

Bayesian Learning

Lecture 5 - Regression and Regularization

Bertil Wegmann

Department of Computer and Information Science
Linköping University



Lecture overview

- The **linear regression** model
- **Non-linear regression**
- **Regularization priors**

Linear regression

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

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

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$
$$X = \begin{pmatrix} x'_1 \\ \vdots \\ x'_n \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix}$$

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

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

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 :

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

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

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

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

- **Marginal posterior** of β :

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

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

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

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

Polynomial regression

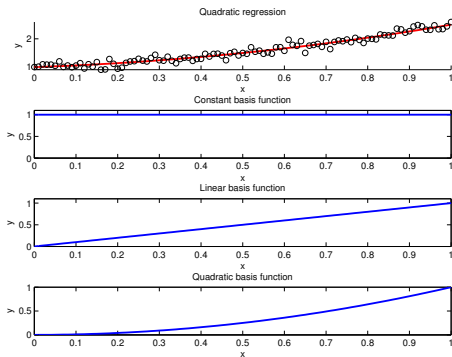
■ Polynomial regression

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

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

where

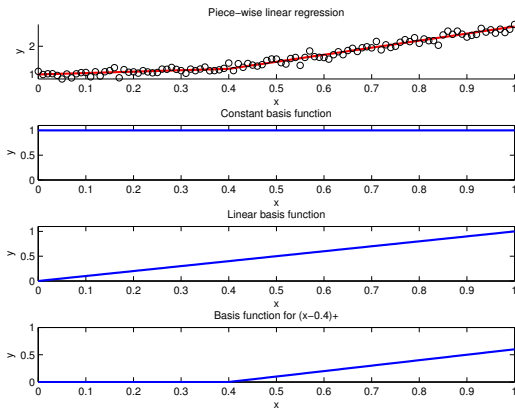
$$X_P = (1, x, x^2, \dots, x^k).$$



Spline regression

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

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



Splines

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

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

where X_b is the **basis matrix**

$$X_b = (b_1, \dots, b_m).$$

- Adding intercept and linear term

$$X_b = (1, x, b_1, \dots, b_m).$$

Smoothness prior for splines

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

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

- Larger λ gives smoother fit. More **shrinkage**. Note: $\mu_0 = 0, \Omega_0 = \lambda I$.
- Equivalent to **penalized likelihood**:

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

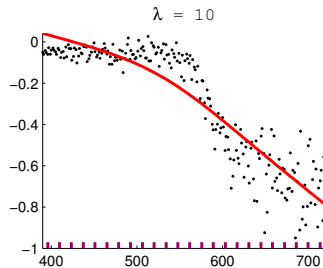
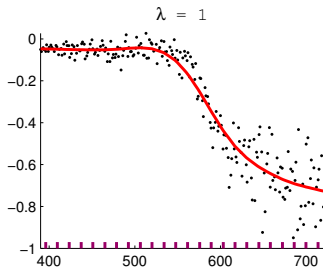
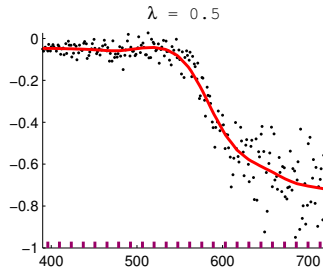
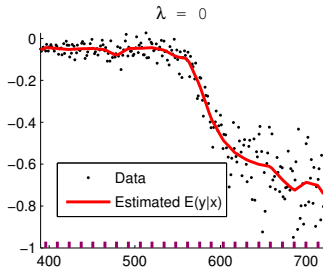
- Posterior mean/mode gives **ridge regression** estimator

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

- When $X'X = I$ (orthogonal, “uncorrelated” features)

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

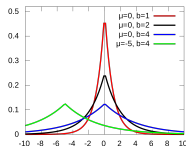
Bayesian spline with smoothness prior



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 prior**:
 - ▶ tails in distribution die off slowly
 - ▶ many β_i close to zero, but some β_i very large.
- **Normal prior**:
 - ▶ tails in distribution die off rapidly
 - ▶ all β_i 's are similar in magnitude.

Estimating the shrinkage

- Cross-validation: determine λ by performance on test data.
- Bayesian: λ is **unknown** \Rightarrow **use a prior** for λ .
- $\lambda \sim \text{Inv} - \chi^2(\eta_0, \lambda_0)$.
- Hierarchical setup:

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

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

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

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

$$\text{so } \mu_0 = 0, \Omega_0 = \lambda I_m.$$

Regression with estimated shrinkage

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

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

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

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

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

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

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

$$v_n = v_0 + n$$

$$v_n \sigma_n^2 = v_0 \sigma_0^2 + \mathbf{y}^T \mathbf{y} - \mu_n^T \Omega_n \mu_n$$

More complexity

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

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

- 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)