Lab6

Andreas Stasinakis & Mim Kemal Tekin

February 28, 2019

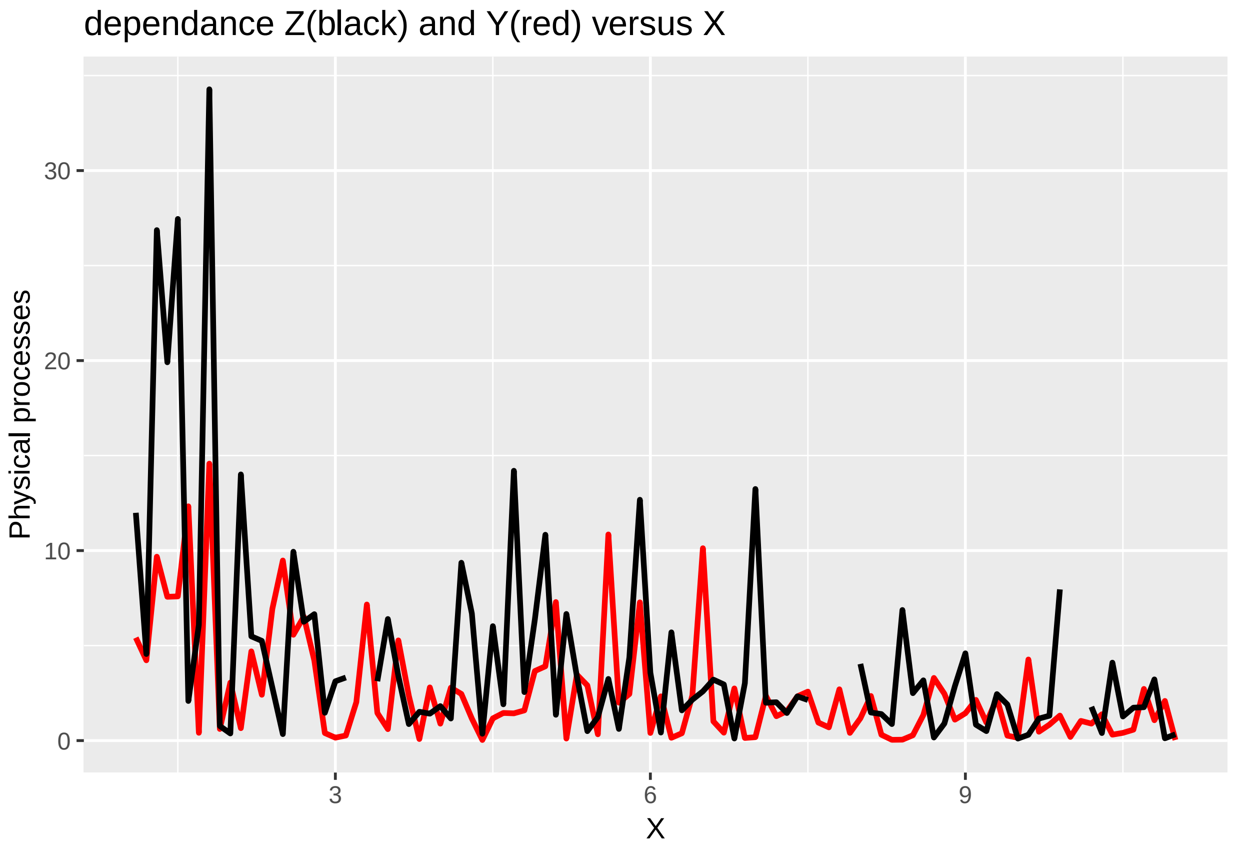
Question 2: EM algorithm (Andreas Stasinakis)

The data file physical.csv describes a behavior of two related physical processes $Y = Y(X)$ and $Z = Z(X)$.

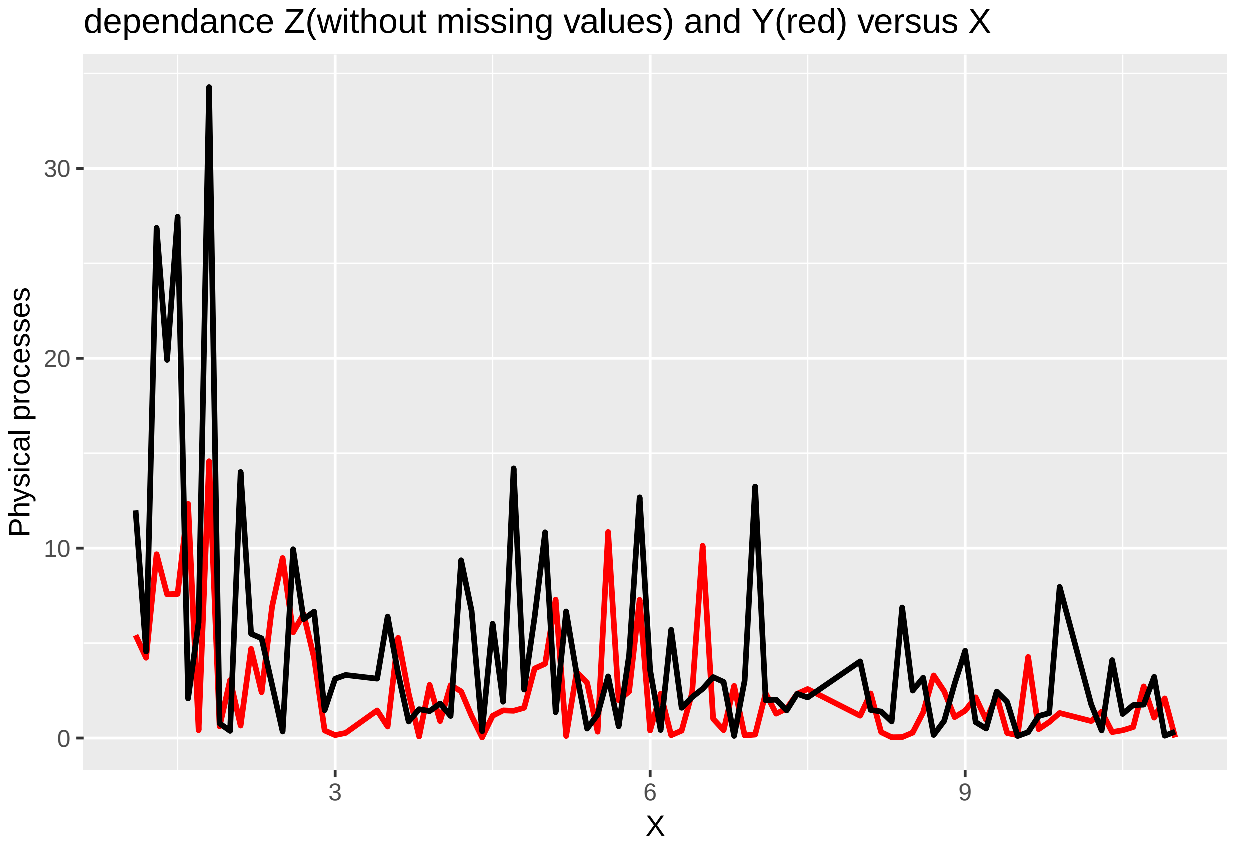
2.1 Import csv file, time series plot.

Make a time series plot describing dependence of Z and Y versus X. Does it seem that two processes are related to each other? What can you say about the variation of the response values with respect to X?

```r
library(ggplot2)
physical = read.csv("../physical1.csv")

#time series plot for the dependence Z,Y versus X

ggplot(physical) +
  geom_line(mapping = aes(x = physical$X,y = physical$Y),color = "red", size =1)+
  geom_line(mapping = aes(x = physical$X,y = physical$Z),color = "black", size =1)+
  labs(title = "dependance Z(black) and Y(red) versus X ",x = "X", y = "Physical processes")
```



```
#exclude the NA observations
data = physical[-c(which(is.na(physical$Z))),]

ggplot(data) +
  geom_line(mapping = aes(x = data$X,y = data$Y),color = "red",size = 1)+
  geom_line(mapping = aes(x = data$X,y = data$Z),color = "black", size =1 )+
  labs(title = "dependance Z(without missing values) and Y(red) versus X ",
       x = "X", y = "Physical processes")
```



In this assignment we have a data set of two related physical processes Y, Z which are a function of variable X. The problem is that in one of them, Z, some missing values occur. In order to understand better the data, we plot the Z and Y variable versus the X variable. In the first plot we use all the data points, in contrast to the second one in which we use only the observations of the non missing values. From the plot we can commend that both X and Y seem to follow the same trend. They increase in the beginning, decreasing after reaching their pick and finally fluctuate between 0 and 5. One other problem we can observe is that the data seems really noise for both X, Y. The fluctuations, especially for $x < 7$, have really high variation. For example, for $x = 5.5$, $y = 0.33$ but for the next value of x the value of y boosted close to 10.85.

2.2 EM step E and M by hand.

Note that there are some missing values of Z in the data which implies problems in estimating models by maximum likelihood. Use the following model

$$Y_i \sim exp(X_i|\lambda), Z_i \sim exp(X_i|(2\lambda))$$

where λ is some unknown parameter.

The goal is to derive an EM algorithm that estimates λ.

We now want to maximize the likelihood of the parameter λ. The problem is that there are some missing values of Z which causes us troubles. For that reason we use the Expectation Maximization(EM) algorithm in order to estimate λ. The algorithm consists of two steps. The E -step and the M -step.

Before starting the steps, we have to do some pre process. We split the Z variable into Z^{obs} and Z^{un}, which is the a split bettween tha observed and the missing values. Now we define the function we want to maximize. As mentioned before we can not work with the log likelihood, so we define the Expected log likelihood, in order to handle the missing values of Z.

Therefore, let

$$Q(\lambda, \lambda^k) = E\big[\text{loglik}(\lambda|Y, Z)|\lambda^k, Y, Z^{\text{obs}}\big] \Rightarrow$$

$$Q(\lambda, \lambda^k) = E\big[\text{loglik}(\lambda|Y)\big] + E\big[\text{loglik}(\lambda|Z^{\text{obs}})\big] + E\big[\text{loglik}(\lambda|Z^{\text{un}})|\lambda^k, Y, Z^{\text{obs}}\big]$$

.

The important think here is that for the non missing values, Y and Z^{obs}, we do not need the expected log likelihood but the log likelihood. So this formula can be also written as:

$$Q(\lambda, \lambda^k) = \text{logL}(\lambda|Y) + \text{logL}(\lambda|Z^{\text{obs}}) + E\big[\text{logL}(\lambda|Z^{\text{un}})|\lambda^k, Y, Z^{\text{obs}}\big]$$

.

As mentioned before, $Y_i \sim exp(\theta = X_i|\lambda), Z_i \sim exp(\theta = X_i|(2\lambda))$. So :

$$P(Y|\theta) = \theta e^{-\theta y} = \frac{x}{\lambda} e^{-\frac{x}{\lambda} y}$$

and

$$P(Z|\theta) = \theta e^{-\theta Z} = \frac{x}{2\lambda} e^{-\frac{x}{2\lambda} z}$$

.

Therefore,using the properties of the logarithm,

$$\text{logL}(\lambda|Y) = log \prod_{i=1}^{n} \frac{x_i}{\lambda} e^{-\frac{x_i}{\lambda} y_i} = \sum_{i=1}^{n} log(xi) - nlog\lambda - \frac{1}{\lambda} \sum_{i=1}^{n} xiyi$$

.

Also,

$$\text{logL}(\lambda|Z^{\text{obs}}) = log \prod_{i=1}^{r} \frac{x_i}{2\lambda} e^{-\frac{x_i}{2\lambda} Z^{\text{obs}}} = \sum_{i=1}^{r} log(xi) - rlog2\lambda - \frac{1}{2\lambda} \sum_{i=1}^{r} xiZ^{\text{obs}}$$

, where r is the observed Z.

Finaly, we have to estimate the expected log likelihood for the missing Z values.

$$E\big[\text{loglik}(\lambda|Z^{\text{un}})|\lambda^k, Y, Z^{\text{obs}}\big] = E\big[\sum_{i=1}^{n-r} log(xi) - (n-r)log2\lambda - \frac{1}{2\lambda} \sum_{i=1}^{n-r} xiZ^{\text{un}}\big] \Rightarrow$$

$$E\big[\text{loglik}(\lambda|Z^{\text{un}})|\lambda^k, Y, Z^{\text{obs}}\big] = E\big[\sum_{i=1}^{n-r} log(xi)\big] - E\big[(n-r)log2\lambda\big] - E\big[\frac{1}{2\lambda} \sum_{i=1}^{n-r} xiZ^{\text{un}}\big]$$

, because the expected value of the sum is the sum of the expected values. It is obvious that the expected value is only valid for the Z^{un}.

Therefore,

$$E\big[\text{loglik}(\lambda|Z^{\text{un}})|\lambda^k, Y, Z^{\text{obs}}\big] = \sum_{i=1}^{n-r} log(xi) - (n-r)log2\lambda - E\big[\frac{1}{2\lambda}\sum_{i=1}^{n-r} xiZ^{\text{un}}\big]$$

.

Moreover

$$E\big[\frac{1}{2\lambda}\sum_{i=1}^{n-r} xiZ^{\text{un}}\big] = \sum_{i=1}^{n-r} \frac{1}{2\lambda} xiE[Z^{\text{un}}]$$

.

In this case we use the expected value of the exponential distribution. So $E[Z^{\text{un}}] = \frac{2\lambda^k}{x_i}$, where λ^k is the parameter of the previous step which help us to estimate the new λ.

As a result,

$$E\big[\frac{1}{2\lambda}\sum_{i=1}^{n-r} xiZ^{\text{un}}\big] = \sum_{i=1}^{n-r} \frac{1}{2\lambda} xi\frac{2\lambda^k}{x_i} = (n-r)\frac{\lambda^k}{\lambda}$$

So the final formula for the function is:

$$Q(\lambda, \lambda^k) = \sum_{i=1}^{n} log(xi) - nlog\lambda - \frac{1}{\lambda}\sum_{i=1}^{n} xiyi + \sum_{i=1}^{r} log(xi) - rlog2\lambda - \frac{1}{2\lambda}\sum_{i=1}^{r} xiZ^{\text{obs}} + \sum_{i=1}^{n-r} log(xi) - (n-r)log2\lambda - (n-r)\frac{\lambda^k}{\lambda} \Rightarrow$$

Now it is time for the E-step. We want to compute the derivative of the function above with respect to λ.

$$\frac{Q(\lambda, \lambda^k)}{d\lambda} = -\frac{2n}{\lambda} + \frac{1}{\lambda^2}\Big[\sum_{i=1}^{n} xiyi + \sum_{i=1}^{r} x_i z_i^{\text{obs}} + (n-r)\lambda^k\Big]$$

For the M-step now, we have to find the λ which maximizes that function. So we need to set this derivative equal to zero.

$$\frac{Q(\lambda, \lambda^k)}{d\lambda} = 0 \Rightarrow$$

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} xiyi + \sum_{i=1}^{r} x_i z_i^{\text{obs}} + (n-r)\lambda^k}{2n}$$

2.3 Implement EM

Implement this algorithm in R, use $\lambda 0 = 100$ and convergence criterion -stop if the change in λ is less than 0.001. What is the optimal λ and how many iterations were required to compute it?

```
#function for findind the optimal lambda

EM.Norm<-function(Y,Z,X,eps,kmax,lambda_k){
  #input Y,Z,X all the data, Z has NA
  # eps is the threshold in order to stop
  #kmax is the number of iterations i want
  #lambda_k is the starting number of the parameter i want to estimate
  #Split the Z data in observed and unobserved

  Zobs = Z[!is.na(Z)]
  Zmiss = Z[is.na(Z)]
  n = length(Y)
```

```
 r = length(Zobs)
 k = 1 # We count the step

 lambda = (sum(X*Y) + sum((X[which(!is.na(Z))]*Zobs)/2) +(n-r)*lambda_k)/(2*n)

 while ((abs(lambda - lambda_k)>eps) && (k<(kmax+1))){
  lambda_k = lambda

   ## E-step
  # For this step we derive the expected log likelihood
  # We did it in the previous task, we do not need to implement it
  # We only need this formula for the M step

   ## M-step
   #Derivative equal to zero, so we have the estimator
   lambda = (sum(X*Y) + sum((X[which(!is.na(Z))]*Zobs)/2) +(n-r)*lambda_k)/(2*n)

   k = k+1

   print(list("lambda" = lambda,"lambda_k  "= lambda_k,k))
  }

  return(lambda)
}

optimal_lambda = EM.Norm(Y = physical$Y,Z =physical$Z,X = physical$X,eps = 0.001,kmax = 1000,lambda_k =
```

```
## $lambda
## [1] 10.83853
##
## $`lambda_k  `
## [1] 14.26782
##
## [[3]]
## [1] 2
##
## $lambda
## [1] 10.70136
##
## $`lambda_k  `
## [1] 10.83853
##
## [[3]]
## [1] 3
##
## $lambda
## [1] 10.69587
##
## $`lambda_k  `
## [1] 10.70136
##
## [[3]]
## [1] 4
##
```

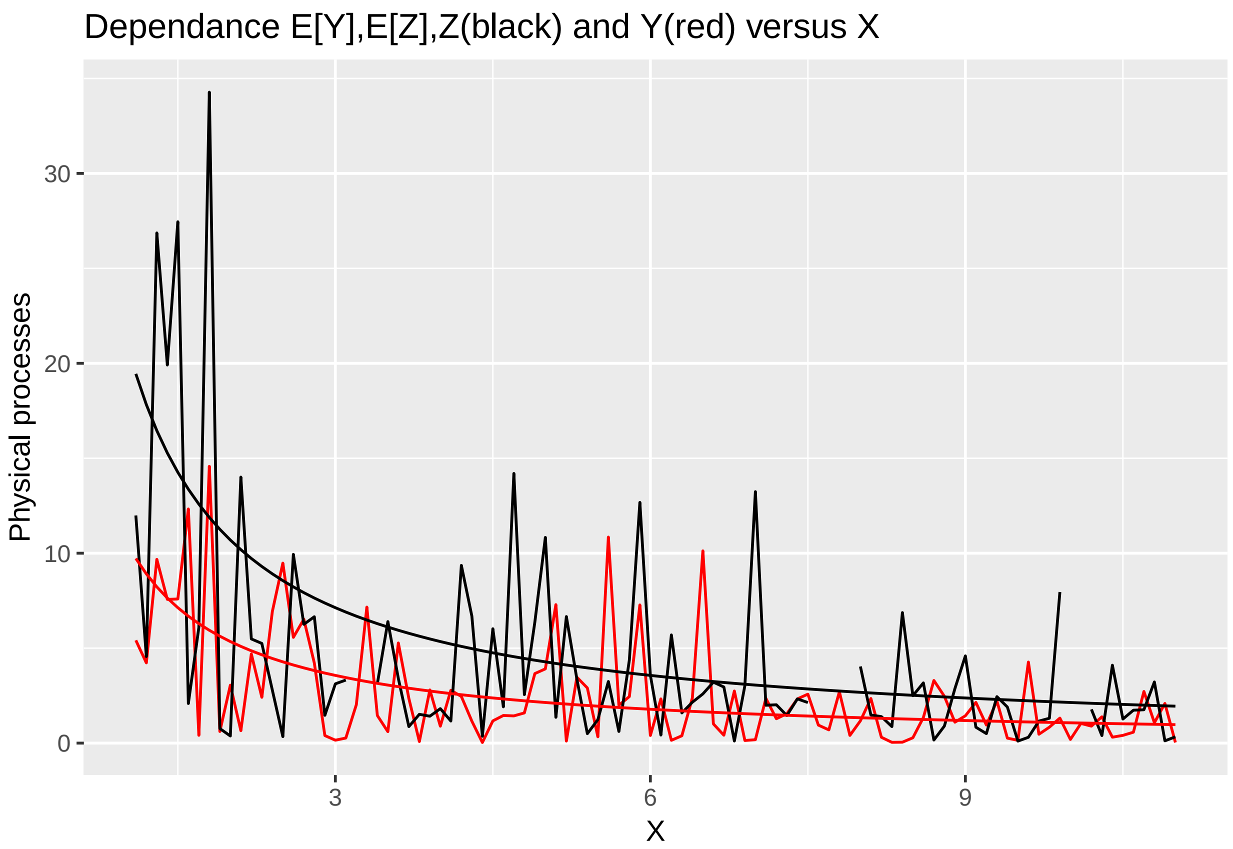
```
## $lambda
## [1] 10.69566
##
## $`lambda_k  `
## [1] 10.69587
##
## [[3]]
## [1] 5
```

In this task we implement the EM algorithms in order to estimate the optimal λ. We use the formulas obtained in task 2.2 and create a function which estimate the optimal λ. For every iteration, we estimate λ, using the previous one. We start with initial $\lambda = 100$. The procedure stops if the difference between the old and the new λ is less than a threshold(in this case 0.001) or after a given number of iterations k(here 100). After runing the algorithm, we obtain the value 10.6956555 as optimal lambda after only 4 iterations.

2.4 Estimate and plot the expected values.

Plot $E[Y]$ and $E[Z]$ versus X in the same plot as Y and Z versus X. Comment whether the computed λ seems to be reasonable.

```r
E_Y = optimal_lambda/physical$X
E_Z = 2*optimal_lambda/physical$X

df = data.frame(x = physical$X, E_y = E_Y,E_z = E_Z, z = physical$Z,y = physical$Y  )



ggplot(df) +
  geom_line(mapping = aes(x = df$x,y = df$y),color = "red")+
  geom_line(mapping = aes(x = df$x,y = df$z),color = "black")+
  geom_line(mapping = aes(x = df$x,y = df$E_y),color = "red")+
  geom_line(mapping = aes(x = df$x,y = df$E_z),color = "black")+
  labs(title = "Dependance E[Y],E[Z],Z(black) and Y(red) versus X ",x = "X",
       y = "Physical processes")
```



After computing the optimal λ parameter, we can estimate the expected values for Y, Z in order to get rid of the missing values. We have to remember here that Y, Z follow exponential distribution, which was not that clear from the first plot in task 2.1. It also known that if X is a random variable and $X \sim \exp(\theta)$, then $E[X] = \frac{1}{\theta}$. In this case $Y_i \sim exp(X_i|\lambda)$ and $Z_i \sim exp(X_i|(2\lambda))$. Therefore,

$$E[Y] = \frac{1}{\theta} = \frac{1}{\frac{x_i}{\lambda}} = \frac{\lambda}{x_i}$$

and

$$E[Z] = \frac{1}{\theta} = \frac{1}{\frac{x_i}{2\lambda}} = \frac{2\lambda}{x_i}$$

, where λ is the optimal λ we obtain using the EM algorithm.

We now plot also the $E[Y]$ and $E[Z]$ versus X in the same plot as Y and Z. It is clear that the optimal λ we estimate before is reasonable. Both of the expected values follow the exponential distibution as they should.