Splines, Knots and Penalties. The Craft of Smoothing

Computer Lab 3

Introduction. The computer labs make you familiar with the practical use of P-splines. There are three labs. The first one shows basic use of P-splines, staying very close to the formulas in the course. The second lab makes uses of a number of complete functions for smoothing of normal and non-normal (counts, binomial) data. The third lab makes use of the powerful *mgcv* library, written by Simon Wood. This is a well-documented and powerful R package; it does not provide P-splines by default, but the documentation shows how to add them.

Tools. The basic tool is the R system. Almost any version will work, but it makes sense to use a recent version. The computer labs were tested with R version 2.5. For computer labs 1 and 2 no additional packages (libraries) are needed. For computer lab 3 the *mgcv* package is needed. At the time of writing these instructions, no Windows binary for R version 2.5 was available, but the one for version 2.4 installs without complaints (first copy the ZIP file to your computer and install the package from this archive). If you work on one of the computers at the University, the software should already be installed.

Software files. The files to be used in the labs will be put in a place on the network. There will also be a USB memory stick that can be passed along. Some files contain functions that you will use. The other files are labeled 'computer_lab1.r' to 'computer_lab3.r'. They contain step-by-step instructions for you to give to R. Make a copy of the files, so that you can change instructions freely, without loosing the original instructions.

How to work. If you are lazy, you copy individual instructions or blocks of instructions to R and see what happens. However, in the initial phase it is advisable to type the instructions yourself. It is better way to learn and to remember. Start R and change to directory in which you stored the files.

Comments. The instruction files contain many comments to document what is going on. A large number of comments contain section numbers, like S1 and S9. These are meant as references for the explanations that follow below.

1. First some scattered data, normally distributed around a trend, are simulated.
2. The *mgcv* library provides the general function *gam()* for fitting of generalized additive models (GAM). But it also works fine with only one independent variable. We specify the formula *y ~ s(x)* to indicate that we model *y* by a smooth function of *x*. With *s(x)* we indicate that the default smoother, the thin plate spline should be used. Notice that *mgcv* automatically optimizes the amount of smoothing. We ask for a summary to be printed on the R console.
3. You have to do some extra work to get a smooth curve on a grid. In a new data frame the *x* grid is placed and the *predict()* is called.
4. To get standard error bands, specify *se = T.* Now a list with two components is returned.
5. In the documentation of *mgcv,* under the heading *smooth.construct*, it is shown how to add P-splines to *mgcv*. The instructions have been copied to the file *ps_construct.r*, which we read in

now. Actually the difference between thin plate splines and P-splines is small. They both uses basis functions and a roughness penalty.

6. Smooth the data with P-splines and take a look at the summary.
7. Smoothing of the ozone data.
8. There is an easier way to get a plot, but it has certain drawbacks. The ozone concentrations have been standardized, so interpretation is less easy. Also the error bands are narrower. They are conditional error bands, conditional on the mean of ozone. So the standard error of that mean is not accounted for. This makes sense in a GAM setting.
9. When fitting a GAM, the power of *mgcv* will be appreciated better. Hit the return key (while in the R console) to get two plots.
10. You can also get two plots on one page.