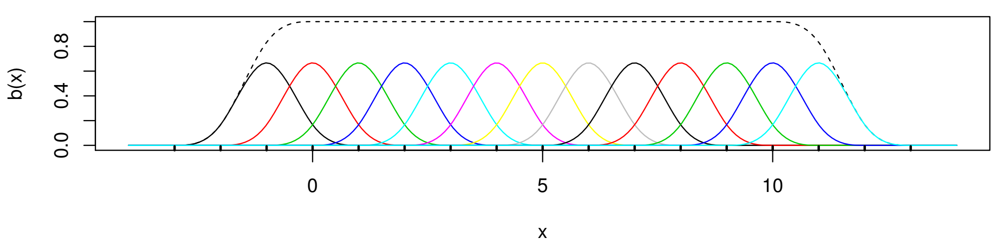
MATHM0019 Assessed Practical, 2018

You should work in groups of 3 for this assignment, handing in one report at the end. It is worth 20% of the marks for this unit.

This project is about smoothing. The aim is to write code for one dimensional 'P-spline' smoothing in R using the `splines` package. The idea is that a smooth function, $f(x)$, can be represented using 'B-spline' basis functions as:

$$f(x) = \sum_{i=1}^{K} \beta_j b_j(x)$$

where the $b_j(x)$ are B spline basis functions[1]. The B-splines are smooth curves defined with reference to a set of 'knots', x_k^*, which are evenly spaced along the x axis. Each B-spline is non-zero over only a limited range of knots. The following plot illustrates a set of 13 B-spline basis functions suitable for representing smooth functions of x over the interval $0 < x < 10$. The dashed line illustrates the sum of the basis functions - it is one over $0 < x < 10$. The 17 knots used to define the basis functions are illustrated with thick black tick marks along the x axis.



The R library `splines` (`library(splines)` loads it into R) contains a function `splineDesign` for creating B-spline bases. In particular, if we want to set up the model

$$y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2) \tag{1}$$

using a B-spline basis for f, then `splineDesign` can be used to produce a matrix, $\mathbf{X}$ such that $X_{ij} = b_j(x_i)$. In that case the model (1) can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

and fitted in R using something like `lm(y~X-1)` (the '`-1`' serving to suppress the redundant extra intercept that would be added by default).

It is possible to control the amount of smoothing by controlling the number of basis functions used, but it is then time consuming and difficult to select the number objectively. An alternative is to use more basis functions than you think are needed, but to put a prior on function 'wiggliness' that assigns higher probability to smooth functions than to wiggly ones. The idea is to impose a prior that penalizes rapid variation in the spline coefficients β_j, which in turn has the effect of penalizing rapid variation in f itself.

A suitable penalty might be

$$\sum_{j=2}^{K-1} (\beta_{j+1} - 2\beta_j + \beta_{j-1})^2$$

It is easy to evaluate this in R...

```
D <- diff(diag(K),differences=2)
Db <- D %*%beta
t(Db)%*%Db
```

[1] be aware that if you look up B splines online you will find alot of discussion of a more general use of the term B-spline for describing general curves in 2 dimensions (useful in computer graphics, for example). Here we are considering a more limited definition

The prior on $\boldsymbol{\beta}$ embodying such a penalty would be $\boldsymbol{\beta} \sim N(\mathbf{0}, \boldsymbol{\tau} = \mathbf{D}^T\mathbf{D}\lambda)$ where λ is a 'smoothing parameter' parameter controlling the prior precision, ($\boldsymbol{\tau}$ is the precision matrix — the inverse covariance matrix).

It turns out that for fixed λ the posterior modes, $\hat{\boldsymbol{\beta}}$, are the minimizers of

$$\sum_{i=1}^{n}(y_i - f(x_i))^2/\sigma^2 + \lambda\boldsymbol{\beta}^T\mathbf{D}^T\mathbf{D}\boldsymbol{\beta} \quad \left(= \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2/\sigma^2 + \lambda\boldsymbol{\beta}^T\mathbf{D}^T\mathbf{D}\boldsymbol{\beta}\right).$$

$\hat{\boldsymbol{\beta}}$ is easily obtained using code like:

```
Xe <- rbind(X,D*lambda^.5*sigma)
ye <- c(y,rep(0,nrow(D)))
coef(lm(ye ~ Xe - 1))
```

The aim of this practical is to produce code implementing the model (1). As a practical example take the `mcycle` data from R library `MASS`, using acceleration as the y variable, and time as the x.

1. Write a function that will produce the model matrix $\mathbf{X}$ for (1), using a B-spline basis, given a vector of values for x, and a required number of knots, K, as arguments. Use the default order 4 B-splines produced by `splineDesign`.

2. As a test, fit your model to the `mcycle` data, using $K = 10$ basis functions, and plot the smooth curve overlaid on the data.

3. Now produce the smoothing penalty and fit a penalized version of your model with $K = 30$, varying the smoothing parameter in order to demonstrate its effect.

4. To *estimate* the appropriate degree of smoothing, as controlled by λ, we can treat $\boldsymbol{\beta}$ as random and seek to maximize the *marginal likelihood* of λ and σ^2. This will be simplest to accomplish if we first modify $\mathbf{D}$ to be full rank, so that the assumed prior distribution for $\boldsymbol{\beta}$ is proper. The following code achieves this while making rather little difference to the model fit.

   ```
   eps <- 1e-3
   D <- rbind(diff(diag(K),differences=2),
         c(rep(0,K-2),eps,0),c(rep(0,K-1),eps))
   ```

   Show that under the smoothing model the marginal distribution of $\mathbf{y}$ is $N(\mathbf{0}, \mathbf{X}\mathbf{D}^{-1}\mathbf{D}^{-T}\mathbf{X}^T/\lambda + \mathbf{I}\sigma^2)$. Hence write an R function to evaluate the negative log marginal likelihood of $\boldsymbol{\theta} = (\log\lambda, \log\sigma^2)^T$, suitable for optimization using `optim`.

5. Hence estimate λ and σ^2 for the motorcycle data (using $K = 30$). Note that the marginal likelihood can be slightly tricky to optimize. You should use `optim`'s default method, and be careful about starting values (initial values of $\log\lambda = -2$ and $\log\sigma^2 = 5$ should be OK).

6. Produce a plot of acceleration against time overlaying the estimated smooth over a plot of the raw data.

7. By considering section 3.7 of the notes, produce a 95% credible interval for the smooth, and overlay this on your plot.

What to hand in: By **12 noon on Friday 20th April 2018**, at the latest, each group should email the following to `simon.wood@bristol.ac.uk` with the subject `M0019 smooth` followed by your surnames:

1. Clear commented R code implementing (1) and carrying out the steps of the assignment given above. A plain text file is best for this.

2. A report of at most 4 pages (normal margins $\geq$ 10pt font), explaining the statistical basis of what has been done, and presenting the plots, along with the evidence that the method is working. This should be a pdf document.

Mark scheme guidance

First class A report without substantive errors which could be used as the basis for a consultancy meeting with the scientists who gathered the original data, and code that could be handed over for others to use, as is. No major omissions. Sensible well backed up conclusions. Work could be repeated on basis of what is written. Reasons for choices are explained and reasonable. Code well structured and commented.

2.1 Would require some relatively minor work before being usable as the basis for a consultancy meeting with the scientists who gathered the original data, and or the code passed on. Some errors or omissions, but basically sound and well put together. Most issues covered well.

2.2 Would require substantial revisions before being usable as the basis for a consultancy meeting with the scientists who gathered the original data, and or the code passed on. Some serious errors or omissions, but some other components of a good standard.

3 Not suitable as the basis for a consultancy meeting with the scientists who gathered the original data, without re-doing. Major errors or omissions, but some understanding of the material demonstrated.