

Bayesian Statistics I

Lecture 5 - Regression and Regularization

Mattias Villani

Department of Statistics
Stockholm University

Department of Computer and Information Science
Linköping University



Lecture overview

- Normal model with conjugate prior
- The linear regression model
- Non-linear regression
- Regularization priors

Linear regression

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

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

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

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

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

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

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 :

$$\begin{aligned}\beta | \sigma^2, \mathbf{y} &\sim N[\hat{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}] \\ \sigma^2 | \mathbf{y} &\sim \text{Inv-}\chi^2(n-k, s^2)\end{aligned}$$

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

- **Simulate** from the joint posterior by simulating from

- ▶ $p(\sigma^2 | \mathbf{y})$
- ▶ $p(\beta | \sigma^2, \mathbf{y})$

- **Marginal posterior** of β :

$$\beta | \mathbf{y} \sim t_{n-k}[\hat{\beta}, s^2(\mathbf{X}'\mathbf{X})^{-1}]$$

Linear regression - conjugate prior

■ Joint prior for β and σ^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

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

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

Polynomial regression

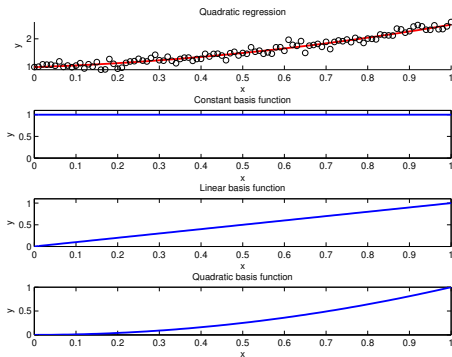
Polynomial regression

$$f(x_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \dots + \beta_k x_i^k.$$

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

where

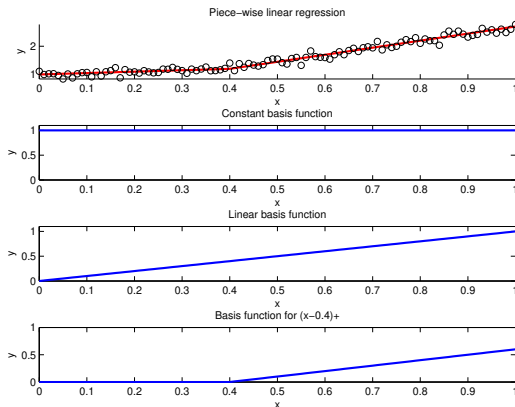
$$\mathbf{X}_P = (1, x, x^2, \dots, x^k).$$



Spline regression

- Polynomials are too global. Need more local basis functions.
- Truncated power splines** given **knot locations** k_1, \dots, k_m

$$b_{ij} = \begin{cases} (x_i - k_j)^p & \text{if } x_i > k_j \\ 0 & \text{otherwise} \end{cases}$$



- Spline regression is linear in the m 'dummy variables' b_j

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

where \mathbf{X}_b is the basis matrix

$$\mathbf{X}_b = (b_1, \dots, b_m).$$

- Adding intercept and linear term

$$\mathbf{X}_b = (1, x, b_1, \dots, b_m).$$

Smoothness prior for splines

- Problem: too many knots leads to **over-fitting**.
- **Smoothness/shrinkage/regularization prior**

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

- Larger λ gives smoother fit. Note: $\Omega_0 = \lambda I$.
- Equivalent to **penalized likelihood**:

$$-2 \cdot \log p(\beta | \sigma^2, \mathbf{y}, \mathbf{X}) \propto \text{RSS}(\beta) + \lambda \beta' \beta$$

- Posterior mean gives **ridge regression** estimator

$$\tilde{\beta} = (\mathbf{X}'\mathbf{X} + \lambda I)^{-1} \mathbf{X}'\mathbf{y}$$

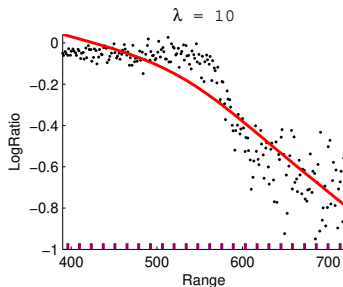
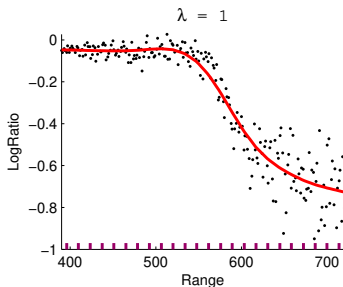
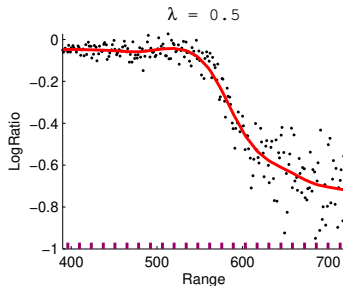
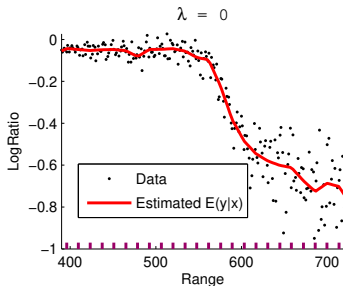
- **Shrinkage** toward zero

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

- When $\mathbf{X}'\mathbf{X} = I$

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

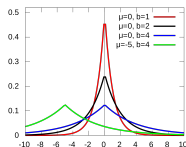
Bayesian spline with smoothness prior



Smoothness prior for splines

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

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



- 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.
- **Normal prior**: all β_i 's are similar in magnitude.

Estimating the shrinkage

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

$$\begin{aligned}\mathbf{y}|\beta, \mathbf{X} &\sim N(\mathbf{X}\beta, \sigma^2 I_n) \\ \beta|\sigma^2, \lambda &\sim N(0, \sigma^2 \lambda^{-1} I_m) \\ \sigma^2 &\sim \text{Inv} - \chi^2(\nu_0, \sigma_0^2) \\ \lambda &\sim \text{Inv} - \chi^2(\eta_0, \lambda_0)\end{aligned}$$

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

Regression with estimated shrinkage

- The **joint posterior** of β , σ^2 and λ is

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

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

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

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

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

$$\Omega_n = \mathbf{X}^T\mathbf{X} + \Omega_0$$

$$\nu_n = \nu_0 + n$$

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

More complexity

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

$$p(\beta, \sigma^2, \lambda, k_1, \dots, k_m | \mathbf{y}, \mathbf{X})$$

- The marginal posterior for λ, k_1, \dots, k_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)