Bayesian Learning

Lecture 6 - Bayesian regularization

Mattias Villani

Department of Statistics
Stockholm University

Department of Computer and Information Science
Linköping University

mattiasvillani.com   @matvil   mattiasvillani

Lecture overview

- Non-linear regression
- Regularization priors

Polynomial regression

- **Polynomial regression**

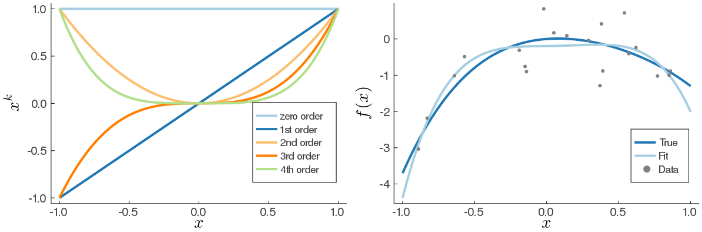
$$f(x_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_k x_i^k, \quad \text{for } i = 1, \ldots, n.$$

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

where ith row of X is

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

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



Spline regression

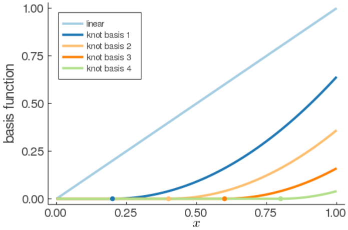
- Polynomials are too global. Need more local basis functions.
- **Truncated quadratic splines** with **knot locations** $\kappa_1, ..., \kappa_m$:

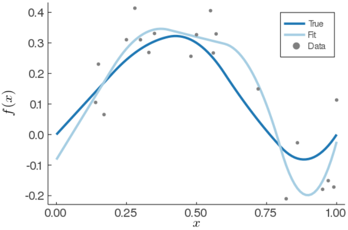
$$b_j(x) = \begin{cases} (x - \kappa_j)^2 & \text{if } x > \kappa_j \\ 0 & \text{otherwise} \end{cases}$$

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

where ith row of X is

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





Regularization prior - Ridge

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

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

- Posterior mean gives **ridge regression** estimator

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

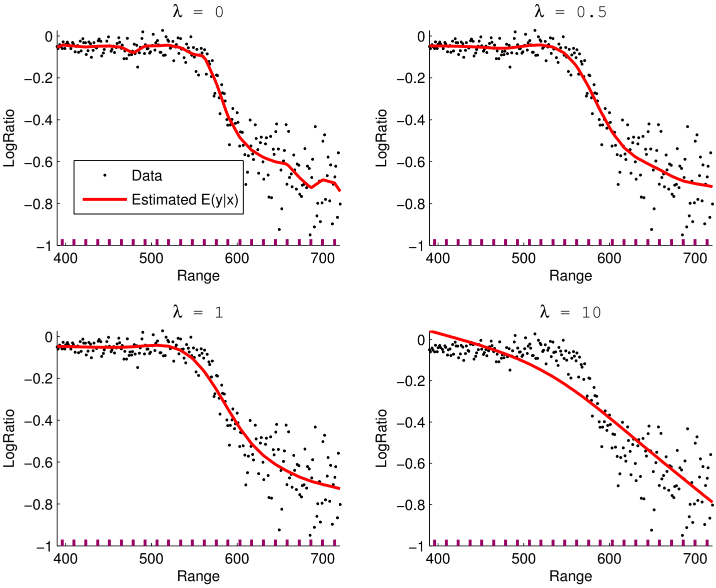
- **Shrinkage** toward zero

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

- When $\mathrm{X}^T \mathrm{X} = I$

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

Bayesian spline with regularization prior



Regularization prior - Lasso

- Lasso is equivalent to posterior mode under Laplace prior

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

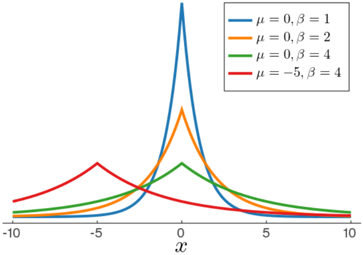
Laplace distribution

$X \sim \text{Laplace}(\mu, \beta)$ 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!
- $\lambda \sim \mathrm{Inv}-\chi^2(\eta_0, \lambda_0)$.
- Hierarchical setup:

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

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

Regression with learned shrinkage

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

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

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

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

More complexity

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

$$p(\boldsymbol{\beta}, \sigma^2, \lambda, \kappa_1, ..., \kappa_m | \mathrm{y}, \mathrm{X})$$

- The marginal posterior for $\kappa_1, ..., \kappa_m$ 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)