Bayesian Learning

Lecture 6 - Bayesian regularization

Mattias Villani 👨

Department of Statistics
Stockholm University

mattiasvillani.com @matvil @matvil mattiasvillani

Lecture overview

- **Non-linear/semiparametric regression**
- **Regularization priors**

Polynomial regression

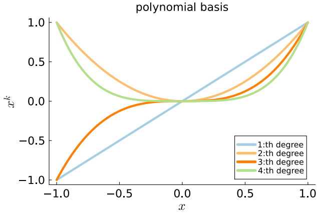
- ## Polynomial regression

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

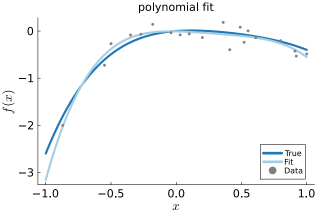
$$\mathbf{y} = \boldsymbol{X}\beta + \varepsilon,$$

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

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

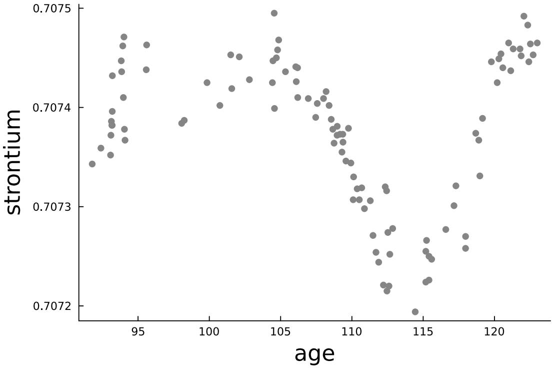


1.00	-1.00	1.00	-1.00	1.00
1.00	-0.90	0.81	-0.73	0.66
1.00	-0.80	0.64	-0.51	0.41
1.00	-0.70	0.49	-0.34	0.24
1.00	-0.60	0.36	-0.22	0.13
1.00	-0.50	0.25	-0.12	0.06
1.00	-0.40	0.16	-0.06	0.03
1.00	-0.30	0.09	-0.03	0.01
1.00	-0.20	0.04	-0.01	0.00
1.00	-0.10	0.01	-0.00	0.00
1.00	0.00	0.00	0.00	0.00
1.00	0.10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



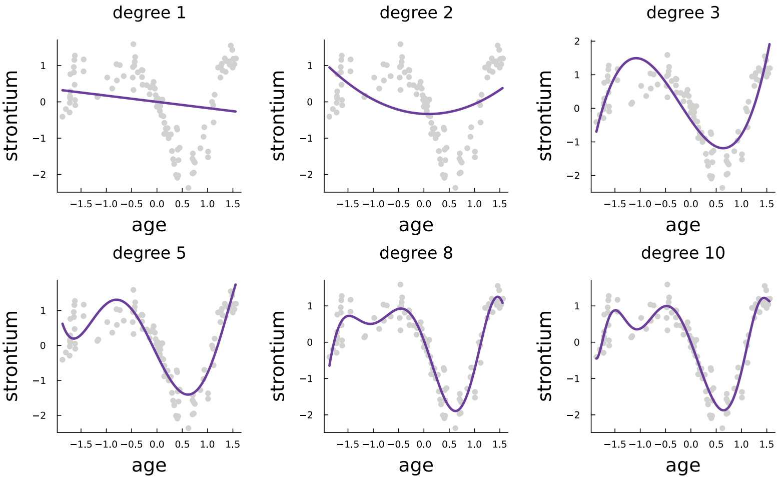
- Polynomials are **global** basis functions. **Local** basis preferred.

Fossil data



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

Polynomial regression - fossil data



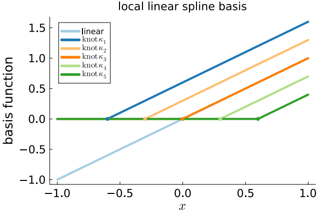
Spline regression - local linear basis

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

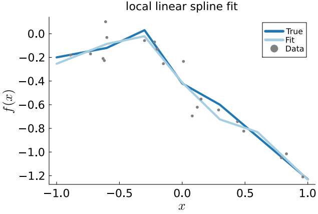
$$b_j(x) = \begin{cases} |x - \kappa_j| & \text{if } x > \kappa_j \\ 0 & \text{otherwise} \end{cases}$$

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

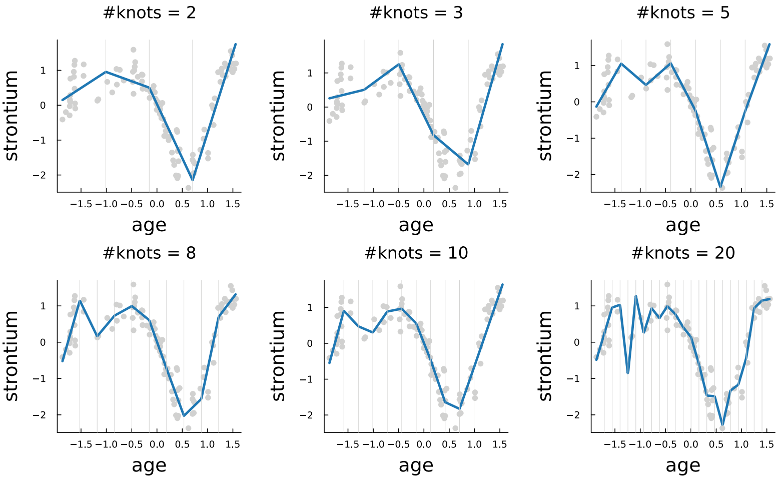
$$\boldsymbol{x}_i = (1, x_i, b_1(x_i), ..., b_m(x_i))^\top$$



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



Linear spline - fossil data



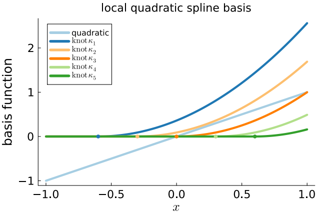
Spline regression - local quadratic basis

- **Truncated quadratic splines** with **knot locations** $\kappa_1, ..., \kappa_m$:

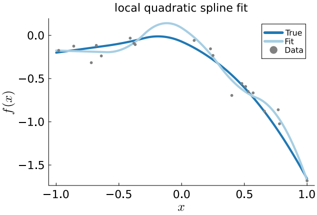
$$b_j(x) = \begin{cases} (x - \kappa_j)^2 & \text{if } x > \kappa_j \\ 0 & \text{otherwise} \end{cases}$$

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

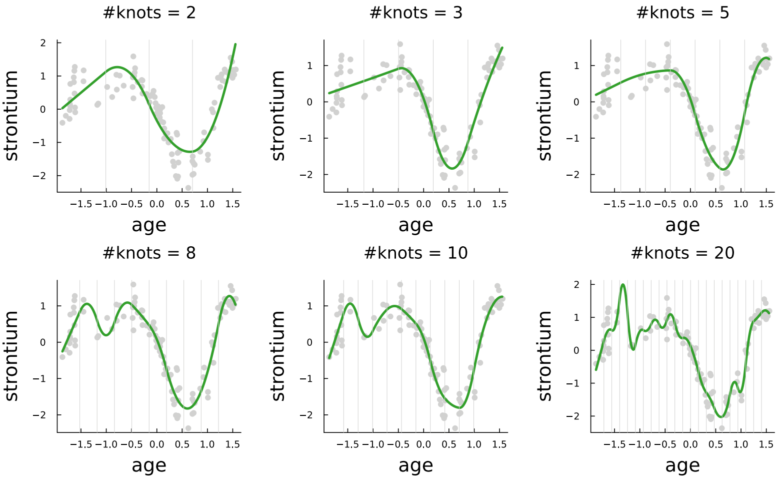
$$\boldsymbol{x}_i = (1, x_i, b_1(x_i), ..., b_m(x_i))^\top$$



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	0.01	0.00	0.00
1.00	-0.30	0.04	0.00	0.00
1.00	-0.20	0.09	0.00	0.00
1.00	-0.10	0.16	0.00	0.00
1.00	0.00	0.25	0.00	0.00
1.00	0.10	0.36	0.01	0.00
1.00	0.20	0.49	0.04	0.00
1.00	0.30	0.64	0.09	0.00
1.00	0.40	0.81	0.16	0.00
1.00	0.50	1.00	0.25	0.00
1.00	0.60	1.21	0.36	0.01
1.00	0.70	1.44	0.49	0.04
1.00	0.80	1.69	0.64	0.09
1.00	0.90	1.96	0.81	0.16
1.00	1.00	2.25	1.00	0.25



Quadratic spline - fossil data



Regularization prior - Ridge

- Splines: too many knots leads to **over-fitting**.
- **Smoothness**/**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

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

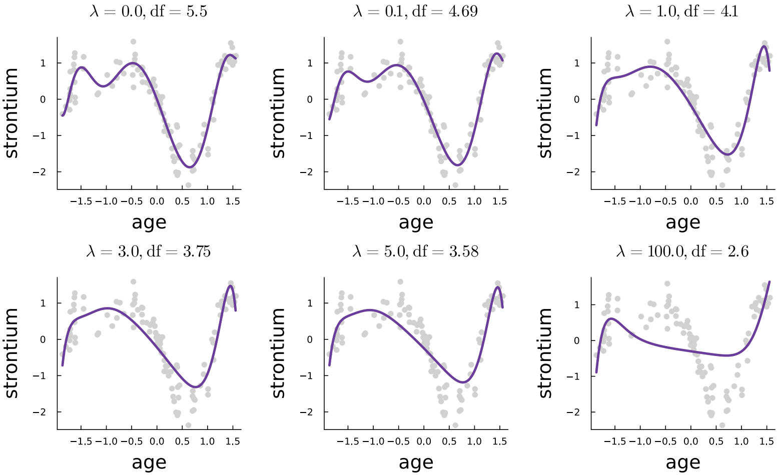
- **Shrinkage** toward zero

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

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

$$\tilde{\boldsymbol{\beta}} = \frac{1}{1+\lambda}\hat{\boldsymbol{\beta}}$$

Polynomial with Gaussian prior - fossil data



Regularization prior - Lasso

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

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

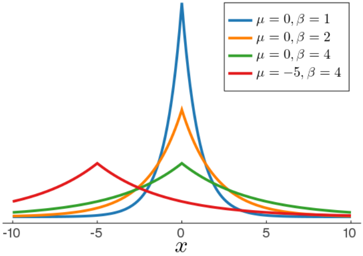
Laplace distribution

$X \sim \text{Laplace}(\mu, \beta)$ 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

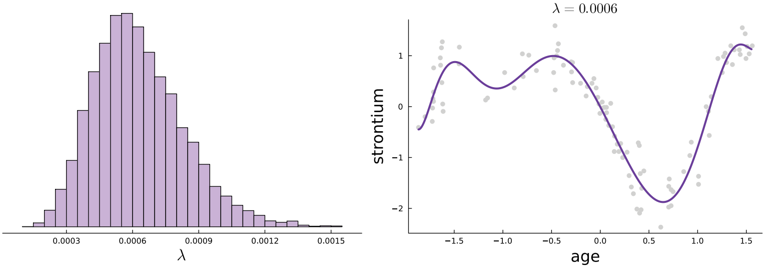
$$\boldsymbol{\beta}|\sigma^2,\lambda,\mathbf{y} \sim N\left(\mu_n,\Omega_n^{-1}\right)$$
$$\sigma^2|\lambda,\mathbf{y} \sim \mathrm{Inv}-\chi^2\left(\nu_n,\sigma_n^2\right)$$
$$p(\lambda|\mathbf{y}) \propto \sqrt{\frac{|\Omega_0|}{|\mathbf{X}^\top\mathbf{X}+\Omega_0|}}\left(\frac{\nu_n\sigma_n^2}{2}\right)^{-\nu_n/2}\cdot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

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

$$\mu_n = \left(\mathbf{X}^\top\mathbf{X}+\Omega_0\right)^{-1}\mathbf{X}^T\mathbf{y}$$
$$\Omega_n = \mathbf{X}^\top\mathbf{X}+\Omega_0$$
$$\nu_n = \nu_0 + n$$
$$\nu_n\sigma_n^2 = \nu_0\sigma_0^2 + \mathbf{y}^T\mathbf{y} - \mu_n^T\Omega_n\mu_n$$

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

Polynomial of 10th degree with regularization prior



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**