Prof. Dr. W. Zucchini
Applied Smoothing Techniques — *Assignment 3 - WS 2003/04*

Part 1: Theory

(a) Show that the minimum value of $\text{MISE}(\hat{f})$ is given by

$$\text{MISE}_{opt}(\hat{f}) = \frac{5}{4}\left(\frac{\beta(f)j_2^4k_2^2}{n^4}\right)^{1/5}$$

(b) Show that

i. for a rectangular Kernel $k_2 = \frac{1}{3}$ and $j_2 = \frac{1}{2}$

ii. for an Epanechnikov kernel $k_2 = 1$ and $j_2 = \frac{3}{5\sqrt{5}}$

iii. the efficiency of the rectangular kernel is $\left(\frac{108}{125}\right)^{1/2} \approx 93\%$.

Part 2: Practical

(a) The function `dmnorm( )` given below computes the mixture of m normal distributions.

```
dmnorm <- function(x,alpha,mean,sd)
{
m<-length(alpha)
pdf<-rep(0,length(x))
for(i in 1:m)
    {
    pdf<-pdf+alpha[i]*dnorm(x,mean[i],sd[i])
    }
return(pdf)
}
```

i. Examine the function in detail to figure out how it works and plot the density for the following parameter values:

$\alpha_i = (0.4, 0.3, 0.3)$; $\mu_i = (10, 20, 30)$; $\sigma_i = (2, 3, 2)$

ii. Write a R function `rmnorm( )` to generate n realizations from the above mixture of m normal distributions and test it by generating a random sample of size $n = 1000$ and comparing the resulting histogram with the pdf.

(b) Generate $n = 50$ observations using `rmnorm( )` with the above parameters. Load the library `sm` and estimate the density using `sm.density( )` and

- the normal-based optimal bandwidth (`hnorm`)
- the cross-validation bandwidth (`hvc`)
- the Shealter-Jones plug-in bandwidth (`hsj`)

In each case compare the estimate with the (known) pdf!

(c) Repeat (b) for $n = 20$ and $n = 100$.
What conslusion can you draw from these experiments?

(d) Load the data in the file `tree.dat`. This contains measurements of diameter (column 1) and height (column 2) for a population of trees. Extract the heights (using the `attach` command) and draw a random sample of size 100 heights. Estimate the pdf of the height from the sample using each of the three bandwidths listed in part (b). Campare these with a histogram of the population of heights. (Note if you try to compute the bandwidth estimates for the entire population you will probably have to wait for several hours to obtain the answer.)