

# Bayesian Statistics I

## Lecture 6 - Bayesian regularization

**Mattias Villani**

Department of Statistics  
Stockholm University

Department of Computer and Information Science  
Linköping University



[mattiasvillani.com](http://mattiasvillani.com)



@matvil



mattiasvillani

# Lecture overview

- Non-linear regression
- Regularization priors

# Polynomial regression

## Polynomial regression

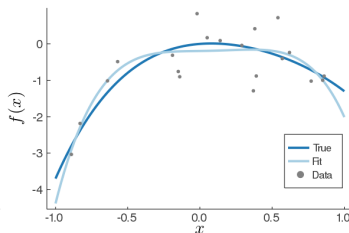
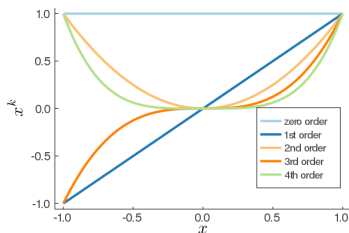
$$f(x_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \dots + \beta_k x_i^k, \quad \text{for } i = 1, \dots, n.$$

$$y = X\beta + \varepsilon,$$

where  $i$ th row of  $X$  is

$$(1, x_i, x_i^2, \dots, x_i^k).$$

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



# Spline regression

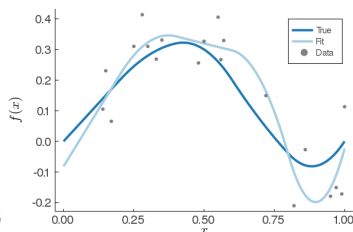
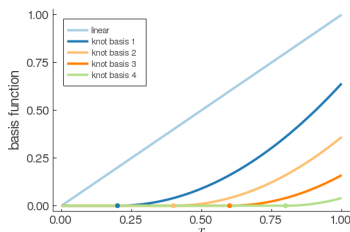
- Polynomials are too global. Need more local basis functions.
- Truncated quadratic splines** with **knot locations**  $\kappa_1, \dots, \kappa_m$ :

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

$$y = X\beta + \varepsilon,$$

where  $i$ th row of  $X$  is

$$(1, x_i, b_1(x_i), \dots, b_m(x_i)).$$



# Regularization prior - Ridge

- Too many knots leads to **over-fitting**.
- **Smoothness/shrinkage/regularization prior**

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

- Larger  $\lambda$  gives smoother fit. Note:  $\mathbf{\Omega}_0 = \lambda I$  in conjugate prior.
- Equivalent to **penalized likelihood**:

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

- Posterior mean gives **ridge regression** estimator

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

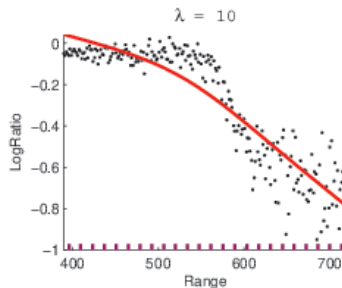
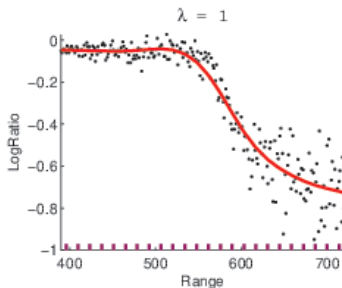
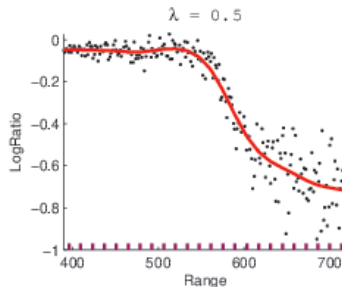
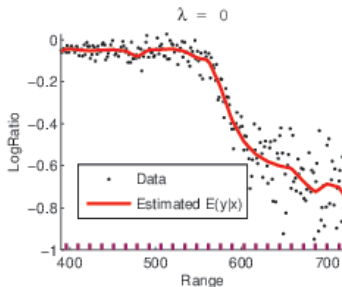
- **Shrinkage** toward zero

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

- When  $X^T X = I$

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

# Bayesian spline with regularization prior



# Regularization prior - Lasso

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

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

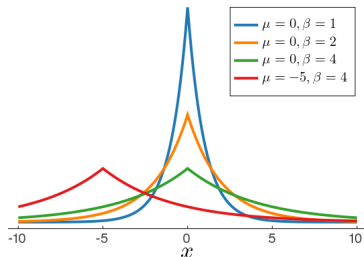
## Laplace distribution

$X \sim \text{Laplace}(\mu, \beta)$  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  $\beta_i$  close to zero, but some  $\beta_i$  very large.
- Normal distribution have light tails.

# Learning the shrinkage

- **Cross-validation** used to determine degree of smoothness,  $\lambda$ .
- Bayesian:  $\lambda$  is **unknown**  $\Rightarrow$  **use a prior** for  $\lambda$ !
- $\lambda \sim \text{Inv-}\chi^2(\eta_0, \lambda_0)$ .
- **Hierarchical** setup:

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

$$\beta|\sigma^2, \lambda \sim N(0, \sigma^2 \lambda^{-1} I_m)$$

$$\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2)$$

$$\lambda \sim \text{Inv-}\chi^2(\eta_0, \lambda_0)$$

$$\text{so } \Omega_0 = \lambda I_m.$$



# Regression with learned shrinkage

- The **joint posterior** of  $\beta$ ,  $\sigma^2$  and  $\lambda$  is

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

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

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

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

$$\mu_n = (X^T X + \Omega_0)^{-1} X^T y$$

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

$$\nu_n = \nu_0 + n$$

$$\nu_n \sigma_n^2 = \nu_0 \sigma_0^2 + y^T y - \mu_n^T \Omega_n \mu_n$$

# More complexity

- The **location of the knots** can be unknown. Joint posterior:

$$p(\beta, \sigma^2, \lambda, \kappa_1, \dots, \kappa_m | y, X)$$

- The marginal posterior for  $\kappa_1, \dots, \kappa_m$  is a nightmare.
- Simulate from joint posterior by MCMC. Li and Villani (2013).
- The basic spline model can be extended with:
  - ▶ **Heteroscedastic errors** (also modelled with a spline)
  - ▶ **Non-normal errors** (student-t or mixture distributions)
  - ▶ **Autocorrelated/dependent errors** (AR process for the errors)