Bayesian Learning

Lecture 5 - Linear Regression. Regularization.

Mattias Villani

Department of Statistics
Stockholm University

mattiasvillani.com @matvil @matvil mattiasvillani

Lecture overview

- Linear regression
- Regularization priors

Linear regression

- The linear regression model in **matrix form**

$$\underset{(n\times 1)}{\mathbf{y}} = \underset{(n\times k)(k\times 1)}{\mathbf{X}\beta} + \underset{(n\times 1)}{\varepsilon}$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \ \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \ \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix}$$

- Usually $x_{i1} = 1$, for all i. β_1 is the intercept.
- **Likelihood**

$$\mathbf{y}|\beta, \sigma^2, \mathbf{X} \sim N(\mathbf{X}\beta, \sigma^2 I_n)$$

Linear regression - uniform prior

- Standard **non-informative prior**: uniform on $(\beta, \log \sigma^2)$

$$p(\beta, \sigma^2) \propto \sigma^{-2}$$

- **Joint posterior** of β and σ^2:

$$\begin{aligned} \beta|\sigma^2, \mathbf{y} &\sim N\left(\hat{\beta}, \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}\right) \\ \sigma^2|\mathbf{y} &\sim \textit{Inv-}\chi^2(n-k, s^2) \end{aligned}$$

where $\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}$ and $s^2 = \frac{1}{n-k}(\mathbf{y} - \mathbf{X}\hat{\beta})^\top(\mathbf{y} - \mathbf{X}\hat{\beta})$.

- **Simulate** from the joint posterior by simulating from
  - $p(\sigma^2|\mathbf{y})$
  - $p(\beta|\sigma^2, \mathbf{y})$
- **Marginal posterior** of β :

$$\beta|\mathbf{y} \sim t_{n-k}\left(\hat{\beta}, s^2(X^\top X)^{-1}\right)$$

Linear regression - conjugate prior

- **Joint prior** for β and σ^2

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

- **Posterior**

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

$$\begin{aligned}\mu_n &= \left(\mathbf{X}^\top\mathbf{X} + \Omega_0\right)^{-1}\left(\mathbf{X}^\top\mathbf{X}\hat{\beta} + \Omega_0\mu_0\right)\\ \Omega_n &= \mathbf{X}^\top\mathbf{X} + \Omega_0\\ \nu_n &= \nu_0 + n\\ \sigma_n^2 &= \left(\nu_0\sigma_0^2 + \mathbf{y}^\top\mathbf{y} + \mu_0^\top\Omega_0\mu_0 - \mu_n^\top\Omega_n\mu_n\right)/\nu_n\end{aligned}$$

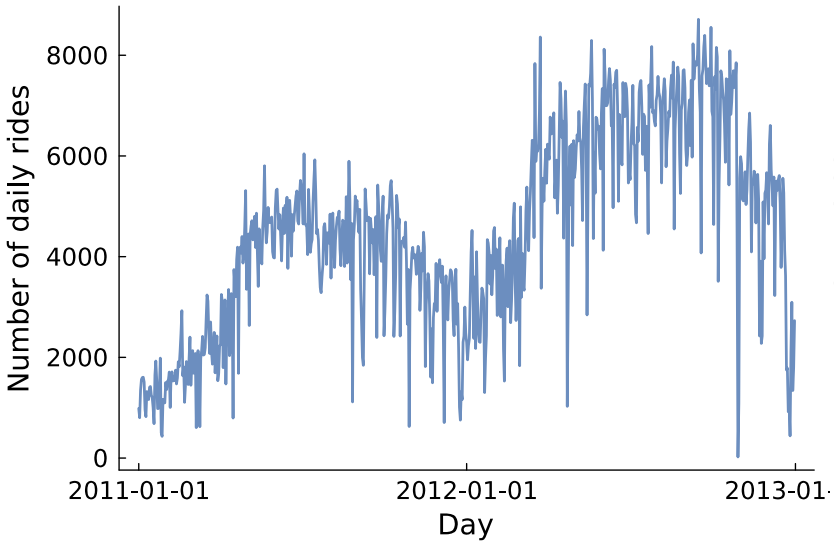
Bike share data

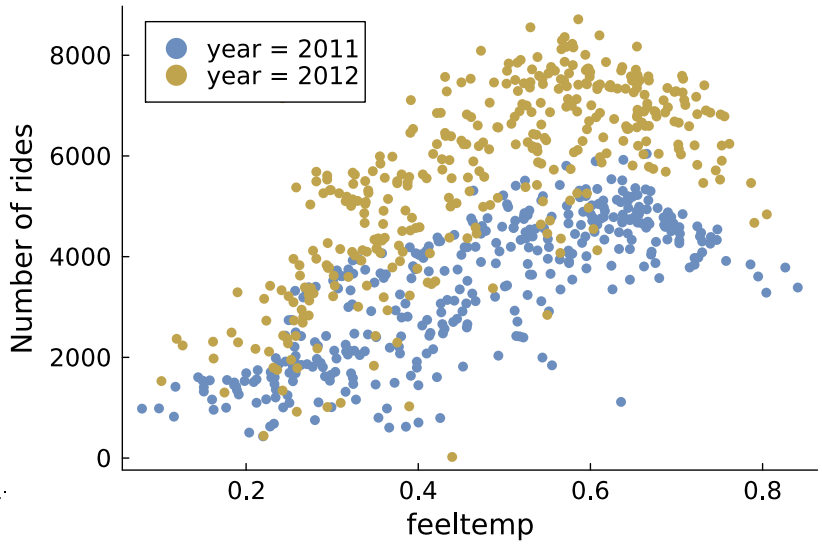
- **Bike share data**. Predict the number of bike rides.
- Response variable: number of rides on 731 days.

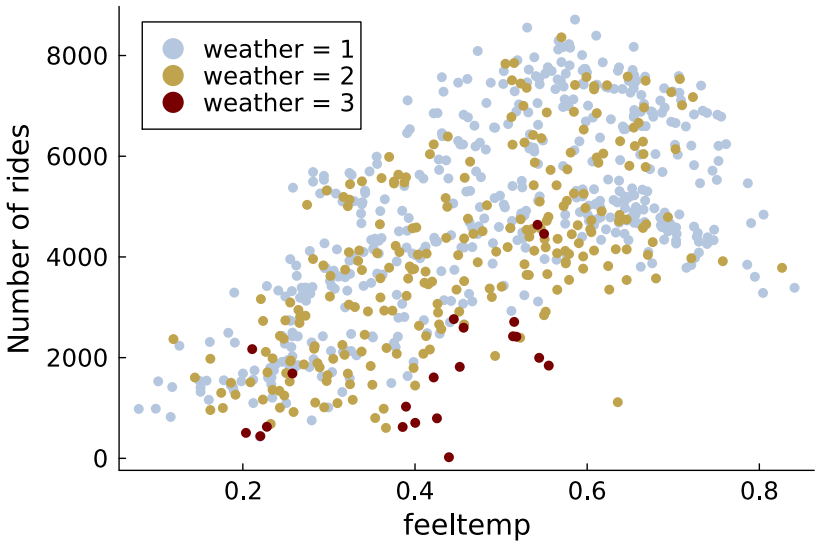
variable	description	type	values	comment
`nrides`	# of rides	counts	$\{0, 1, ...\}$	min$= 22$, max$= 8714$
`feeltemp`	perceived temp	cont.	$[0, 1]$	min$= 0.07$, max$= 0.85$
`hum`	humidity	cont.	$[0, 1]$	min$= 0.00$, max$= 0.98$
`wind`	wind speed	cont.	$[0, 1]$	min$= 0.02$, max$= 0.51$
`year`	year	binary	$\{0, 1\}$	year 2011 $= 0$
`season`	season	cat.	$\{1, 2, 3, 4\}$	winter $\rightarrow$ fall
`weather`	weather	ordinal	$\{1, 2, 3\}$	clear $\rightarrow$ rain/snow
`weekday`	day of week	cat.	$\{0, ..., 6\}$	sunday $\rightarrow$ saturday
`holiday`	holiday	binary	$\{0, 1\}$	holiday $= 1$

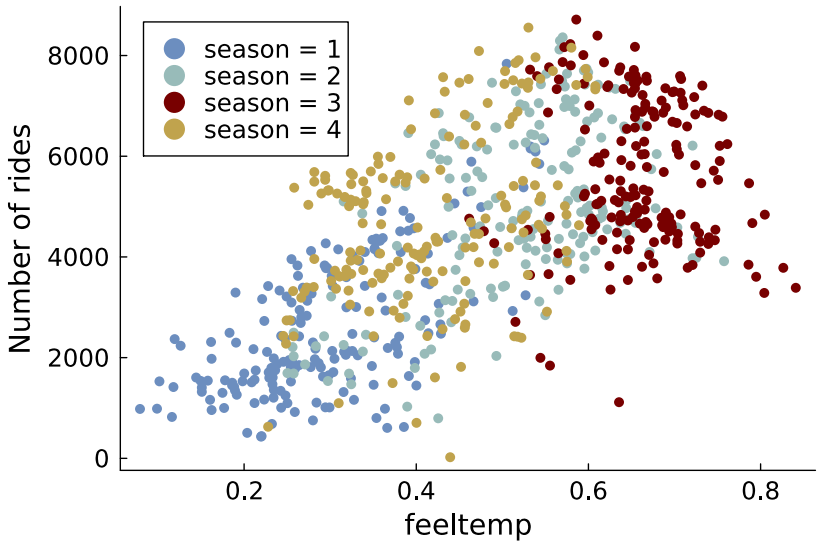
- Prior:
  - $\mu_0 = (1000, 0, \ldots, 0)^\top$
  - $\Omega_0 = \frac{\kappa_0}{n} \boldsymbol{X}^\top \boldsymbol{X}$ with $\kappa_0 = 1$ (unit information prior)
  - $\sigma_0^2 = 1000^2$ and $\nu_0 = 5$.

Bike share data

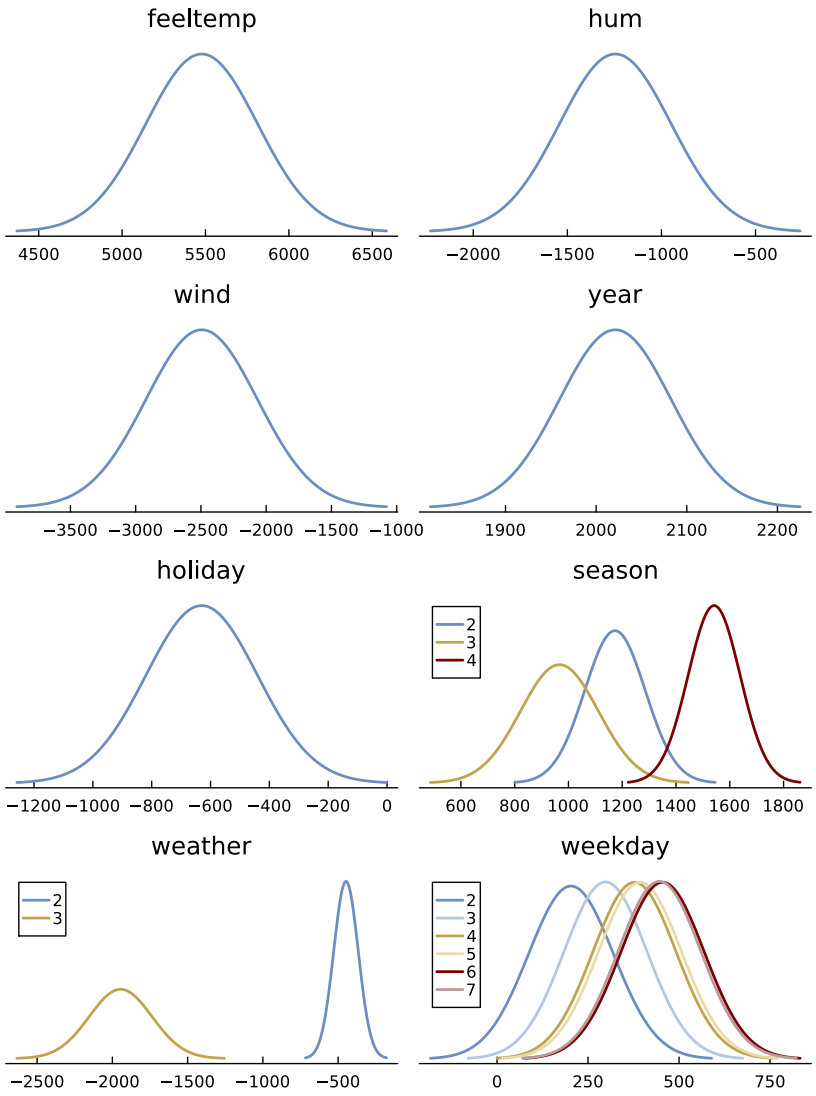




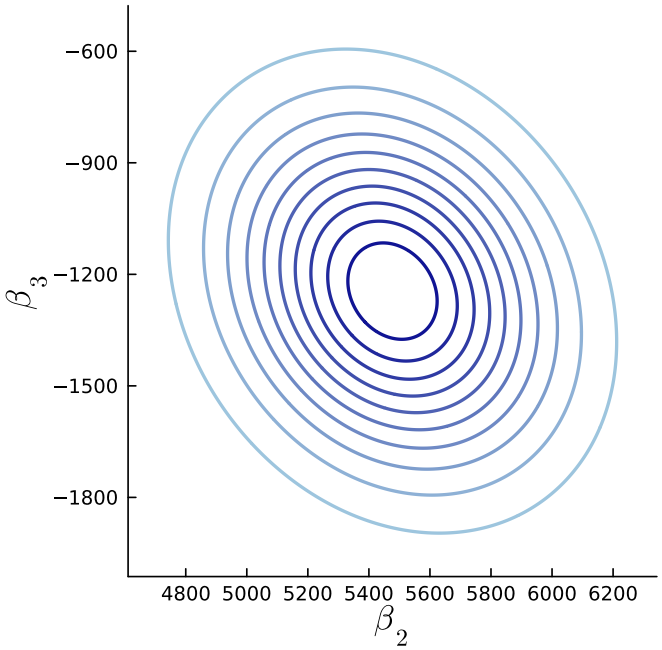


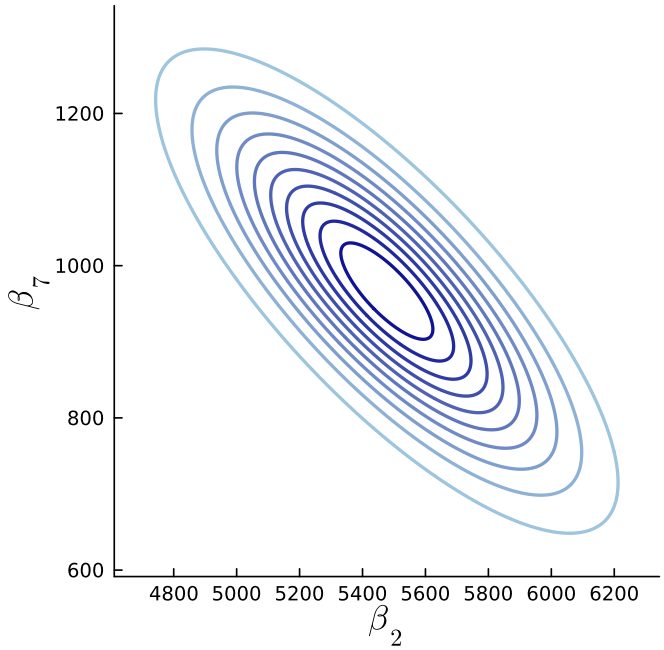


Bike share data - marginal posteriors of β

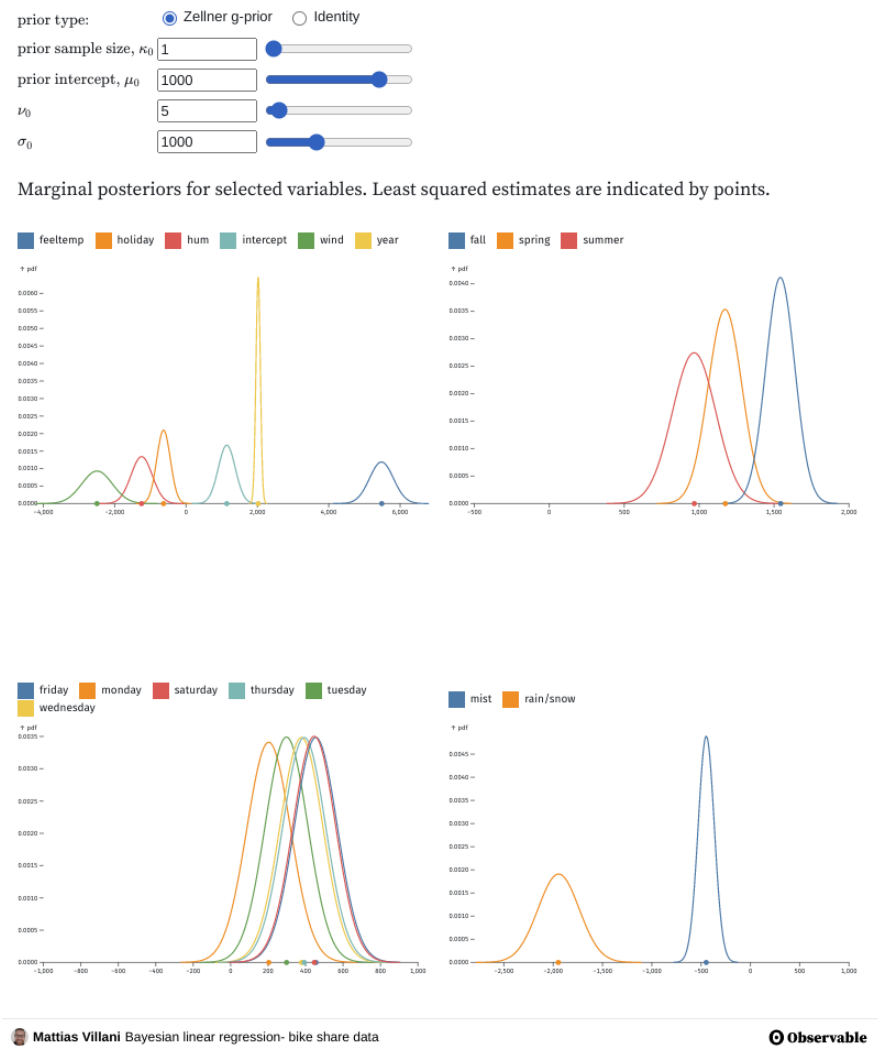


Bike share data - joint posteriors of β





Interactive - Bayesian regression



Polynomial regression

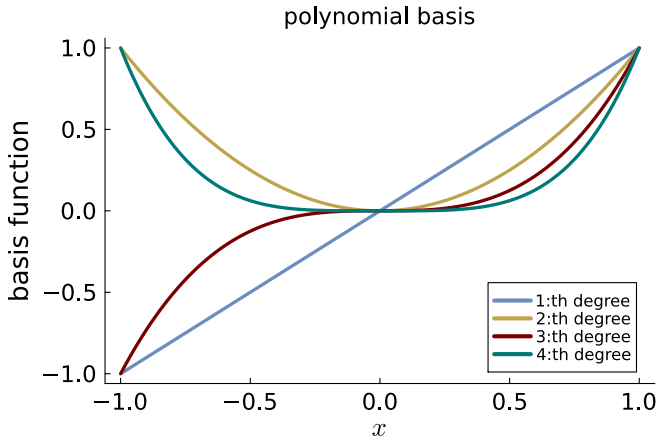
Polynomial regression

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

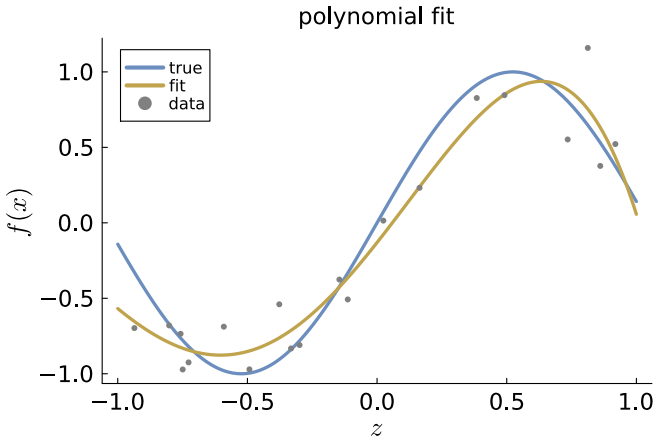
$$\mathbf{y} = \boldsymbol{X}\beta + \varepsilon,$$

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

- Still **linear in** β and $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$. Bayes unchanged.

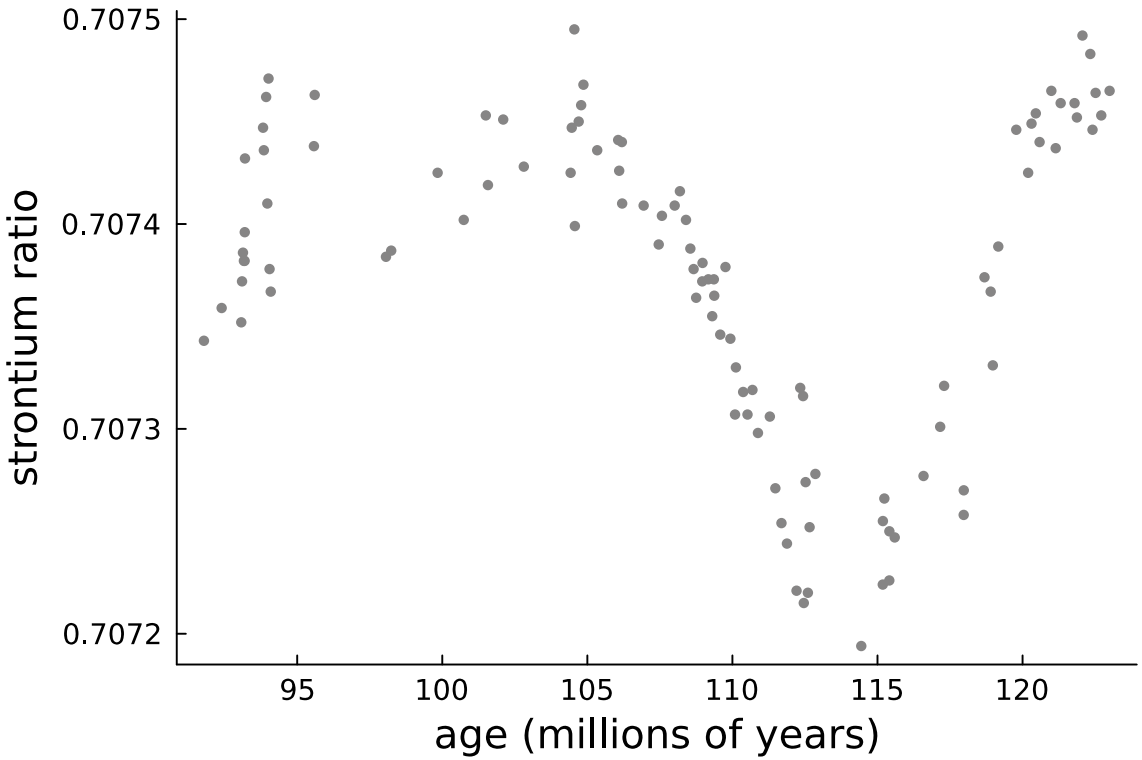


1.00	-1.00	1.00	-1.00	1.00
1.00	-0.90	0.81	-0.73	0.66
1.00	-0.80	0.64	-0.51	0.41
1.00	-0.70	0.49	-0.34	0.24
1.00	-0.60	0.36	-0.22	0.13
1.00	-0.50	0.25	-0.12	0.06
1.00	-0.40	0.16	-0.06	0.03
1.00	-0.30	0.09	-0.03	0.01
1.00	-0.20	0.04	-0.01	0.00
1.00	-0.10	0.01	-0.00	0.00
1.00	0.00	0.00	0.00	0.00
1.00	0.10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



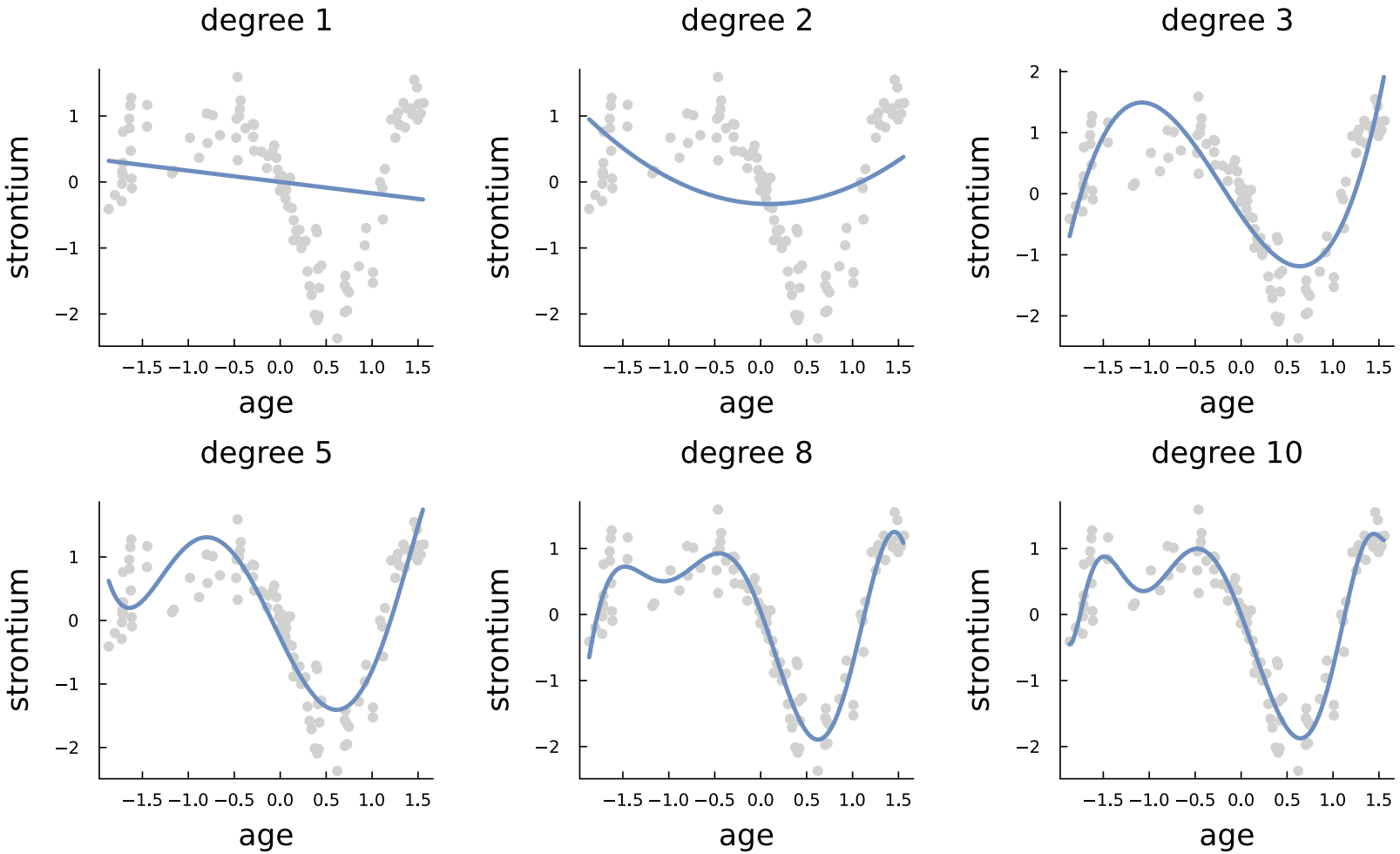
- Polynomials are **global** basis functions. **Local** basis preferred.

Fossil data

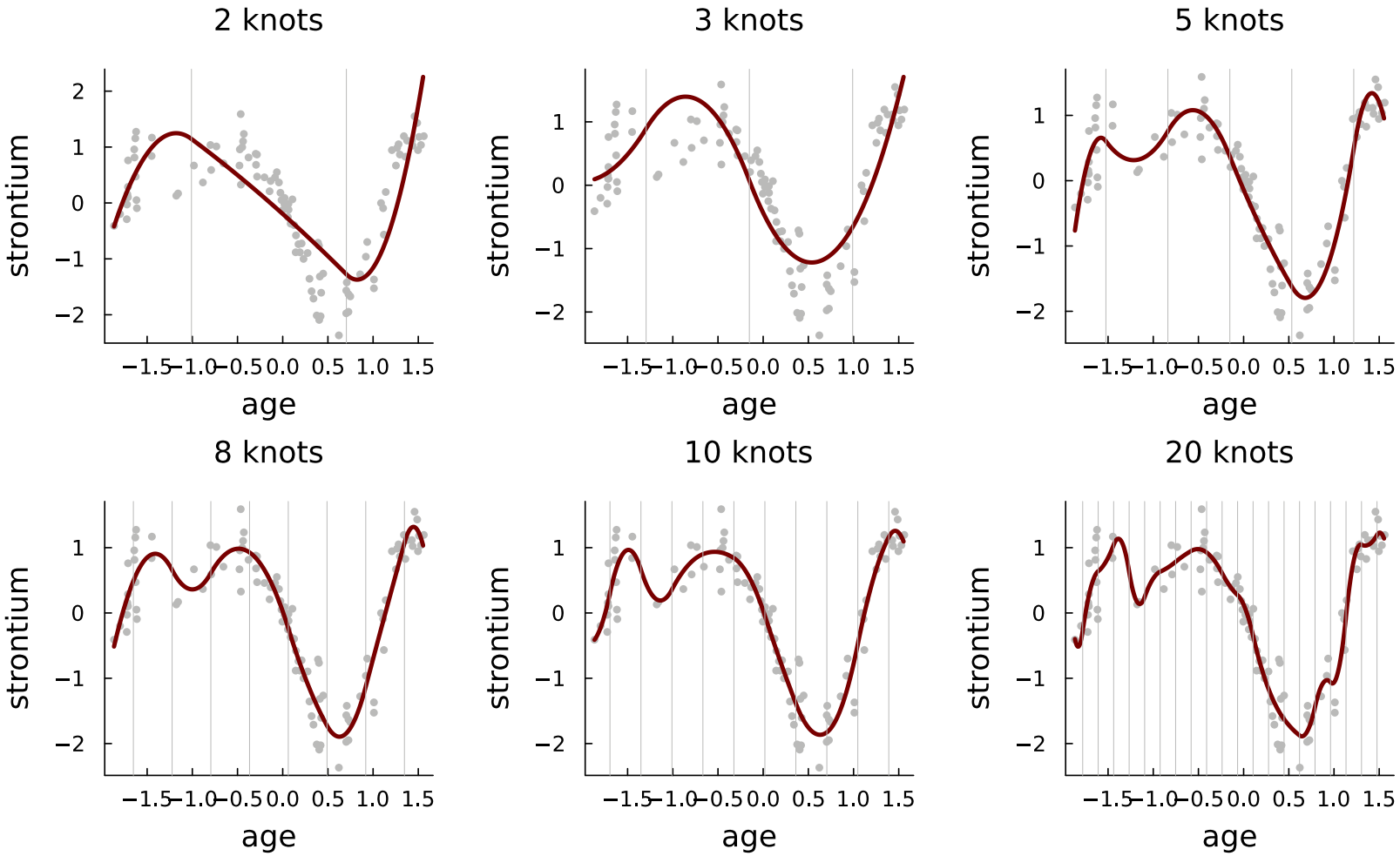


From Ruppert, Wand and Carroll (2003). Semiparametric regression.

Polynomial regression - fossil data



Quadratic spline - fossil data



Regularization prior - Ridge

- Splines: too many knots leads to **over-fitting**.
- **Shrinkage/regularization prior**

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

- Larger λ gives smoother fit. Note: $\boldsymbol{\Omega}_0 = \lambda I$ in conjugate prior.
- **Prior** acts like penalty in **penalized likelihood**:

$$-2 \cdot \log p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}, \boldsymbol{X}) \propto (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}$$

- Posterior mean gives **ridge regression** estimator

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

Ridge shrinks toward zero

- **Ridge regression** estimator

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

- **Shrinkage** toward zero

$$\text{As } \lambda \to \infty,\ \tilde{\boldsymbol{\beta}} \to 0$$

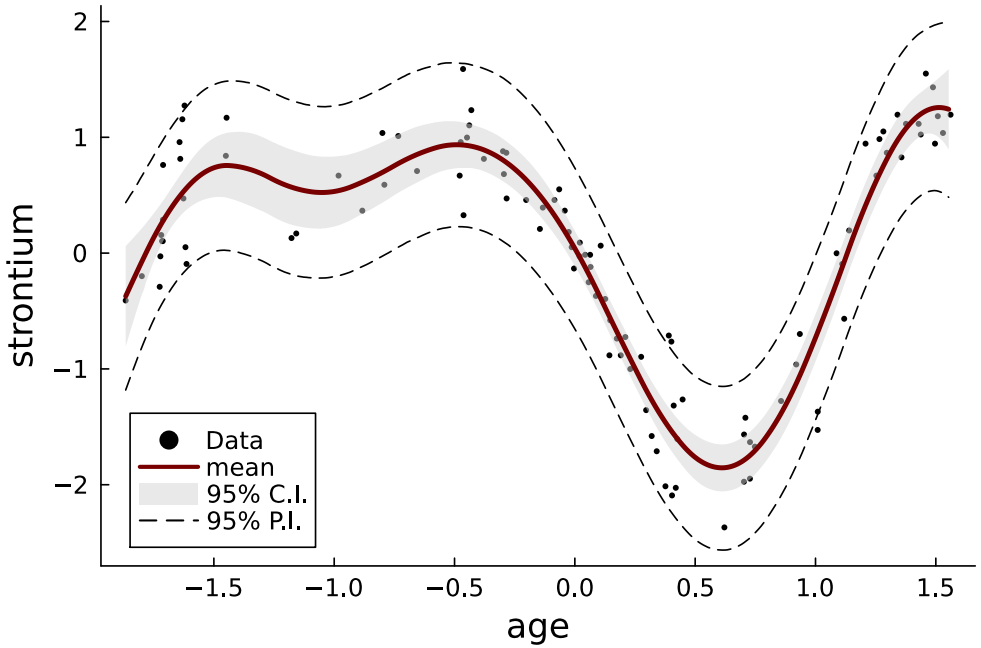
- When $\boldsymbol{X}^\top\boldsymbol{X} = I_p$

$$\tilde{\boldsymbol{\beta}} = \frac{1}{1+\lambda}\hat{\boldsymbol{\beta}} = (1-\phi)\hat{\boldsymbol{\beta}}$$

- **Shrinkage factor**

$$\phi = \frac{\lambda}{1+\lambda}$$

Spline with Gaussian prior - fossil data



Regularization prior - Lasso

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

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

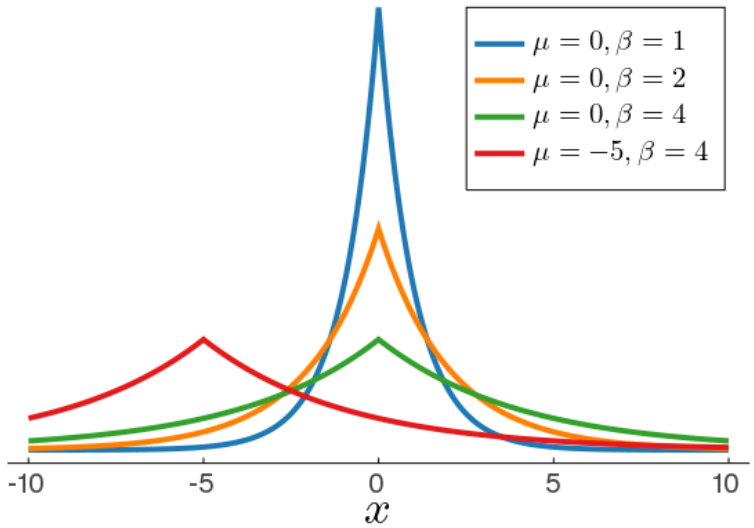
Laplace distribution

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

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

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

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



- The **Bayesian shrinkage** prior is **interpretable**. **Not ad hoc**.
- Laplace distribution have heavy tails.
- **Laplace prior**: many β_i close to zero, but some β_i very large.
- Normal distribution have light tails.

Learning the shrinkage

- **Cross-validation** used to determine degree of smoothness, λ.
- Bayesian: λ is **unknown** $\Rightarrow$ **use a prior** for λ!
- $\lambda^{-1} \sim \text{Inv}{-}\chi^2(\omega_0, \psi_0^2)$.
- **Hierarchical** setup:

$$\begin{aligned} \mathbf{y}|\boldsymbol{\beta}, \sigma^2, \mathbf{X} &\sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I_n) \\ \boldsymbol{\beta}|\sigma^2, \lambda &\sim N\left(0, \sigma^2 \lambda^{-1} I_m\right) \\ \sigma^2 &\sim \text{Inv}{-}\chi^2(\nu_0, \sigma_0^2) \\ \lambda^{-1} &\sim \text{Inv}{-}\chi^2(\omega_0, \psi_0^2) \end{aligned}$$

so $\boldsymbol{\Omega}_0 = \lambda I_m$.

Regression with learned shrinkage

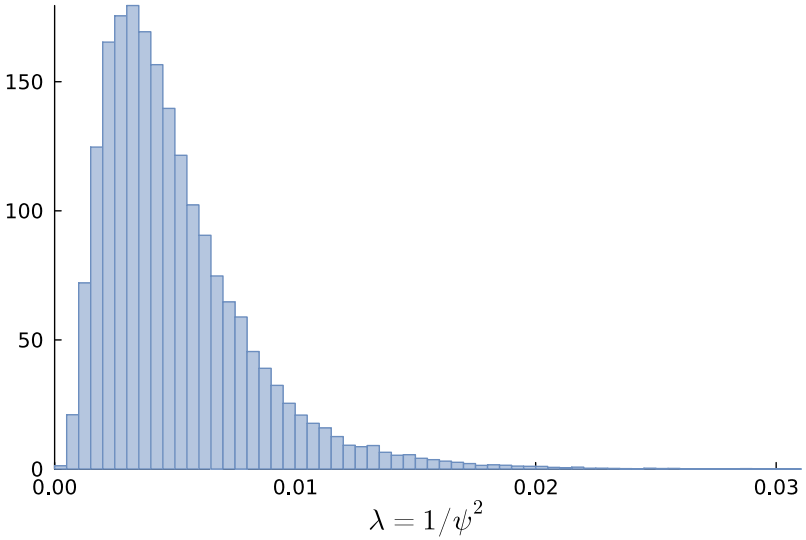
- The **joint posterior** of $\boldsymbol{\beta}$, σ^2 and λ is

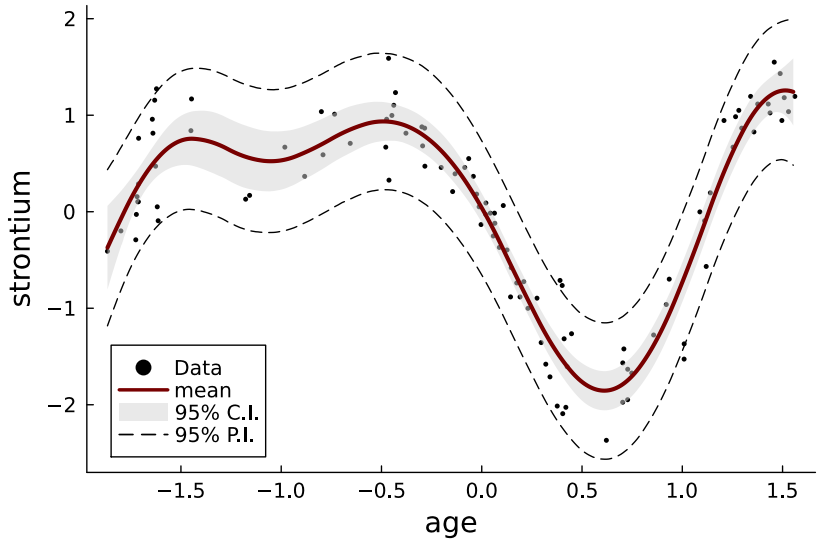
$$\begin{aligned}
\boldsymbol{\beta}|\sigma^2, \lambda, \mathbf{y} &\sim N\left(\mu_n, \Omega_n^{-1}\right) \\
\sigma^2|\lambda, \mathbf{y} &\sim \text{Inv} - \chi^2\left(\nu_n, \sigma_n^2\right) \\
p(\lambda|\mathbf{y}) &\propto \sqrt{\frac{|\Omega_0|}{|\mathbf{X}^\top\mathbf{X} + \Omega_0|}} \left(\frac{\nu_n\sigma_n^2}{2}\right)^{-\nu_n/2} \cdot p(\lambda)
\end{aligned}$$

  where $\Omega_0 = \lambda I_m$, and $p(\lambda)$ is the prior for λ.

- Or simulate from $p(\boldsymbol{\beta}, \sigma^2, \lambda|\boldsymbol{y}, \boldsymbol{X})$ using Gibbs sampling (L7).

Spline with Gaussian prior - fossil data





Horseshoe prior

- Normal and Laplace - only one global shrinkage parameter λ.
- **Global-Local shrinkage**: global + local shrinkage for each β_j.
- **Horseshoe prior**:

$$\beta_j|\lambda_j^2, \tau^2 \sim N\left(0, \tau^2\lambda_j^2\right)$$
$$\lambda_j \sim C^+(0,1)$$
$$\tau \sim C^+(0,1)$$

- When $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{I}_p$, posterior mean for $\boldsymbol{\beta}$ satisfies approximately

$$\mu_{n,j} \approx (1-\phi_j)\hat{\boldsymbol{\beta}}_j, \text{ where } \phi_j = \frac{1}{1+(n/\sigma^2)\tau^2\lambda_j^2}$$

- Implied **prior on shrinkage**