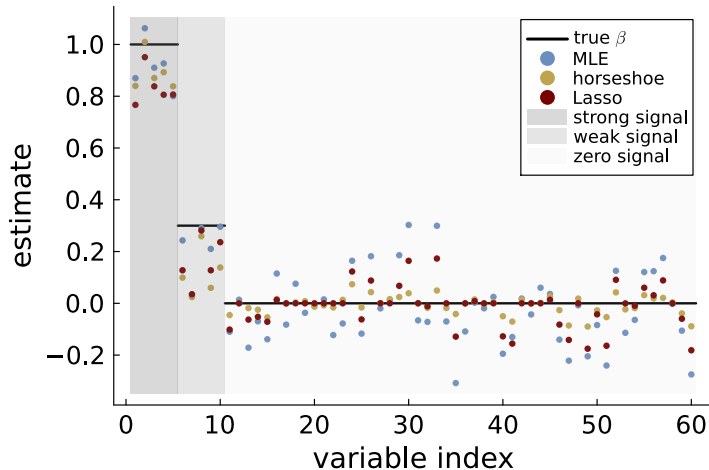
$$\tau^2 | \xi, \lambda_1, \dots, \lambda_p, \beta, \sigma \sim \text{Inv-}\chi^2\left(p + 1, \frac{\frac{2}{\xi} + \sum_{j=1}^p \left(\frac{\beta_j}{\sigma\lambda_j}\right)^2}{p + 1}\right)$$



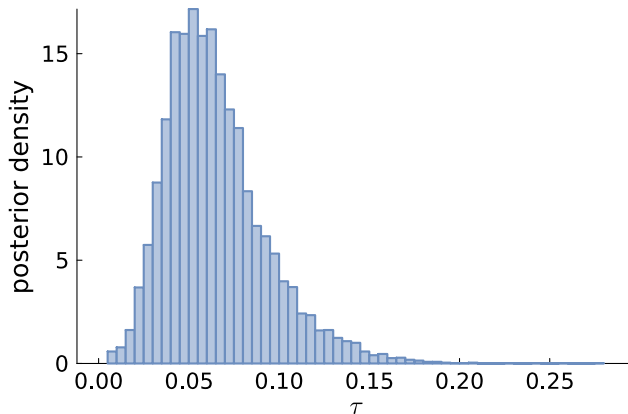
# Horseshoe prior on simulated data



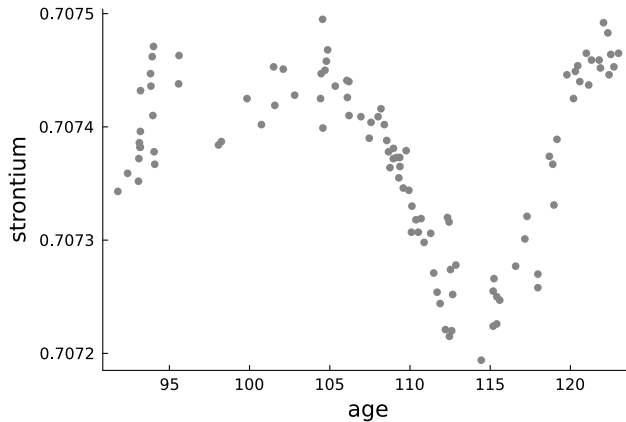
# Horseshoe prior on simulated data



# Horseshoe prior on simulated data



# Spline regression - fossil data case study



# Polynomial regression

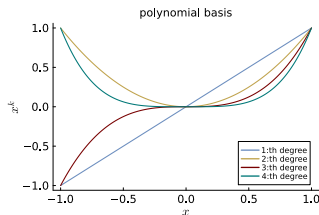
## Polynomial regression

$$f(x_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \dots + \beta_k x_i^k, \quad \text{for } i = 1, \dots, n.$$

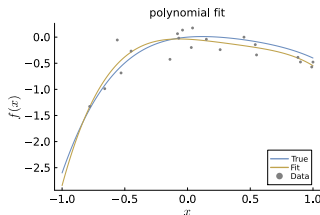
$$y = \mathbf{X}\beta + \varepsilon,$$

$$\mathbf{x}_i = (1, x_i, x_i^2, \dots, x_i^k)^\top$$

Still linear in  $\beta$  and  $\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ . Bayes unchanged.

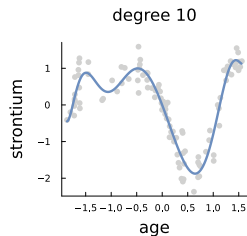
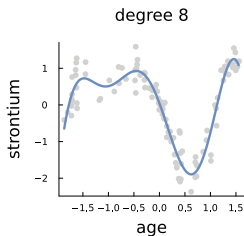
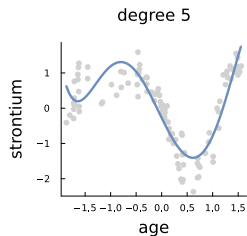
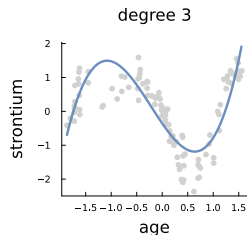
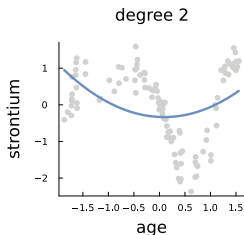
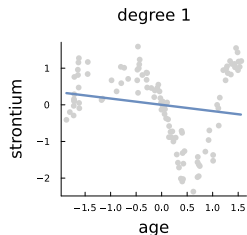


1.00	-1.00	1.00	-1.00	1.00
1.00	-0.90	0.81	-0.73	0.66
1.00	-0.80	0.64	-0.51	0.41
1.00	-0.70	0.49	-0.34	0.24
1.00	-0.60	0.36	-0.22	0.13
1.00	-0.50	0.25	-0.12	0.06
1.00	-0.40	0.16	-0.06	0.03
1.00	-0.30	0.09	-0.03	0.01
1.00	-0.20	0.04	-0.01	0.00
1.00	-0.10	0.01	-0.00	0.00
1.00	0.00	0.00	0.00	0.00
1.00	0.10	0.01	0.00	0.00
1.00	0.20	0.04	0.01	0.00
1.00	0.30	0.09	0.03	0.01
1.00	0.40	0.16	0.06	0.03
1.00	0.50	0.25	0.12	0.06
1.00	0.60	0.36	0.22	0.13
1.00	0.70	0.49	0.34	0.24
1.00	0.80	0.64	0.51	0.41
1.00	0.90	0.81	0.73	0.66
1.00	1.00	1.00	1.00	1.00



Polynomials are global basis functions. Local basis preferred.

# Polynomial regression - fossil data



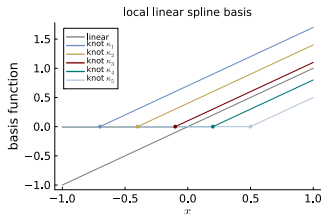
# Spline regression - local linear basis

- Truncated linear splines with knot locations  $\kappa_1, \dots, \kappa_m$ :

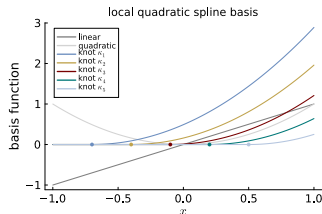
$$b_j(x) = \begin{cases} |x - \kappa_j|^p & \text{if } x > \kappa_j \\ 0 & \text{otherwise} \end{cases}$$

$$y = \mathbf{X}\beta + \varepsilon,$$

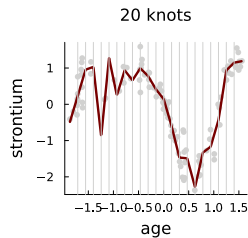
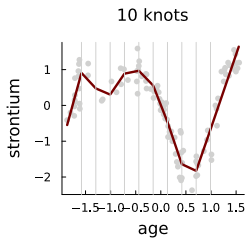
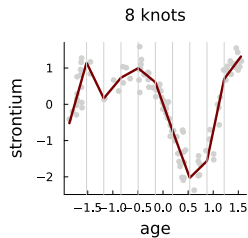
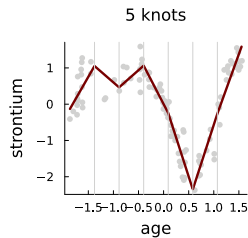
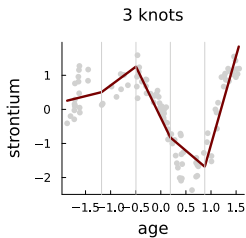
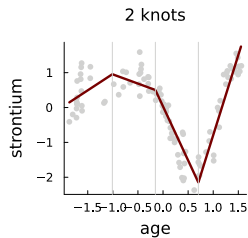
$$\mathbf{x}_i = (1, x_i, b_1(x_i), \dots, b_m(x_i))^T$$



1.00	-1.00	-0.00	-0.00	-0.00
1.00	-0.90	-0.00	-0.00	-0.00
1.00	-0.80	-0.00	-0.00	-0.00
1.00	-0.70	-0.00	-0.00	-0.00
1.00	-0.60	-0.00	-0.00	-0.00
1.00	-0.50	0.00	-0.00	-0.00
1.00	-0.40	0.10	-0.00	-0.00
1.00	-0.30	0.20	-0.00	-0.00
1.00	-0.20	0.30	-0.00	-0.00
1.00	-0.10	0.40	-0.00	-0.00
1.00	0.00	0.50	0.00	-0.00
1.00	0.10	0.60	0.10	-0.00
1.00	0.20	0.70	0.20	-0.00
1.00	0.30	0.80	0.30	-0.00
1.00	0.40	0.90	0.40	-0.00
1.00	0.50	1.00	0.50	0.00
1.00	0.60	1.10	0.60	0.10
1.00	0.70	1.20	0.70	0.20
1.00	0.80	1.30	0.80	0.30
1.00	0.90	1.40	0.90	0.40
1.00	1.00	1.50	1.00	0.50

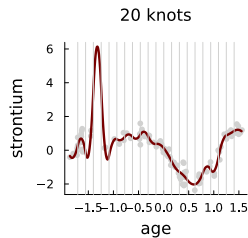
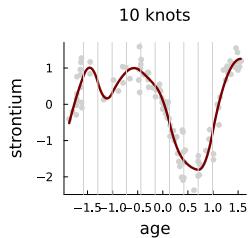
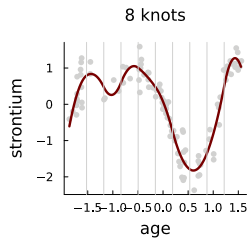
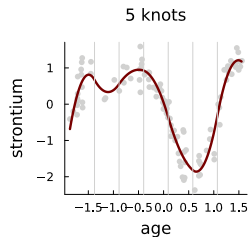
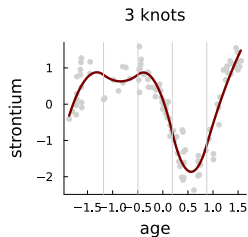
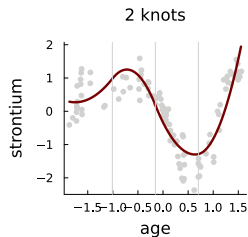


# Spline regression - local linear basis

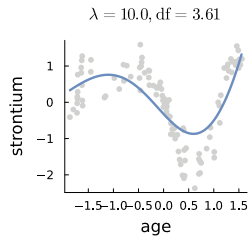
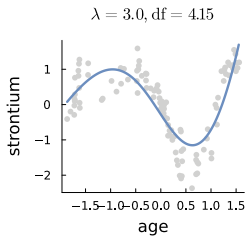
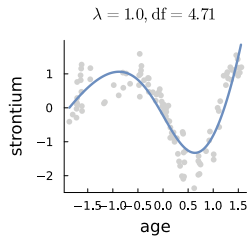
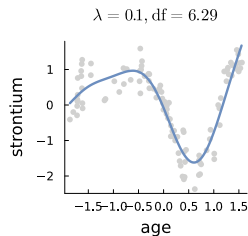
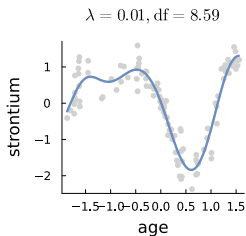
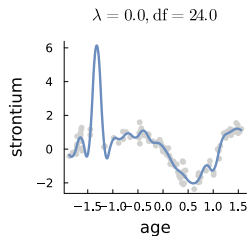




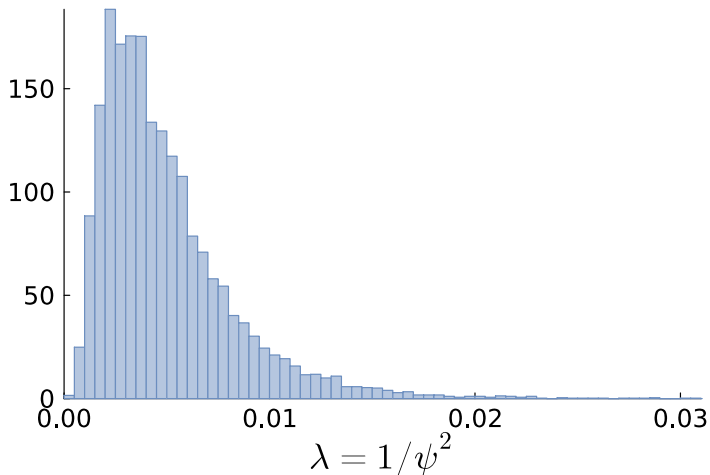
# Spline regression - local quadratic basis



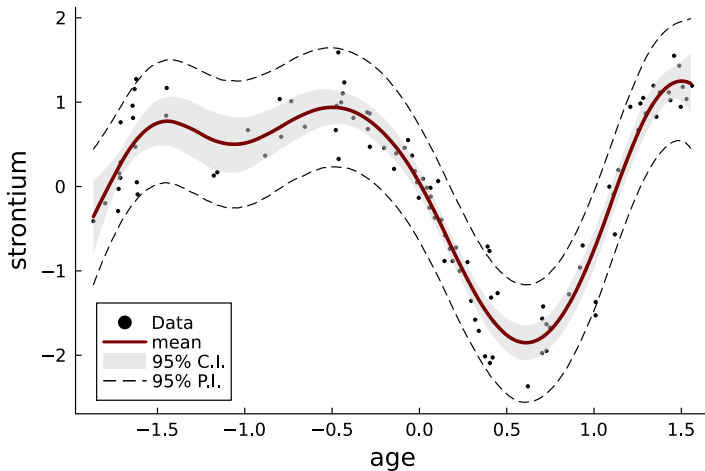
# Spline regression - L2-regularization



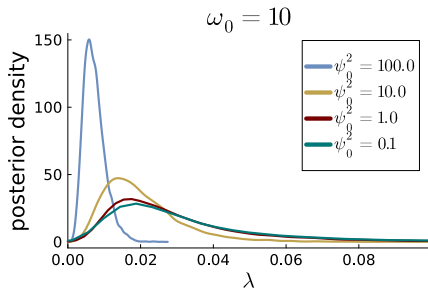
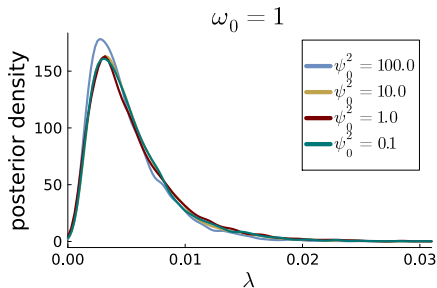
## Spline regression - posterior for $\lambda$



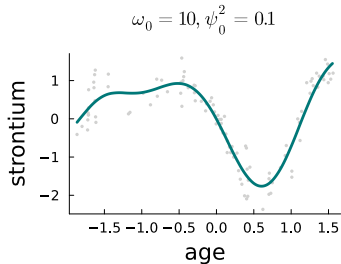
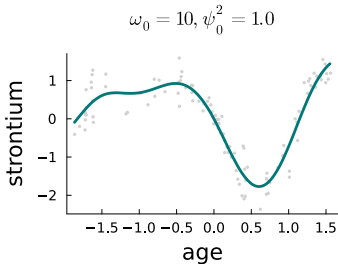
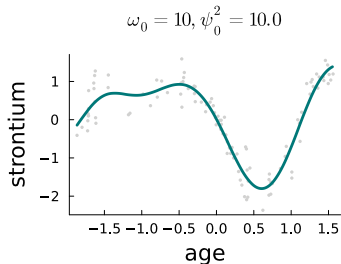
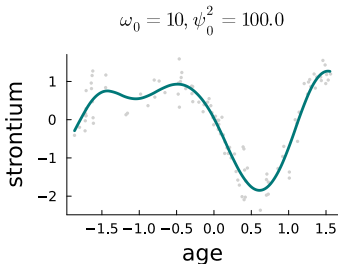
# Posterior and predictive distribution



# Prior sensitivity $\lambda = 1/\psi^2$ for $\psi^2 \sim \text{Inv-}\chi^2(\omega_0, \psi_0^2)$



# Prior sensitivity fit $\psi^2 \sim \text{Inv-}\chi^2(\omega_0, \psi_0^2)$



# Regularization in state-space models

- **Local level model** (state-space) for time series

$$\begin{aligned}x_t &= \mu_t + \varepsilon_t, & \varepsilon_t &\sim N(0, \sigma_\varepsilon^2) \\ \mu_t &= \mu_{t-1} + \nu_t, & \nu_t &\sim N(0, \sigma_\nu^2)\end{aligned}$$

- **Innovation variance**  $\sigma_\nu^2 \Rightarrow$  how fast the mean evolves.
- Same normal  $N(0, \sigma_\nu^2)$  for all  $\nu_t$ . Compare Ridge regression.
- Restrictive parameter evolution. Can't get all of this:
  - 1  $\nu_t \approx 0$  for some  $t$  (parameters stand still)
  - 2 large  $\nu_t$  for some  $t$  (jumps)
  - 3 persistent periods of rapid changes

# Dynamic horseshoe process prior

- **Horseshoe prior** for time series

$$\mu_t = \mu_{t-1} + \nu_t, \quad \nu_t \sim N(0, \tau^2 \lambda_j^2)$$

$$\lambda_t \stackrel{\text{iid}}{\sim} C^+(0, 1)$$

$$\tau \sim C^+(0, 1)$$

- This gives us Property 1 and 2 above. 😊
- Local variances  $\lambda_t^2$  are independent. No Property 3. 😞
- **Dynamic horseshoe process** [1]

$$\mu_t = \mu_{t-1} + \nu_t, \quad \nu_t \sim N(0, \tau^2 \exp(h_t))$$

$$h_t = \phi h_{t-1} + \eta_t, \quad \eta_t \sim Z(1/2, 1/2, 0, 1)$$

$$\tau \sim C^+(0, 1)$$

- The horseshoe prior is the special case with  $\phi = 0$

$$\eta_t \sim Z(1/2, 1/2, 0, 1) \iff \lambda_t = \exp(\eta_t/2) \sim C^+(0, 1)$$

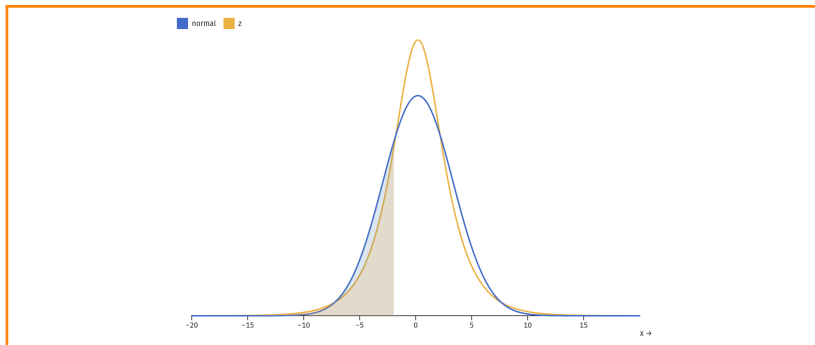


# Z-distribution

- Also called **Logistic-Beta distribution** since

$$X \sim \text{Beta}(\alpha, \beta) \implies \log \left( \frac{X}{1-X} \right) \sim Z(\alpha, \beta, 0, 1)$$

- The  $Z(\alpha, \beta, 0, 1)$  distribution is heavy tailed.
- Linearly decaying log density.





D. R. Kowal, D. S. Matteson, and D. Ruppert, “Dynamic shrinkage processes,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 81, no. 4, pp. 781–804, 2019.