Exercise 4

Philipp Drebes
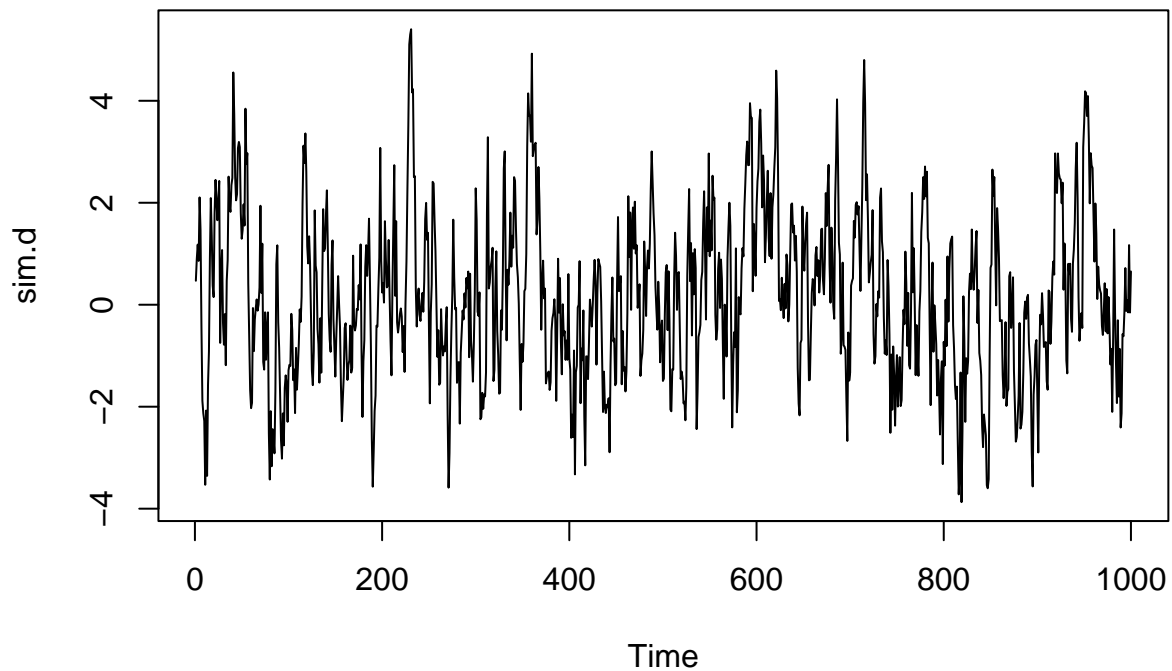
22.03.2023

Exercise 4.1

In this exercise, we would like to investigate the properties of an AR(1) process.

    a) Simulate a realization of the process $X_t = 0.8 \cdot X_{t-1} + e_t$ with e_t an innovation process of length 1000.

```
set.seed(99999)
sim.d <- arima.sim(list(ar = c(0.8)), n = 1000)
plot(sim.d)
```
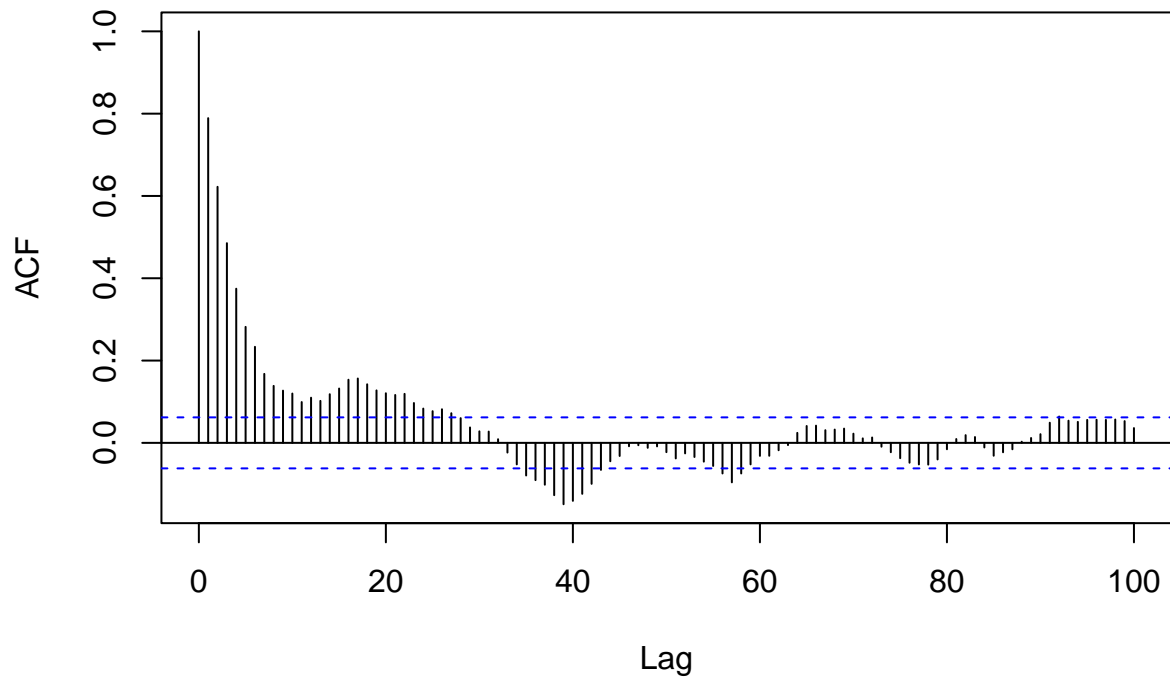

    b) Calculate the theoretical autocorrelation function and the plug-in estimator of the autocorrelation of the simulation results in a) and plot both curves for lags from 0 to 100.

```
# theoretical autocorrelation function
theo.acf <- ARMAacf(ar = c(0.8), lag.max = 100)
plot(theo.acf)
```
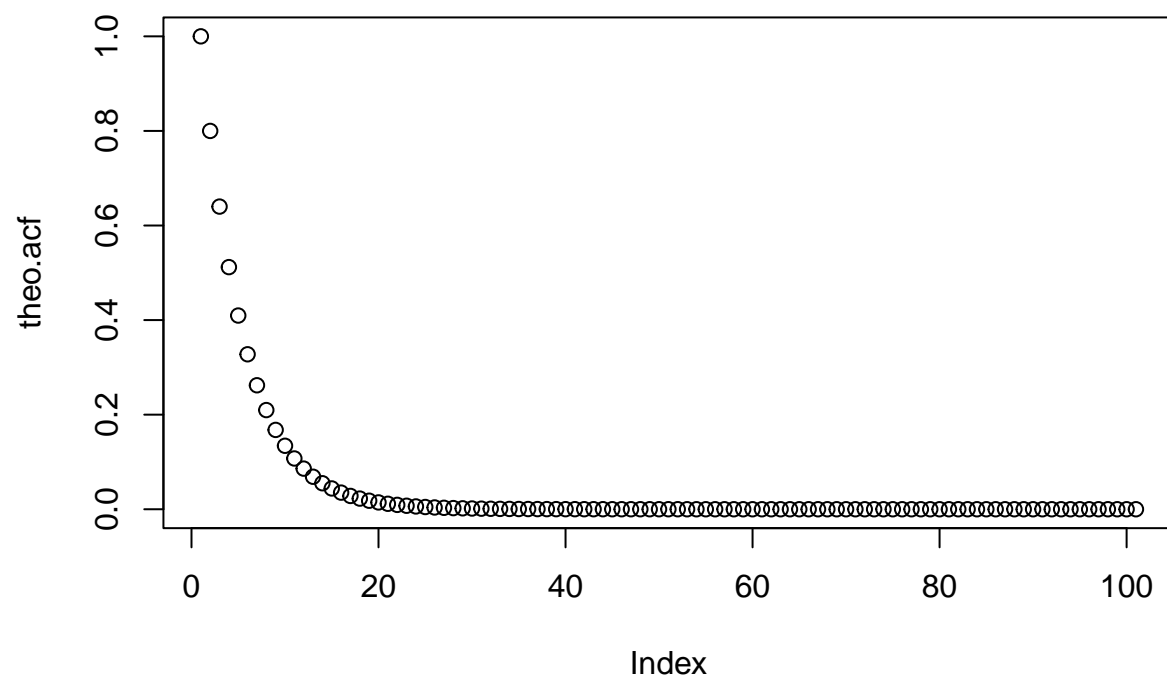
```
# plug-in estimator of the autocorrelation of the simulation results
acf(sim.d, lag.max = 100)
```
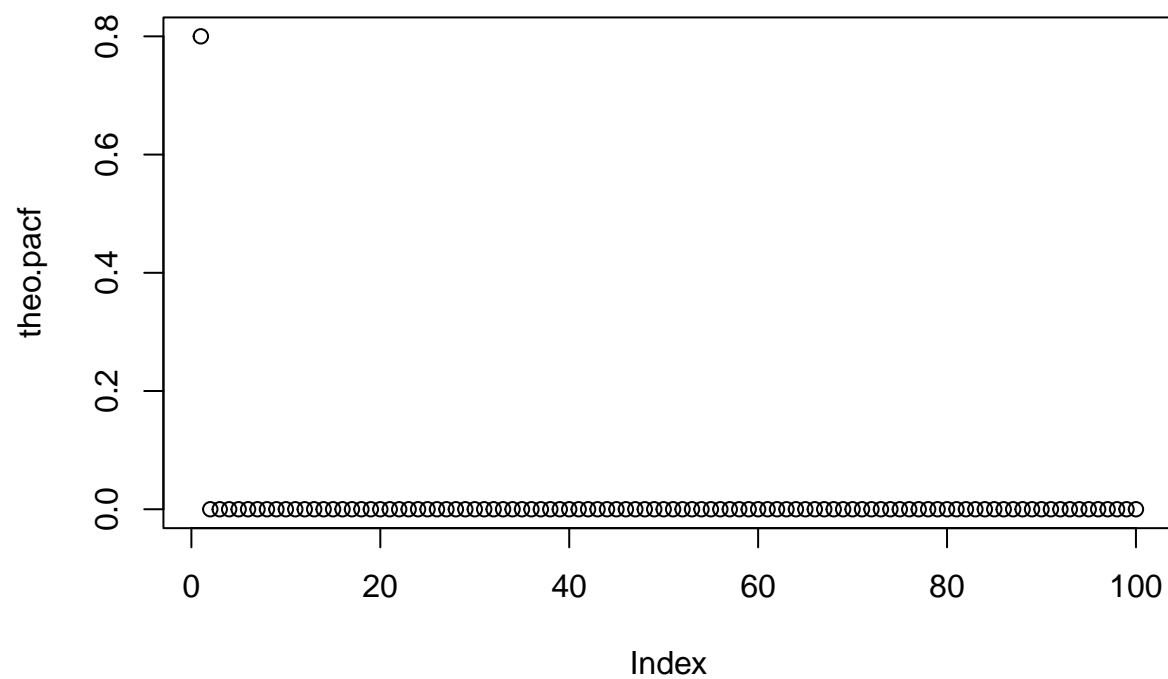
Series sim.d

c) What is the functional dependence of the theoretical autocorrelation function on the lag k and $\alpha_1 = 0.8$?

```
Box.test(sim.d, lag=100, type="Ljung-Box")
```

```
##
##  Box-Ljung test
##
## data:  sim.d
## X-squared = 2058.2, df = 100, p-value < 2.2e-16
```
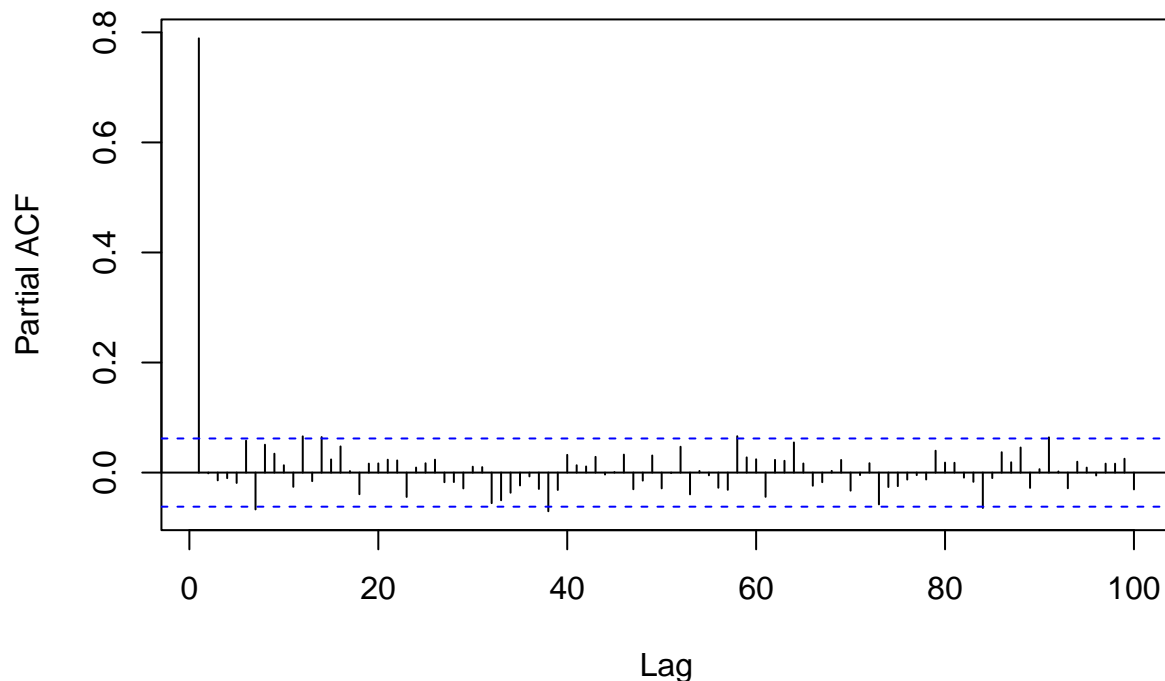
d) Now compare the theoretical partial autocorrelation function with the estimated version for the simulated process. Which particularity do you observe for the two representations?

```
theo.pacf <- ARMAacf(ar = c(0.8), lag.max = 100, pacf = TRUE)
plot(theo.pacf)
```

```
pacf(sim.d, lag.max = 100)
```

Series sim.d

The partial autocorrelation function of the simulated AR(1) process is identical to the theoretical partial autocorrelation function. Whereas, the autocorrelation function of the simulated process differs quite a bit from the theoretical autocorrelation function.

Therefore, I tend to conclude, that the partial autocorrelation function is the better tool for distinguishing an AR(1) process.

Exercise 4.2

An analytic device measures the creatine concentration of human muscular tissue. In this exercise, we would like to check whether it is operating correctly, i.e. the measured values does not depend on the measuring instance.

A sample with known concentration is split into 157 samples and measured by the device one after the other. You can find them in the data under http://stat.ethz.ch/Teaching/Datasets/WBL/kreatin.dat

In this exercise, we focus only on the variable "gehalt" (content) in the data.

```
creatine <- read.csv('http://stat.ethz.ch/Teaching/Datasets/WBL/kreatin.dat', sep = ' ')
creatine.ts <- ts(creatine$gehalt)
plot(creatine.ts)
```
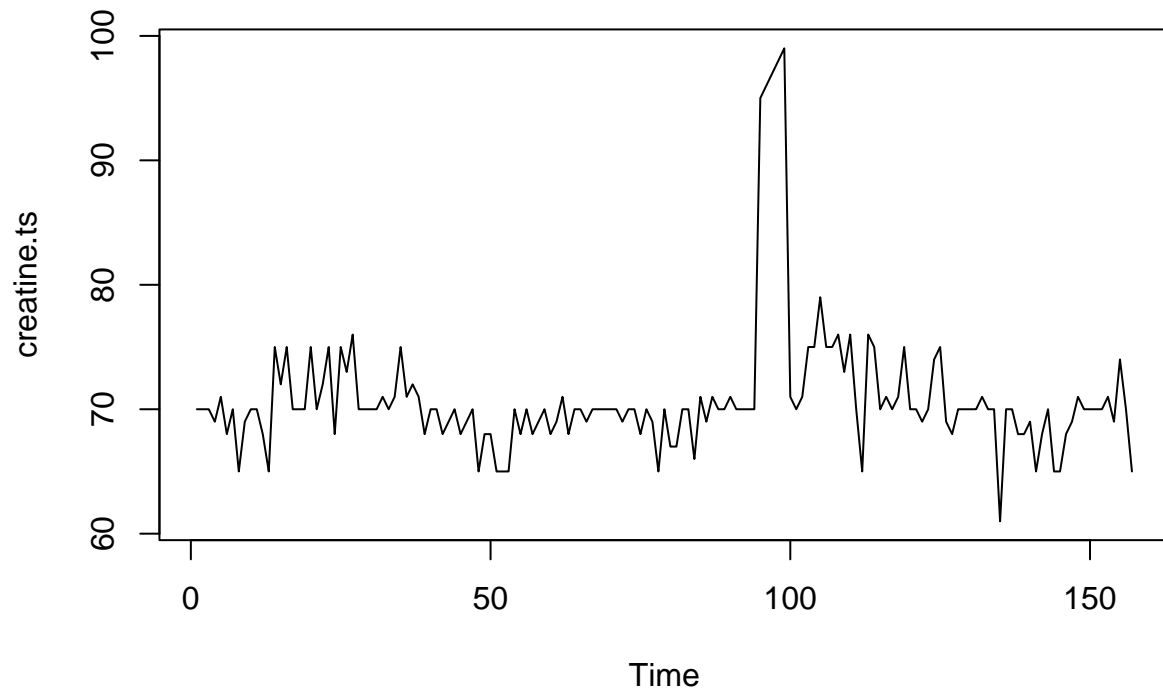
a) Which stochastic model should this series of data follow if the machine is working correctly?

It should be an AR model, because the current value will most likely be a linear combination of the previous time steps, plus an additional innovation. The creatine content in the muscle tissue either decays or increases over time. Additionally, the body will react to outside influences and regulate the creatine content, which in this model is represented by the innovation.

b) Use the time series plot, the autocorrelations (and the partial autocorrelations) to determine whether these data fit the ideal model found in Part a) or not.
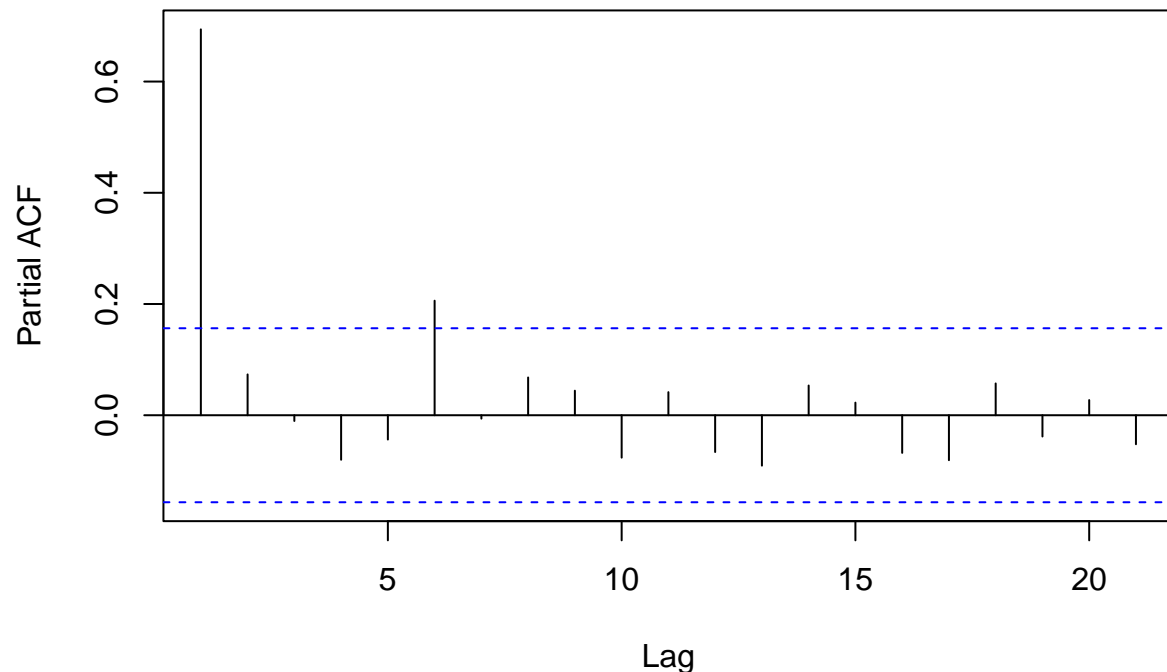
```
acf(creatine.ts)
```

Series creatine.ts
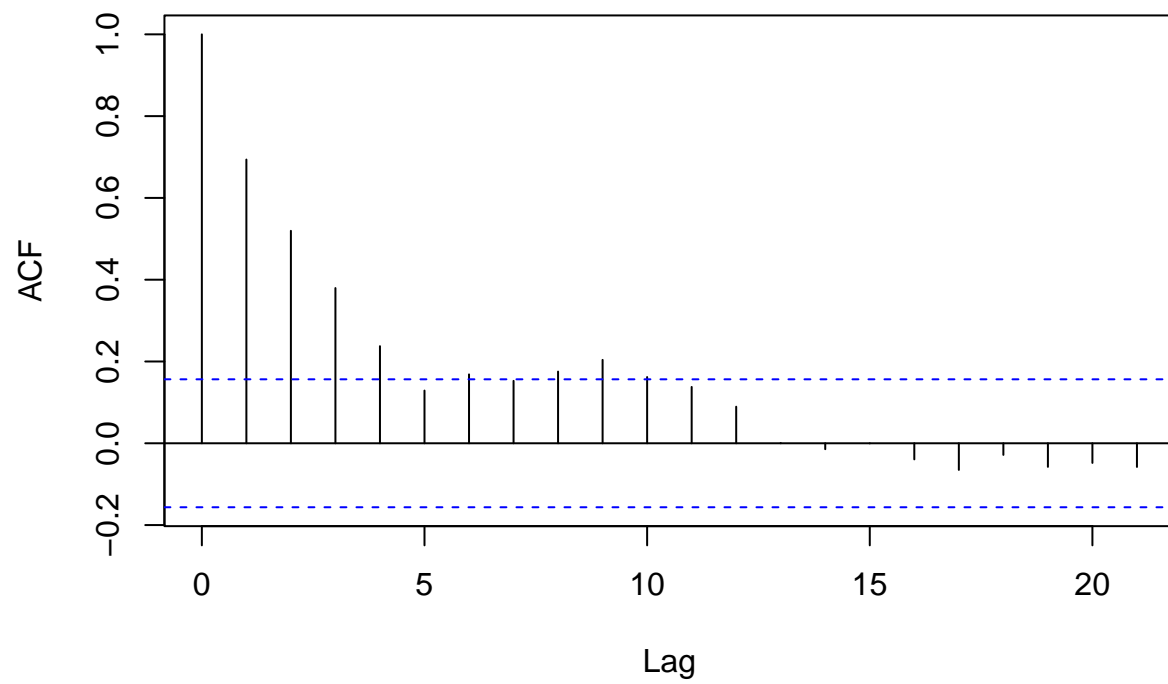

```
pacf(creatine.ts)
```

Series creatine.ts

The plot shows a large spike between time step 90 and 100. A measuring error could be suspected. However, the autocorrelation function and the partial autocorrelation function appear to be very similar to the theoretical ACF and PACF. The ACF has a fast decay, which is an indicator for an AR process. The PACF should have no significant values after a lag of 1. However, this condition is violated at lag 6. Therefore, I would argue that the spike at around step 90 to 100 is due to a measuring error.

Exercise 4.3

In this exercise, we consider two time series ts1 and ts2, which putatively were created by an AR process. You may download the data from http://stat.ethz.ch/Teaching/Datasets/WBL/ts_S3_A2.dat

```
data <- read.csv('http://stat.ethz.ch/Teaching/Datasets/WBL/ts_S3_A2.dat', sep = ' ')
data.ts1 <- ts(data$ts1)
data.ts2 <- ts(data$ts2)
```

a) Visualize both time series. Are both time series stationary? What is their mean?

```
plot(data.ts1)
```
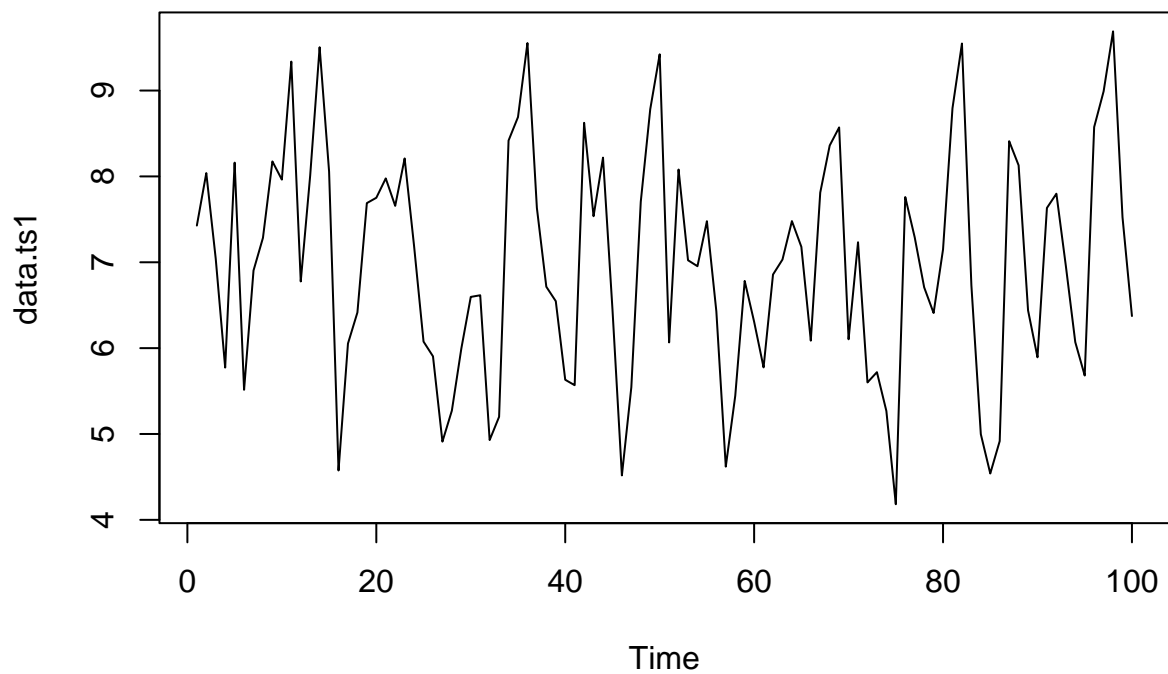
```
mean(data.ts1)
```

```
## [1] 6.97997
```

```
Box.test(data.ts1, lag = 10, type="Ljung-Box")
```

```
##
##  Box-Ljung test
##
## data:  data.ts1
## X-squared = 32.62, df = 10, p-value = 0.0003155
```
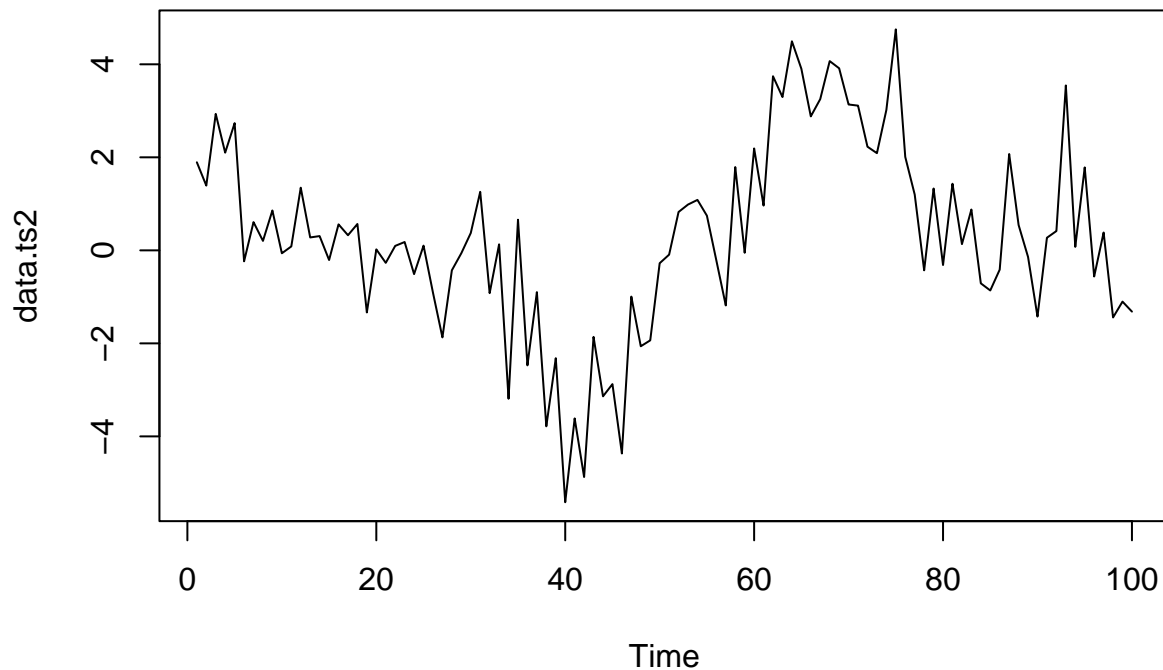
```
plot(data.ts2)
```

```
mean(data.ts2)
```

```
## [1] 0.3037681
```

```
Box.test(data.ts2, lag = 10, type="Ljung-Box")
```

```
##
##  Box-Ljung test
##
## data:  data.ts2
## X-squared = 295.85, df = 10, p-value < 2.2e-16
```

The means of both time series are not 0, but they could be shifted. So, they can still be stationary.
The Ljung-Box test for both time series has a p-value far below 0.05, which means that the null-hypothesis is rejected and the time series are not stationary.

We do not know the coefficients of the putative AR processes, otherwise we could also check if the absolute values of the roots of the characteristic polynomial are all larger than 1.

b) Consider the (partial) autocorrelation function and decide whether the two time series can be generated by an AR process. If yes, what is the order of the respective AR process?
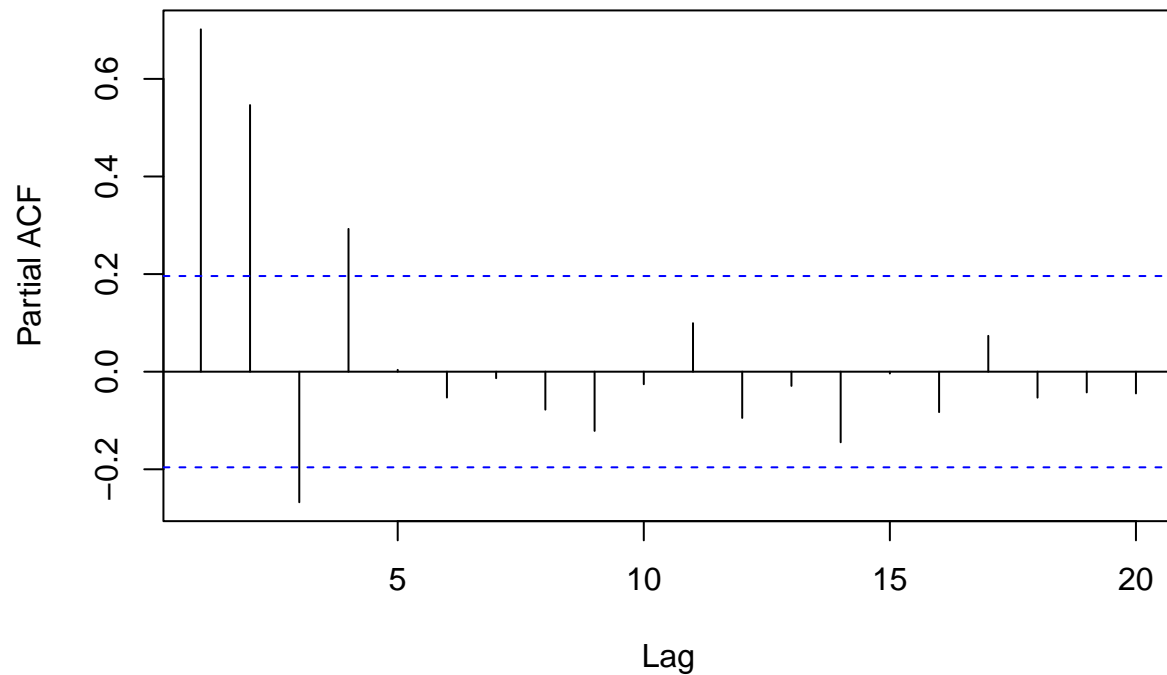
Hint: The partial auto correlation function of an AR(p) process displays a sudden drop for lags larger than p.

```
pacf(data.ts1)
```
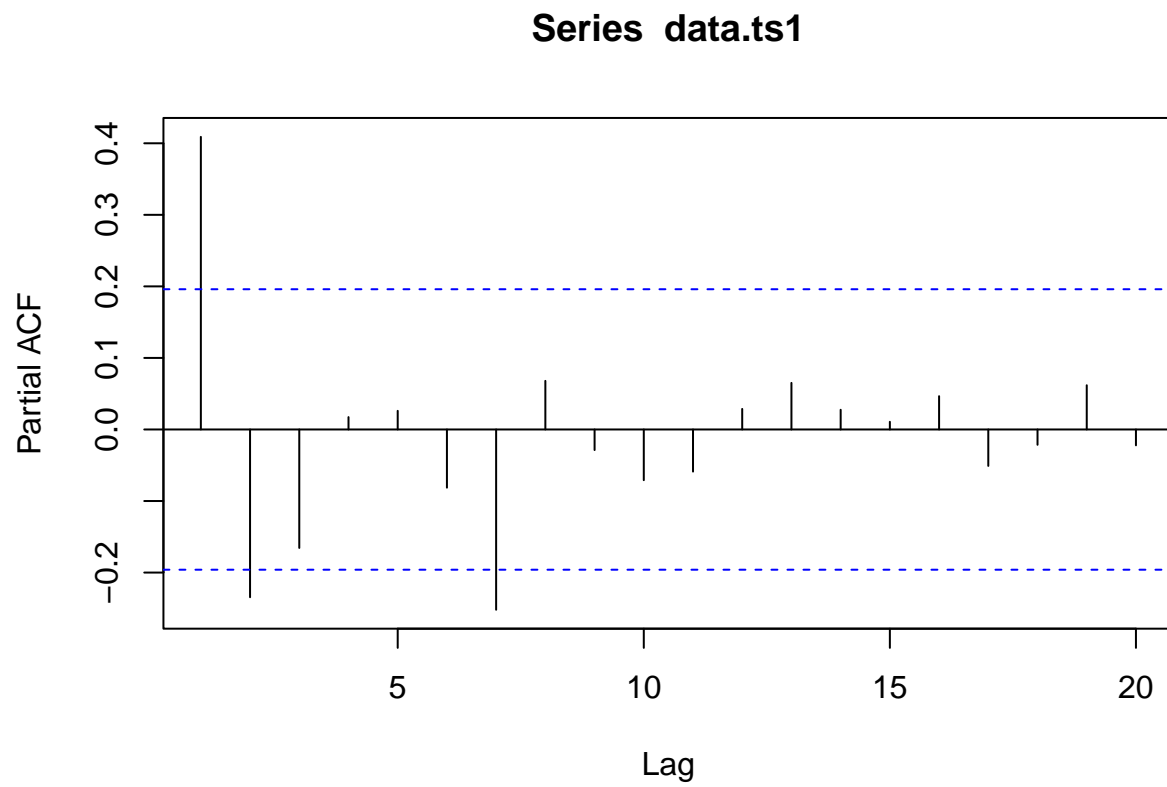
Series data.ts1


```
pacf(data.ts2)
```

Series data.ts2

I would conclude that ts1 was not generated by an AR(p) process because after lag 2 there are a few not significant values until lag 7, when there is another significant one. There is no sudden drop after a specific lag.

The PACF of ts2 shows a drop at lag 3, but this value is still significant and negative correlations do exist. I assume this is a possibility in AR models. Lag 4 has the last significant value. Afterwards, no more significant values are recorded. My conclusion: ts2 was in fact generated by an AR(p) process with $p = 4$.