Prof. Dr. W. Zucchini
Applied Smoothing Techniques *Assignment 6 - WS 2003/04*

Part 1: Theory

1. Consider the spline with knots $\xi_1, \xi_2, ..., \xi_K$ given by

$$m(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{k=1}^{K} \theta_k (x - \xi_k)_+^3 \ ,$$

where the notation a_+ is defined as $a_+ = \begin{cases} a & \text{if} \quad a > 0 \\ 0 & \text{if} \quad a \le 0 \end{cases}$

Show that $m(x)$ has the properties

a) m is a cubic polynomial in each subinterval $[\xi_k, \xi_{k+1})$, $k = 1, 2, ..., K-1$.

b) m has two continuous derivatives.

c) The third derivative of m is a piecewise constant function with jumps at the knots.

2. Semi–parametric regression: Suppose that we have a set of n observations (with p predictors) arranged in a $n \times p$ matrix X, an additional covariate z, and a response vector y. Denote the i-th row of X by $x^{(i)}$. We wish to fit a model

$$y_i = x^{(i)}\beta + f(z_i) + e_i \ , \quad i = 1, 2, ..., n$$

using penalized least squares (as one uses in the context of splines).

Construct the appropriate penalized residual sum of squares, and show that the minimizers satisfy the following pair of "estimating equations".

$\hat{\beta} = (X^T X)^{-1} X^T (y - \hat{f})$

$\hat{f} = \mathcal{S}(y - X\hat{\beta})|z)$

where $\mathcal{S}$ is the spline smoothing operator, and f represents the function f evaluating at the n values z_i.

Part 2: Practical

1. a) The function `complete.cases` is usefull when dealing with incomplete data sets. Figure out what the function does and how to use it. (You will need it Question 3.)

b) Load the R library `modreg` and figure out how to use the function `smooth.spline` [Hint: Run the examples given in the help documentation for that function.]

2. a) Write an R function `genmod` that generates observations from the following model

$$y_i = m(x_i) + e_i \ , \quad i = 1, 2, ..., n$$

where $x_i \overset{iid}{\sim} U(0, 100)$, $e_i \overset{iid}{\sim} N(0, \sigma^2)$ and $m(x) = \alpha \cos\left(\frac{2\pi x}{100}\right) + \beta \sin\left(\frac{2\pi x}{100}\right)$ for given parameters σ^2, α, β.

b) Generate $n = 100$ observations from this model with $\alpha = 10$, $\beta = 5$, $\sigma^2 = 1$.

c) Plot the observations obtained in b) and fit cubic smoothing splines to these using different degrees of freedom. Plot the fitted splines and $m(x)$.

d) Repeat b) and c) for different parameter values. The point is to figure out the effect on the fit when one changes the degrees of freedom.

3. The file `corruption.dat` contains the *Transparency International Corruption Perception Index (CPI)* and other economic indicators for 106 countries. Read the data with: `X<-read.table("<path>corruption.dat")`.

Focus on the relationship between CPI (1st column of `X`) and GDP/pop. (4th column): Examine the relationship between *CPI* and *GDP/pop* with kernel regression using different bandwidths (using the function `sm.regression` in the R library `sm`) and with smoothing splines with different degrees of freedom (using the function `smooth.spline` in the R library `modreg`).