Bayesian Learning - Lecture 6

Mattias Villani

Division of Statistics
Department of Computer and Information Science
Linköping University

Lecture overview

- Flexible nonlinear regression and splines
- Smoothness/shrinkage priors
- Bayesian variable selection

Non-parametric/Non-linear regression

- Recall the linear regression model with a single covariate

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

- Extension to non-linearity:

$$y_i = f(x_i) + \varepsilon_i, \qquad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2),$$

  where $f(\cdot)$ is a non-linear function.
- Polynomial regression:

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

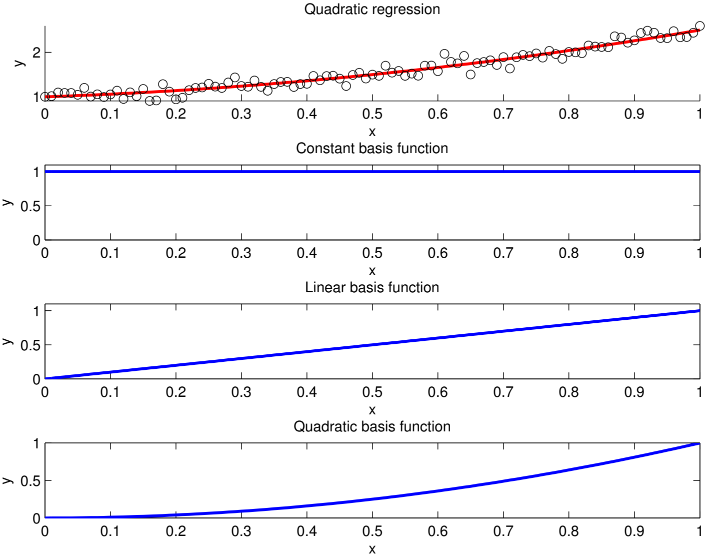
  This can be written as a linear regression

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

  where

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

Polynomial basis functions



Smooth interpolation

- Another approach treats all n ordinates as unknown parameters:

$$f(x_i) = \gamma_i.$$

- Problem: too many parameters. Estimated curve wiggles way too much.
- Solution: use a (multivariate) prior on $\gamma = (\gamma_1, ..., \gamma_n)'$ that carries the info that the regression curve is smooth:

$$\text{if } x_i \text{ and } x_k \text{ are close then } \gamma_i \text{ is close to } \gamma_k$$

- Order the data with respect to covariates and assign the prior

$$p(\gamma_i | \gamma_{i-1}) \sim N\left(\gamma_{i-1}, \tau_0^2 \cdot |x_i - x_{i-1}|\right), \text{ for } i = 2, ..., n.$$

- The hyperparameter τ_0^2 controls the degree of prior smoothness.

Splines

- Warm-up: change-point analysis using piecewise constant dummies.
- Use m change-points (knots) $k_1 < k_2 < ... < k_m$. Construct a 'dummy variable' for each change-point:

$$b_{ij} = \begin{cases} 1 & \text{if } x_i > k_j \\ 0 & \text{otherwise} \end{cases}$$

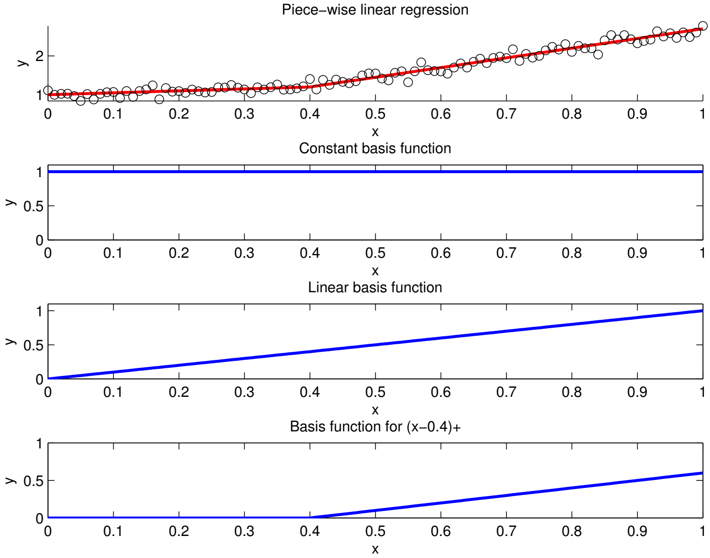
  Not smooth, the regression line has sudden jumps.
- Smoother: tunctated linear splines

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

- Generalization: *truncated power splines*

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

Truncated polynomial basis functions



Splines, cont.

- Note: given the knots, the non-parametric spline regression model is a linear regression of y on the m 'dummy variables' b_j

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

where X_b is the basis regression matrix

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

- It is also common to include an intercept and the linear part of the model separately. In this case we have

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

Smoothness prior for splines

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

$$\beta_i \overset{iid}{\sim} N(0, \lambda^{-1})$$

- Larger λ gives smoother fit.
- Equivalent to a penalized likelihood:

$$-2 \cdot LogPost \propto RSS(\beta) + \lambda \beta' \beta$$

- Posterior mean gives **ridge regression** estimator

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

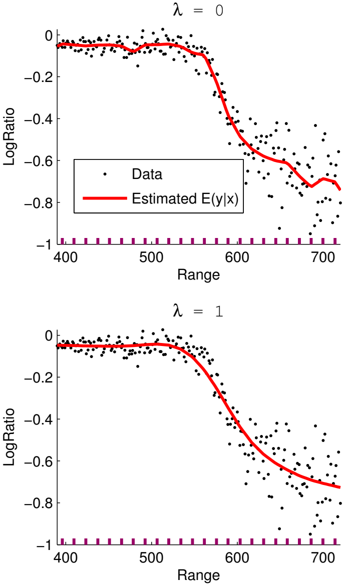
- **Shrinkage** toward zero

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

- When $X'X = I$

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

Bayesian spline with smoothness prior



Smoothness prior for splines, cont.

- The famous **Lasso** variable selection method is equivalent to using the posterior mode estimate under the prior:

$$\beta_i \overset{iid}{\sim} \text{Laplace}(0, \lambda^{-1})$$

where the Laplace density is

$$p(\beta_i) = \frac{1}{2b} \exp\left(-\frac{|\beta_i - \mu|}{b}\right)$$

- The Bayesian shrinkage prior is **interpretable**, and the regularization is **not ad hoc**.
- Laplace distribution have heavy tails.
- Laplace prior: we believe in many β_i close to zero, but some β_i may be very large.
- Normal distribution have light tails.
- Normal prior: most β_i are fairly equal in size, and no single β_i can be very much larger than the other ones.

ESTIMATING THE SHRINKAGE

- How do we determine the degree of smoothness, λ? Cross-validation is one possible approach.
- Bayesian: I cannot specify $\lambda \Rightarrow \lambda$ is unknown $\Rightarrow$ use a prior for λ.
- One possibility: $\lambda \sim Inv-\chi^2(\eta_0, \lambda_0)$. The user specifies η_0 and λ_0.
- Alternative approach: specify the prior on the *degrees of freedom*.
- Hierarchical setup:

$$
\begin{aligned}
y|\beta, x &\sim N(x'\beta, \sigma^2) \\
\beta|\sigma^2 &\sim N(0, \sigma^2 D^{-1}) \\
\sigma^2 &\sim Inv-\chi^2(\nu_0, \sigma_0^2) \\
\lambda &\sim Inv-\chi^2(\eta_0, \lambda_0)
\end{aligned}
$$

where

$$
D = \begin{pmatrix} \delta_0 I_q & 0 \\ 0 & \lambda I_m \end{pmatrix}
$$

Note: different shrinkage on the original q covariates (δ_0) and the covariates that comes from the knots (λ).

Estimating the shrinkage, cont.

- Joint posterior

$$p(\beta, \sigma^2, \lambda | y, x) = p(\beta, \sigma^2 | \lambda, y, x) p(\lambda | y, x)$$

  where

$$p(\lambda | y, x) = \int \int p(\beta, \sigma, \lambda | y, x) d\beta d\sigma^2$$

  is the marginal posterior of λ.
- The conditional posterior $p(\beta, \sigma^2 | \lambda, y, x)$ is a special case of our previous results for linear regression with a conjugate prior. Here $\mu_0 = 0$ and $\Omega_0 = \lambda I$.

Estimating the shrinkage, cont.

- The conditional posterior of β and σ^2 is therefore

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

where

$$\mu_n = \left(X'X + D\right)^{-1} X'y$$
$$\Omega_n = X'X + D$$
$$\nu_n = \nu_0 + n$$
$$\nu_n\sigma_n^2 = \nu_0\sigma_0^2 + \left(y'y - \mu_n'\Omega_n\mu_n\right)$$

- The marginal posterior of λ can be shown to be

$$p(\lambda|y, x) \propto \sqrt{\frac{|D|}{|X'X + D|}} \frac{1}{\left(\frac{\nu_n\sigma_n^2}{2}\right)^{\nu_n/2}} \cdot p(\lambda),$$

where $p(\lambda)$ is the prior for λ.

Summary of the posterior with normal shrinkage prior

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

$$\beta|\sigma^2,\lambda,y \sim N\left(\mu_n,\Omega_n^{-1}\right)$$
$$\sigma^2|\lambda,y \sim Inv-\chi^2\left(\nu_n,\sigma_n^2\right)$$
$$p(\lambda|y) \propto \sqrt{\frac{|D|}{|X'X+D|}}\left(\frac{\nu_n\sigma_n^2}{2}\right)^{-\nu_n/2}\cdot p(\lambda)$$

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

$$\mu_n = \left(X'X+D\right)^{-1}X'y$$
$$\Omega_n = X'X+D$$
$$\nu_n = \nu_0+n$$
$$\nu_n\sigma_n^2 = \nu_0\sigma_0^2+y'y-\mu_n'\Omega_n\mu_n$$

Regularization through Bayesian variable selection

- Selecting the knots in a spline regression is exactly like variable/covariate selection in linear regression.
- Bayesian variable selection is ideal here.
- Introduce variable selection indicators, I_j such that

$$\begin{aligned} \beta_j &= 0 && \text{if } I_j = 0 \\ \beta_j &\sim N(0, \sigma^2 \lambda^{-1}) && \text{if } I_j = 1 \end{aligned}$$

- Need a prior on $I_1, ..., I_K$. Simple choice: $I_1, ..., I_K | \theta \overset{iid}{\sim} Bernoulli(\theta)$.
- Simulate from the posterior distribution:

$$p(\beta, \sigma^2, I_1, ...I_K | \mathbf{y}) = p(\beta, \sigma^2 | I_1, ..., I_K, \mathbf{y}) p(I_1, ..., I_K | \mathbf{y}).$$

- Simulate from $p(I_1, ..., I_K | \mathbf{y})$ using Gibbs sampling [More later].
- Automatic model averaging, all in one simulation run.

Taking it all the way - estimating knot locations

- The location of the knots can be treated as unknown, and estimated from the data. This gives a joint posterior

$$p(\beta, \sigma^2, \lambda, \xi_1, ..., \xi_q | y, x)$$

where ξ_i is the location of the ith knot.
- Posterior is complex but can be sampled from by Markov Chain Monte Carlo (MCMC).