02417: Time Series Analysis

Lasse Engbo Christiansen
Fall 2019

DTU Compute
Technical University of Denmark

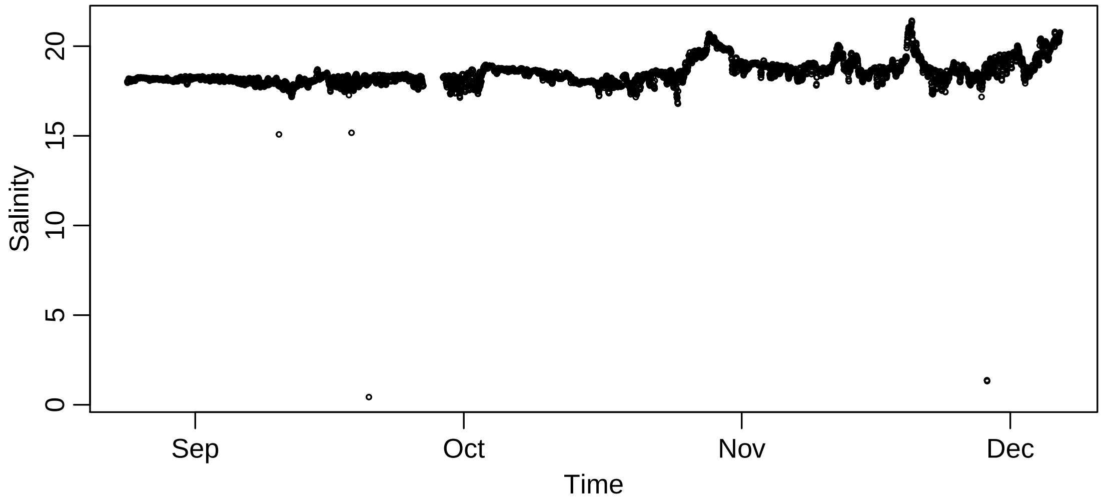
Assignment 4: Kalman Filter

Sensor platforms that can be deployed for extended periods of time and preferably sending real time measurements is a great tool. Some sensors requires regular maintenance which may be rather costly due to the time spent. This assignment will focus on a bouy that is located in the water near Kulhuse - in opening of Roskilde Fjord.
Every half an hour the sensor platform records the following:

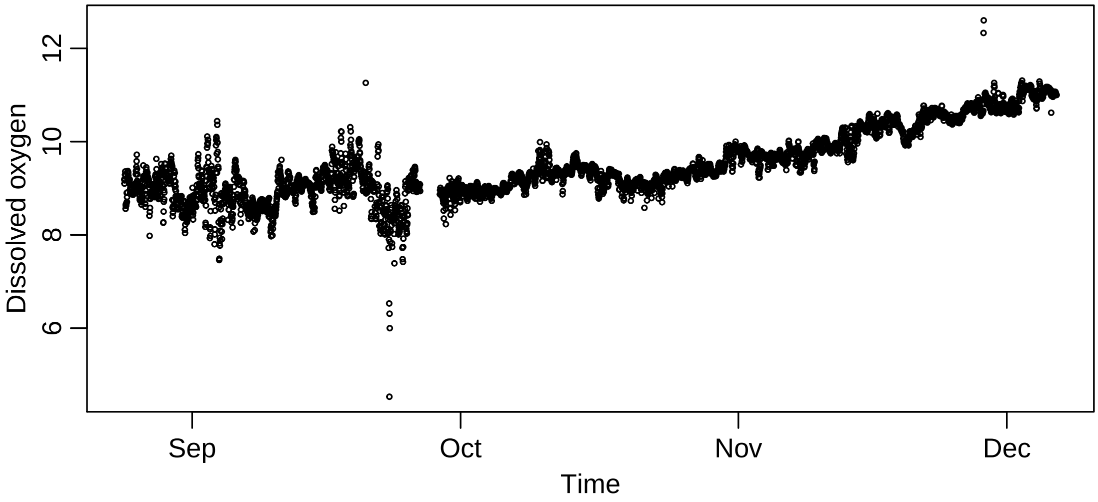
- `Temp` is the water temperature [degC]
- `Sal` is the water salinity (Amount of salt) [PSU] = [g/kg]
- `Depth` is the depth of the sensor platform [m]
- `pH` is the water pH
- `Chl` is the water cholorphyll concentration [mg/L]
- `ODOsat` is the percent dissolved oxygen saturation [%]
- `ODO` is the dissolved oxygen [mg/L]
- `Battery` is the voltage of the battery in the sensor platform [V]
- `DateTime` is the time stamp

Measurements from a period of time is made available in the file `A4_Kulhuse.csv`.
This assignment will only look into the salinity and the dissolved oxygen. The focus is on an initial filtering to remove outliers in the data.

Question 4.1: Presenting data *Load the data and present it by making some plots for salinity and dissolved oxygen. Comment on what you see.*



In the above plot the measured salinity is plotted as a function of time. It is clearly seen that there are a few outliers with low values in short periods. It is also noticed that there is a gap in the data around the beginning of October. The rest can be considered as natural variation due to mixing of water bodies with different salinities.



The plot above shows the evolution of the concentration of dissolved oxygen. In the first half of the data the process seems more volatile than in the last half. However, the last half shows a near linear increase. Again there are a few outliers and the same gap as was seen for the salinity.

(These plots can also be made with lines - it is recommended to look at both. Here it was chosen to plot points to emphazise that the outliers are single points.)

Here (the expected) it is chosen not to include further plots, but it would make sense to show a zoom that shows the daily variations (E.g. showing data for a couple of weeks) - *count that as constrictive extra in the end*.

Question 4.2: Random walk state-space model *To begin with the focus is on the salinity. At first it is assumed that the salinity can be described by a random walk process that is observed at equidistant time points.*

Write the system as a state-space model according to section 10.1. It is sufficient if you define the relevant matrices.

This model is a univariate state space model with univariate observations. The task is to define the following five matrices:

$$\begin{aligned} \mathbf{A} &= [1] & (1)\\ \mathbf{B} &= \text{NULL (no input)} & (2)\\ \mathbf{C} &= [1] & (3)\\ \Sigma_1 &= [\sigma_1^2] & (4)\\ \Sigma_2 &= [\sigma_2^2] & (5) \end{aligned}$$

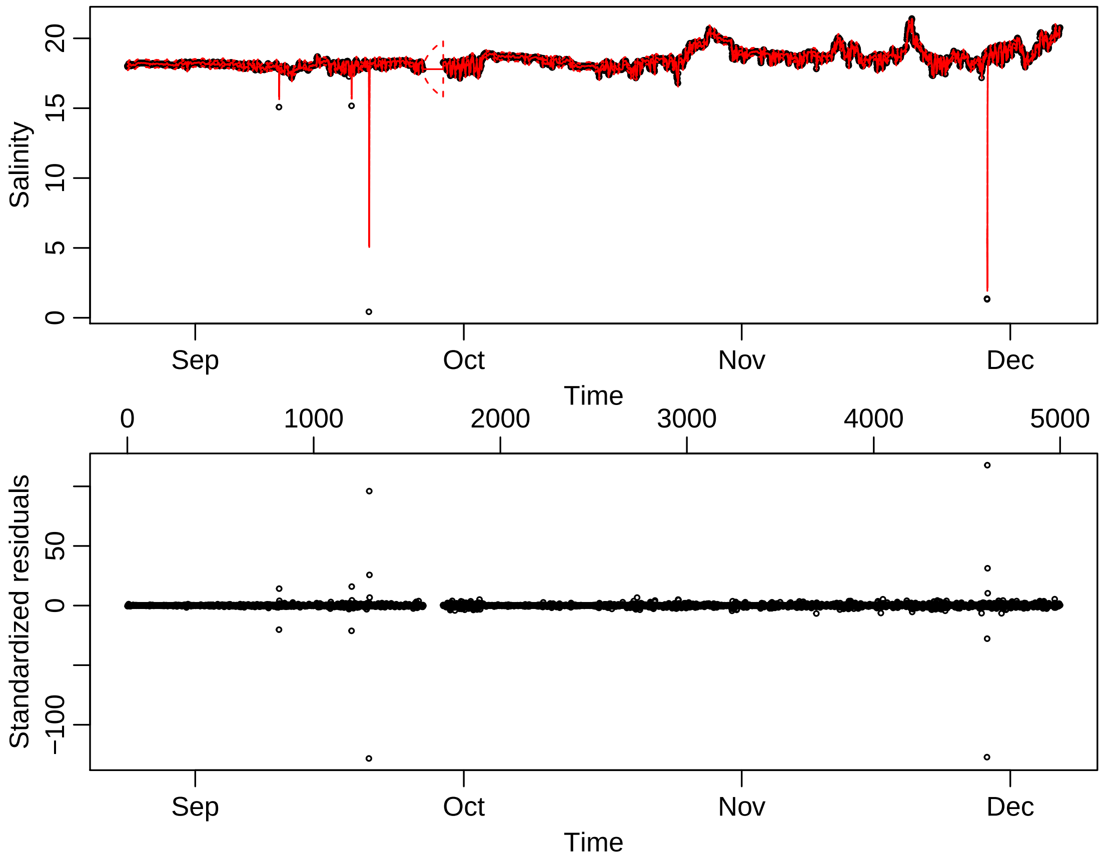
The $\mathbf{B}$ can be mentioned in several ways. It should somehow be mentioned that there is no input in this model.

Question 4.3: Pure Kalman filter *Filter the salinity with a Kalman filter (Either your own implementation or a build in function). Use the first observation as initial value and use the* *system variance as initial variance. Fix the system variance to* 0.01 *and the observation variance to* 0.005*.*

1. Plot the one step predictions along the data. Do include 95% PI.
2. Plot the standardized one step prediction errors. (Prediction errors normalized with the standard error of the prediction.)
3. Repeat the two plots with the same content but zooming into observations 800 to 950.
4. Report the values that defines the final state of the filter (at observation 5000).

Comment on what you see.

Q4.3.1-2 Presenting the filtered observations

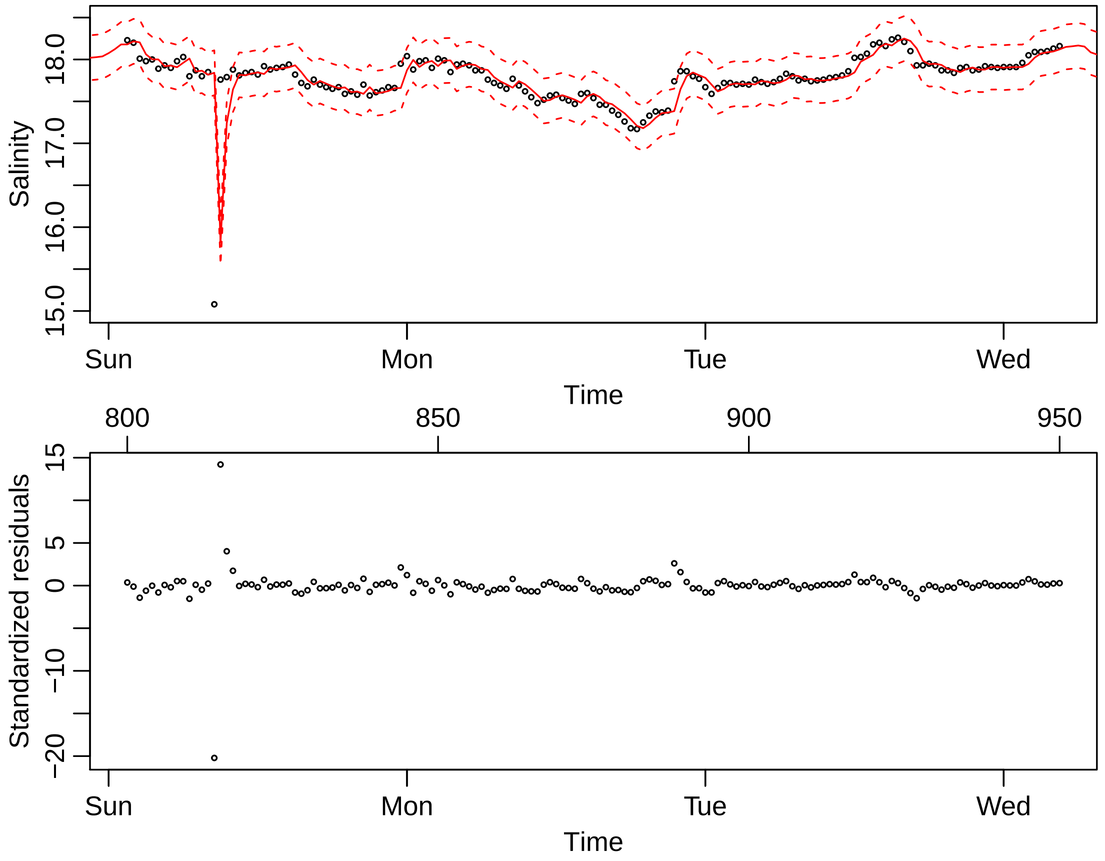


The top shows the data and the 95 % PI. It is not easy to see how the prediction interval performs. What can be seen is that the large outliers (Values below 15) drags the predictions far away from the non-outliers. It is also noticed that the large gap in the end of September leads to an increase in the prediction variance.

The bottom plot shows the standardized one step prediction errors. It should be noticed that each of the outliers causes a large negative residual followed by a slightly smaller positive residual (need to get back to the normal range).

The x-axis is given both as months and as observation index - either is fine.

Q4.3.3 Zoom



Same setup as in the previous plot. One outlier is observed and here it is easily seen that the outlier causes the prediction of the following observation to be too low and thus resulting in a positive residual - actually it takes a couple of observations before the residuals are as before (this is easier to see in the bottom plot.)

Except for the outlier then the PI includes almost all observations so it is probably a little too wide.

Q4.3.4 Final state of the filter

The final state can be given as either the last reconstruction: $\hat{X}_{5000|5000} = 20.76$ and $\hat{\Sigma}^{xx}_{5000|5000} = 0.00366$

or as the prediction of observation 5001: $\hat{X}_{5001|5000} = 20.76$ and $\hat{\Sigma}^{xx}_{5001|5000} = 0.01366$.

It is OK if Y is also included but it is the second order moment representation of the state that is expected.

Question 4.4: Skipping outliers when filtering *Some observations are very unlikely. Make your own implementation of the Kalman filter. The implementation should treat observations that are more than six standard deviations away from the prediction as missing (as in don't perform a reconstruction step at these points in time).*

Include the code for your filter in the main report. (Include comments in your code)

Filter the data with your version of the filter.

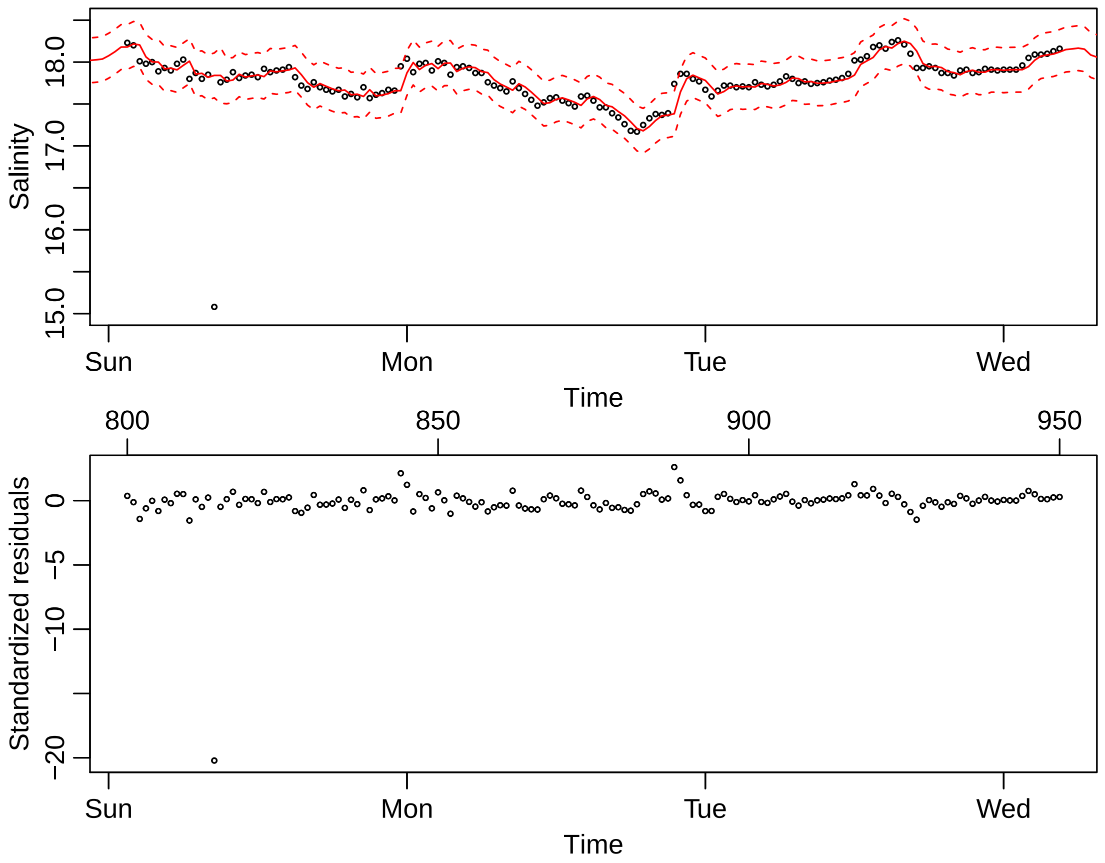
1. Present the 1-step predictions for index 800 to 950 as in question 4.3.3.
2. Report the indexes for the first five detected outliers (observations that are skipped because they are more than six standard deviations away).
3. Do also report the number of observations that are skipped.
4. Report the values that defines the final state of the filter (at observation 5000).

Comment on your results

Q4.4.0 Implementation

```
## This implementation is overly simplified as A=C=1 is used to
## simplify the code.
## However, it still returns a lot of information
kalmanS <- function(Y, Sigma.1, Sigma.2, threshold = Inf){
  ##
  V0=Sigma.1
  Xhat0=Y[1]

  ## Y has to be a column matrix.
  if(class(Y)=="numeric"){
    Y <- matrix(Y,ncol=1)
  }
  dim.Y <- dim(Y)

  ## Initializing variables
  X.hat <- Xhat0
  Sigma.xx <- V0
  Sigma.yy <- Sigma.xx + Sigma.2

  ## for saving everything - initializes as NAs per default
  X.rec <- array(dim=c(dim.Y[1] , 1))
  X.pred <- array(dim=c(dim.Y[1] + 1, 1))
  K.out <- array(dim=c(dim(Sigma.xx%*%solve(Sigma.yy)), dim.Y[1]))
  Sigma.xx.rec <- array(dim=c(dim(Sigma.xx), dim.Y[1] ))
  Sigma.xx.pred <- array(dim=c(dim(Sigma.xx), dim.Y[1] + 1))
  Sigma.yy.pred <- array(dim=c(dim(Sigma.yy), dim.Y[1] + 1))

  ## Running the filter
  for(tt in 1:dim.Y[1]){
    eps <- (t(Y[tt,])- as.matrix(X.hat))/sqrt(Sigma.yy)
    ## Reconstruction
    if( (!is.na(eps)) & abs(eps)<threshold ) {
      K <- Sigma.xx%*%solve(Sigma.yy)                 ## (10.75)
      X.hat <- X.hat + K%*%(t(Y[tt,])- as.matrix(X.hat))##(10.73)
      X.rec[tt,] <- X.hat
      Sigma.xx <- Sigma.xx - K%*%Sigma.xx             ## (10.74)
      Sigma.yy <- Sigma.xx + Sigma.2
      K.out[,,tt] <- K
    }
    Sigma.xx.rec[,,tt] <- Sigma.xx

    ## Predicting
    X.pred[tt+1,] <- X.hat    ##(10.76) Here X.pred == Y.pred
    Sigma.xx <- Sigma.xx + Sigma.1               ##(10.77)
    Sigma.yy <- Sigma.xx + Sigma.2               ##(10.78)

    #### these are the prediction error variance-covariances
    Sigma.xx.pred[,,tt+1] <- Sigma.xx
    Sigma.yy.pred[,,tt+1] <- Sigma.yy
  }
  out <- list(rec=X.rec, pred=X.pred, K=K.out,
              Sigma.xx.rec =  Sigma.xx.rec,
              Sigma.xx.pred=Sigma.xx.pred,
              Sigma.yy.pred=Sigma.yy.pred)
  return(out)
}
```

Q4.4.1 Presenting zoom



It is seen that the filter now skips the outlier. Observations that are not labeled as outliers can now be used for further modelling.

Q4.4.2-3 Outliers

Ten outliers are detected. You are only asked to provide the first five but here all are presented:

```
##  [1]  814 1203 1296 2733 3692 4038 4578 4607 4609 4686
```

Q4.4.4 Final state of the filter

The final state can be given as either the last reconstruction: $\hat{X}_{5000|5000} = 20.76$ and $\hat{\Sigma}^{xx}_{5000|5000} = 0.00366$

or as the prediction of observation 5001: $\hat{X}_{5001|5000} = 20.76$ and $\hat{\Sigma}^{xx}_{5001|5000} = 0.01366$.

It is worth noting that the final state is the same when rounded. And checking the difference R returns a zero meaning that it is less machine precision.

Question 4.5: Optimizing the variances *Maximum likelihood estimation of the two variance parameters.*

1. What is a sensible lower bound for the observation variance?
2. Find the ML estimates of the two parameters using the first 800 observations.
3. Filter the data with the optimal parameters.
4. Plot as in Q4.3.3 for observations 800 through 950.
5. Report the values that defines the final state of the filter (at observation 5000).

Comment on your results

Q4.5.1 Lower bounds

Σ_1 should be different from zero to allow the state to change as a random walk. However, a lower bound is not well defined. (*And you are not asked about* Σ_1.)

Σ_2 on the other hand should have a lower bound that reflects the precision of the measurements. The performance of the sensor is not given but it is recorded with two decimals so that rounding to two decimals should be reflected as a lower bound.

Any value that is recorded, say 18.03 is actually 18.03 ± 0.005. The question is how to translate ± 0.005 to a lower bound for the observation variance. Three arguments:

- Choosing 0.005 as one standard deviation in a normal distribution yields $0.005^2 = 2.5 \times 10^{-5}$
- Choosing 0.005 as two (or 1.96) standard deviations in a normal distribution yields $0.0025^2 = 6.25 \times 10^{-6}$
- It is fair to assume that the true values are uniform and thus it would make sense to use the variance of a uniform variable on an interval with a width of 0.01: $\frac{1}{12}0.01^2 \approx 8.33 \times 10^{-6}$

The largest is a factor of four greater than the smallest. So accept values within a factor of two of the ones presented here.

In reality one should always record numbers with enough significant digits to ensure that the measurement uncertainty dominates. Assuming that is done in the present case then it is chosen to proceed with the largest value: 2.5×10^{-5}.

Q4.5.2 ML

In this solution parameters are optimized in a log-transformed space. The optimal values are $[\Sigma_1, \Sigma_2] = [0.001839, 2.5 \times 10^{-5}]$.

	est.exp	est	sd.exp	lower	upper
Sigma1	-6.299E+00	1.839E-03	6.420E-02	1.621E-03	2.086E-03
Sigma2	-1.060E+01	2.500E-05	1.537E+00	1.230E-06	5.082E-04

If the lower bound had been set very low then the optimal values are $[\Sigma_1, \Sigma_2] = [0.001887, 2.402 \times 10^{-9}]$. So Σ_1 is almost invariant to Σ_2 being reduced quite a bit.

	est.exp	est	sd.exp	lower	upper
Sigma1	-6.273E+00	1.887E-03	5.004E-02	1.710E-03	2.081E-03
Sigma2	-1.985E+01	2.402E-09	3.617E+02	3.165E-317	1.823E+299

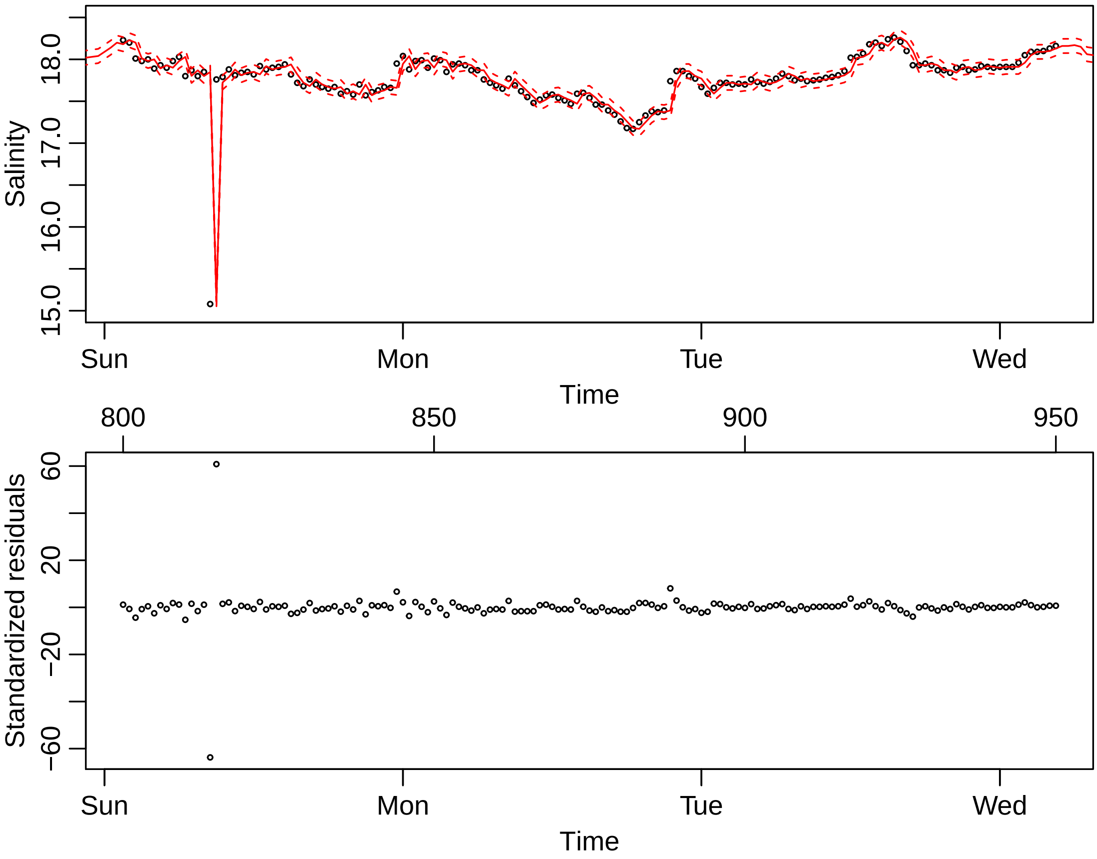
It is noticed that Σ_2 is no longer identifiable when it is allowed to be much smaller than the precision in the provided observations. This does make some intuitive sense.

Both of the above should be considered correct. It is not expected that you get exactly the same numbers so judge what is provided. If in doubt then give a too low score and write a note like ”Consider flagging this”.

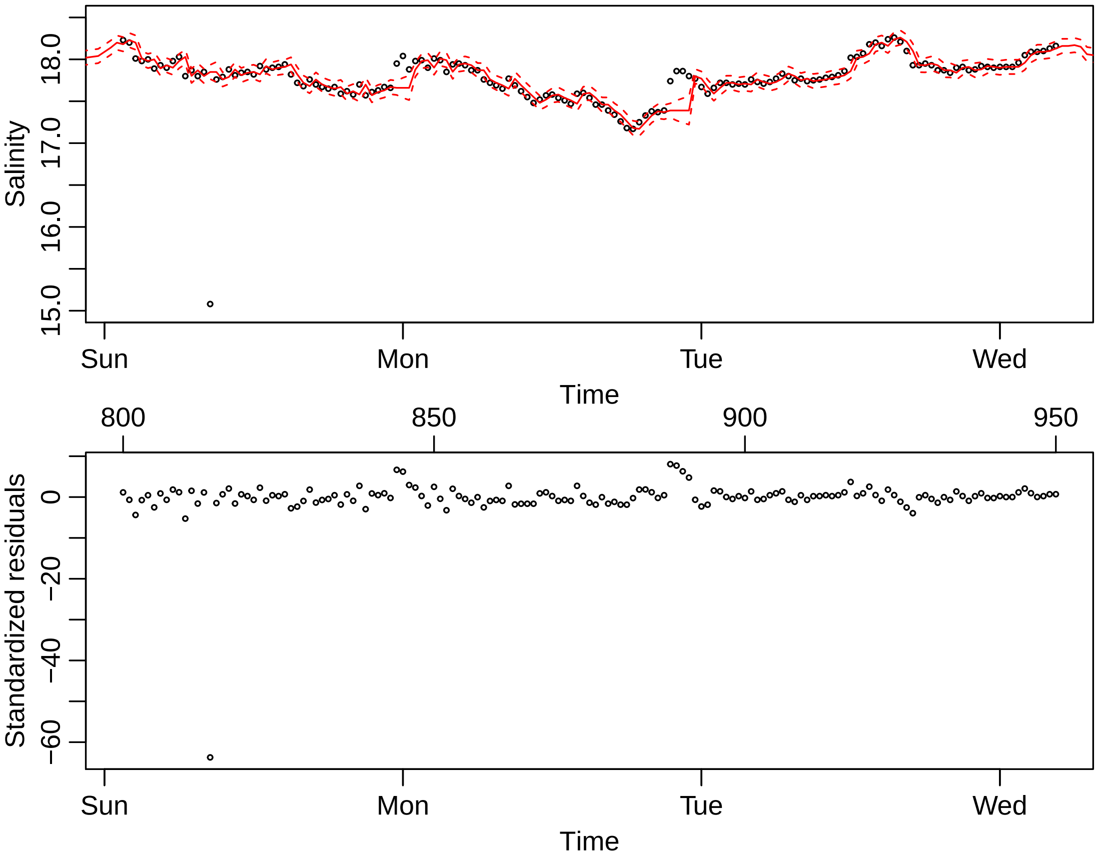
Q4.5.4 Presenting for a zoom

It is not given if this should done with a threshold at six sigma so both will be presented here.

First without a threshold:



The main differences relative to Q4.3.3 are that the PI in general is narrower. It is also relevant to mention that the impact of the outlier is reduced to two observations.



The main difference relative to Q4.4.1 is that the PI in general is narrower. It is also noticed that several additional observations are labeled as outliers.

With these parameters 463 outliers are detected and six of those are in the interval from 800 to 950.

You are not asked for the following: However, if you look closer at the entire data then you can see that the variance seems smaller in the beginning (where we optimized) and is slightly higher in the later part. So it would make sense to estimate parameters based on a larger part of the data.

Q4.5.5 Final state of the filter

The final state can be given as either the last reconstruction: $\hat{X}_{5000|5000} = 20.77$ and $\hat{\Sigma}^{xx}_{5000|5000} = 2 \times 10^{-5}$

or as the prediction of observation 5001: $\hat{X}_{5001|5000} = 20.77$ and $\hat{\Sigma}^{xx}_{5001|5000} = 0.00186$.

It is worth noting that the final state is almost same when rounded. However, the variance of the state is much smaller for the optimal parameters.

Question 4.6: Model for dissolved oxygen *The dissolved oxygen concentration varies during the day due to oxygen production (Called primary production) from chlorophyll and sunlight and oxygen consumption from all biological matter (Called respiration). Furthermore, there is exchange of oxygen with the atmosphere so that the dissolved oxygen concentration approaches the saturation concentration.*

The saturation concentration for dissolved oxygen depends on the temperature and the salinity. Here you can treat it as a known input.

Specify a state-space model for dissolved oxygen that includes primary production as a linear function of the sun intensity and exchange of oxygen with the atmosphere. The model should include respiration as a random walk which is to be used to identify the amount of biomass.

Use the following naming of variables in your model formulation:

DO_t *for dissolved oxygen at time t*
I_t *for intensity of sunlight at time t*
$DOsat_t$ *for saturation concentration of dissolved oxygen at time t*
R_t *for respiration at time t*

This model is a bivariate state space model with univariate observations.

First, we'll state the anticipated state equations for DO_t and R_t:

$$\begin{aligned} DO_t &= DO_{t-1} + a_{atm}(DOsat_{t-1} - DO_{t-1}) + a_I I_{t-1} + R_{t-1} + \varepsilon_{DO,t} && (6) \\ R_t &= R_{t-1} + \varepsilon_{R,t} && (7) \end{aligned}$$

Defining the state as

$$X_t = \begin{bmatrix} DO_t \\ R_t \end{bmatrix}$$

and the input as

$$u_t = \begin{bmatrix} I_t \\ DOsat_t \end{bmatrix}$$

The the observation equation becomes

$$Y_t = \begin{bmatrix} 1 & 0 \end{bmatrix} X_t + \varepsilon_{2,t}$$

The task is to define the following five matrices:

$$\begin{aligned} \mathbf{A} &= \begin{bmatrix} 1 - a_{atm} & 1 \\ 0 & 1 \end{bmatrix} && (8) \\ \mathbf{B} &= \begin{bmatrix} a_I & a_{atm} \\ 0 & 0 \end{bmatrix} && (9) \\ \mathbf{C} &= \begin{bmatrix} 1 & 0 \end{bmatrix} && (10) \\ \Sigma_1 &= \begin{bmatrix} \sigma^2_{DO} & 0 \\ 0 & \sigma^2_R \end{bmatrix} && (11) \\ \Sigma_2 &= [\sigma^2_2] && (12) \end{aligned}$$

It is OK to use a different ordering of the states and inputs. It is also OK to use other names for the parameters that are different from zero and one.

It is important to have the correct coupling between DO_{t-1} and $DOsat_{t-1}$. Therefore, a_{atm} should be both in the A and B matrices with opposite signs.

It is not possible to both identify σ^2_R and a multiplyer for how R_{t-1} influences DO_t. Therefore, $A_{1,2}$ should be one but it is a minor error if an extra constant is introduced there.

You don't have any information about the off diagonal elements of Σ_1 so one would typically start with zeros but it is OK if these are given as non-zero.