SDA 2021 — Assignment 3

For these exercises you can use the functions `CV` and `h_opt` (in the file "functions_Ch4.txt") and the function `bootstrap` (in the file "functions_Ch5.txt") on the Canvas page. Investigate these functions before you use them. Anyhow, here is some preliminary information on the functions `h_opt` and `CV`:

`h_opt` uses formula (4.3) in the syllabus. It is based on the sample standard deviation as an estimator of the standard deviation and, to find some hopefully reasonable value for the involved integral $\int (f'')^2(t)dt$ the normal location scale family has been used.

Note: for any location scale family $\{G_{\mu,\sigma} : \mu \in \mathbb{R}, \sigma > 0\}$ w.r.t. the distribution function G with density g, we have $g_{\mu,\sigma}(t) = \frac{d}{dt}G_{\mu,\sigma}(t) = \frac{d}{dt}G(\frac{t-\mu}{\sigma}) = g(\frac{t-\mu}{\sigma})/\sigma$. Thus, one can show that $\int (g''_{\mu,\sigma})^2(t)dt = \int (g'')^2(t)dt/\sigma^5$. The following table is a list of values of $\int (g''_{\mu,\sigma})^2(t)dt$ for different unimodal densities g with variance 1:

density	$g(t)$	$\int (g'')^2(t)dt$
standard normal	$(2\pi)^{-1/2}\exp(-t^2/2)$	$\frac{3}{8}\pi^{-1/2}$
logistic (with scale parameter $s = \sqrt{3}/\pi$)	$\frac{\exp(-t/s)}{s(1+\exp(-t/s))^2}$	$\pi^5\frac{13}{3^{7/2}\cdot 35}$
Double exponential (with scale parameter $b = 1/\sqrt{2}$)	$(2b)^{-1}\exp(-\frac{\lvert t-\mu\rvert}{b})$	$\sqrt{2}$
exponential (with rate parameter $\lambda = 1$)	$\lambda\exp(-\lambda t)1\{t > 0\}$	0.5

The function `CV` computes for a given bandwidth, sample, and kernel the cross-validation criterion $\hat{R}(\hat{f})$. Thus, in order to find the minimizing bandwidth value, you should apply `CV` to multiple bandwidths and then select the one that led to the smallest value of $\hat{R}(\hat{f})$. In R this can be achieved via, e.g.

```
cv_crit <- sapply(h_vec, CV, sample=sample, kernel="gauss")
h_min <- h_vec[which(cv_crit == min(cv_crit))]
```

where `h_vec` is a vector of bandwidths for each of which the cross-validation criterion shall be computed.

Kernel density estimators in R can be obtained by using the function `density`. Use the `help`-function to find out how to use this function.

Make a concise report of *all* your answers in *one single PDF file*, with only *relevant R* code *in an appendix*. It is important to make clear in your answers how you have solved the questions. Graphs should look neat (label the axes, give titles, use correct dimensions etc.). Multiple graphs can be put into one figure using the command `par(mfrow=c(k,l))`, see `help(par)`. Sometimes there might be additional information on what exactly has to be handed in.
Read the file AssignmentFormat.pdf on Canvas carefully.

Exercise 3.1 The file `sample31.txt` contains a sample of $n = 100$ observations. Assume that the true density is bimodal. By varying the bandwidth h, use a trial and error approach to find a kernel density estimator that matches a (suitable) histogram of the data quite well. Conclude what the obtained value of h tells you about the value of $\int (f'')^2(t)dt$.

Hint: consider the formulas on pp. 43 and 45 of the syllabus and suppose that the bandwidth you have found is "optimal": $h = h_{opt}$.

Warning: normally, i.e. in other exercises, one should approach density estimation in a somewhat systematic way. It is good though to compare the resulting kernel density estimators with histograms and possibly re-think the estimation strategy.

Exercise 3.2 The file `sample32.txt` contains a sample of $n = 60$ *positive* observations. Find a suitable kernel density estimate based on this sample.

Hint: it seems appropriate to assign no mass to $\hat{f}(x)$ *for* $x < 0$. *Take a look at Lecture 4 to find out how this can be achieved in a reasonable way.*

Hint: if you would like to use the log*-transformation and first find a suitable kernel density estimate* $\hat{f}_y$ *based on the* log*-transformed sample* $\mathbf{y}$, *i.e.* $y_1 = \log(x_1), ..., y_n = \log(x_n)$, *you can obtain the density estimate* $\hat{f}_x$ *for the original sample based on*

```
yrange <- seq(min(y), max(y), length.out=512)
lines(exp(yrange), density(y, ..., from=min(yrange) , to=max(yrange))$y/exp(yrange))
```

Exercise 3.3 The file `sample33.txt` contains a sample of $n = 110$ observations. Find two kernel density estimates based on this sample: for the first, use the bandwidth obtained from the function `h_opt`, for the second, use the bandwidth obtained from the cross-validation criterion. Compare these estimates to the true density function, which is a gamma density (with shape parameter 3 and scale parameter 0.4). Argue which kernel density estimate seems preferable.

Tip: first explore the functions `h_opt` & `CV`.

Hand in for Exercises 3.1–3.3: answers to the questions, your estimation strategy, and relevant plots that helped you to find the final kernel density estimates. Motivate your choices of kernel functions, bandwidths, and transformations, if any are used.

Exercise 3.4 One sample drawn from a t-distribution with unknown degrees of freedom $k > 0$ is stored in the file `t-sample.txt`. With the help of this sample, we would like to estimate the distribution of the sample (excess) kurtosis statistic.[1]

a. Write a function `kurtosis(x)` that returns the sample excess kurtosis of a sample `x`. Use it to estimate the unknown distribution's excess kurtosis.
NB. if you found an R package that offers such a function, you may use it instead of programming it from scratch. In this case, report which library you used and what the function's name is; this should also be apparent from your R code in the appendix.

b. Use the empirical bootstrap method applied to our t-sample to generate $B = 5000$ bootstrap estimates of the excess kurtosis statistic. Store these in a vector `empBS` in your R environment.

c. Repeat the steps of part b. but with the parametric bootstrap instead of the empirical bootstrap. Use $\hat{k} = 2s^2/(s^2 - 1)$ as an estimator of the degrees of freedom k where s^2 denotes the sample variance. Call the vector which containes the obtained bootstrap statistics `parBS`.

d. Plot two separate histograms of the bootstrap samples obtained in b. and c.
Compare them to another histogram for the true distribution of the excess kurtosis statistic. One can obtain this in the following way: Generate 5000 independent samples of size 120 from the t-distribution with 10 degrees of freedom (which is the true underlying distribution by the way!), and compute the excess kurtosis for each sample; then plot their histogram.
Based on these comparisons, which bootstrap method seems preferable in the present context? Motivate your answer.

e. Use the empirical and the parametric bootstrap samples to find estimates of the variance of the excess kurtosis statistic. Compare these two estimates with an approximation of the true variance of that statistic, which could be obtained as the sample variance of the realizations from part d.

Hand in: answers to the questions and relevant plots.

[1] The sample excess kurtosis for a given sample $x_1, \ldots, x_n$ is given by $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4/s^4 - 3$, where $\bar{x}$ is the sample mean and s^2 is the sample variance. It can be used to get an impression of the heaviness of the tails of the data distribution.