Danmarks Tekninske Universitet

Advanced Time Series Analysis
Course number: 02427

# Computer Exercise 1

*Authors:*
*Mads Esben Hansen, s174434*
*Karl Emil Takeuchi-Storm, s130377*

October 6, 2020

# Introduction

This is answer to "Computer Exercise 1" in the course in advanced times series analysis at DTU. The exercise is divided into 5 parts that will be answered independently. The subject of the exercise is non-linear models and identifying functional dependencies using non-parametric estimations. Whenever there is a referral to *the lecture notes*, the notes in mention is *Modelling Non-Linear and Non-Stationary Time Series*, by Henrik Madsen and Jan Holst, December 2006.

# Part 1

## Self-Exiting Threshold Autoregressive model

A Self-Exciting Threshold Autoregessive model (SETAR) divides the space with a set of thresholds, the thresholds determines when the model change the regime. In each regime an AR model determine the model behaviour. Once the model exceeds the threshold, the regime changes.

The SETAR $(l; d; k_1, k_2, \ldots, k_l)$ model is given by :

$$X_t = a_0^{(J_t)} + \sum_{i=1}^{k_{J_t}} a_i^{(J_t)} X_{t-i} + \epsilon_t^{(J_t)} \tag{1}$$

where

$$J_t = \begin{cases} 1 & X_{t-d} \in R_1 \\ 2 & X_{t-d} \in R_2 \\ \vdots & \vdots \\ l & X_{t-d} \in R_l \end{cases} \tag{2}$$

An example of the SETAR(2,1,1) is plotted below, using;

$$X_t = \begin{cases} 2 + X_{t-1} + \epsilon_t^1 & X_{t-1} \leq 0 \\ -2 - X_{t-1} + \epsilon_t^2 & X_{t-1} > 0 \end{cases} \tag{3}$$

where both $\epsilon \sim N(0, 1)$.

From the first plot in figure 1 it is clear that the regime changes very frequently, while the two groups in the second plot shows the behaviour of each model. In the regime where $X_{t-1} > 0$, the regime switches nearly every time, while the other regime is slightly more stable.

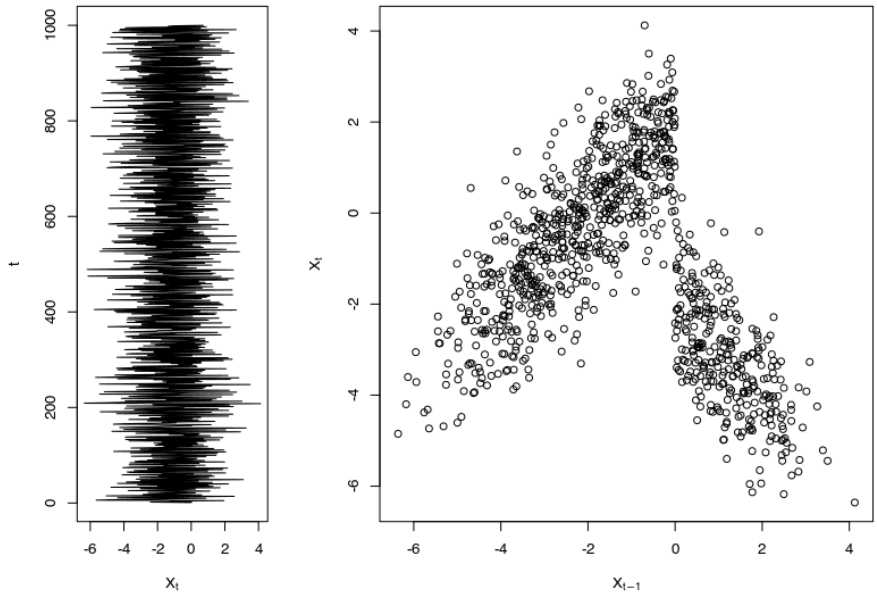


Figure 1: Raw data of simulation of SATAR(2,1,1) model with n=1000.

As discussed above the SETAR model in figure 1 switches between regimes often, while a different implementation of the SETAR model, causes fewer state switches as the two regime are much more stable in each regime. To illustrate this, the following SETAR(2,1,1) is used:

$$X_t = \begin{cases} -1.5 + 0.5 \cdot X_{t-1} + \epsilon_t^1 & X_{t-1} \leq 0 \\ 1.5 + 0.5 \cdot X_{t-1} + \epsilon_t^2 & X_{t-1} > 0 \end{cases} \tag{4}$$

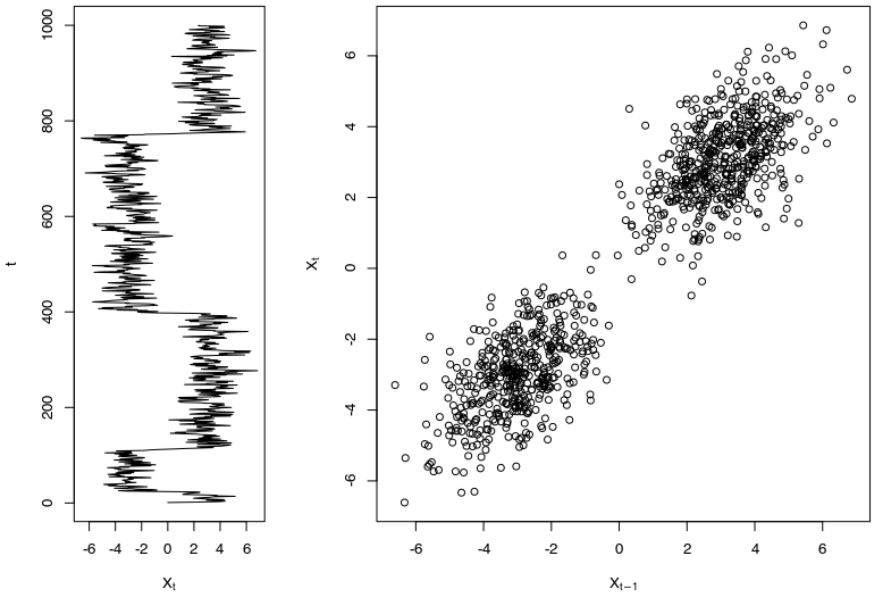


Figure 2: Raw data of simulation of SATAR(2,1,1) model with n=1000.

The SETAR model is a good tool to model behaviour, where the states changes depending on previous values.

## Independently Governed AR model

An Independently Governed AutoRegressive (IGAR) model, is an AR model with different regimes. Each regime is connected with some probability such that there is a given probability, $p_i$, of being in regime $i$.
IGAR $(l; k)$ model is given by :

$$X_t = a_0^{(J_t)} + \sum_{i=1}^{kJ_t} a_i^{(J_t)} X_{t-i} + \epsilon_t^{(J_t)} \tag{5}$$

where

$$J_t = \begin{cases} 1 & \text{with prob } p_1 \\ 2 & \text{with prob. } p_2 \\ \vdots & \vdots \\ l & \text{with prob. } 1 - \sum_{i=1}^{l-1} p_i \end{cases} \tag{6}$$

An example of the IGAR(2,1) is plotted below, using;

$$X_t = \begin{cases} 3 + 0.5 \cdot X_{t-1} + \epsilon_t^1 & \text{with prob. } 0.5 \\ -3 - 0.5 \cdot X_{t-1} + \epsilon_t^2 & \text{with prob. } 0.5 \end{cases} \tag{7}$$

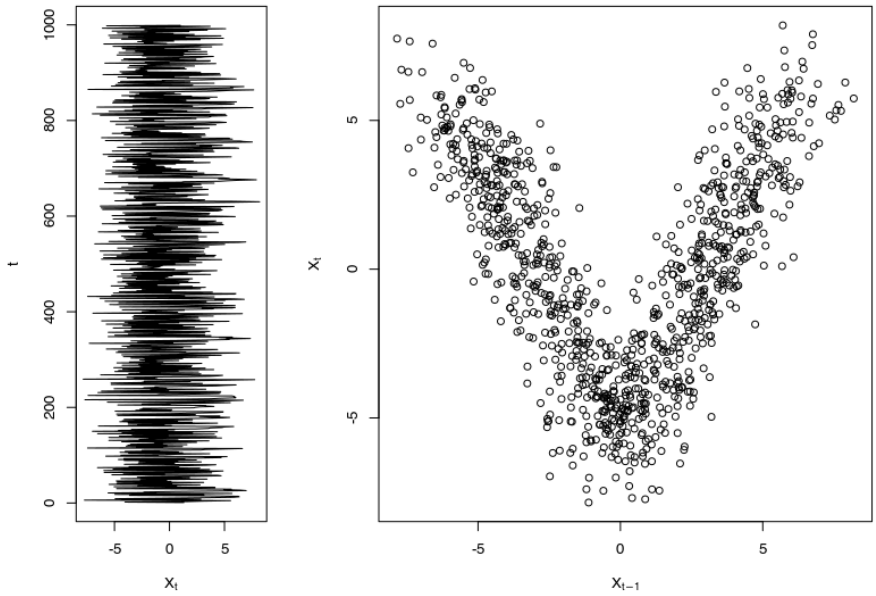


Figure 3: Raw data of simulation of IGAR(2,1) model with n=1000.

A Markov Modulated AR (MMAR) model is is a special case of IGAR. For MMAR $(l; k)$, $J_t$ is given by :

$$P = \begin{pmatrix} p_{11} & p_{12} & \dots & (1 - p_{11\cdots 1n}) \\ p_{21} & p_{22} & \dots & (1 - p_{21\cdots 2n}) \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \dots & (1 - p_{n1\cdots nn}) \end{pmatrix} \tag{8}$$

Such that a regime switch is governed by the rows in $P$, where $p_{11}$ is the probability for staying in regime 1 and $p_{12}$ the probability for switching from regime 1 to 2. Which in turn exhibits the behavior of a Markov chain.

An example of the MMAR(2,1) is plotted below, using regimes given by:

$$X_t = \begin{cases} 3 + 0.5 \cdot X_{t-1} + \epsilon_t^1 \\ -3 + 0.5 \cdot X_{t-1} + \epsilon_t^2 \end{cases} \tag{9}$$

and transition matrix given by:

$$P = \begin{pmatrix} 0.95 & 0.05 \\ 0.05 & 0.95 \end{pmatrix} \tag{10}$$

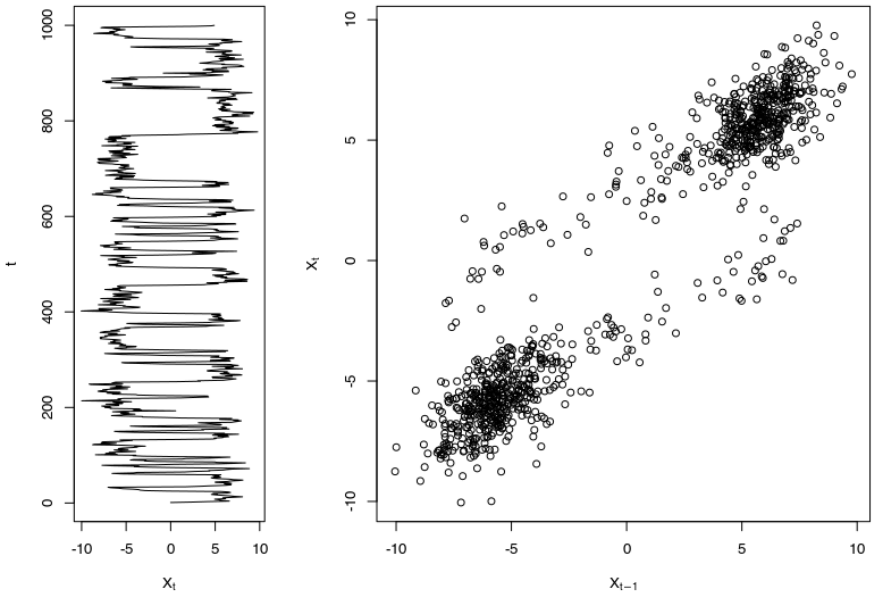


Figure 4: Raw data of simulation of MMAR(2,1) model with n=1000.

## Smooth Threshold AR model

A Smooth Threshold AR (STAR) model, is a model structure that allows for smooth transitions between regimes using a transition function e.g. a Sigmoid function. If we consider the SETAR model from before given by equation 4, we can model this using a Sigmoid function by:

$$X_t = (2 + y_{t-1} + \epsilon_t^1) + (-4 - 2y_{t-1} - \epsilon_t^1 + \epsilon_t^2) \cdot S(y_{t-1}) \tag{11}$$

Where $S(x)$ is a sigmoid function given by $S(x) = \frac{1}{1+e^{-(x-\beta)\cdot\alpha}}$, where $\beta$ controls the threshold and $\alpha$ controls how smooth/abrupt the transition is. We can now simulate this model using $\alpha = 8$ and $\beta = 0$:

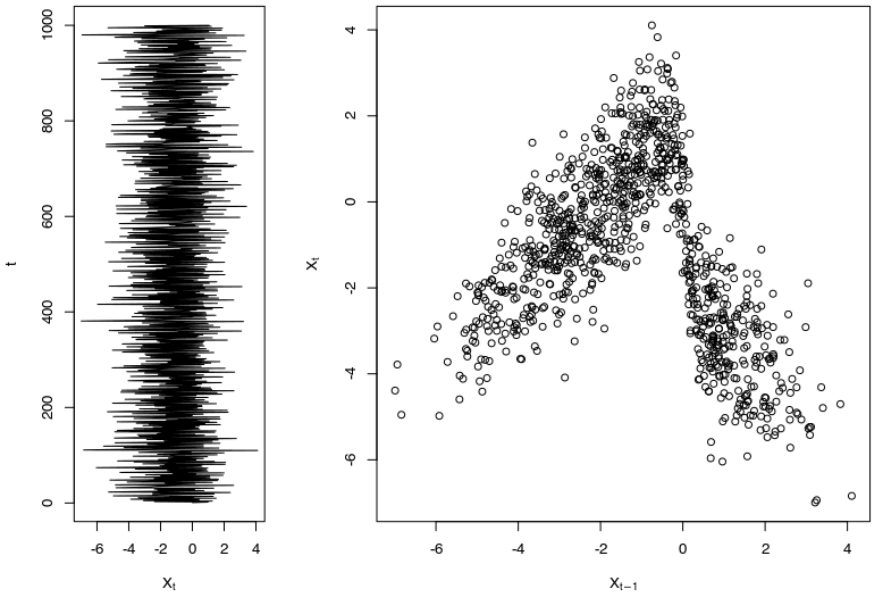


Figure 5: Simulation of STAR model with n=1000.

We can see that this look a lot like the SETAR(2,1,1) model displayed in figure 1, however the transition between the regimes is smooth.

# Part 2

We are asked to compute the conditional mean $M(x) = E[X_t|X_{t-1} = x]$ of an arbitrary SETAR(2,1,1)-model. We use the following model:

$$X_t = \begin{cases} 2 + X_{t-1} + \epsilon_t^1 & X_{t-1} \leq 0 \\ -2 - X_{t-1} + \epsilon_t^2 & X_{t-1} > 0 \end{cases} \tag{12}$$

With $\epsilon^1 \sim N(0,1)$ and $\epsilon^2 \sim N(0,1)$. We will now compute its conditional mean:

$$E[X_t|X_{t-1} = x] = \begin{cases} E[2 + X_{t-1} + \epsilon_t^1 & X_{t-1} \leq 0|X_{t-1} = x] \\ E[-2 - X_{t-1} + \epsilon_t^2 & X_{t-1} > 0|X_{t-1} = x] \end{cases} \tag{13}$$

$$E[X_t|X_{t-1} = x] = \begin{cases} 2 + x & x \leq 0 \\ -2 - x & x > 0 \end{cases} \tag{14}$$

We are now asked to simulate 1000 values from the SETAR model, and use these simulated data and a local regression model to estimate the the conditional mean $\hat{M}(x) = E[X_t|X_{t-1} = x]$. We will use a non-parametric kernel method to estimate the conditional mean as described in *the lecture notes* section 2.3. We will use the "loess" function in r to do so.

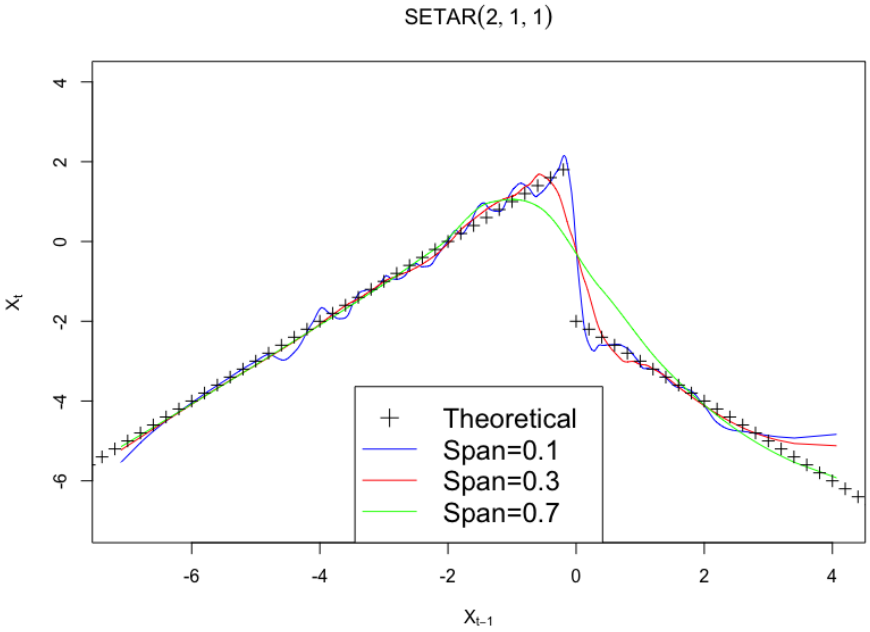


Figure 6: Conditional means of SATAR(2,1,1) model with n=1000.

From 6 we can see that a larger span (meaning larger bandwidth) means more aggressive smoothing and vice versa. In this example a span of 0.1 seems to approximate the SETAR model fairly well. If we want to be more systematic about it, we can do a cross validation on the span value. Doing so with a leave one out method yields the following result:

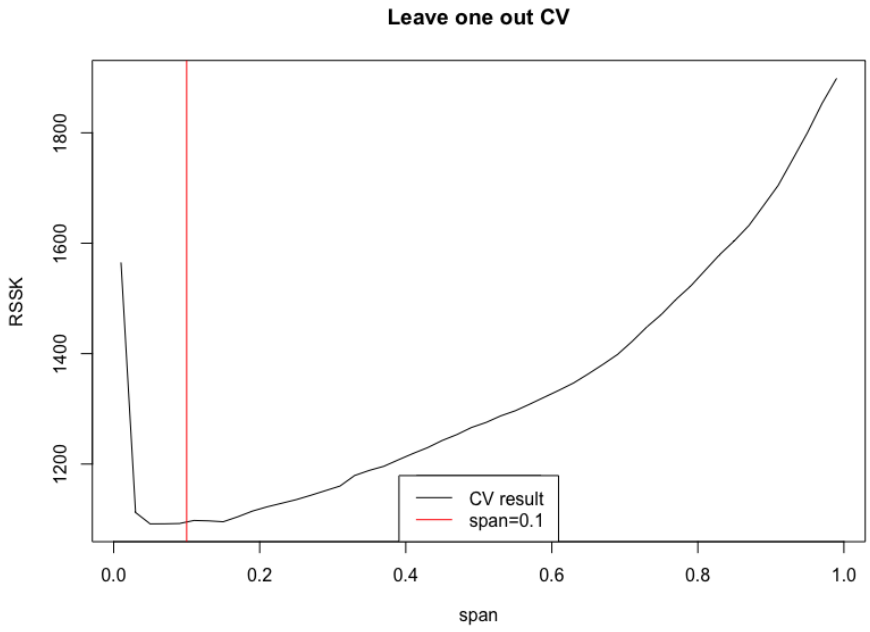


Figure 7: Cross validation on span with n=1000.

We can see that all span value in the region 0.05 to 0.15 seem to perform best according to the cross validation.

# Part 3

We are asked to compute the cumulative conditional mean of our chosen SETAR model and the theoretical cumulative conditional mean, and compare their results. We will find the cumulative conditional mean as described on page 70 in *the lecture notes*. We start by finding the theoretical cumulative conditional mean:

$$\lambda(X) = E[X_t|X_{t-1} = x] = \begin{cases} 2+x & x \leq 0 \\ -2-x & x > 0 \end{cases} \quad \Rightarrow \tag{15}$$

$$\Lambda(\beta) = \int_a^\beta \lambda(x)\, dx = \begin{cases} -0.5a^2 + 0.5\beta^2 - 2a + 2\beta & \beta \leq 0 \\ -0.5a^2 - 0.5\beta^2 - 2a - 2\beta & \beta > 0 \end{cases} \tag{16}$$

Note we have used $\beta$ instead of $(\cdot)$, because it was very hard to interpret the indefinite integral otherwise. The function $\Lambda$ should be taken point-wise in some interval $[a; b]$. We can now plot the theoretical cumulative conditional means together with our simulated data (we will use span=0.1, since it yielded the best fit as discussed in part 2):

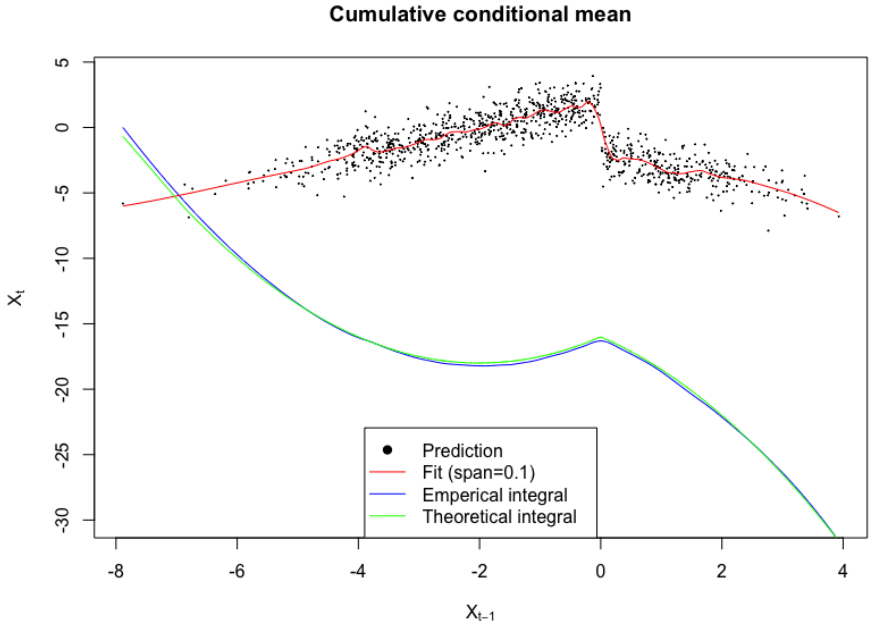


Figure 8: Cumulative conditional means of SATAR(2,1,1) model with n=1000.

From figure 8 it is quite obvious that the theoretical cumulative conditional means and the simulated ones lie quite close to each other. We can also see that their shapes are almost identical. We can see that simulating data and then computing the cumulative conditional means can be a powerful way of estimating cumulative conditional means that would otherwise have been difficult to find.

# Part 4

When calculating the heat-loss of building the coefficient $U_a$ is often regarded as constant. Provided a set of data for heat-loss in a building, a non-parametric fit of the heat-loss $U_a$ as a function of $W_t$ must be carried out, to accommodate the dependency of wind speeds.

$$\Phi_t = U_a \left(T_t^i - T_t^e\right) + \epsilon_t \tag{17}$$

$$U_a(W_t) = \frac{\Phi_t}{T_t^i - T_t^e} \tag{18}$$

The non-parametric fit is carried out using a span width, where a too large span will result in under-fitting and a too small in over-fitting. Thus the optimal span width is found using cross-validation method *Leave-one-out*.

$$CV(h) = \frac{1}{N}\sum_{i=1}^{N}\left[Y_i - \hat{g}_{(i)}(\boldsymbol{X}_i)\right]^2 \pi(\boldsymbol{X}_i) \tag{19}$$

The cross-validation method should be combined with a visual interpretation of the quality.

## Gaussian Kernel Estimate

Using the build in function loess from r, a gaussian kernel estimate is carried out. Plotting the resulting RSS-values over spans, gives two interesting values for the span, namely 0.478 and 0.778.

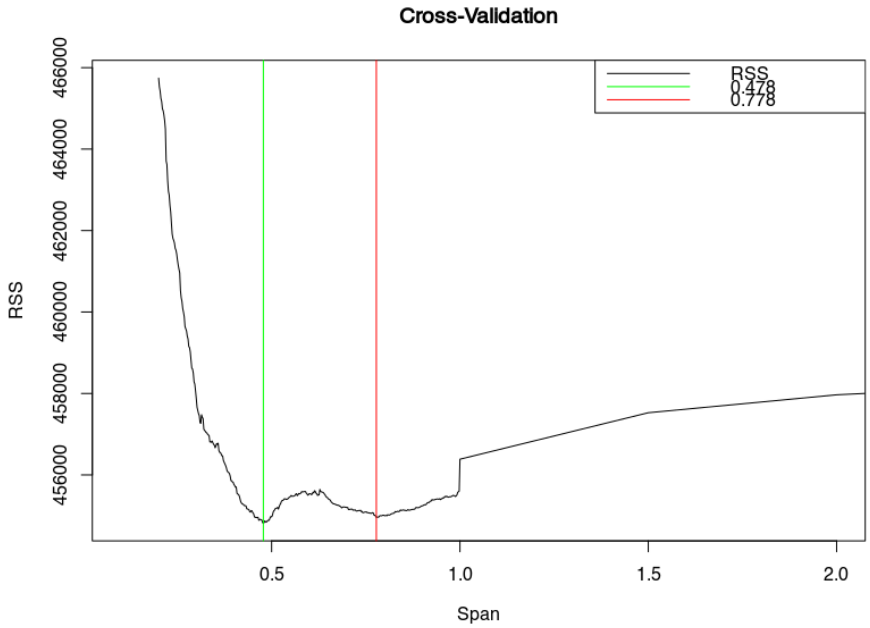


Figure 9: Residuals sum of squares for different span widths

However cross-validation is not set in stone, thus additional two fits are made using 0.628 and 0.928 span width. As seen in figure 10 a span width of 0.928 clearly under-fit much more than 0.778, while 0.478 have significant over-fitting. As the CV-value for 0.778 is better than 0.628, a span width of 0.778 is chosen.

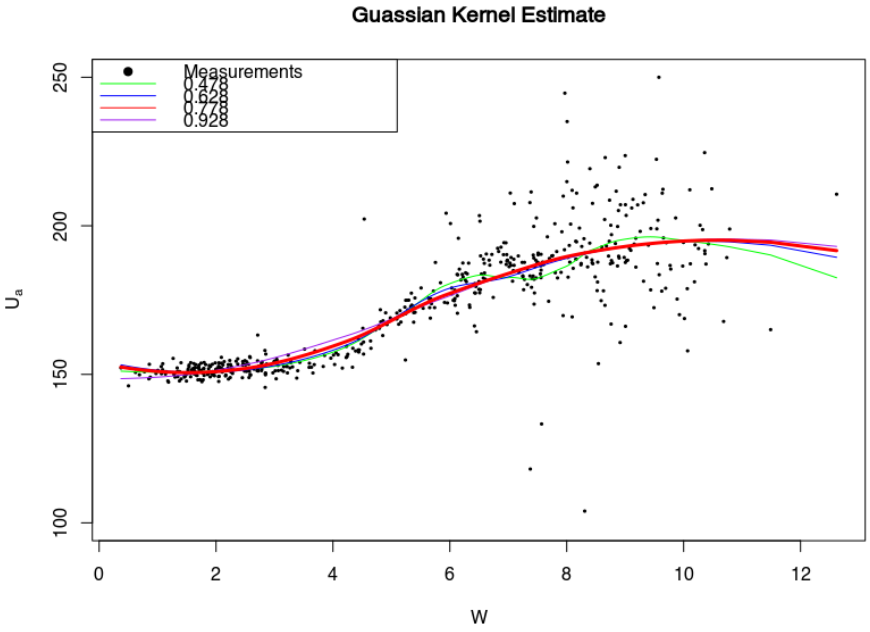


Figure 10: Non-parametric fit of the heat-loss coefficient as a function of wind speed

In figure 10 a nice smooth fit, with little under-fitting is achieved. As can be seen at high wind speeds the $U_a$-value drops as speeds increase, which is unexpected. This is likely due to the relatively low amount of data points, with a high spread at these wind speeds. Clearly the non parametric estimate is better than the assumed constant, however the computation is very demanding.

## Epanechnikov Kernel Estimate

As to achieve better performance an Epanechnikov kernel estimate is carried out. This will enable a better understanding of the cost of the performance increase.

Just like the Guassian estimate a cross-validation check for best span is performed. As indicated in figure 11 the smallest RSS values are achieved at 0.931.

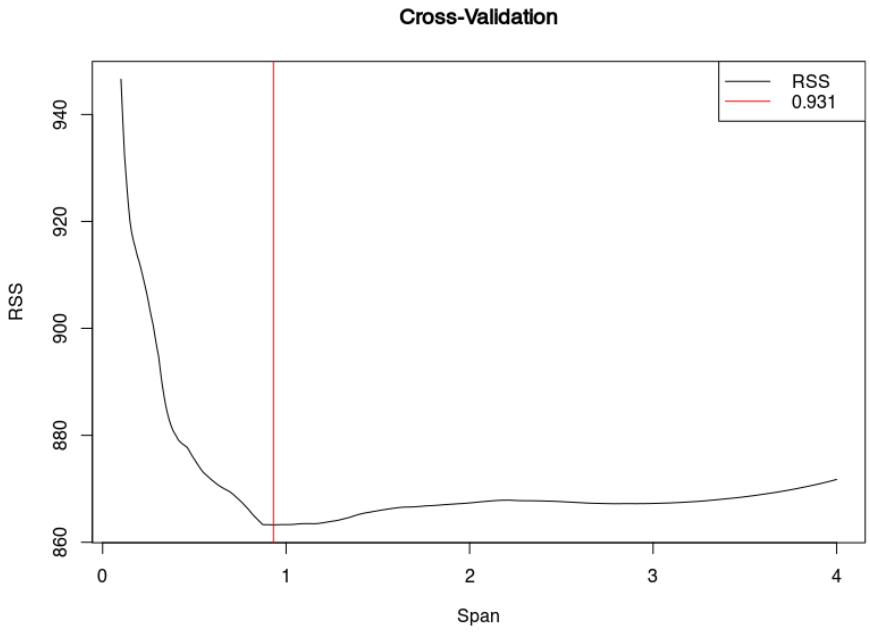


Figure 11: Residuals sum of squares for different span widths

Further the increase in RRS-value is relatively small for increasing span widths. In figure 12 four fits are plotted, where it is very clear that the RSS-result from above over-fit to a very high degree. In conclusion the fit with span 2.5 is more preferable, as it is the best compromise between smoothness and under-fitting.

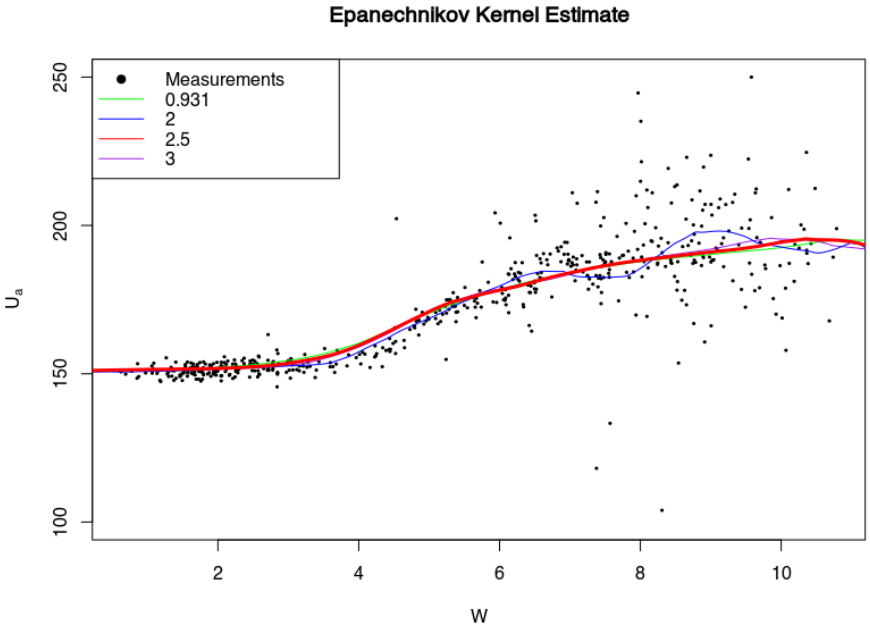


Figure 12: Non-parametric fit of the heat-loss coefficient as a function of wind speed

Comparing the Gaussian- and Epanechnikov fits, the Gaussian does a better job, as it is smoother and does not

under-fit as much as the selected Epanechnikov fit. However the differences are subtle, and the performance of the Epanechnikov kernel estimate computations are an order of magnitude faster for this data set.

# Part 5

In order for us to fit an ARMA model to the data, we must first determine the number of AR and MA terms. We therefore start by looking at the autocorrelation function and the partial autocorrelation function.

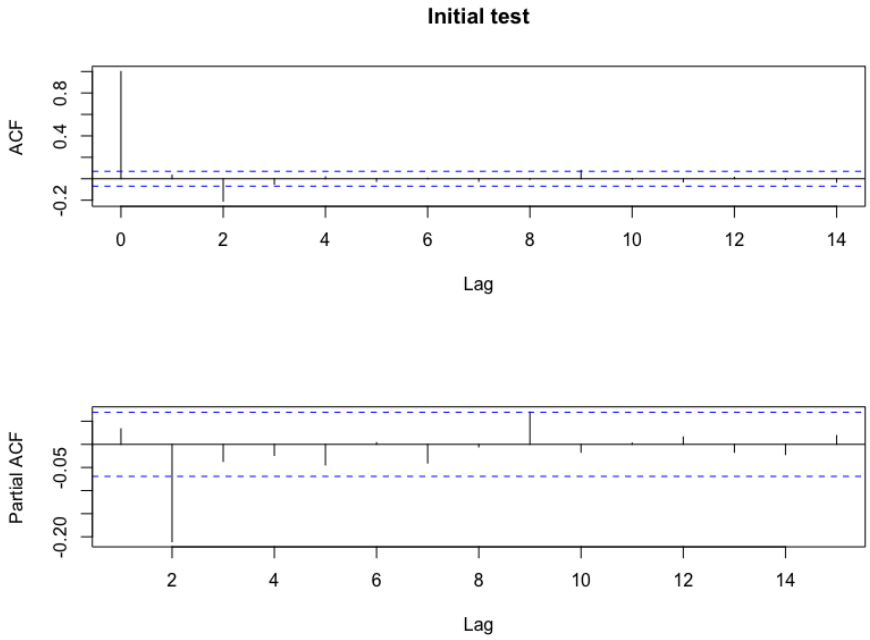


Figure 13: Initial ACF and PACF of the time series data.

We see a signal at lag 2 for both the ACF and PACF, therefore we propose an initial (0,0,0)(1,0,0)[2], i.e. a seasonal AR(1) with seasonality 2. In other words we use the initial linear model given by:

$$X_t = \phi_1 X_{t-2} + \epsilon_t \tag{20}$$

We will now look at the ACF and PACF again to see if there is a need for other terms.

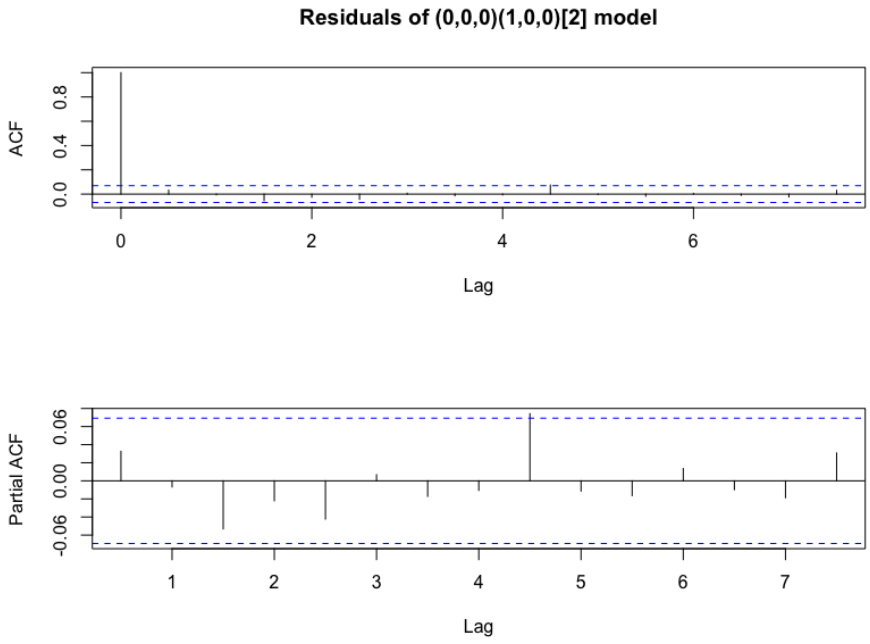


Figure 14: ACF and PACF of the model given in line 20.

We can now see that there are no obvious signals but before we can be sure let us look at the Ljung Box test and a qq-plot of the residuals.

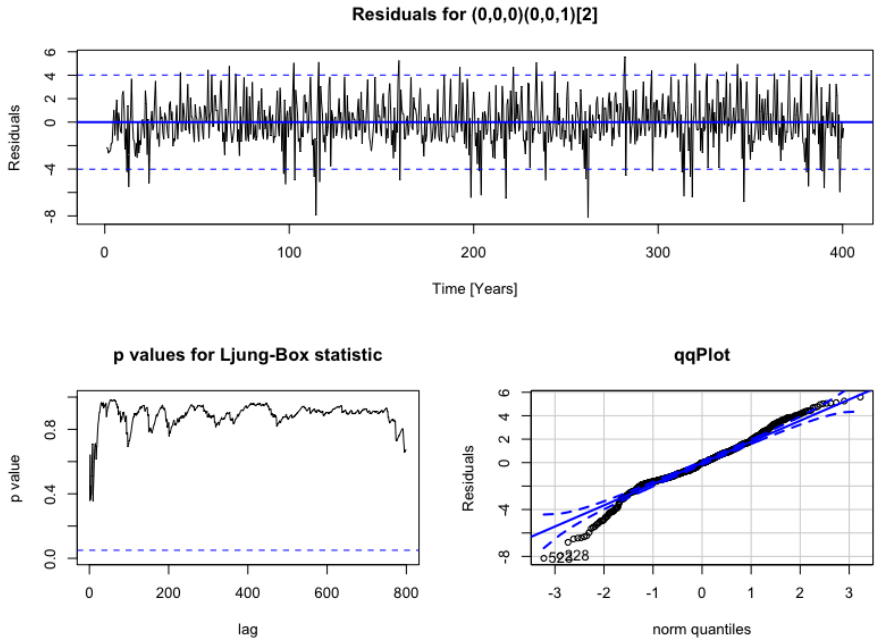


Figure 15: Residuals of the model given in line 20.

From figure 15 we can see that the residuals look relatively white. It does look like the residuals have heavy tails, i.e. there are a concerning amount of outliers especially in the negative direction. However, since the ACF and PACF

are both neutral there is nothing more we can do using an ARMA model.
We will now compute the Lag Dependency Function (LDF) of the residuals from lag 1 to lag 6 to look for non-linear dependencies.

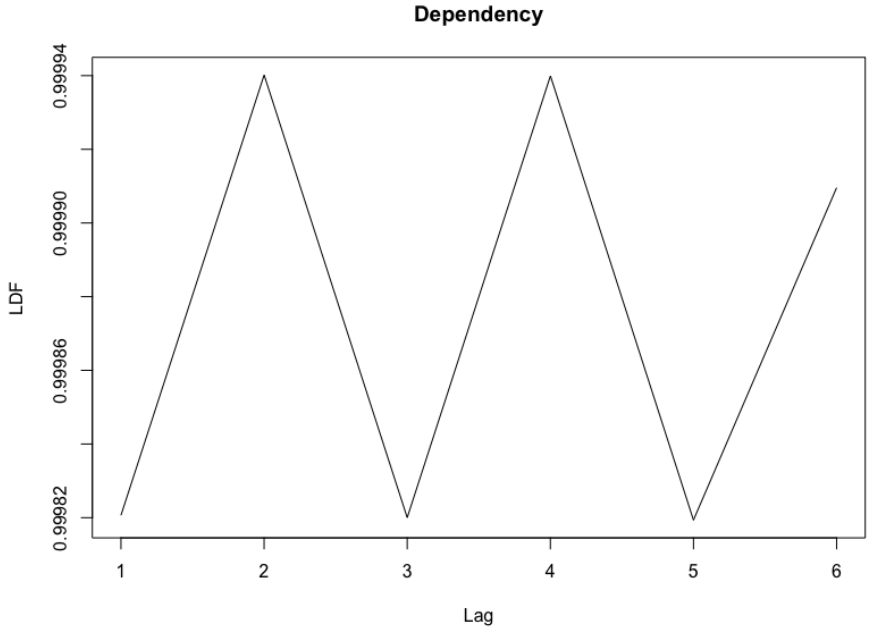


Figure 16: LDF of residuals.

From figure 16 we can see a spike in the LDF value at lag 2 and 4. This indicates that there is some dependency between $R_t$ and $R_{t-2}$, where $R_t$ is the residual at time t, and/or between $R_t$ and $R_{t-4}$. Let us start by plotting $R_t$ as a function of $R_{t-2}$:

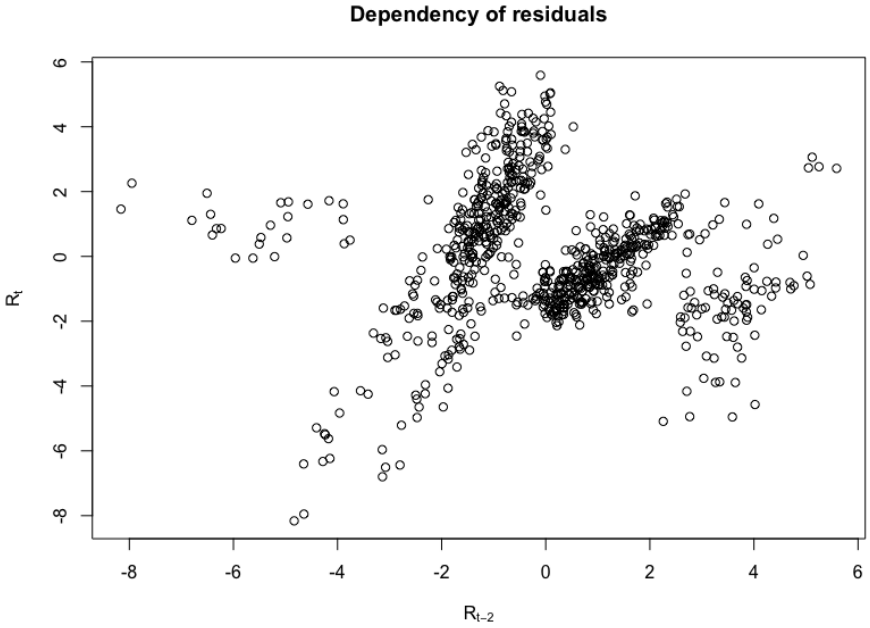


Figure 17: $R_t$ as a function of $R_{t-2}$.

In figure 17 we can see that the residuals seem to follow multiple linear models depending on the interval of $R_{t-2}$. This behaviour looks a lot like the behavior of a SETAR model. It could therefore be interesting to plot $X_t$ as a function of $X_{t-2}$, i.e. the same plot but using the original data rather than the residuals of the ARMA model.

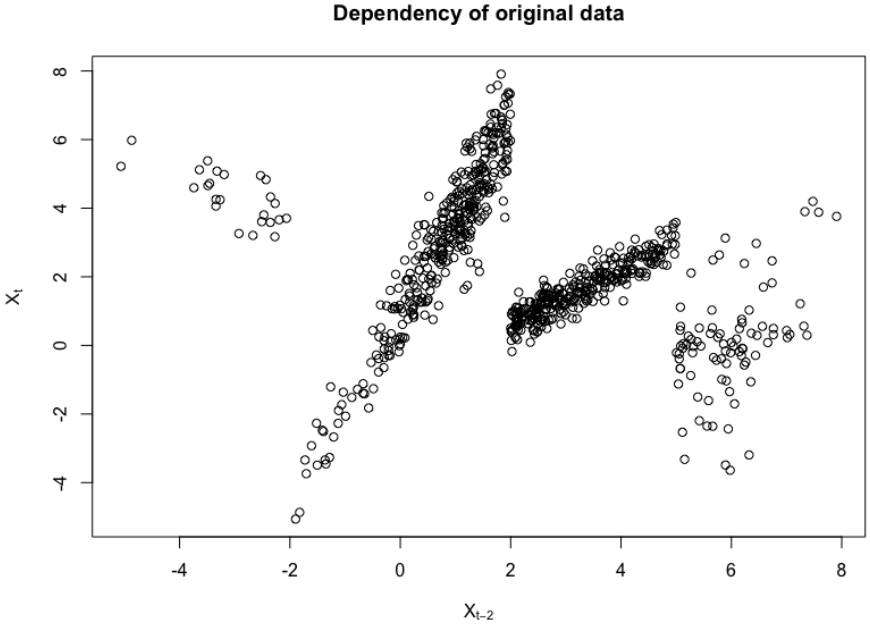


Figure 18: $X_t$ as a function of $X_{t-2}$.

From this it is quite clear that the data follows a SETAR model. It seems there are 4 paradigms, the $X_t$ is dependent

of $X_{t-2}$, and each model is of 1 order, i.e. a linear model. This means a SETAR(4,2,1) model structure, which is the model structure we propose. We can now try to approximate this model using a non-parametric method, and try to compute the individual parameters so we can compare the two. We compute the parameters in the SETAR model using simple least squares regression. It is important to note that parameters for the outer paradigms are quite uncertain since they contain few measurements. We find the model to be:

$$X_t = \begin{cases} -0.680X_{t-2} + 2.281 + \epsilon_t & X_{t-2} < -2 \\ 2.7342X_{t-2} + 0.9009 + \epsilon_t & -2 \leq X_{t-2} < 2 \\ 0.804X_{t-2} - 1.002 + \epsilon_t & 2 \leq X_{t-2} < 5 \\ 1.020X_{t-2} - 5.996 + \epsilon_t & 5 \leq X_{t-2} \end{cases} \tag{21}$$

We can now plot the SETAR model together with the non-parametric estimate:

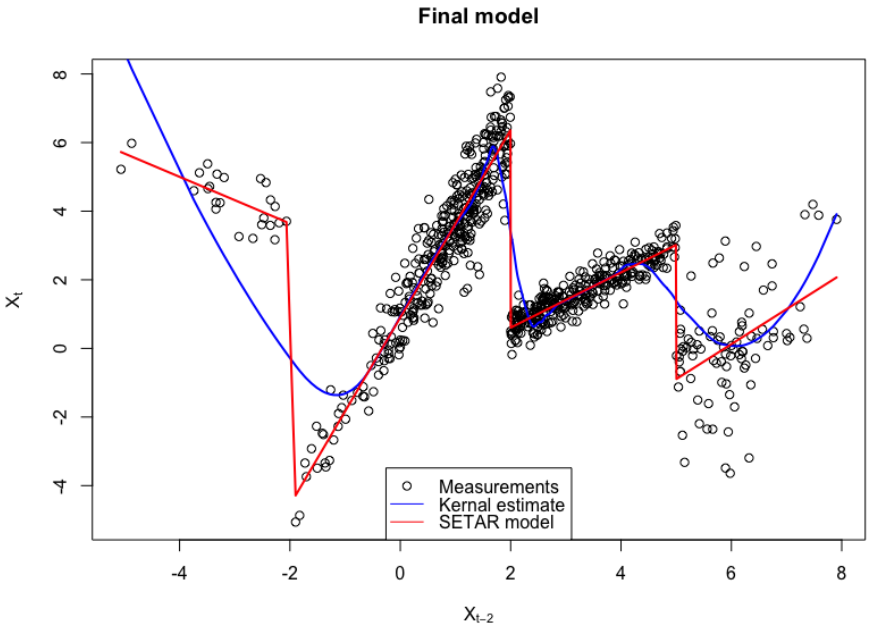


Figure 19: Estimated SETAR model and non-parametric approximation.

Visually it seems that the SETAR fits the data very well. The kernal estimate does an OK job of approximating the model, but have some relative large issues when the regime change is large, and especially the change between the 1st and 2nd regime.