APTS Course
Nonparametric Smoothing
How to be flexible

September 2013

Practical II

Adrian Bowman & Ludger Evers

The sm and mgcv packages in R

The sm package, written by Adrian Bowman and Adelchi Azzalini, has been used at various points in the lectures. There are two main functions.

```
sm.density(y)
sm.regression(x, y)
```

to implement density estimation (where y is a vector or a matrix with 2 or 3 columns) and nonparametric regression (where x is a vector or a matrix with 2 columns and y is a vector of responses). Nonparametric regression is implemented by local linear estimation. In both functions, the argument panel = TRUE creates a control panel from which various aspects of the estimation process can be controlled interactively.

The mgcv package, written by Simon Wood, is a powerful set of tools for fitting and analysing generalised additive models. The book, Wood (2006), referenced in the lecture notes, makes extensive reference to Rand the mvcv package so this is a good way of learning both about additive models and how to use them in practice. The principal function is mgcv which is called, for example, as

```
model <- gam(y ~ s(x1) + s(x2) + s(x3))
```

where the vector of responses y is modelled through a set of smooth functions in x1, x2 and x3. The s(x1) notation identifies that a smooth function should be fitted. Linear terms can be fitted in the usual way simply through the variable name. The syntax and operation of the gam function follows that of lm and vlm as closely as possible.

In the exercises below, scripts are provided to get you started on each problem, and in some cases to lead you through. Of course, feel free to experiment. There should be someone to call on if you get stuck.

Getting started

Please log into the computer using the credentials given to you in the delegate pack. Please note that each account is specific to one computer.

In this lab you should start by loading a prepared script found at

```
p:/data/practical-two.r
```

Cartoons

In Lectures 1 and 2, some `rpanel` 'cartoons' were used to illustrate some concepts and explore some datasets. These are all available and you may like to experiment with them a little. The first line of the script give you a command which will launch a menu from which the cartoons can be selected.

The Clyde data

In lecture 4, we looked at the relationship of DO to Temperature, Salinity and Year at a single sampling station in the River Clyde. Repeat this analysis at one or two other stations, to see whether the same features are apparent at there locations.

The mackerel egg survey

A multi-country survey of mackerel eggs was carried out in the Eastern Atlantic in 1992. The aim was to estimate the total biomass of spawning mackerel. However, a first step is to construct a model which describes the pattern of egg counts in the water samples collected.

Use the script to try the following operations.

1. Plot the sampling region
2. Explore the (marginal) relationship between the egg density (log scale) and latitude, longitude, depth (log scale) and temperature.
3. Fit and interpret a simple additive model
4. Consider whether the Temperature variable is really needed.

Can you identify the preferred depth at which mackerel lay their eggs?

The data collected by Spanish vessels differs from that of most other countries, in that it is largely in the form of presence or absence of eggs in each water sample. The script will lead you through some plots of this. Also use the script to try the following operations.

1. Explore the (marginal) effect of depth and temperature. In particular, is there any difference from the earlier dataset in the indication of preferred depth at which mackerel lay their eggs?
2. Fit and interpret an additive model

Water quality in Loch Leven

Loch Leven is situated in lowland Scotland in the Perth and Kinross area. It is the largest shallow lake in Great Britain with an area of 13.3km^2, mean depth 3.9m and a maximum depth 25.5m. There are approximately 150 variables measured at Loch Leven (chemical, physical, biological and meteorological) and most of this monitoring is carried out by the Centre for Ecology & Hydrology in Edinburgh.

One of the features of interest is the water quality and hence the relationship between chlorophyll$_a$ (as an indicator of water quality) and Soluble Reactive Phosphorus (a nutrient) is very important. This case study explores this relationship.

The data provided are the natural logarithm of the monthly means for chlorophyll$_a$ (`lchla`) and SRP (`lsrp`) from January 1988 to December 2007. Natural log transforms of the data are used to stabilize the variance and there are some missing values. Columns of data for year and month are also provided.

The script contains commands to investigate the following questions.

1. Plot the data to examine the relationship between `lsrp` and `lchla`.
2. Is the relationship between `lsrp` and `lchla` affected by month?
3. A possible model is a *varying coefficient* model. This fits a linear regression between the two variables but allows the parameters of the regression to change smoothly with time of year. This can be fitted by smoothing over `month` and `lchla`, with a very large smoothing parameter for `month` to make this term linear.

SO_2 over Europe

In the 1970's and 1980's there was considerable concern about SO2 air pollution. This was emitted by power stations and other installations and the material rises high in the atmosphere and can travel long distances, causing pollution problems in neighbouring countries. The `SO2` dataset documents values of SO_2, on a log scale, from monitoring stations across Europe from 1990 to 2001. The aim of the monitoring stations was to assess whether increasing European regulatory control of SO_2 emissions was effective.

The data were collected through the *European monitoring and evaluation programme* (EMEP) and they are available at `www.emep.int`. The data recorded here have been organised into a convenient form for analysis. The data file consists of six variables:

`site`	a site code for the monitoring station
`longitude`	longitude of the monitoring station
`latitude`	latitude of the monitoring station
`year`	year of measurement
`month`	month of measurement
`logSO2`	SO2 measurement on a log scale

Here are some things for you to consider.

1. The script gives commands which organise and plot the data over space and time. See whether you can identify spatial and temporal patterns from this.
2. One of the roles of a model is to clarify the nature and size of different effects. The script shows how to fit and plot a model which is additive in space and time. What do the results show?
3. A more realistic model would allow space-time interaction. A command is given to fit this (which may take a little time). Use the earlier code to plot this new model and consider the difference it makes.

Analysis of these data is reported in *Spatiotemporal smoothing and sulphur dioxide trends over Europe*, A. W. Bowman, M. Giannitrapani and E. M. Scott; *Applied Statistics*, 58 (2009), 737-752.