Prof. Dr. W. Zucchini
Applied Smoothing Techniques *Assignment 7 - WS 2003/04*

Part 1: Theory

Consider the following two optimization problems:

Interpolation problem: Given data points $(x_1, y_1), , (x_n, y_n)$ (with the x_i unique and increasing), the function minimizing

$$\int_{x_1}^{x_n} \{f''(t)\}^2 dt$$

subject to $f(x_i) = y_i$, $i = 1, , n$, is a natural cubic spline with knots at the values of x_i. This is known as the solution to the *interpolation problem.*

Smoothing problem: Among all functions $f(x)$ with two continous derivatives, find the one that minimizes the penalized residual sum of squares

$$\sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda \int_a^b \{f''(t)\}^2 dt$$

where λ is a fixed constant, and $a \leq x_1 \leq \cdots \leq x_n \leq b$.

Use the result of the interpolation problem to provide an heuristic argument to show that any solution to the smoothing problem must be a cubic spline.

Part 2: Practical

1. Some useful **R** functions.

   Figure out what the following functions do and how to use them: `approxfun, jitter, gam` (in the library `mgcv`), `smooth.spline` (in the library `modreg`). Browse through the help documentation for the function `par`.

2. Fitting GAMS using the library `mgcv`.

   Load the data `Assignment07data.txt`, which contains a 100-row by 3-column table of observations. The first columns refers to the response variable, y, and second and third to two covariates, x_1 and x_2, respectively.

   Load the library `mgcv` and use the function `gam` to fit a GAM that models the conditional expectation of y as the sum of smooth functions of x_1 and x_2, respectively.

Initially use the degrees of freedom (df) estimated by the method of cross-validation (which is used by default when one applies `gam`). After that experiment with the dfs. Finally, find out how to draw a three-dimensioanl display of the regression of y on x_1 and x_2.

3. Applying fack-fitting to fit GAMS.

   The object of this exercise is for you to learn how to estimate the smooth functions in a GAM "by hand" using the method of back-fitting. Use the data set that you loaded to answer question 2. Use the function `smooth.spline` in the library `modreg` to compute the smoothing splines. Carry out the following steps:

   - Specify the degrees of freedom to be used for f_1 and f_2
   - Standardize the y vector: subtract the mean, $\hat{\beta}$.
   - Set $f_1 = f_2 = 0$.
   - Iterate:
     Fit a smooth spline $f_1(x)$ to the residuals of $y - (\hat{\beta} + f_2(x_{2i})$,
     fit a smooth spline $f_2(x)$to the residuals of $y - (\hat{\beta} + f_1(x_{1i})$.
     Compute the residual sum of squares and continue to iterate until this sum convergers. (Hint: Using the function `approxfun` can simplify the programing.)

4. An application og GAMs.

   The data set `trees` in the library `sm` contains measurements on the volume, girth and height of a sample of trees. Fit a suitable GAM to model the volume as a function of the girth and height.