Bayesian Statistics - Lecture 5

Lecture 5: Regression. Regularization priors.

Mattias Villani

Department of Statistics
Stockholm University
and
Department of Computer and Information Science
Linköping University

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

- **Model**

$$y_1, \ldots, y_n | \theta, \sigma^2 \overset{iid}{\sim} N(\theta, \sigma^2)$$

- **Conjugate prior**

$$\theta | \sigma^2 \sim N\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right)$$
$$\sigma^2 \sim Inv\text{-}\chi^2(\nu_0, \sigma_0^2)$$

- **Posterior**

$$\theta|y,\sigma^2 \sim N\left(\mu_n, \frac{\sigma^2}{\kappa_n}\right)$$
$$\sigma^2|y \sim Inv\text{-}\chi^2(\nu_n, \sigma_n^2).$$

where

$$\begin{aligned}
\mu_n &= \frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\bar{y} \\
\kappa_n &= \kappa_0 + n \\
\nu_n &= \nu_0 + n \\
\nu_n\sigma_n^2 &= \nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0+n}(\bar{y}-\mu_0)^2.
\end{aligned}$$

- **Marginal posterior**

$$\theta \sim t_{\nu_n}\left(\mu_n, \sigma_n^2/\kappa_n\right)$$

The Linear Regression Model

- The ordinary **linear regression** model:

$$\begin{aligned} y_i &= \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i \\ & \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2). \end{aligned}$$

- Parameters $\theta = (\beta_1, \beta_2, \ldots, \beta_k, \sigma^2)$.

- **Assumptions**:
    - $E(y_i) = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$ (linear function)
    - $Var(y_i) = \sigma^2$ (homoscedasticity)
    - $Corr(y_i, y_j | X, \beta, \sigma^2) = 0$, $i \neq j$.
    - Normality of ε_i.
    - The x's are assumed known (non-random).

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

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

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

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

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

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

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

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

- **Joint posterior** of β and σ^2:

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

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 :

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

- **Joint prior** for β and σ^2

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

- **Posterior**

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

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

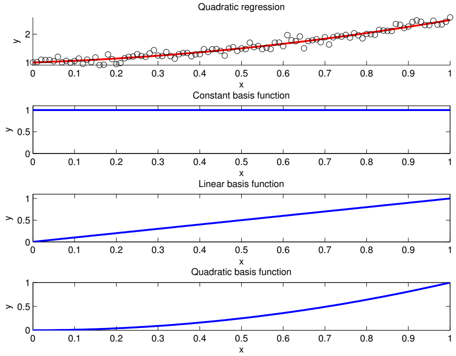
Polynomial regression

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

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

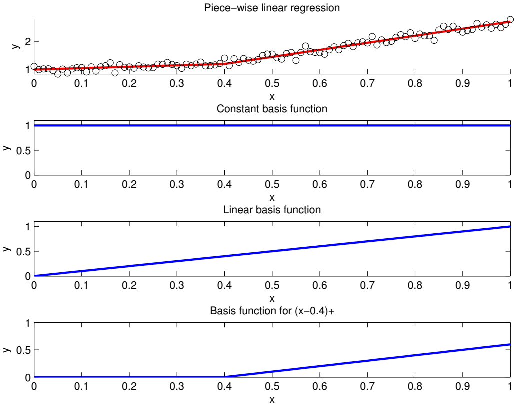
where

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



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

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



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

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

where X_b is the **basis matrix**

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

- Adding intercept and linear term

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

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

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

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

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

- Posterior mean gives **ridge regression** estimator

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

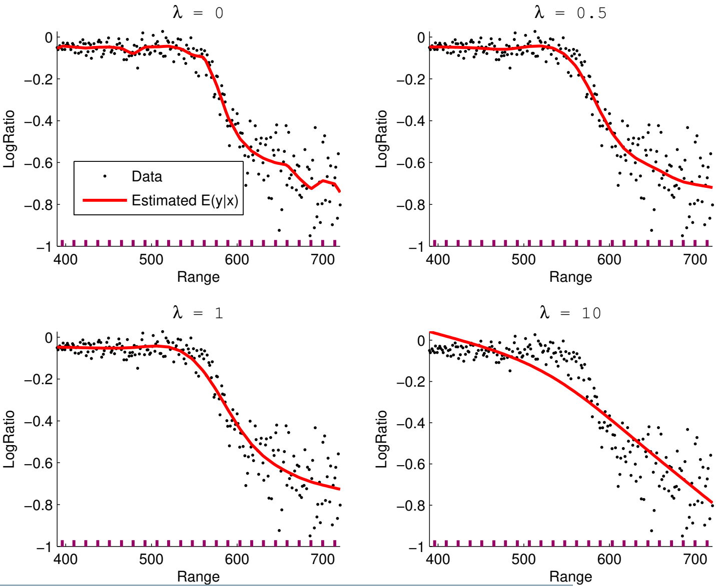
- **Shrinkage** toward zero

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

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

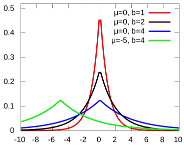
$$\tilde{\beta} = \frac{1}{1+\lambda}\hat{\beta}_{OLS}$$

Bayesian spline with smoothness prior



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

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



- 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.
- $\lambda \sim \text{Gamma}\left(\frac{\eta_0}{2}, \frac{\eta_0}{2\lambda_0}\right)$. The user specifies η_0 and λ_0.
- Hierarchical setup:

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

so $\Omega_0 = \lambda I_m$.

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

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

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

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

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

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

- The marginal posterior for $\lambda, k_1, ..., k_m$ 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)