6. Computer exercises

6.1 The file `Const.txt` contains monthly data of the number of started construction projects in a neighborhood in the USA between the years 1966 and 1974.

(a) Visualize the data. Does the time series look stationary? Into which components could one try to decompose the time series?

(b) Use the function `stl` to decompose the time series. In other words, decompose the time series into a trend component, a seasonal component and a random component.

(c) Use the following filter to estimate the trend:

$$y_t = \frac{1}{24}\left(x_{t-6} + 2x_{t-5} + 2x_{t-4} + \ldots + 2x_t + \ldots + 2x_{t+4} + 2x_{t+5} + x_{t+6}\right).$$

Plot the obtained estimate y_t, the estimate given by the function `stl` and the original time series into a single figure. Are there differences between the estimates?

(d) Remove the trend and seasonal component from the time series by using difference operations. Use the function `stl` to decompose the obtained time series.

Solution.

(a) By Figure 2, the time series does not seem to be stationary. We decompose the time series into trend (m_t), seasonal (s_t) and random (e_t) components:

$$x_t = m_t + s_t + e_t,$$

where $m_t = \beta_0 + \beta_1 \cdot t + \ldots + \beta_k t^k$ is a polynomial of order k.

(b) The decomposition can be performed as follows.

```
CONST<- read.table("Const.txt",header=T,sep=",",row.names=1)

const <- ts(CONST,start=1966, frequency=12)
const.stl <- stl(const[,1], s.window="periodic")
# with s.window-parameter it is possible to set the method for
# estimating the seasonal component.

plot(const.stl)
```

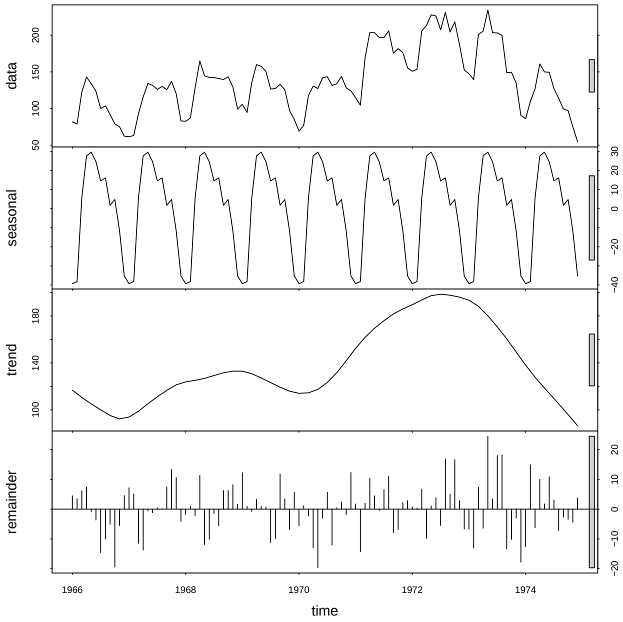

Figure 1: Decomposition of the time series **Const**.

(c) The filtering can be applied conveniently by utilizing the function `filter`. From Figure 2, we see that the trend given by `stl` is almost identical to the one given by filtering. However, the trend given by the filter is a bit rougher than the trend given by `stl`, which can be seen by zooming the figure.

```
const.filt <- filter(const, c(1,rep(2,11),1)/24 )
trend <- const.stl$time.series[,2]

plot(const,lty=3)
lines(trend, col="blue")
lines(const.filt, lty=2, col="red")
legend("topleft", legend=c("Time series","Filter","STL"),
       col=c(1,"red","blue"), lty=c(3,2,1))
```

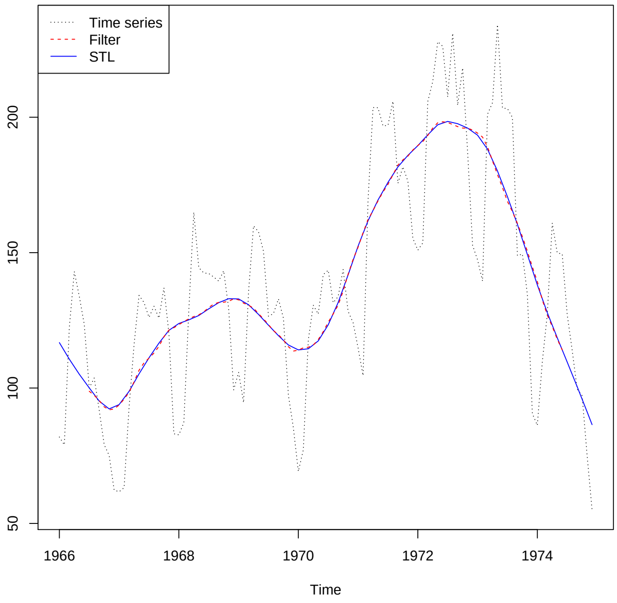

Figure 2: The original time series as gray, the trend given by **stl** function as blue and the trend given by the filter as red.

(d) Next, we calculate the difference DD^{12} and decompose the obtained time series. By the top subfigure of Figure 3, the time series obtained by taking the differences could be stationary.

```
const.diff <- diff(diff(const, lag=12))

const.diff.stl <- stl(const.diff[,1], s.window="periodic")

plot(const.diff.stl)
```

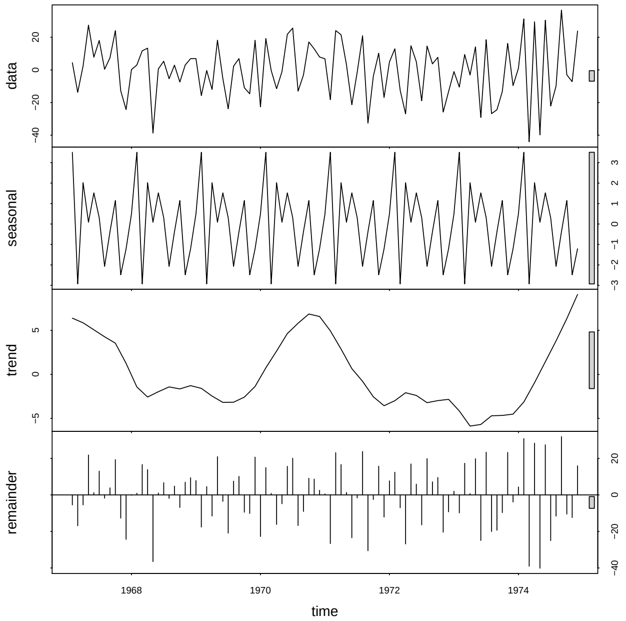

Figure 3: Decomposition of the time series DD^{12}Const.

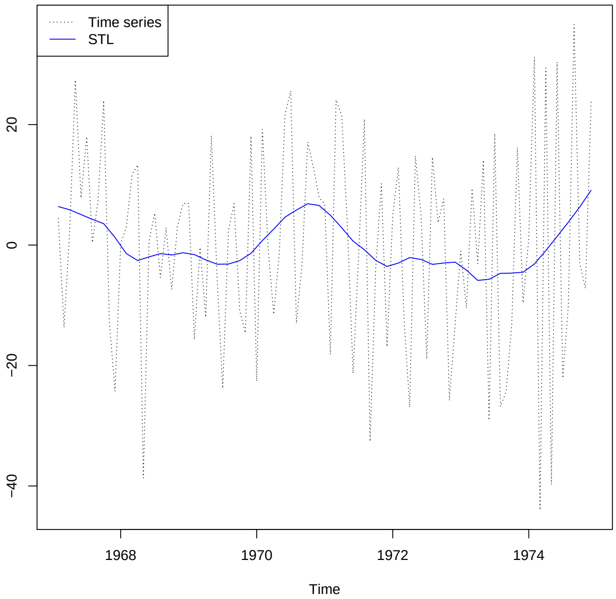

Figure 4: DD^{12}Const as gray and the trend given by **stl** function as blue.

6.2 The file `arsimulation.txt` contains realizations at 100 distinct points of time for the following three processes,

$$x_t = \varepsilon_t - \frac{x_{t-1}}{2}, \tag{1}$$
$$y_t = \nu_t - \frac{y_{t-1}}{2}, \tag{2}$$
$$z_t = \eta_t - \frac{z_{t-1}}{2}, \tag{3}$$

where for every $t \in \{2, 3 \ldots, 100\}$, ε_t was generated independently from the standard Cauchy distribution, ν_t was generated independently from the Student's t-distribution with 3 degrees of freedom and η_t was generated independently from the Student's t-distribution with 30 degrees of freedom. The starting point for the time series was chosen to be deterministically $x_1 = y_1 = z_1 = 0$.

(a) Visualize the three time series.

(b) Fit an AR(1) process to each of the three time series by using the function `Arima` from the package **forecast**. Use `Arima` with the arguments `include.mean = FALSE` and `method="ML"`. Does the AR(1) parameter estimates given by `Arima` match the true parameter values $-1/2$?

(c) Bootstrap 95% confidence intervals for the AR(1) model parameters. Use `Arima` with the arguments `include.mean = FALSE` and `method="ML"`.

(d) Assume that for all $s \geq 1$, we have that $x_{t-s} \perp\!\!\!\perp \varepsilon_t$, $y_{t-s} \perp\!\!\!\perp \nu_t$, and $z_{t-s} \perp\!\!\!\perp \eta_t$, where $\perp\!\!\!\perp$ is used to denote stochastic independence. In addition, assume that the elements of the set $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ are i.i.d., the elements of $\{\nu_t\}_{t\in\mathbb{Z}}$ are i.i.d. and the elements of $\{\eta_t\}_{t\in\mathbb{Z}}$ are i.i.d. Under these assumptions, which of the theoretical processes (1)–(3) are weakly stationary?

Solution.

(a) We import the data and plot the three time series, see Figure 5 and the following code.

```
#install.packages("forecast")
library(forecast)

simu <- read.table("arsimulation.txt",header=TRUE)

ts.plot(simu[,1], main="Time series (1)",ylab="x")
ts.plot(simu[,2], main="Time series (2)",ylab="y")
ts.plot(simu[,3], main="Time series (3)",ylab="z")
```

In Figure 5, there is a visible peak for time series (1) with $t = 43$. Note that, the standard Cauchy distribution has considerably heavier tails when compared to, e.g., any Normal distribution.

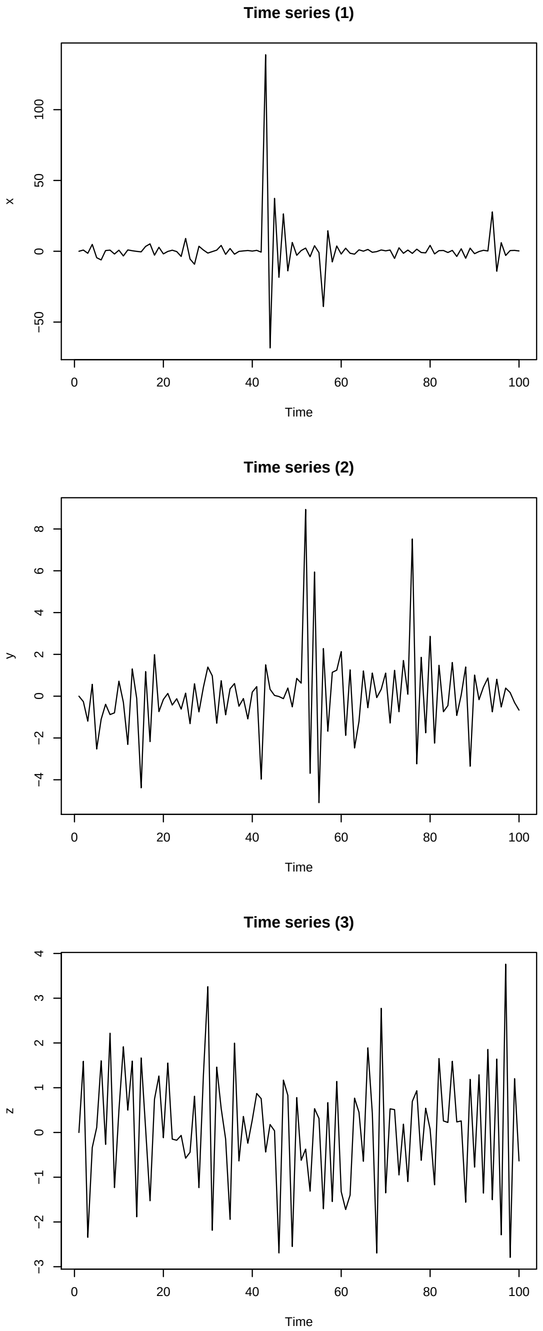

Figure 5: Length 100 realizations of the stochastic processes (1) - (3).

(b) We estimate the AR(1) parameters using the three time series.

```
fitX <- Arima(X,order=c(1,0,0), include.mean = FALSE, method="ML")
fitY <- Arima(Y,order=c(1,0,0), include.mean = FALSE, method="ML")
fitZ <- Arima(Z,order=c(1,0,0), include.mean = FALSE, method="ML")
```

```
fitX$coef # -0.499
fitY$coef # -0.494
fitZ$coef # -0.492
```

The AR(1) parameters estimated from the first, second and third time series are approximately -0.499, -0.494 and -0.492, respectively. Hereby, the estimates are all relatively close to the true parameter value $-1/2$.

(c) We implement steps 1-5 from the lecture slides, see the following code. Here, the theoretical minimum time series length required for estimating the model parameters is $w = 2$. However, estimating the AR(1)-parameter from only two observations makes the maximum-likelihood (ML) procedure unstable. We ensure more fluent computations by setting $w = 10$.

```
set.seed(3141)

# Choose number of repetitions
m <- 5000 # if slow, make this smaller
w <- 10 # Theoretical minimum would be w =2
# However, estimating the AR(1)-parameter from 2 observations
# makes the ML-estimation procedure unstable
# Thus, we set w = 10

n <- length(X) # The length of the time series is 100
resX <- rep(NA,m) # Initialize an empty vector for the results

for(i in 1:m){
  # Step 1: Select two time points s and u, 0 < s < u <= n, u-s >= w
  #         uniformly

  # Keep choosing the time points, until u-s >= w is satisfied
  u <- 0 # initialize u and s
  s <- 0
  while(u-s < w){
    # The same point cannot be chosen twice
    su <- sample(1:n,2,replace=FALSE)
    s <- min(su)
    u <- max(su)
  }# Keep repeating until u-s >= w = 10

  # Step 2: Calculate a new parameter vector esimate from the series
  #         x_s, x_(s+1),..., x_u

  # print(i) #uncomment this, if you want to track progress
  resX[i] <- Arima(X[s:u],order=c(1,0,0),include.mean = FALSE,
                    method="ML")$coef[1]
} # Step 3: Repeat m-1 times

# Step 4: Order the obtained m estimates from the smallest to the
#         largest

resXsort <- sort(resX)

# Step 5: Set lower end of the boostrap confidence interval to be
#         smaller than or equal to the 125th ordered estimate and
#         set the upper end of the bootstrap confidence interval
#         to be larger than or equal to the 4875th ordered estimate.

confintX <- c(resXsort[125],resXsort[4875])

# Repeat the same for Y and Z
```

The requested confidence intervals estimated from the first, second and third time series are approximately $[-0.94, -0.17]$, $[-0.75, -0.26]$ and $[-0.75, -0.23]$, respectively.

(d) The standard Cauchy distribution is exactly the Student's t-distribution with 1 degree of freedom. Student's t-distribution with k degrees of freedom has $k-1$ theoretical moments. As the variance of x_t and the variance of ε_t are undefined, the process (1) is not weakly stationary. Recall that, weak stationarity contains the assumption that the variance is time invariant and finite.

Consequently, the AR(1) processes (2) and (3) have finite variances. Under the assumptions of this exercise and by previous theoretical exercises, we have that the theoretical processes (2) and (3) are weakly stationary.

6.3 The file `alcoholdeaths.txt` contains a univariate time series of the number of yearly alcohol related deaths in Finland per 100 000 individuals belonging to the age group 40–49. The data covers the years 1969–2007 and the data set is available in the homepage of Statistics Finland.

Denote the observed number of deaths per 100 000 individuals in the year t as y_t. Assume that the underlying stochastic process that has generated the time series y_t is normally distributed for every t such that $y_t \sim \mathcal{N}(\mu_t, \sigma_\varepsilon^2)$, that is, the variance is the same for every t and the mean depends on t. Assume that the mean process μ_t is a random walk with drift such that,

$$\mu_{t+1} = \mu_t + \nu + \eta_t, \tag{4}$$

where $\eta_t \sim \mathcal{N}(0, \sigma_\eta^2)$ for every t. Assume that we have no prior information regarding the initial state μ_1 and the constant slope ν. Using Kalman filter, estimate the model parameter ν and calculate the one year prediction for the time series.

Solution. In order to apply the Kalman filter, we first need to construct the state–space representation for the process. In this exercise, we use the R-package **KFAS**. The notation used by the corresponding R-package differs a little from the notation used in the lecture slides. In order to keep the estimation steps more clear, we follow the notation of the R-package in this exercise.

Using the notation of the package **KFAS**, the state-space representation of a dynamical system is of the from:

$$\begin{aligned} y_t &= Z_t\alpha_t + \varepsilon_t, && \text{(observation equation)} \\ \alpha_{t+1} &= T_t\alpha_t + R_t\eta_t && \text{(state equation),} \end{aligned}$$

where $\varepsilon_t \sim \mathcal{N}(0, H_t)$, $\eta_t \sim \mathcal{N}(0, Q_t)$ for every t and $\alpha_1 \sim \mathcal{N}(\alpha_1, P_1)$ and ε_t, η_s and α_1 are mutually independent of each other for every t and s. In general, the system matrices Z_t, T_t and R_t can be time dependent. However, in this exercise all of the system matrices are time invariant, that is, constant over time.

The state–space representation for the process μ_t, given in Equation (4), is obtained by defining,

$$\begin{aligned} &Z = \begin{pmatrix} 1 & 0 \end{pmatrix}, && H = \sigma_\varepsilon^2, && T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \\ &\alpha_t = \begin{pmatrix} \mu_t \\ \nu_t \end{pmatrix}, && R = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, && Q = \sigma_\eta^2, \\ &\alpha_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, && P_{*,1} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, && P_{\infty,1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \end{aligned}$$

where the matrices $P_{*,1}$ and $P_{\infty,1}$ are related to the estimation of the unknown variances σ_ε^2 and σ_η^2. Again, the notation used by the package **KFAS** deviates slightly from the notation of the lecture slides. Defining the matrices $P_{*,1}$ and $P_{\infty,1}$ as above, corresponds to the initial guess that the covariance matrix is an identity matrix.

The state–space representation for the process is then,

$$\begin{aligned} y_t &= \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} \mu_t \\ \nu_t \end{pmatrix} + \varepsilon_t, \\ \begin{pmatrix} \mu_{t+1} \\ \nu_{t+1} \end{pmatrix} &= \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_t \\ \nu_t \end{pmatrix} + \begin{pmatrix} \eta_t \\ 0 \end{pmatrix}. \end{aligned}$$

Note that even though the slope term ν is time invariant in the model above ($\nu_{t+1} = \nu_t$), it is still recursively estimated by the Kalman filter for every t. In the recursive Kalman filter estimation process, when the new observation y_t becomes available, the estimates are updated to take account of the new information given by y_t. Thus, the final estimate for the slope term will be the one given after the information of every y_t has been utilized.

The estimation procedure in R can be performed as follows. Note that all of the unknown model parameters are set to be NA below. After estimating the model, the unknown model parameters σ_ε^2 and σ_η^2 can be estimated using the function fitSSM.

```
#install.packages("KFAS")
library(KFAS)

alko <-ts(read.table("alcoholdeaths.txt"),start=1969)

a1 <- c(0,0) # Initial guess for mu and nu
Zt <- matrix(c(1, 0), 1, 2)
Ht <- matrix(NA)
Tt <- matrix(c(1, 0, 1, 1), 2, 2)
Rt <- matrix(c(1, 0), 2, 1)
Qt <- matrix(NA)
P1 <- matrix(0, 2, 2)
P1inf <- diag(2)

# -1 sets that no constant is estimated in the model.

model_gaussian <- SSModel(alko~-1+SSMcustom(a1=a1,Z=Zt,T=Tt,R = Rt,Q=Qt,
                                            P1=P1,P1inf=P1inf),H=Ht)

fit_gaussian <- fitSSM(model_gaussian, inits = c(0, 0))
# above inits-parameter is related to the estimation procedure of the
# unknown variances

fit_gaussian$model$Q # ML-estimate for the variance of eta_t
fit_gaussian$model$H # ML-estimate for the variance of epsilon_t

out_gaussian <- KFS(fit_gaussian$model)

plot(alko)
lines(out_gaussian$a[,1],col="red")
```

Thus, the obtained estimate for the parameter ν is 0.84. The estimates for ν at every t is given by:

```
out_gaussian$a[,2]
```

The one step predictions for the state variable μ_t given by the Kalman filter are presented in Figure 6. These predictions also serve as predictions for alcohol related deaths. The prediction for the year 2008 is visible from the Figure 6. Note that longer term predictions produced by Kalman filter involve generally relatively wide confidence intervals. In other words, long term predictions are generally unreliable.

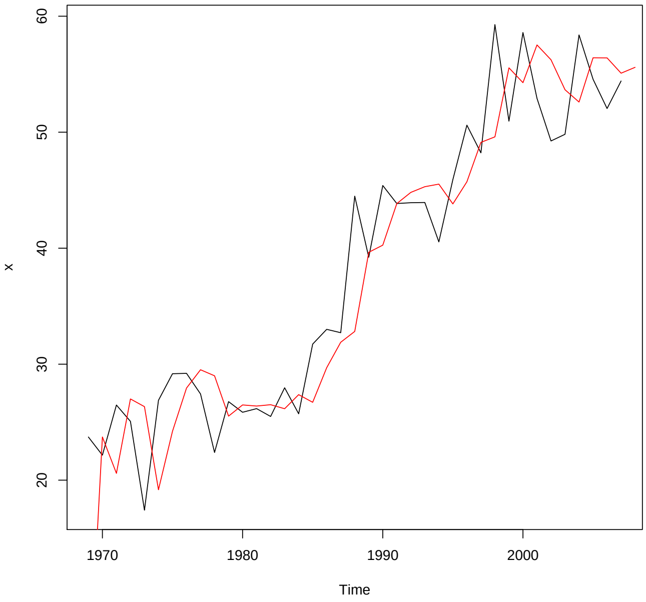

Figure 6: The original time series as black and the one step predictions as red.