# APTS course 21st August – 25th August 2017

## Nonparametric Smoothing

*Preliminary Material*

Adrian Bowman & Ludger Evers

University of Glasgow | School of Mathematics & Statistics

This APTS course will cover a variety of methods which enable data to be modelled in a flexible manner. As preparation, it would be helpful to revise the following topics covered in earlier APTS courses:

- linear models, including the Bayesian version;
- generalised linear models;
- `R` programming;
- Taylor series expansions and standard asymptotic methods.

The main emphasis will be on regression settings, because of the widespread use and application of this kind of data structure. However, the preliminary material also covers various aspects of density estimation, to introduce some of the main ideas of nonparametric smoothing and to highlight some of the main issues involved. It is likely that many people will have come across these ideas in one form or another. The preliminary material aims to

- explore simple kernel methods and show how these can be used to construct smooth density estimates and regression curves;
- explore spline approaches to constructing regression curves;
- investigate some simple theoretical properties;
- experiment with software available in `R`;
- consider some illustrations of their use in analysing data.

Exercises are provided to assist in engaging with the material.

# 1 Density estimation

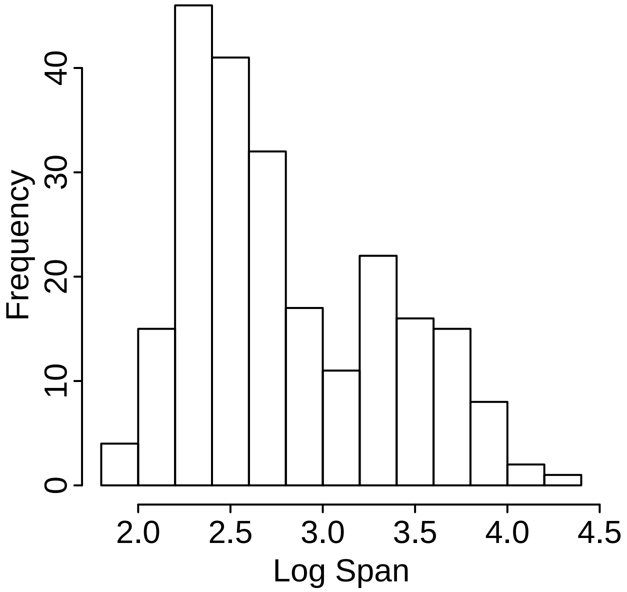
A probability density function is a key concept through which variability can be expressed precisely. In statistical modelling its role is often to capture variation sufficiently well, within a model where the main interest lies in structural terms such as regression coefficients. However, there are some situations where the shape of the density function itself is the focus of attention. The example below illustrates this.

Example: Aircraft data

These data record six characteristics of aircraft designs which appeared during the twentieth century. The variables are:

| | |
|---|---|
| `Yr` | year of first manufacture |
| `Period` | a code to indicate one of three broad time periods |
| `Power` | total engine power (kW) |
| `Span` | wing span (m) |
| `Length` | length (m) |
| `Weight` | maximum take-off weight (kg) |
| `Speed` | maximum speed (km/h) |
| `Range` | range (km) |

A brief look at the data suggests that the six measurements on each aircraft should be expressed on the log scale to reduce skewness. Span is displayed on a log scale below, for Period 3 which corresponds to the years after the Second World War. The pattern of variability shown in the histogram exhibits some skewness. There is perhaps even a suggestion of a subsidiary mode at high values of log span, although this is difficult to evaluate.



## 1.1 A simple density estimate

The histogram is a very familiar object. It can be written as

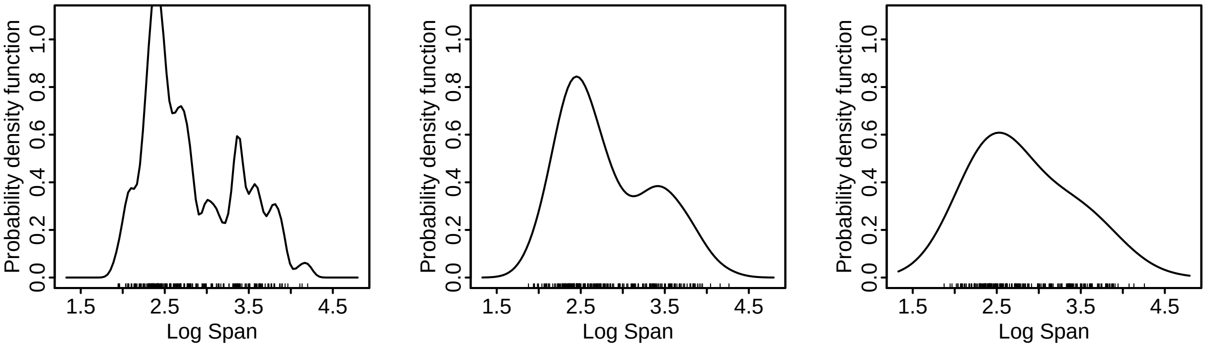
$$\tilde{f}(y) = \sum_{i=1}^{n} I(y - \tilde{y}_i; h),$$

where $\{y_1, \ldots, y_n\}$ denote the observed data, $\tilde{y}_i$ denotes the centre of the interval in which $y_i$ falls and $I(z; h)$ is the indicator function of the interval $[-h, h]$. (Notice that further scaling would be required to ensure that $\tilde{f}$ integrates to 1.)

The form of the construction of $\tilde{f}$ highlights some features which are open to criticism if we view the histogram as an estimator of the underlying density function. Firstly the histogram is not smooth, when we expect that the underlying density usually will be. Secondly, some information is lost when we replace each observation $y_i$ by the bin mid-point $\tilde{y}_i$. Both of these issues can be addressed by using a density estimator in the form

$$\hat{f}(y) = \frac{1}{n} \sum_{i=1}^{n} w(y - y_i; h),$$

where $w$ is a probability density, called here a *kernel function*, whose variance is controlled by the *smoothing parameter* $h$. The middle panel in the plots below shows the effects of doing this with the aircraft data. Large changes in the value of the smoothing parameter have large effects on the smoothness of the resulting estimates, as the left and right hand plots below illustrate.

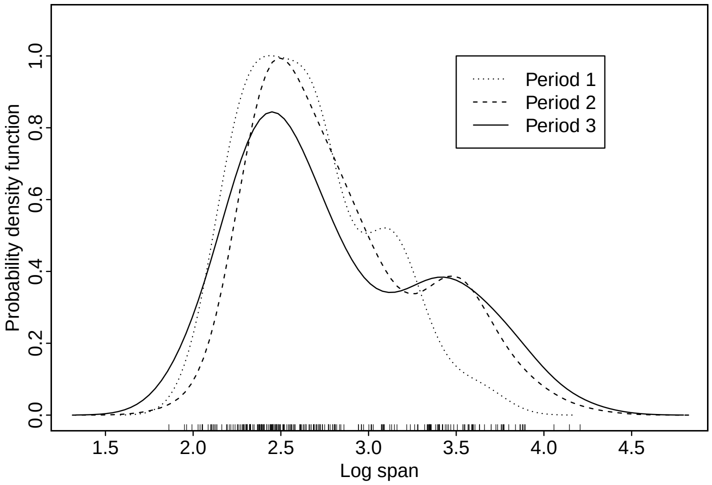


Exercise: The effect of the smoothing parameter

To experiment with density estimates, download the `sm`, `rpanel` and `tkrplot` packages for R. The code below should launch a new window with interactive controls. Try altering the smoothing parameter through the slider. Does this help you assess whether the subsidiary mode is a genuine feature or an artefact of random variation?

```
library(sm)
y <- log(aircraft$Span[aircraft$Period == 3])
sm.density(y, panel = TRUE)
```

One advantage of density estimates is that it is a simple matter to superimpose these to allow different groups to be compared. Here the groups for the three different time periods are compared. It is interesting that the 'shoulder' appears in all three time periods.



Exercise: Computing density estimates

Write some `R` code which will take a vector of data `y` and return a density estimate. See if you can do this without using a `for` loop!

## 1.2 Extending the idea

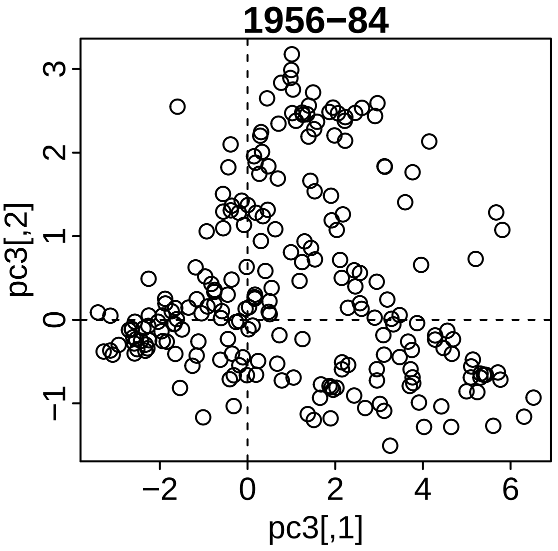
The simple idea of density estimation is to place a kernel function, which in fact is itself a density function, on top of each observation and average these functions. This extends very naturally to a wide variety of other types of data and sample spaces.

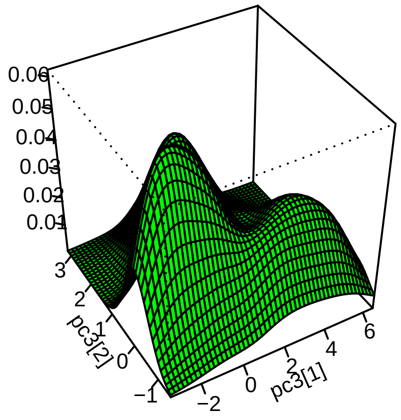
For example, a two-dimensional density estimate can be constructed from bivariate data $\{(y_{1i}, y_{2i}) : i = 1, \ldots, n\}$ by employing a two-dimensional kernel function in the form

$$\hat{f}(y_1, y_2) = \frac{1}{n} \sum_{i=1}^{n} w(y_1 - y_{1i}; h_1)\, w(y_1 - y_{2i}; h_2).$$

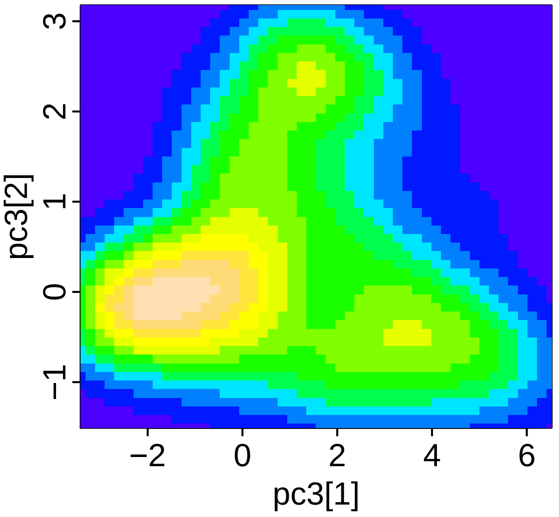
Notice that there are now two smoothing parameters, $(h_1, h_2)$. A more general two-dimensional kernel function could be used, but the simple product form is very convenient and usually very effective.

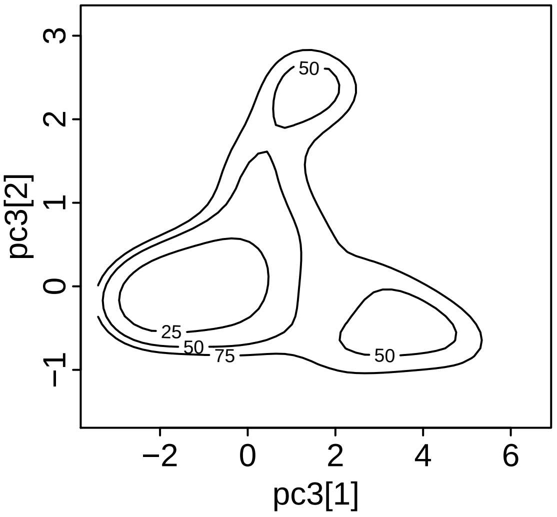
Here is an example which uses the scores from the first two principal components of the aircraft data, again focussing on the third time period. The left hand scatterplot shows the individual scores while the right hand plot shows a density estimate, which suggests three separate modes. This feature is not so easily seen from the raw scatterplot.





Here are alternative ways of presenting a two-dimensional estimate, using a coloured image on the left and contour lines on the right. Notice that the contours on the right have been chosen carefully to contain the quarters of the data with successively higher density, in a manner which has some similarities with a boxplot.





This principle extends to all kinds of other data structures and sample spaces by suitable choice of an appropriate kernel function.

Exercise: Circular data

If you were given a sample of circular data, consisting of angles (in radians) between 0 and $2\pi$ (where, of course, these two end points are coincident), how could you construct a smooth density estimate by choice of an appropriate kernel function? (*Hint*: the *von Mises* distribution provides a natural model for a distribution on a circle.) Can you write some code in R which will construct a density estimate from a sample of circular data. (*Hint*: you might find the R function `besselI` useful.)

## 1.3 Simple properties of density estimates

Without any real restriction, we can assume that the kernel function can be written in the simple form $w(y-y_i;h)=\frac{1}{h}w\left(\frac{y-y_i}{h}\right)$. The mean of a density estimator can then be written as

$$\mathbb{E}\left\{\hat{f}(y)\right\}=\int\frac{1}{h}w\left(\frac{y-z}{h}\right)f(z)dz=\int w(u)f(y-hu)du,$$

where the last expression simply involves a change of variable $u=\frac{y-z}{h}$. A Taylor series expansion of the term involving $f$ in the last expression gives

$$f(y-hu)=f(y)-huf'(y)+\frac{1}{2}h^2u^2f''(y)+o(h^2)$$

and, on insertion into the expression for the mean, this produces the approximation

$$\mathbb{E}\left\{\hat{f}(y)\right\}\approx f(y)+\frac{h^2}{2}\sigma_w^2 f''(y),$$

where we assume that the kernel function is symmetric so that $\int uw(u)du=0$, and where $\sigma_w^2$ denotes the variance of the kernel, namely $\int u^2w(u)du$.

The variance of the density estimate can be written as

$$\begin{aligned}\mathsf{var}\left\{\hat{f}(y)\right\} &= \frac{1}{n}\mathsf{var}\left\{\frac{1}{h}w\left(\frac{y-Y}{h}\right)\right\}\\ &= \frac{1}{n}\left\{\mathbb{E}\left\{\left[\frac{1}{h}w\left(\frac{y-Y}{h}\right)\right]^2\right\}-\mathbb{E}\left\{\frac{1}{h}w\left(\frac{y-Y}{h}\right)\right\}^2\right\}\end{aligned}$$

A similar change of variable and Taylor series expansion produces the approximation

$$\mathsf{var}\left\{\hat{f}(y)\right\}=\frac{1}{nh}f(y)\,\alpha(w)+o\left(\frac{1}{nh}\right),$$

where $\alpha(w)=\int w^2(u)du$.

These expressions capture the essential features of smoothing. In particular, bias is incurred and we can see that this is controlled by $f''$, which means that where the density has peaks and valleys the density estimate will underestimate and overestimate respectively. This makes intuitive sense.

Similar expressions can be derived in higher dimensions.

Exercise: Asymptotic properties

A useful global measure of performance is the *mean integrated squared error* (MISE) which balances squared bias and variance.

$$\begin{aligned}\text{MISE}(\hat{f}) &= \mathbb{E}\Big\{\int[\hat{f}(y)-f(y)]^2dy\Big\}\\ &= \int\Big[\mathbb{E}\Big\{\hat{f}(y)\Big\}-f(y)\Big]^2dy+\int\mathsf{var}\Big\{\hat{f}(y)\Big\}\,dy.\end{aligned}$$

(i) Verify the expression given for $\mathsf{var}\Big\{\hat{f}(y)\Big\}$ above by employing a change of variable and a Taylor series expansion.

(ii) Use the Taylor expansions discussed above to show that MISE can be expressed as

$$\text{MISE}(\hat{f}) \approx \frac{1}{4}\,h^4\sigma_w^4\int f''(y)^2dy+\frac{1}{nh}\,\alpha(w).$$

(iii) Now show that the value of $h$ which minimizes MISE in an asymptotic sense is

$$h_{\text{opt}}=\left\{\frac{\gamma(w)}{\beta(f)n}\right\}^{1/5},$$

where $\gamma(w)=\alpha(w)/\sigma_w^4$, and $\beta(f)=\int f''(y)^2dy$.

(iv) Show that when $f$ is a normal density this yields the simple formula

$$h=\left(\frac{4}{3n}\right)^{1/5}\sigma,$$

where $\sigma$ denotes the standard deviation of the distribution.

(v) Considering $\hat{f}(y)$ as a density function in $y$, what is the mean and variance of the distribution which this density function represents? Consider what this says about the operation of smoothing.

## 1.4 Deciding how much to smooth

The theory sketched in the exercise above shows that an optimal smoothing parameter can be defined as the value which minimises MISE. This has the form

$$h_{\text{opt}}=\left\{\frac{\gamma(w)}{\beta(f)n}\right\}^{1/5}.$$

Of course, this is of rather limited use because it is a function of the unknown density. However, there are two practical approaches which can be taken to deciding on a suitable smoothing parameter to use. One is to construct an estimate of MISE and minimise this. Another is to estimate the optimal smoothing parameter. These two approaches are outlined below.

### Cross-validation

The integrated squared error (ISE) of a density estimate is

$$\int \{\hat{f}(y) - f(y)\}^2 dy = \int \hat{f}(y)^2 dy - 2\int f(y)\hat{f}(y)dy + \int f(y)^2 dy.$$

Only the first two of these terms involve $h$ and these terms can be estimated by

$$\frac{1}{n}\sum_{i=1}^{n}\int \hat{f}_{-i}^2(y)dy - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{-i}(y_i),$$

where $\hat{f}_{-i}(y)$ denotes the estimator constructed from the data without the observation $y_i$. The value of $h$ which minimises this expression is known as the *cross-validatory* smoothing parameter.

### Plug-in methods

By inserting suitable estimates of the unknown quantities in the formula for the optimal smoothing parameter, a *plug-in* choice can be constructed. The difficult part is the estimation of $\beta(f)$ as this involves the second derivative of the density function. Sheather & Jones (JRSSB 53, 683–90) came up with a good, stable way of doing this. The Sheather-Jones remains one of the most effective strategies for choosing the smoothing parameter.

A very simple plug-in approach is to use the normal density function in the expression for the optimal smoothing parameter. An earlier exercise showed that this produces the simple formula $\left(\frac{4}{3n}\right)^{1/5}\sigma$, where the standard deviation $\sigma$ can be estimated from the data. This approach is surprisingly effective, in large part because it is very stable.

Exercise: Smoothing parameter selection in practice

The R expression `sm.density(y, panel = TRUE)` allows interactive experimentation in the construction of density estimates from a data vector `y`. Try this function out with data of your own choice and look at the comparative performances of the methods of smoothing parameter section outlined here.

Try this out with two-dimensional data, using the aircraft data or some simple randomly generated data, for example by `y <- cbind(rnorm(50), rnorm(50))`. The `sm.density` function operates as before, although the Sheather-Jones plug-in method is not available because it is trickier to implement beyond the one-dimensional setting.

## 1.5 Some simple inferential tools

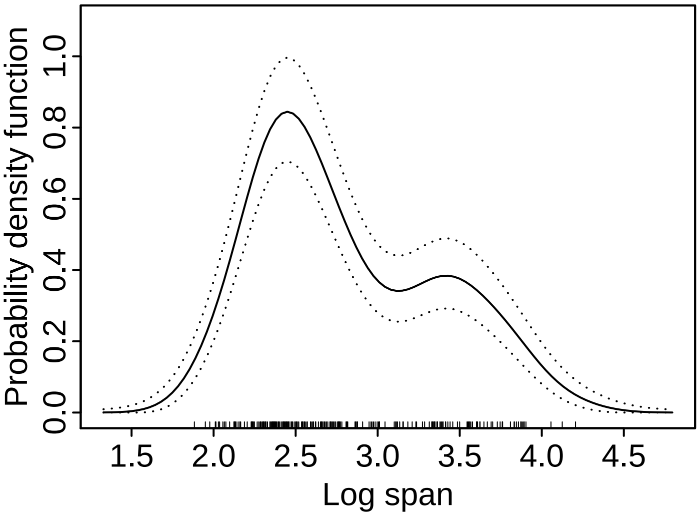
Once an estimate has been constructed, a natural next step is to find its standard error. The earlier result on the variance of $\hat{f}$ is a natural starting point, but this expression involves the unknown density. A helpful route is to consider a ‘variance stabilising’ transformation. For any transformation $t(\cdot)$, a Taylor series argument shows that

$$\mathsf{var}\left\{t(\hat{f}(y))\right\} \approx \mathsf{var}\left\{\hat{f}(y)\right\}\left[t'\left(\mathbb{E}\left\{\hat{f}(y)\right\}\right)\right]^2.$$

When $t(\cdot)$ is the square root transformation, the principal term of this expression becomes

$$\mathsf{var}\left\{\sqrt{\hat{f}(y)}\right\} \approx \frac{1}{4}\frac{1}{nh}\,\alpha(w),$$

which does not depend on the unknown density $f$. This forms the basis of a useful *variability band.* We cannot easily produce proper confidence intervals because of the bias present in the estimate. However, if the standard error is constructed and the intervals corresponding to two s.e.'s on the square root scale are transformed back to the origin scale, then a very useful indication of the variability of the density estimate can be produced. This is shown below for the aircraft span data from period 3.



A useful variation on this arises when the true density function is assumed to be normal with mean $\mu$ and variance $\sigma^2$, and the kernel function $w$ is also normal. If the standard normal density function is denoted by $\phi$, then the mean and variance of the density estimate at the point $y$ are then

$$\begin{aligned}
\mathbb{E}\left\{\hat{f}(y)\right\} &= \phi\left(y-\mu;\sqrt{h^2+\sigma^2}\right)\\
\mathsf{var}\left\{\hat{f}(y)\right\} &= \frac{1}{n}\phi\left(0;\sqrt{2}\,h\right)\,\phi\left(y-\mu;\sqrt{\sigma^2+\frac{1}{2}h^2}\right)-\frac{1}{n}\phi\left(y-\mu;\sqrt{\sigma^2+h^2}\right)^2
\end{aligned}$$

These expressions allow the likely range of values of the density estimate to be calculated, under the assumption that the data are normally distributed. This can be expressed graphically through a *reference band.*

Exercise: Icelandic tephra layer

Data on the percentages of aluminium oxide found in samples from a tephra layer resulting from a volcanic eruption in Iceland around 3500 years ago are available in the `tephra` dataset in the `sm` package. To deal with the percentage scale, apply the logit transformation

```
logit <- log(tephra$Al2O3/(100-tephra$Al2O3))
```

Now use the `sm.density` function with the interactive controls provided by `panel = TRUE` to add standard errors and a reference band for a normal model. Is there convincing evidence that the tephra data exhibit non-normal features?

Exercise: Means and variance again

Verify the results stated above on the approximate variance of $\sqrt{\hat{f}(y)}$ and the exact mean and variance of $\hat{f}(y)$ when the underlying distribution is normal with mean $\mu$ and variance $\sigma^2$.

The bootstrap is sometimes a useful way of gaining information about the properties and sampling variation of quantities of interest. Here is a simple version of the bootstrap for density estimation.

1. Construct a density estimate $\hat{f}$ from the observed data $\{y_1, \ldots, y_n\}$.
2. Resample the data with replacement to produce a bootstrap sample $\{y_1^*, \ldots, y_n^*\}$.
3. Construct a bootstrap density estimate $\hat{f}^*$ from the bootstrap data $\{y_1^*, \ldots, y_n^*\}$.
4. Repeat steps 2 and 3 a large number of times to create a collection of bootstrap density estimates $\{\hat{f}_1^*, \ldots, \hat{f}_B^*\}$.
5. Use the empirical distribution of $\hat{f}^*$ about $\hat{f}$ to mimic the distribution of $\hat{f}$ about $f$.

However, we need to be careful about the interpretation of this. Since the distribution of $y_i^*$ is uniform over $\{y_1, \ldots, y_n\}$, it follows that

$$\mathbb{E}_*\left\{\hat{f}^*(y)\right\} = \mathbb{E}_*\{w(y - y_i^*; h)\} = \hat{f}(y)$$

and so the bias which we know is present in the distribution of $\hat{f}$ is absent in the bootstrap version. However, the bootstrap does usefully mimic the variance of $\hat{f}$.

Exercise: Bootstrapped variability bands

Write a few lines of R code which will construct and plot bootstrap density estimates of the aircraft log span data. Compare the results with the variability band shown near the start of Section 5.

# 2 Nonparametric regression

Regression is one of the most widely used modelling paradigms and this will be the main focus in the course. Here is an example which will be used to illustrate the initial discussion.

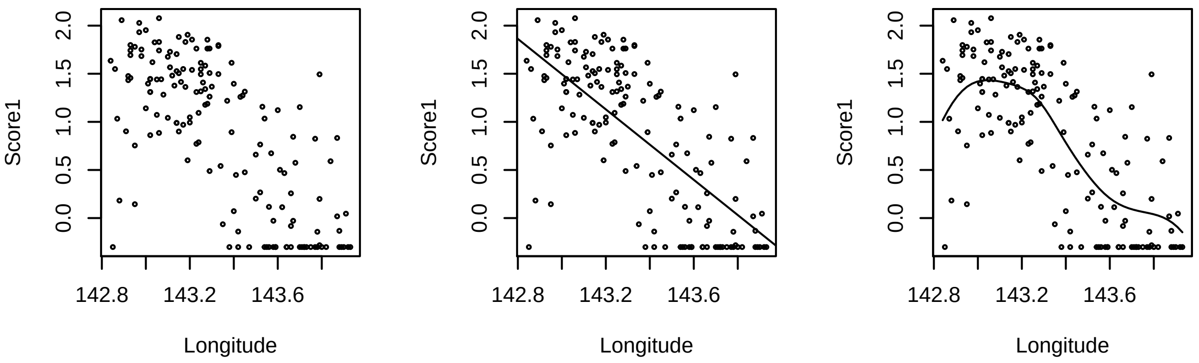
Example: Great Barrier Reef data

A survey of the fauna on the sea bed lying between the coast of northern Queensland and the Great Barrier Reef was carried out. The sampling region covered a zone which was closed to commercial fishing, as well as neighbouring zones where fishing was permitted. The variables are:

| | |
|---|---|
| `Zone` | an indicator for the closed (1) and open (0) zones |
| `Year` | an indicator of 1992 (0) or 1993 (1) |
| `Latitude` | latitude of the sampling position |
| `Longitude` | longitude of the sampling position |
| `Depth` | bottom depth |
| `Score1` | catch score 1 |
| `Score2` | catch score 2 |

The details of the survey and an analysis of the data are provided by Poiner et al. (1997), *The effects of prawn trawling in the far northern section of the Great Barrier Reef*, CSIRO Division of Marine Research, Queensland Dept. of Primary Industries.

The relationship between catch score (Score1) and longitude is of particular interest because, at this geographical location, the coast runs roughly north-south and so longitude is a proxy for distance offshore. We might therefore reasonably expect the abundance of marine life to change with longitude. The first of the three panels below shows that there is indeed a strong underlying negative relationship, with considerable variability also present. The middle panel summarises this in a simple linear regression which captures much of this relationship. However, if we allow our regression model to be more flexible then a more complex relationship is suggested in the right hand panel, with a broadly similar mean level for some distance offshore followed by a marked decline, possibly followed by some levelling off thereafter. This gives valuable informal and graphical insight into the data but how can flexible regression models can be constructed and how can we use them to evaluate whether there is really evidence of non-linear behaviour in the data?



## 2.1 A local fitting approach

A simple nonparametric model has the form

$$y_i = m(x_i) + \varepsilon_i,$$

where the data $(x_i, y_i)$ are described by a smooth curve $m$ plus independent errors $\varepsilon_i$. One approach to fitting this is to take a model we know and fit it locally. For example, we can construct a *local linear regression*. This involves solving the least squares problem

$$\min_{\alpha,\beta} \sum_{i=1}^{n} \{y_i - \alpha - \beta(x_i - x)\}^2 \, w(x_i - x\, ; h)$$

and taking as the estimate at $x$ the value of $\hat{\alpha}$, as this defines the position of the local regression line at the point $x$. This has an appealing simplicity and it can be generalised quite easily to other situations. This was the approach used to produce the nonparametric regression of the Reef data in the plot above.

An even simpler approach is to fit a local mean. Specifically, at any point of interest $x$, we choose our estimator of the curve there as the value of $\mu$ which minimises

$$\sum_{i=1}^{n} \{y_i - \mu\}^2 w(x_i - x; h)$$

and this is easily shown to produce the 'running mean'

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} w(x_i - x; h)\, y_i}{\sum_{i=1}^{n} w(x_i - x; h)}.$$

If we do the algebra to minimise the sum-of-squares in the local linear approach, then an explicit formula for the local estimator can be derived as

$$\hat{m}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{\{s_2(x;h) - s_1(x;h)(x_i - x)\} w(x_i - x; h) y_i}{s_2(x;h) s_0(x;h) - s_1(x;h)^2},$$

where $s_r(x;h) = \{\sum (x_i - x)^r w(x_i - x; h)\}/n$.

In both the local mean and the local linear cases, the estimator is seen to be of the form $\sum_i \kappa_i y_i$, where the weights $\kappa_i$ sum to 1. There is a broad sense then in which even the local linear method is 'locally averaging' the data. In fact, many other forms of nonparametric regression can also be formulated in a similar way.

## 2.2 Some simple properties

One question which immediately arises is whether it matters very much which form of nonparametric smoothing is used. Sometimes computational and other issues may constrain what choices are practical. However, if we take the simple local mean and local linear examples, what principles can we use to guide our choice? Deriving expression which capture simple properties such as bias and variance is an obvious place to start.

We will start with the local mean estimator. The exploration will be a little informal, without the full technicality of formal proofs (although these could be added if time permitted). The aim is to identify the properties of the estimator in conceptual form. If the numerator and denominator of the local mean estimator are both scaled by $1/n$, then the denominator has a familiar form, namely a kernel density estimator. As we saw earlier, this has expectation

$$\mathbb{E}\left\{\frac{1}{n}\sum_i w(x_i - x; h)\right\} = f(x) + \frac{h^2}{2} f''(x) + o(h^2),$$

where, as before, we assume for convenience that the kernel function can be rewritten as $\frac{1}{h}w((x_i - x)/h)$ and $w$ is a symmetric probability density function around 0 with variance 1. Turning now to the numerator, we have

$$\begin{aligned}
\mathbb{E}\left\{\frac{1}{n}\sum_i w(x_i - x; h)\, y_i\right\} &= \frac{1}{n}\sum_i \frac{1}{h} w\left(\frac{x_i - x}{h}\right) m(x_i) \\
&\approx \int \frac{1}{h} w\left(\frac{z - x}{h}\right) m(z) f(z) \quad \text{[integral approximation]} \\
&= \int w(u) m(x + hu) f(x + hu) du \quad \text{[change of variable]}
\end{aligned}$$

Now apply a Taylor series expansion to the terms involving $x + hu$, to give

$$\begin{aligned}
m(x + hu) &= m(x) + hu\, m'(x) + \frac{(hu)^2}{2} m''(x) + o(h^2), \\
f(x + hu) &= f(x) + hu\, f'(x) + \frac{(hu)^2}{2} f''(x) + o(h^2).
\end{aligned}$$

Substituting these in and integrating over $u$ gives

$$\mathbb{E}\left\{\frac{1}{n}\sum_i w(x_i - x; h)\, y_i\right\} \approx m(x)f(x) + h^2 \left\{\frac{1}{2} f(x) m''(x) + m'(x) f'(x) + \frac{1}{2} f''(x) m(x)\right\} + o(h^2).$$

Dividing both numerator and demoninator by $f(x)$ gives

$$\begin{aligned}
\text{numerator:} \quad & m(x) + h^2 \left\{\frac{1}{2} m''(x) + m'(x)\frac{f'(x)}{f(x)} + \frac{1}{2}\frac{f''(x)}{f(x)} m(x)\right\} + o(h^2) \\
\text{denominator:} \quad & 1 + \frac{h^2}{2}\frac{f''(x)}{f(x)} + o(h^2)
\end{aligned}$$

The dominant term in the mean of the ratio of numerator and denominator is the ratio of the means. Applying the series expansion for $(1+x)^{-1}$ allows the reciprocal of the denominator to be written as

$$1 - \frac{h^2}{2}\frac{f''(x)}{f(x)} + o(h^2).$$

Multiplying the different terms out, we have

$$\begin{aligned}
\mathbb{E}\{\hat{m}(x)\} &\approx \left\{ m(x) + h^2 \left\{ \frac{1}{2}m''(x) + m'(x)\frac{f'(x)}{f(x)} + \frac{1}{2}\frac{f''(x)}{f(x)}m(x) \right\} + o(h^2) \right\} \\
&\quad \left\{ 1 - \frac{h^2}{2}\frac{f''(x)}{f(x)} + o(h^2) \right\} \\
&= m(x) + h^2 \left\{ \frac{1}{2}m''(x) + \frac{m'(x)f'(x)}{f(x)} \right\} + o(h^2).
\end{aligned}$$

Phew!

A similar sequence of manipulations (which you might like to try on your own) gives an asymptotic expression for the variance as

$$\mathsf{var}\{\hat{m}(x)\} \approx \frac{1}{nh}\left\{ \int w(u)^2 du \right\} \sigma^2 \frac{1}{f(x)},$$

where $\sigma^2$ denotes the variance of the error terms $\varepsilon_i$.

In the local linear case, the estimator can be written as $\sum_i a_i y_i / \sum_i a_i$, where $a_i = \frac{1}{n}\frac{1}{h}w(\frac{x_i - x}{h})\{s_2 - (x_i - x)s_1\}$. Consider first $s_1$, which can be written as

$$\begin{aligned}
s_1 &= \frac{1}{n}\sum_j \frac{1}{h} w\left(\frac{x_i - x}{h}\right)(x_j - x) \\
&\approx \int \frac{1}{h} w\left(\frac{x - x}{h}\right) f(z)(z - x)dz \\
&= \int w(u)hu\{f(x) + huf'(x) + o(h)\}du \\
&= h^2 f'(x) + o(h^2)
\end{aligned}$$

By a similar argument,

$$s_2 \approx h^2 f(x) + o(h^2).$$

The weights $a_i$ can then be approximated by

$$a_i \approx \frac{1}{n}\frac{1}{h} w\left(\frac{x_i - x}{h}\right) h^2\{f(x) - (x_i - x)f'(x)\}.$$

The mean of the estimator is $\mathbb{E}\{\hat{m}(x)\} = \sum_i a_i m(x_i) / \sum_i a_i$. Ignoring the term $h^2$ which cancels in the ratio, the numerator can be expressed as

$$\left\{ f(x)^2 + \frac{h^2}{2} f(x) f'(x) - h^2 f'(x)^2 \right\} m(x) + \frac{h^2}{2} f(x)^2 m''(x)^2,$$

after an integral approximation, a change of variable and a Taylor series expansion. By a similar argument, the denominator of $\mathbb{E}\{\hat{m}(x)\}$ can be approximated by

$$f(x)^2 + \frac{h^2}{2} f(x) f''(x) - h^2 f'(x)^2.$$

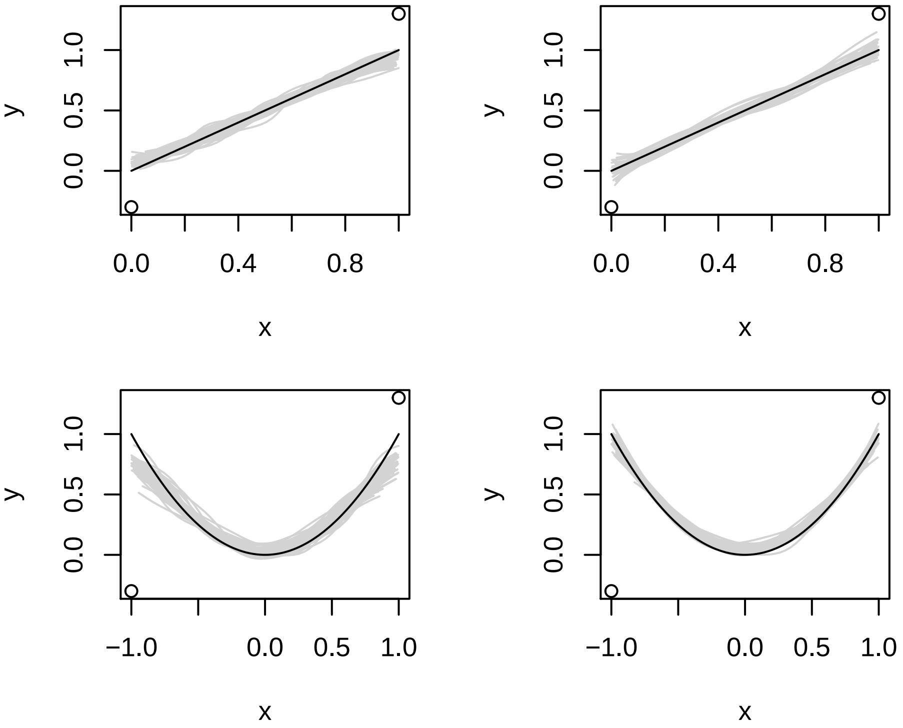
The principal term of the ratio then gives

$$\mathbb{E}\{\hat{m}(x)\} \approx m(x) + \frac{h^2}{2} m''(x).$$

So, after considerable work, a very simple expression has been achieved. Similar manipulations for the variance produces an expression which is exactly the same as that for the variance of the local mean estimator.

A comparison of the expressions for the local mean and local linear estimators is interesting. For example, the principal terms in the expression for the mean of the local linear estimator is not only simpler but also does not involve $f(x)$, both of which are attractive properties. This is one of the reasons that the local linear estimator is generally preferred over the local mean.

However, another issue concerns edge effects. These require more careful analysis to identify so, instead, we will use a simple illustration based on simulation. The figures below show the results of repeatedly simulating 50 data points, equally spaced over $[0, 1]$, from the model $y = x + \varepsilon$, where the standard deviation of the error terms is 0.1. For each set of simulated data, a nonparametric regression curve is plotted, using local mean (left) and local linear (right) estimators. Notice that at the ends of the sample space the local mean has strong bias, because there is data only on one side of the estimation point of interest. In contrast, the local linear method is unaffected. The same pattern is displayed in the lower plots, using the model $y = x^2 + \varepsilon$ over the range $[-1, 1]$.



With a little more theoretical work, a central limit theorem can be constructing to show that

$$\frac{\hat{m}(x) - m(x) - b(x)}{\sqrt{v(x)}} \to N(0,1),$$

where $b(x)$ and $v(x)$ denote the bias band variance of $\hat{m}(x)$.

Following on from the discussions in density estimation, the performance of a nonparametric estimator can be summarised in the *mean integrated squared error*, defined as

$$MISE = \int \mathbb{E}\{\hat{m}(x) - m(x)\}^2 f(x)dx$$

and an optimal smoothing parameter can be defined as the value of $h$ which minimises the asymptotic approximation of MISE, namely

$$h_{\text{opt}} = \left\{ \frac{\gamma(w)\sigma^2}{\int [m''(x)]^2 f(x)dx} \right\}^{1/5} n^{-1/5}.$$

If we use this optimal smoothing parameter, then both the bias and the square root of the variance, which determines the rate of convergence, are of order $n^{-2/5}$. Notice that this rate of convergence is slower than the $n^{-1/2}$ which applies for parametric models.

Fan & Gijbels (1996) give further details on the theoretical aspects of local polynomial smoothing.

## 2.3 Basis function approaches

Basis function approachaes are not based on local weights, but based on expanding the design matrix used in linear regression. To fix notation, we quickly state the simple linear regression model

$$\mathbb{E}\{\}\left(Y_i\right) = m(x_i) = \beta_0 + \beta_1 x_i \qquad \text{for } i = 1, \ldots, n,$$

or equivalently, in matrix-vector notation,

$$\mathbb{E}\{\}\left(\mathbf{y}\right) = \mathbf{B}\boldsymbol{\beta} \qquad \text{with } \mathbf{y} = (Y_1, \ldots, Y_n)^\top \text{ and } \mathbf{B} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

Basis function approaches effectivly consist of introducing functions of $\mathbf{x}$ (other than just the identity) into the design matrix $\mathbf{B}$.

Basis approaches to function approximation have a very long history. One of the first was Fourier series, which is based on the expansion

$$m(x_i) \approx \frac{a_0}{2} + \sum_{j=1}^{r} a_j \cos\left(\frac{2\pi j x_i}{P}\right) + b_j \sin\left(\frac{2\pi j x_i}{P}\right),$$

where $x_i \in (0, P)$. This approximation corresponds to using the design matrix

$$\mathbf{B} = \begin{pmatrix} \frac{1}{2} & \cos\left(\frac{2\pi x_1}{P}\right) & \sin\left(\frac{2\pi x_1}{P}\right) & \ldots & \cos\left(\frac{2\pi r x_1}{P}\right) & \sin\left(\frac{2\pi r x_1}{P}\right) & \\ & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{1}{2} & \cos\left(\frac{2\pi x_n}{P}\right) & \sin\left(\frac{2\pi x_n}{P}\right) & \ldots & \cos\left(\frac{2\pi r x_1}{P}\right) & \sin\left(\frac{2\pi r x_1}{P}\right) & \end{pmatrix}$$

with $\boldsymbol{\beta} = (a_0, a_1, b_1, \ldots, a_r, b_r)$.

Panel (a) below shows the basis functions of a Fourier basis with $r = 3$.

Each basis function in a Fourier expansion has effects across the entire range of the data. Another approach with this – as we will find out unfortunate – property is polynomial regression, which correponds to the expansion

$$m(x_i) \approx \beta_0 + \beta_1 x_i + \ldots + \beta_r x_i^r,$$

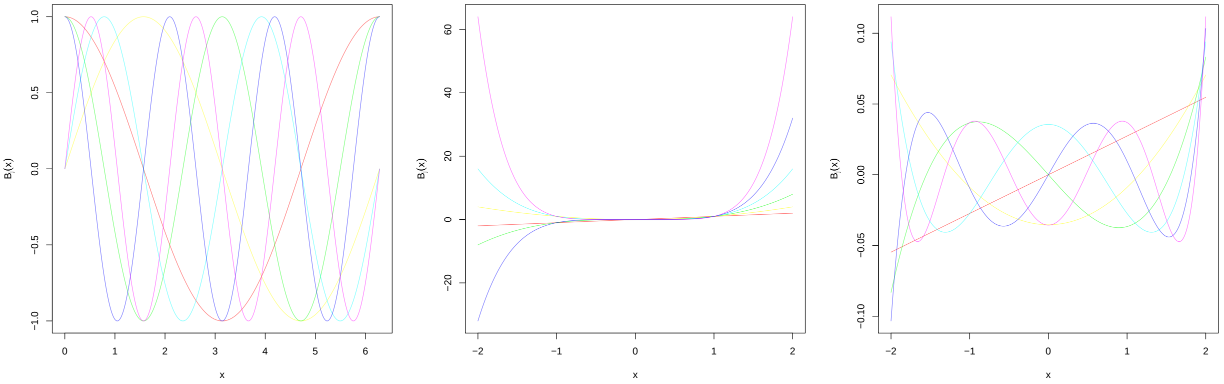
corresponding to the design matrix

$$\mathbf{B} = \begin{pmatrix} 1 & x_1 & \ldots & x_1^r \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \ldots & x_n^r \end{pmatrix}.$$

Instead of using monomials it is numerically more stale to use so-called Tchebychev polynomials (as produced for example by the R function `poly`). Both sets of basis functions are equivalent, i.e. they span the same linear subspace and thus yield identical predictions. Panel (b) below shows the basis functions of monomial basis, where as panel (c) shows the equivalent Tchebychev polynomial basis.

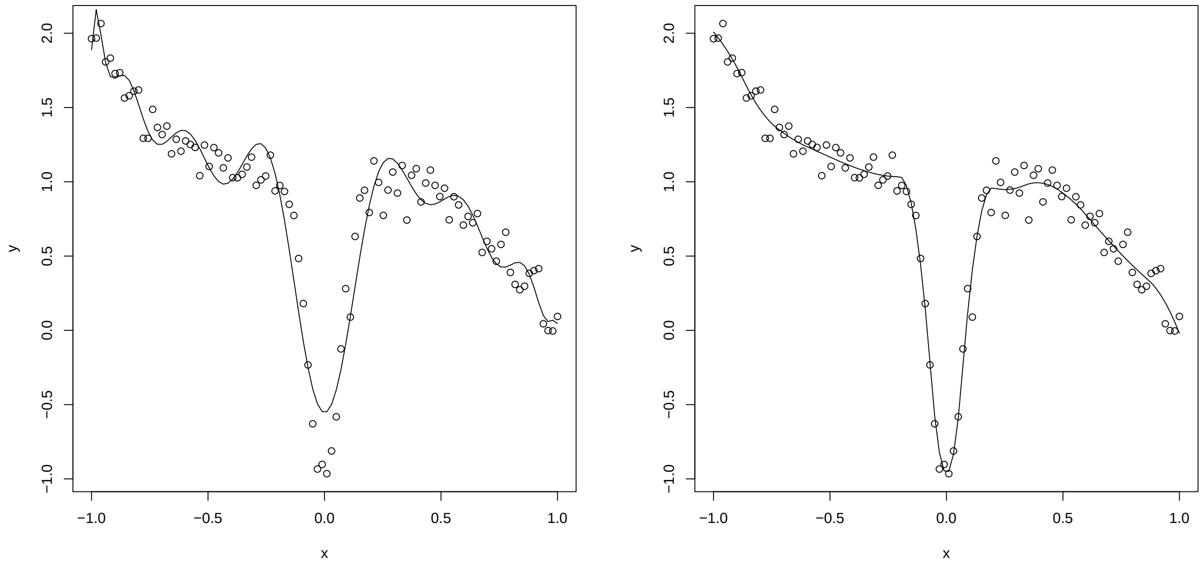
One key advatage of basis-expansion methods is that we can then estimate $\boldsymbol{\beta}$ using the same techniques as used in multiple linear regression, i.e. the least-squares estimator is

$$\hat{\boldsymbol{\beta}} = (\mathbf{B}^\top \mathbf{B})^{-1} \mathbf{B}^\top \mathbf{y}$$



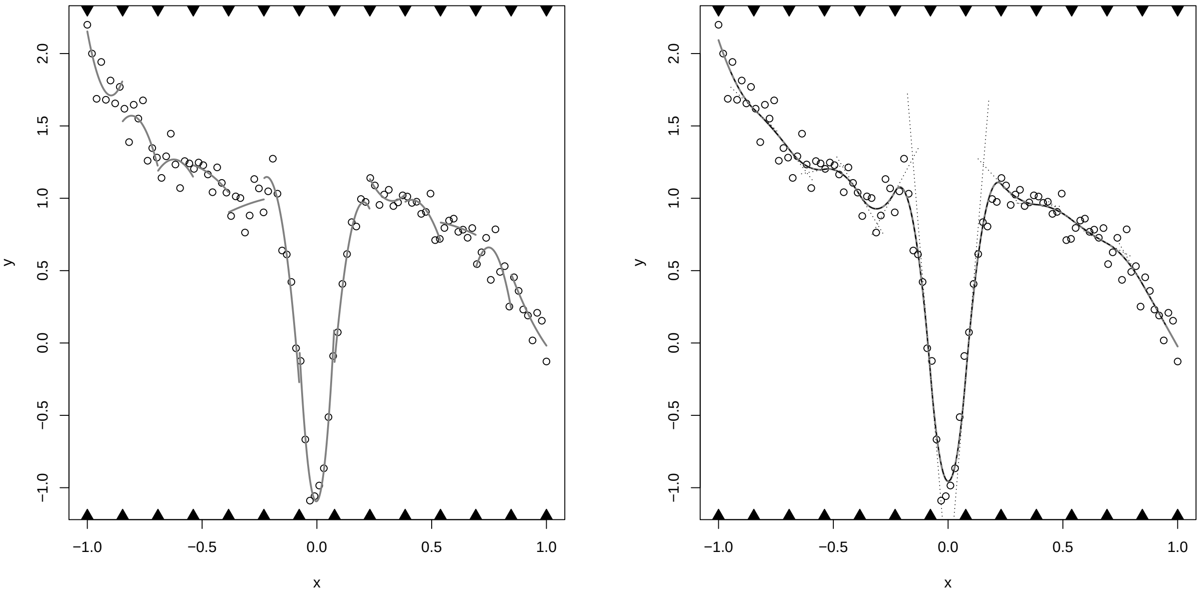
(a) Fourier basis (b) Polynomial basis (monomials) (c) Polynomial basis (Tchebychev)

Polynomial regression can be a useful tool if a polynomial of very low order yields a sufficient fit to the data. However, polynomial regression is not very well suited for modelling more complex relationships. As each basis function acts "globally" rather than "locally", fitting data well in one part of the sample space can create artefacts elsewhere. The figure below shows a polynomial regression model (panel (a)) and a so-called spline-based model (panel (b)) fitted to an artificial toy data set. The polynomial does not capture the sharp dip particularly well and the capturing the dip introduces oscillations which are not supported by the data. The spline model does not suffer from this problem.



(a) Polynomial regression of degree 17 (b) Quadratic-spline-based regression

An alternative approach is to use a set of basis functions which are more local in their effects. Polynomial splines are the most popular such model. Polynomial spline models are based on piecewise polynomial models of low order. As we can see from panel (a) overleaf



(a) Discontinuous piecewise polynomials

(b) Piecewise polynomials which form a continuously differentiable function (derivatives at knots shown as dashed lines)

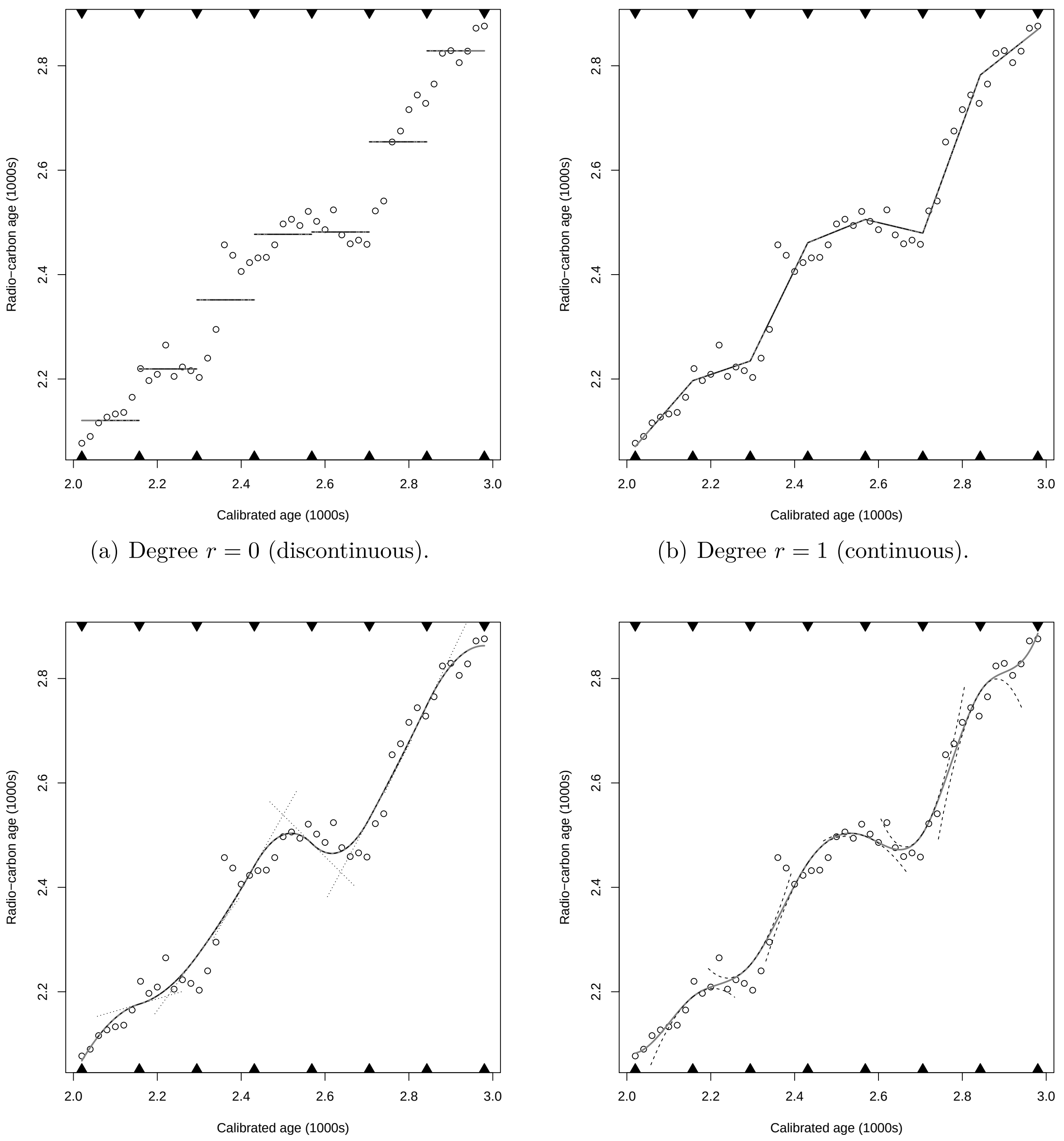
fitting polynomials independently of each other would not yield satisfactory results as the resulting function is discontinuous where the polynomial pieces meet. Thus polynomial splines are based on introducing additional constraints which make the function continuous and (potentially) differentiable (cf. panel (b)).

The key advandtage of splines is that only low-order derivatives are continuous, thus we can avoid the problem of propagating high-order derivariates which created the oscillations in the above figure.

When fitting a polynomial spline model we need to choose the degree $r$ of the polynomials as well as the number $l$ of so-called knots. Knots are the points where two neighbouring polynomials meet, i.e. where higher-order derivatives are discontinuous.

The degree $r$ of the spline controls the smoothness in the sense of controlling its differentiability. For $r = 0$ the spline is a discontinuous step function. For $r = 1$ the spline is a polygonal line. For larger values of $r$ the spline is increasingly smooth, but also behaves more and more like one global polynomial. In practice it is rarely necessary to go beyond $r = 3$. The figure overleaf shows the effect of the degree $r$ for a data set relating radiocarbon age and the calendar age of samples of Irish oak.

The number of knots $l$ acts as a smoothing parameter. The more knots are used, the more flexible the regression function can become. A more flexible regression function has a lower bias, but a higher variance. If number of knots $l$ is too low the regression function struggles to follow the data (high bias). However if $l$ is chosen too large, then the regression function will overadapt to the sample at hand, i.e. also adapt to artefacts in the noise process (high



(a) Degree $r = 0$ (discontinuous).

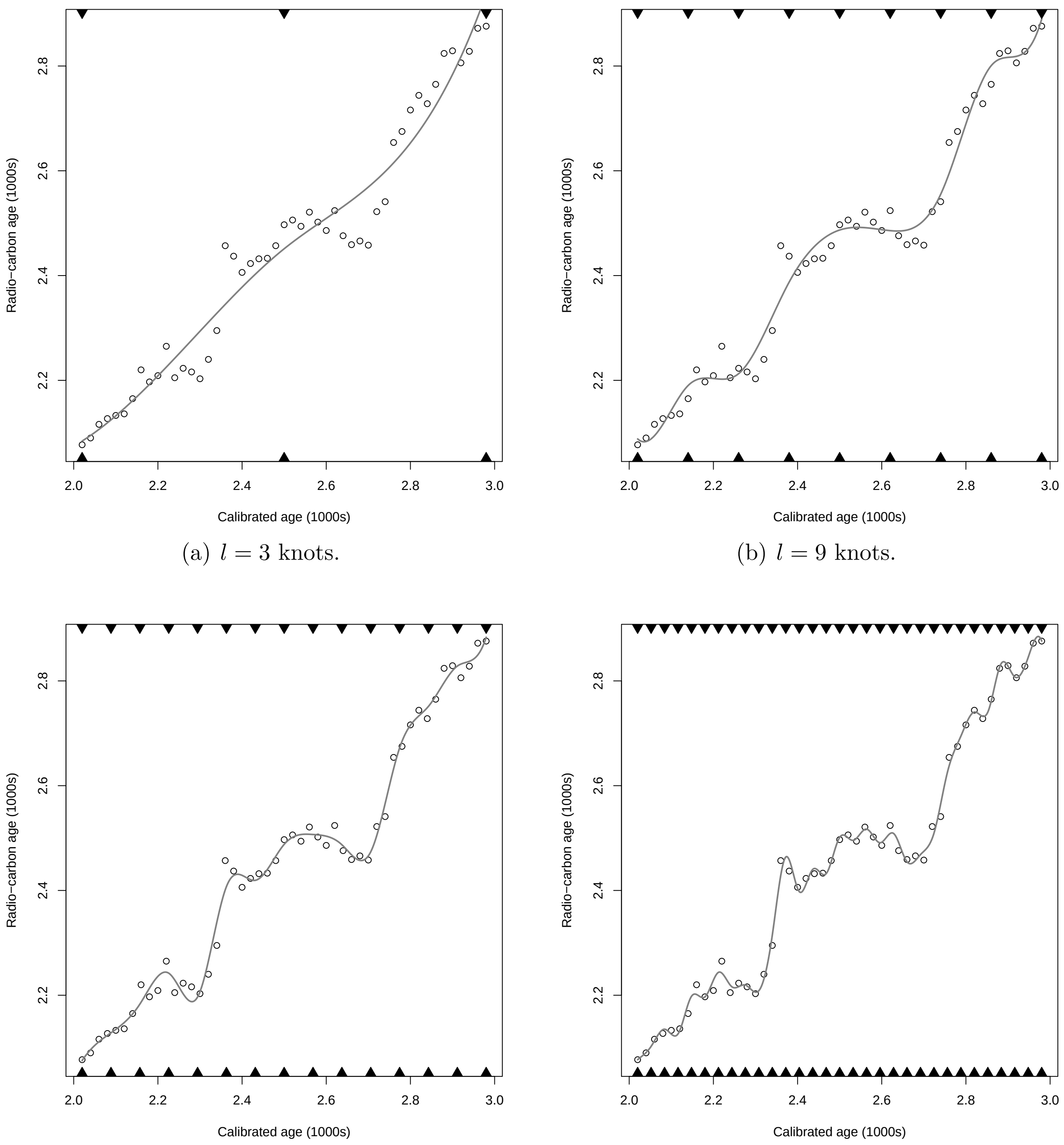
(b) Degree $r = 1$ (continuous).

(c) Degree $r = 2$ (continuous first derivative).

(d) Degree $r = 3$ (continuous second derivative).

variance).[1] The figure on the subsequent page shows the effect of the number of knots of the fitted regression function.

[1] In the course we will also look at P-splines, which introduce a penalty, so that the number of knots does not have to be used as a smoothing parameter.



(a) $l = 3$ knots.

(b) $l = 9$ knots.

(c) $l = 15$ knots.

(d) $l = 31$ knots.

The easiest basis for polynomial splines is the so-called truncated power basis, shown in the figure below. Given a set of knots $\kappa_1 < \ldots < \kappa_l$ the truncated power basis of degree $r$ is based on the expansion

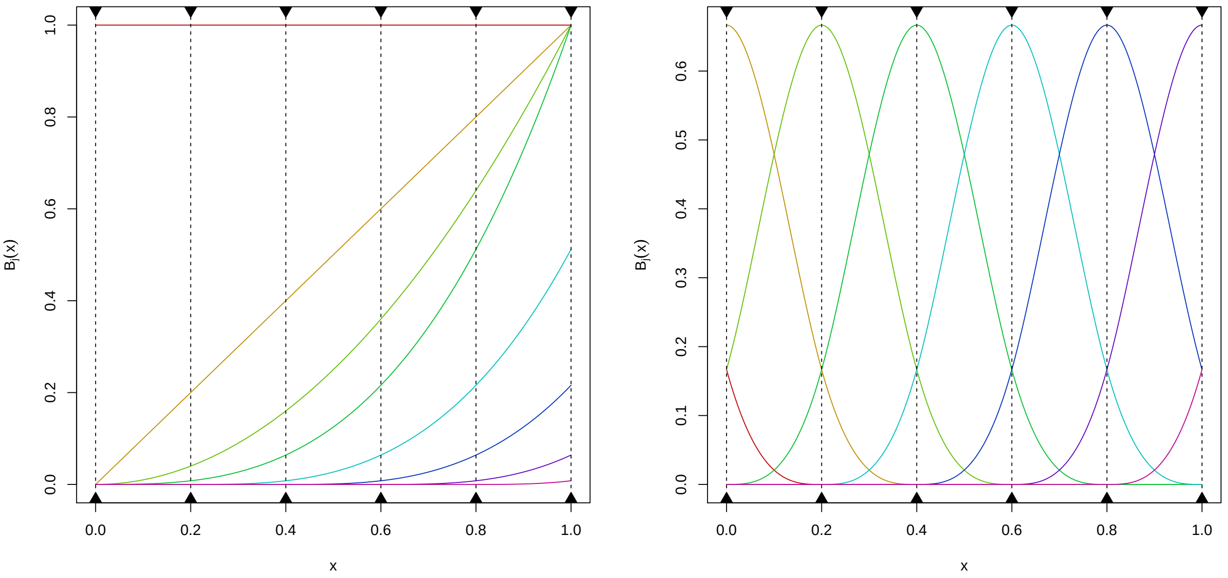
$$m(x_i) \approx \beta_0 + \beta_1 x_i + \ldots + \beta_{r-1} x_i^{r-1} + \beta_r (x_i - \kappa_1)_+^r + \ldots + \beta_{r+l-2} (x_i - \kappa_{l-1})_+^r,$$

where $(z)_+^r = \begin{cases} z^r & \text{for } z > 0 \\ 0 & \text{otherwise.} \end{cases}$

This corresponds to the design matrix

$$\mathbf{B} = \begin{pmatrix} 1 & x_1 & \dots & x_1^{r-1} & (x_1-\kappa_1)_+^r & \dots & (x_1-\kappa_{l-1})_+^r \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \dots & x_n^{r-1} & (x_n-\kappa_1)_+^r & \dots & (x_n-\kappa_{l-1})_+^r \end{pmatrix}$$

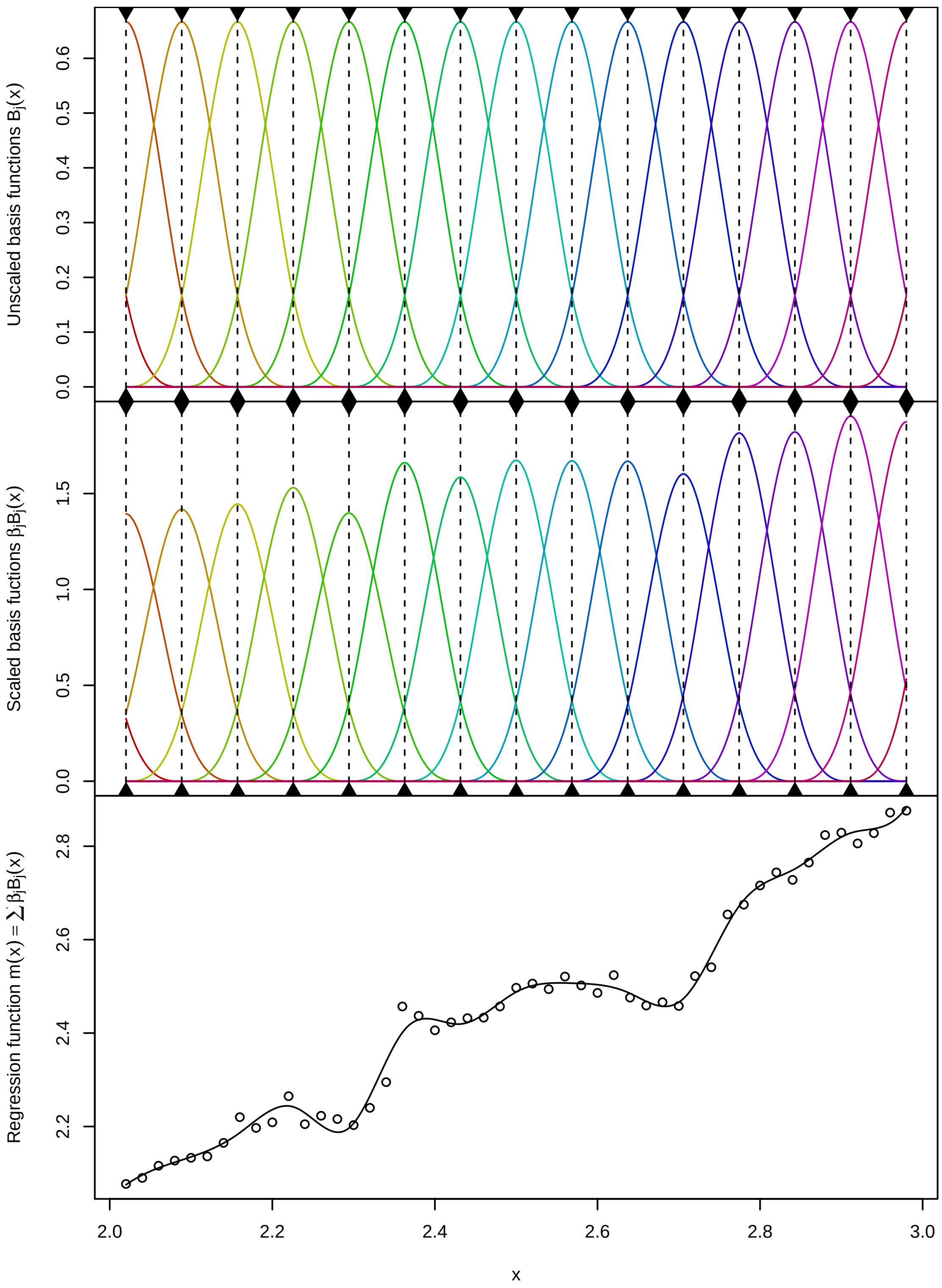
The figure below shows the truncated power basis of degree 3 for six equally spaced knots, together with the so-called B-spline basis. B-splines form an equivalent basis (in the sense of yielding the same fitted function) and are numerically more stable than the truncated power basis. Each basis function in the B-spline basis is non-zero only for $r+1$ intervals, illustrating the local nature of polynomial splines. In the course we will look in more detail at how B-spline bases are constructed.



(a) Truncated power series

(b) B-splines

The figure below shows how a B-spline basis can be used to construct a smooth curve fitting the radiocarbon data from above. The top panel shows the unscaled basis functions $B_j(x)$. The middle panel shows the scaled basis functions $\hat{\beta}_j B_j(x)$. The bottom panel shows a scatter plot of the data together with the fitted function $\hat{f}(x) = \sum_j \hat{\beta}_j B_j(x)$.}



A useful connection can be made here with wavelets, where local basis functions are also used to construct sophisticated function estimators.

## 2.4 Degrees of freedom and standard errors

It is helpful to express the fitted values of the nonparametric regression as

$$\hat{\mathbf{m}} = \mathbf{S}\mathbf{y},$$

where $\hat{\mathbf{m}}$ denotes the vector of fitted values, $\mathbf{S}$ denotes a *smoothing matrix* whose rows consist of the weights appropriate to estimation at each evaluation point, and $\mathbf{y}$ denotes the observed responses in vector form. [2] This linear structure is very helpful.

For example, it gives us a route to defining *degrees of freedom* by analogy with what happens with the usual linear model, where the number of parameters is the trace of the projection matrix. An approximate version of these can be constructed for nonparametric models as

$$\text{df} = \text{tr}\left\{\mathbf{S}\right\}.$$

Similarly, we can construct an estimate of the error variance $\sigma^2$ through the residual sum-of-squares, which in a nonparametric setting is simply

$$\text{RSS} = \sum \{y_i - \hat{m}(x_i)\}^2.$$

This leads to the estimator of the error variance

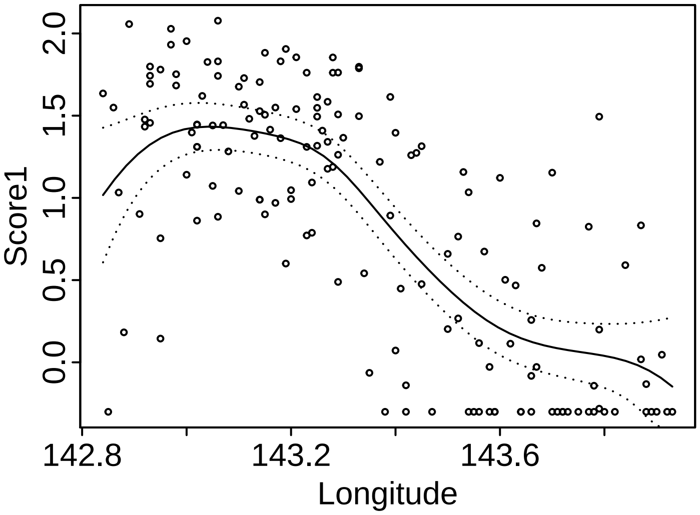
$$\hat{\sigma}^2 = \text{RSS/(n - df)}.$$

The linear structure of the fitted values also makes it very easy to produce standard errors which quantify the variability of the estimate at any value of $x$. Unfortunately, we can't easily produce confidence intervals for the curve because of the bias mentioned above. However, by adding and subtracting two standard errors at each point on the curve we can produce *variability bands* which express the variation in the curve estimate. In fact, we don't need to rely on the asymptotic formula for variance. If $\hat{m}$ denotes the estimated values of $m$ at a set of evaluation points then

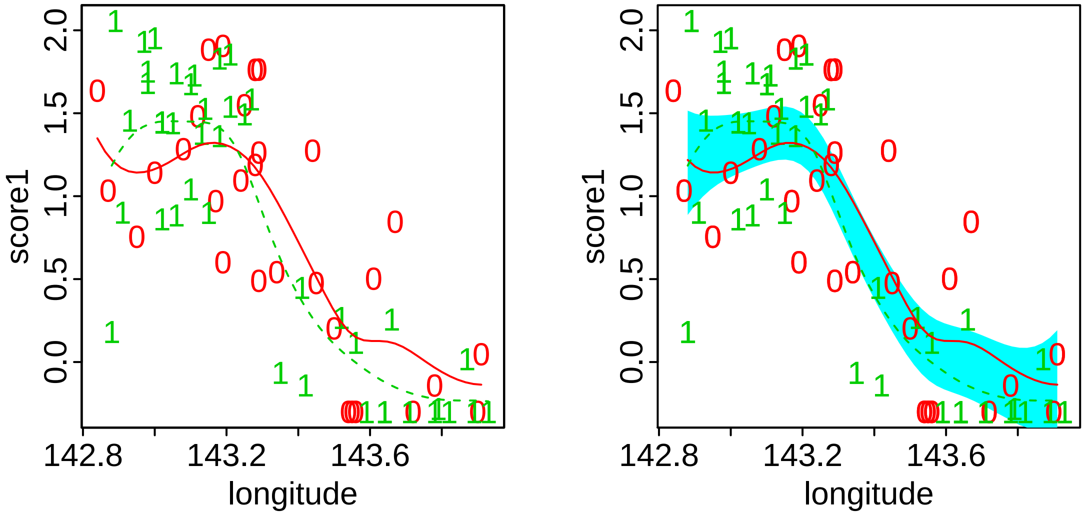
$$\mathsf{var}\{\hat{\mathbf{m}}\} = \mathsf{var}\{\mathbf{S}\mathbf{y}\} = \mathbf{S}\mathbf{S}^\top\sigma^2$$

and so, by plugging in $\hat{\sigma}^2$ and taking the square root of the diagonal elements, the standard errors at each evaluation point are easily constructed. The plot below illustrates this on the Reef data.

[2] For a model based on basis expansions $\mathbf{S}$ is calculated in the same way as the hat matrix of the linear model, $\mathbf{S} = \mathbf{B}(\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{B}^\top\mathbf{y}$ with the design matrix $\mathbf{B}$ containing the evaluated basis functions.



Sometime we want to compare curves, at least informally, and we can use the standard errors from each curve to do that. At any point $x$, the standard error of the difference between the curves is $se_d(x) = \sqrt{se_1(x)^2 + se_2(x)^2}$, where $se_1(x)$ and $se_2(x)$ denote the standard errors of each curve at that point. A neat trick is to plot a band whose width is $2se_d(x)$. By centring this band at the average of the two curves we can see where they are more than two standard errors apart. We can see this with the Reef data, for a single year, separated into two groups corresponding to open and closed fishing zones.



## 2.5 How much to smooth

One of the key questions with nonparametric models is how much smoothing to apply to the data. For exploratory work, it can often be helpful simply to experiment with different degrees of smoothing. One appealing way to do that is to specify how many *degrees of freedom* (see discussion above) you would like to have. This puts things on a natural scale.

However, in more complicated situations that can be difficult and it is helpful to have an automatic way of producing a suitable level of smoothing. There are several ways to do this, some of which are carefully tailored to particular models. Here we will outline a method called *cross-validation* which, although it has some difficulties, has the advantage that the generality of its definition allows it to be applied to quite a wide variety of settings. In the present setting, the idea is to choose $h$ to minimise

$$\text{CV:} \quad \sum_{i=1}^{n} \{y_i - \hat{m}_{-i}(x_i)\}^2.$$

The subscript $-i$ denotes that the estimate of the smooth curve at $x_i$ is constructed from the remainder of the data, excluding $x_i$. The aim then is to evaluate the level of smoothing through the extent to which each observation is predicted from the smooth curve produced by the rest of the data. The value of $h$ which minimises the expression above should provide a suitable level of smoothing.

It is often convenient to use an approximation known as *generalised cross-validation* (GCV) which has the efficient computational form

$$\text{GCV:} \quad n\text{RSS}/\{\text{tr}\,\{I - S\}^2\}.$$

The degree of smoothing can also be selected automatically by minimising a quantity based on *Akaike's information criterion*, namely

$$\text{AIC:} \quad \frac{\text{RSS}}{n} + 1 + \frac{2(\nu + 1)}{(n - \nu - 2)}.$$

# 3 Broad concepts and issues of smoothing

The simple cases of density estimation and nonparametric regression highlight features and issues which are common to a wide range of problems involving the estimation of functions, relationships or patterns which are *nonparametric* but *smooth*. The term nonparametric is used in this context to mean that the relationships or patterns of interest cannot be expressed in specific formulae which involved a fixed number of unknown parameters. This means that the parameter space is the space of functions, whose dimensionality is infinite. This takes us outside of the standard framework for parametric models and the main theme of the course will be to discuss how we can do this while producing tools which are highly effective for modelling and analysing data from a wide variety of contexts and exhibiting a wide variety of structures.

On a side note, the term nonparametric is sometimes used in the narrower setting of simple statistical methods based on the ranks of the data, rather than the original measurements. This is not the sense in which it will be used here.

Further details on density estimation and nonparametric regression and available in the books listed below.

The issues raised by our brief discussion in this preliminary material include

- how to construct estimators which match the type of data we are dealing with;
- how to find a suitable balance between being faithful to the observed data and incorporating the underlying regularity or smoothness which we believe to be present;
- how to construct and make use of suitable inferential tools which will allow the models to weigh the evidence for effects of interest, in a setting which takes us outside of standard parametric methods.

These broad issues will be explored in a variety of contexts in the remainder of the course.

# 4 Further reading

**Classic texts on density estimation:**

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman & Hall, London.

Scott, D.W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization.* John Wiley, New York.

**A variety of texts on flexible regression:**

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models.* Chapman & Hall: London.

Wand, M.~P. and Jones, M.~C. (1995). *Kernel smoothing.* Chapman and Hall, London.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications.* Chapman and Hall: London.

Simonoff, J.~S. (19916). *Smoothing methods in statistics.* Springer, New York.

Bowman, ~A.W. & Azzalini, ~A. (1997). *Applied Smoothing Techniques for Data Analysis.* OUP: Oxford.

Ruppert, D., Wand, M.P. & Carroll, R.J. (2003). *Semiparametric Regression.* CUP: Cambridge.

Wood, S. (2006). *Generalized additive models: an introduction with* R. Chapman & Hall, London.