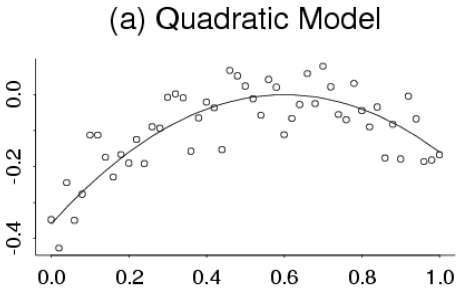
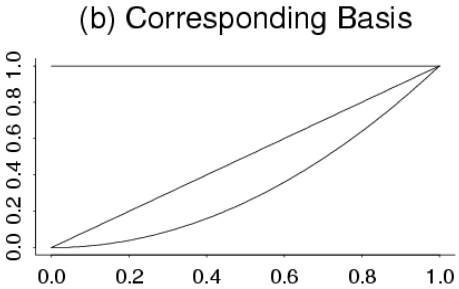
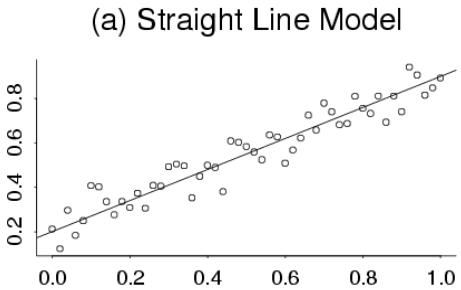
37457

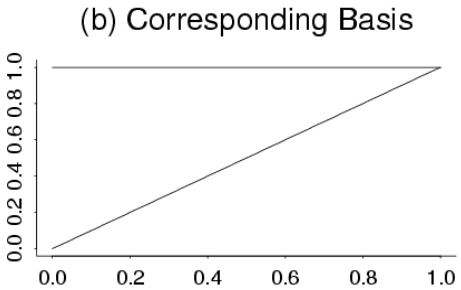
Advanced Bayesian Methods

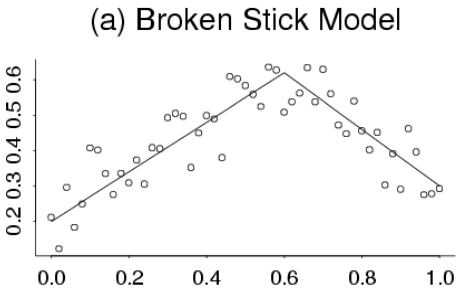
Penalized Splines

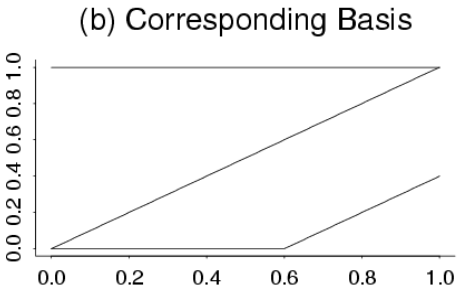


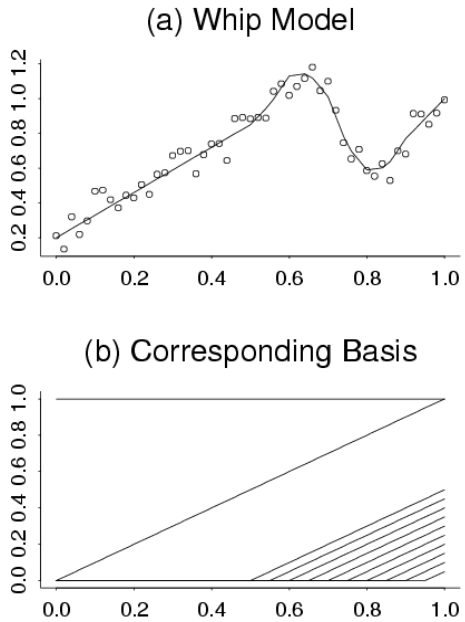












The truncated line spline basis functions for these knots are shown in the next graphic.

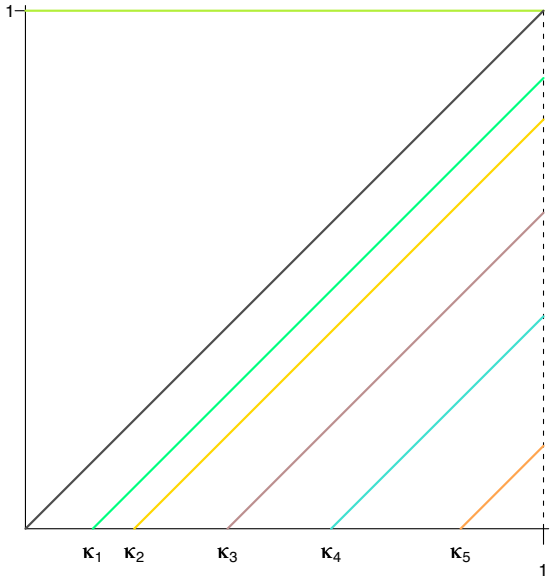
Consider the nonparametric regression setting

$$y_i = f(x_i) + \varepsilon_i$$

where the $x_i \in [0, 1]$.

To keep discussion simple, we work with 5 knots

$$\kappa_1, \kappa_2, \kappa_3, \kappa_4, \kappa_5 \quad \text{in } [0, 1].$$



Note: All spline functions in this lecture are defined over $[0, 1]$.

They equal zero in regions apart from where the functions are shown as non-zero.

The model is:

$$y_i = \beta_0 + \beta_1 x_i + \sum_{k=1}^{5} \beta_{1+k}(x_i - \kappa_k)_+ + \varepsilon_i$$

where

$$(x - \kappa)_+ = \begin{cases} x - \kappa, & x > \kappa \\ 0, & x \leq \kappa. \end{cases}$$

More generally, $x_+ = x$ if $x > 0$
while $x_+ = 0$ if $x \leq 0$.

e.g. $7_+ = 7, \qquad (-3)_+ = 0.$

Ordinary least squares minimises

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n} \left\{ y_i - \beta_0 - \beta_1 x_i - \sum_{k=1}^{5} \beta_{1+k}(x_i - \kappa_k)_+ \right\}^2 \\ &= \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 \end{aligned}$$

where

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 & (x_1 - \kappa_1)_+ & (x_1 - \kappa_2)_+ & (x_1 - \kappa_3)_+ & (x_1 - \kappa_4)_+ & (x_1 - \kappa_5)_+ \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & (x_n - \kappa_1)_+ & (x_n - \kappa_2)_+ & (x_n - \kappa_3)_+ & (x_n - \kappa_4)_+ & (x_n - \kappa_5)_+ \end{bmatrix}$$

But to avoid overfitting we use penalized least squares:

Minimise RSS subject to $\sum_{k=1}^{5} \beta_{1+k}^2 < C$

for some $C > 0$.

This is mathematically equivalent to minimising: (for some $\lambda > 0$):

$$\text{RSS} + \lambda \sum_{k=1}^{5} \beta_{1+k}^2.$$

Note that we only penalise the coefficients of the spline functions.

For $\lambda = 0$ we get ordinary least squares.

$\lambda > 0$ is often referred to as ridge regression.

We end up with

$$\widehat{\boldsymbol{\beta}}_\lambda = (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{D})^{-1}\boldsymbol{X}^T\boldsymbol{y} \quad \text{where}$$

$$\boldsymbol{D} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{0}_{2\times 2} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{5\times 5} \end{bmatrix}.$$

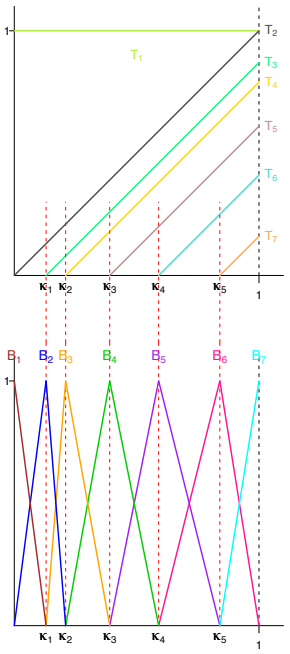
Alternative Spline Bases

The graphic on the following page shows the original basis functions

$$T_1, T_2, \ldots, T_7$$

and an alternative set of basis functions:

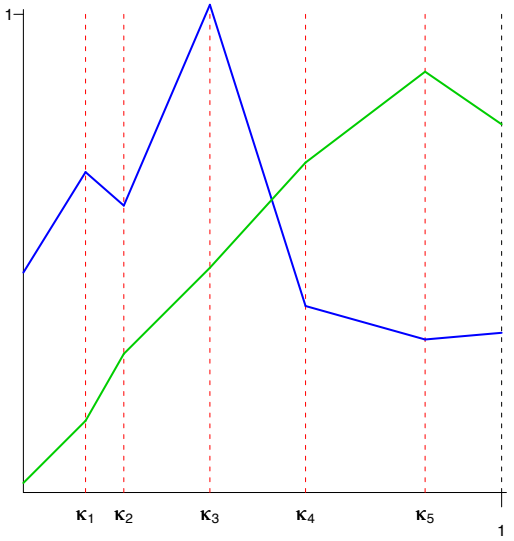
$$B_1, B_2, \ldots, B_7.$$



$\{T_1, \ldots, T_7\}$ is a basis for the vector space

$$\mathcal{V} = \{\text{continuous piecewise lines on } [0,1] \text{ with knots at } \kappa_1, \kappa_2, \kappa_3, \kappa_4 \text{ and } \kappa_5\}.$$

Two typical members of $\mathcal{V}$ are as shown in the next graphic:



One can show that

$$\{B_1, B_2, \ldots, B_7\}$$

is an invertible linear transformation of

$$\{T_1, T_2, \ldots, T_7\}.$$

i.e. $\{B_1, B_2, \ldots, B_7\}$ is an alternative basis for $\mathcal{V}$, known as the B-spline basis.

B-splines are:

- bounded,
- have compact support,
- closer to being orthogonal.

$\Longrightarrow$ better numerical properties.

Recall the $\boldsymbol{X}$ matrix from earlier:

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 & (x_1-\kappa_1)_+ & (x_1-\kappa_2)_+ & (x_1-\kappa_3)_+ & (x_1-\kappa_4)_+ & (x_1-\kappa_5)_+ \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & (x_n-\kappa_1)_+ & (x_n-\kappa_2)_+ & (x_n-\kappa_3)_+ & (x_n-\kappa_4)_+ & (x_n-\kappa_5)_+ \end{bmatrix}.$$

Let $T_j(x_i)$ be the (i,j) entry of $\boldsymbol{X}$.

Also, give $\boldsymbol{X}$ the new name $\boldsymbol{X}_T$ to emphasise its use of Truncated line basis functions.

i.e.

$$\boldsymbol{X}_T = \begin{bmatrix} T_1(x_1) & T_2(x_1) & T_3(x_1) & T_4(x_1) & T_5(x_1) & T_6(x_1) & T_7(x_1) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ T_1(x_n) & T_2(x_n) & T_3(x_n) & T_4(x_n) & T_5(x_n) & T_6(x_n) & T_7(x_n) \end{bmatrix}.$$

From earlier

$$\widehat{\boldsymbol{\beta}}_\lambda = (\boldsymbol{X}_T^T\boldsymbol{X}_T + \lambda\boldsymbol{D})^{-1}\boldsymbol{X}_T^T\boldsymbol{y}.$$

$\Longrightarrow$ the vector of fitted values is

$$\widehat{\boldsymbol{y}}_\lambda = \boldsymbol{X}_T\widehat{\boldsymbol{\beta}}_\lambda$$

$\Longrightarrow$

$$\widehat{\boldsymbol{y}}_\lambda = \boldsymbol{X}_T(\boldsymbol{X}_T^T\boldsymbol{X}_T + \lambda\boldsymbol{D})^{-1}\boldsymbol{X}_T^T\boldsymbol{y}.$$

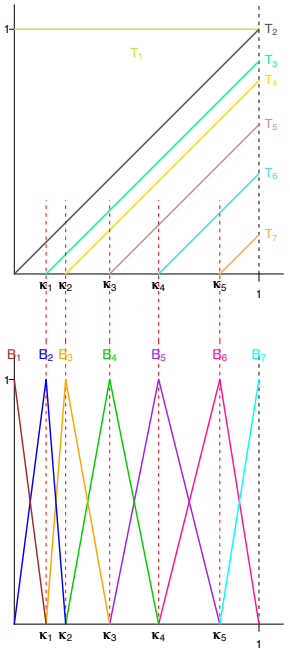
Let

$$\boldsymbol{X}_B = \begin{bmatrix} B_1(x_1) & B_2(x_1) & B_3(x_1) & B_4(x_1) & B_5(x_1) & B_6(x_1) & B_7(x_1) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ B_1(x_n) & B_2(x_n) & B_3(x_n) & B_4(x_n) & B_5(x_n) & B_6(x_n) & B_7(x_n) \end{bmatrix}$$

Then

$$\boldsymbol{X}_B = \boldsymbol{X}_T \boldsymbol{L}$$

where $\boldsymbol{L}$ is a 7×7 invertible matrix.



With some algebra we can show that

$$\widehat{\boldsymbol{y}}_\lambda = \boldsymbol{X}_B(\boldsymbol{X}_B^T\boldsymbol{X}_B + \lambda \boldsymbol{L}^T\boldsymbol{D}\boldsymbol{L})^{-1}\boldsymbol{X}_B^T\boldsymbol{y}$$

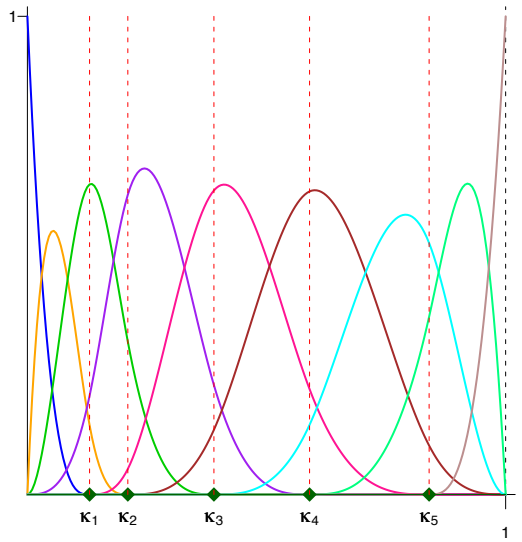
Changing from one basis to another involves an adjustment to the penalty term.

$$\text{i.e. } \lambda\boldsymbol{D} \rightarrow \lambda\boldsymbol{L}^T\boldsymbol{D}\boldsymbol{L}.$$

Cubic B-splines

In practice it is more common to use cubic rather than linear splines.

The cubic B-spline basis for this set of knots is shown in the next graphic:



The R function smooth.spline() uses penalized cubic B-splines.

The mathematics for cubic penalized splines is similar to that given in this lecture for linear penalized splines.

Even though cubic penalized splines are more common in practice, linear penalized splines allow us to explain the basic ideas fairly simply.

Bayesian Penalized Splines

In earlier slides $\beta_3, \ldots, \beta_{K+2}$ are the spline coefficients.

If we treat these coefficients as random with

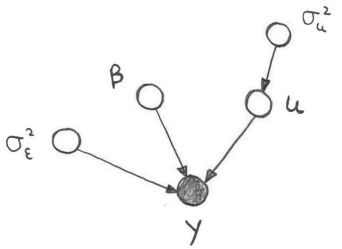
$$\text{e.g. } \beta_3, \ldots, \beta_{K+2} | \sigma_u^2 \sim N(0, \sigma_u^2)$$
$$\sigma_u \sim \text{Half-Cauchy}(100000)$$

then Bayesian inference automatically takes care of choosing a (hopefully) good smoothing parameter.

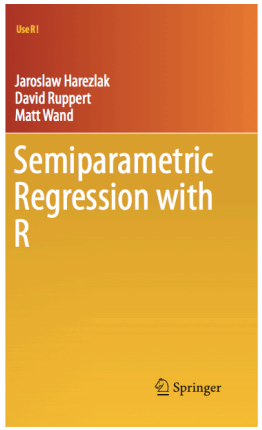
If we re-label as follows:

$$u_1 = \beta_3, \; u_2 = \beta_4, \; \ldots, \; u_K = \beta_{K+2}$$

then we have a Bayesian model analogous to the linear mixed model for grouped data:



One could write an entire book on the uses of penalized splines, such as:



♦ ♦ ♦