Assignment 3

Group3 - DSE

2020/11/24

Problem

Consider the stochastic process of spread of an epidemics (of COVID-19, influenza, or other similar viruses) in a population composed by *N* individuals.
Assume that each individual at each time step can be in one of the following 3 states:

• S = susceptible = not infected and not immune (then s/he can be infected)

• I = infected

• R = removed = recovered and immune, or dead

Assume also that at each time step each susceptible individual becomes infected with probability 0.07, each infected individual is removed with probability 0.25 and each removed individual returns back to be susceptible with probability 0.04.
Represent the evolution of the epidemics as a Markov chain and write the transition matrix.
Imagine now that the number *N* of individuals in the population is *N* = 10000.
If it is possible, find what will be the number of individuals in the three states S,I,R in the long run, after many time steps.

INDEX:

Introduction

In the problem proposed, we want to analyze the spread of a virus across a population of N individuals. In order to do so we identify the possible states in which any individual can be at any time, from the moment in which the virus starts circulating:

• **S:** *susceptible* = not infected and not immune (then s/he can be infected)

• **I:** *infected*

• **R**: *removed* = recovered and immune, or dead

To understand the evolution of the virus we create a random variable (X_n) indicating the state in which each individual is going to be at time n. We can then represent the spread of the virus as a *stochastic process,* and define the set of the aforementioned states, in which its random variable can take values, as the finite state space: $S = \{ S, I, R \}$.

The instructions of the problem give no information on whether the epidemic caused by the virus will ever be eradicated, thus we assume that the stochastic process is represented by the infinite sequence of a random variable $(X_0, X_1, X_2, ...)$. Also, we assume that the process moves at discrete time steps.

From the definitions of the three states of this problem a fundamental consideration can be made. We notice that in order to determine the probability of an individual to be “infected” in the future, the only information we need is whether today he or she is susceptible, immune or already infected, regardless of the state the individual was at yesterday. Similarly, to determine the probability of someone to be “removed” in the future, we just need to know the state (S, I, R) at which the individual is today. The same consideration holds to determine the probability for someone to be “susceptible” tomorrow. In other words, the value that X_{n+1} can take at any point in time is determined solely by X_n , regardless of the full history of the values it took in the past. Therefore we deduce the following:

$$P(X_{n+1} = s_j | X_0 = s_{i_0}, X_1 = s_{i_1}, ..., X_{n-1} = s_{i_{n-1}}, X_n = s_i) = P(X_{n+1} = s_j | X_n = s_i \ ) \ \forall i,j \ \in \{1, ..., k\}$$

At any point in time X_{n+1}, the probability that an individual finds him or herself in a certain state depends solely on what happens today. We recognize this aspect of the stochastic process as the defining property of a Markov Chain (*Markov property)*. The process is, indeed, memoryless.

Preliminary considerations and Methodology

From the data we are given, we know that when we move one step forward in the future the following events can happen:

- A *susceptible* individual gets *infected* with probability 0.07
- An *infected* individual is *removed* with probability 0.25
- A *removed* individual goes back to being *susceptible* with probability 0.04

And we notice the following:

- The process of infection follows a **cycle**, since there is 0 probability to move from S to R (i.e. of being removed without having contracted the virus first), from I to S (i.e. of being susceptible again without having been removed first), and from R to I (i.e. of being infected again without being susceptible again first).
- The **initial state** of any individual should be S, as everybody is susceptible before the virus spreads.
- The **final state** of any individual should be R, as at a certain point the virus will be extinguished or a vaccine will be found.

From this information we understand that at each time step any individual can rather remain in its state, rather move to the next one, with a constant probability across time. These probabilities are called *transition probabilities*, since they represent the chance of moving from one state to another in a single time step.

From the *Markov property* we know that the probability to be in s_j at time n+1 only depends on the position s_i at time n. In this case the resulting transition probability (p_{ij}) does not depend on the time n, thus the MC is ***homogeneous*** and we can write:

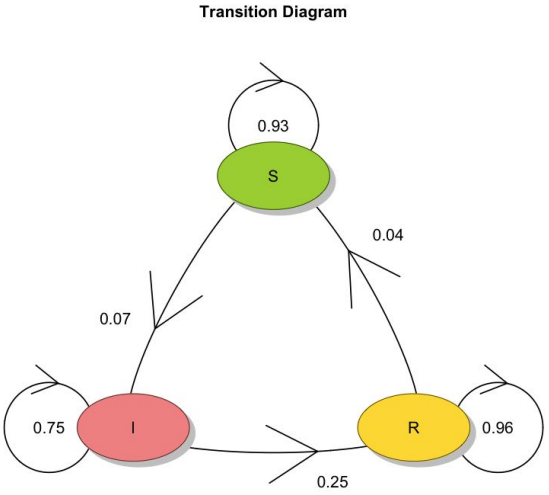
$$\forall i,j \in \{1, ..., k\} \qquad P(X_{n+1} = s_j \mid X_n = s_i) = p_{ij}$$

and create the ***transition matrix*** (P) with entries corresponding to p_{ij}, so that:

	[S]	[I]	[R]
[S]	0.93	0.07	0.00
[I]	0.00	0.75	0.25
[R]	0.04	0.00	0.96

- $p_{ij} \in [0, 1] \quad \forall i,j$
- $\sum_{j=1}^{k} p_{ij} = 1 \quad \forall i$ i.e. each row is the conditional distribution of a starting state.

We can represent the process through a ***transition diagram*****,** which is a *directed graph* having the states of the MC as the set of Vertices, and an edge for each non zero transition probability.



- $V = \{S, I, R\}$
- $E = \{p_{SI}, p_{SS}, p_{IR}, p_{II}, p_{RS}, p_{RR}, p_{RS}\}$
- The degree of each vertex is $d(v) = 4$
- The graph is *complete*: it contains all possible edges between the vertices

The spread of the virus represents a *walk* on this graph, where a *walk* is a finite sequence which alternates vertices and edges.
We can make some considerations about this walk:

- Recollecting our first assumptions, the walk is *open*, as the initial state $v_{i_0} = S$ is different from the final state $v_{i_k} = R$.
- Since the walk connects all the vertices with one another, the graph is *connected.*

We notice that this MC has some nice *properties*:

❖ It is ***irreducible:***

From the transition diagram depicted above, we can observe that each state of the diagram intercommunicates with the others, meaning that from each starting state s_i there is a positive probability of ever reaching s_i and vice versa. (We will provide proof of the irreducibility of this MC in the following section).

❖ It is ***aperiodic:***

Let's first define the *period* of a state as $d(s_i) = gcd\{n \geq 1 : (P^n)_{ii} > 0\}$
We say that a state is aperiodic if $d(s_i) = 1$, and we say that a MC is *aperiodic* when all its states are aperiodic.
This means that there is a positive probability of coming back to s_i, if we started from s_i, in a finite number of steps. (We will provide proof of the periodicity of this MC in the following section).

If a MC is both irreducible and aperiodic, there is a positive probability to reach any state s_i *starting from any any state* s_j *, in a finite number of steps.*

The importance of this information consists in the fact that we can exploit it to study the asymptotic behaviour of the process; in fact, if a MC is *irreducible* and *aperiodic,* its values will keep fluctuate, but its distribution will settle down to a limit i.e. the chain will converge to a unique ***stationary distribution***.

$\pi = (\pi_1, \pi_2, \pi_3)$ is a *stationary distribution* if:

- $\pi_i \in [0,1] \quad \forall i = 1, ..., k$
- $\pi P = \pi$

It is possible to prove that this stationary distribution exists, is unique, and in the long run, whatever is the initial distribution, this will converge to it.
From the definition of stationary distribution we understand that π is the ***normalized eigenvector*** corresponding to *eigenvalue 1* of the transition matrix P of the process. We are going to use this finding to compute the stationary distribution of the spread of the virus after many time steps.

An important consideration to make is that being an irreducible, finite state MC with at least one aperiodic state (here we also have aperiodicity of the whole process), ***ergodicity*** is guaranteed. This means that to find the distribution between the three states of *N* independent and identically distributed individuals we can observe the distribution of an *N*-dimensional time series.

For this reason, in our computations we will analyze the spread of the virus across *N* steps in the future, and the obtained distribution will tell us the proportion of *N* individuals arriving in each state at that time.

Computation and interpretation of the results

To address our problem, we want to simulate the Markov chain to find what will be the distribution of N individuals across the state space in a given number of time steps; in order to do so we use the inverse transformation method. We start our computation by generating a function to implement this method. Through this invtransform function we create some intervals and set their length corresponding to the probability of observing each state, so that the sum of the intervals is 1; through a cumulative summation we create their boundaries. Then we generate a uniform random number in the interval [0,1] so that the probability of the number falling in a particular state is equal to the length of the related interval. We ask as an output an index that shows in which interval the uniform random number has fallen (the first for which the distance between the number and the lower boundary is positive).

```
-----------------------------------------------------------------------------------------------
library(stats)
invtransform<-function(p,s){
# generate a random number in the states s, with discrete distribution p with the method of
the inverse transform
# s=vector of states
# p=probability of each state. Vector of the same length as s
n<-length(p)
cs<-cumsum(p)
u<-runif(1,min = 0,max = 1)
d<-cs-u
i<-min(which(d>0))
s[i]
}
```

By stating that our initial distribution is (1,0,0), we are imposing the length of the first interval to be 1, and the other two to be 0. So, the first state will definitely (by construction) be S, the state that we have regarded as the starting point for the Epidemic spread model. We want to observe the distribution of 10000 individuals; to do so, because of the ergodicity of the MC, we simulate 10000 time steps; in this way we can observe its asymptotic behaviour.

To visualize the output of the MC as a state vector, we create the matrix x, with one column of N rows, where N is equal to the number of steps (in this case, 10000 rows). We then call the different states “S, I, R” and build the transition matrix P. To verify if we can eventually find a convergence to a unique stationary distribution, we check for aperiodicity and irreducibility of the chain using two functions of the library “markovchain”. We then plot the transition diagram displayed in the methodology section.

```
# n. of simulated steps
nsteps<-10000

# initial distribution
p0<-c(1,0,0)

x<-matrix(0,nsteps,1)

# Epidemic spread
states<-c("S","I","R")

# starting state
x[1]<-invtransform(p0,states)

P<-matrix(c(0.93,0,0.04,0.07,0.75,0,0,0.25,0.96),nrow = 3,ncol=3)


P
     [,1] [,2] [,3]
[1,] 0.93 0.07 0.00
[2,] 0.00 0.75 0.25
[3,] 0.04 0.00 0.96

# we check for aperiodicity and irreducibility of the MC
library(markovchain)
mc<-new("markovchain",states=states,transitionMatrix=P)
is.irreducible(mc)
[1] TRUE
period(mc)
[1] 1

# plot the transition diagram
library(diagram)
library(RColorBrewer)

tP<-t(P)
palette<-colors()[c(496,404,143)]

plotmat(tP,pos=c(1,2),name=c("S","I","R"),curve=0.1,box.lwd=0.5,box.size=0.1,box.type="ellips
e",box.prop=0.6,box.col=palette,dtext=0.5,arr.lwd=1.3,self.lwd=1.3,arr.length=1.5,arr.width=0
.8,arr.type="simple",arr.pos=0.6,self.arrpos=c(1.4,1.4,1.4),self.shiftx=c(-0.00,-0.15,0.15),s
elf.shifty=c(0.09,0.00,0.00),self.cex=0.8,main="Transition Diagram")
```

After having proved that the MC is irreducible and aperiodic , we update the function to study its evolution. Starting from the initial distribution at the first time step, we watch where the random number falls in subsequent time steps.

```
# evolution
for (j in c(2:nsteps)){
 if (x[j-1]=="S")
 nrow<-1
 else if (x[j-1]=="I")
 nrow<-2
 else
 nrow<-3
 x[j]<-invtransform(P[nrow,],states)
}
```

We print x to have a visualization of what a walk of 10000 time steps on this MC would look like. To see what is the distribution of the N individuals between the three states and have a compact way of looking at the data, we create a table in which we count the number of times the random number has fallen in the three different states and we divide it by the number of steps (i.e. individuals) we considered.

```
x
## [1,]  "S"
## [2,]  "S"
## [3,]  "S"
## [4,]  "S"
## [5,]  "S"
## [6,]  "S"
## [7,]  "S"
## [8,]  "S"
## [9,]  "S"
## [10,] "S"
## [11,] "S"
## [12,] "S"
## [13,] "S"
## [14,] "S"
## [15,] "S"
## [16,] "S"
## [17,] "S"
## [18,] "I"
## [19,] "I"
## [20,] "R"

# omitted 9960 rows


## [9980,] "R"
## [9981,] "R"
## [9982,] "R"
## [9983,] "R"
## [9984,] "R"
## [9985,] "R"
## [9986,] "S"
## [9987,] "S"
## [9988,] "S"
## [9989,] "S"
## [9990,] "S"
## [9991,] "S"
## [9992,] "S"
## [9993,] "I"
## [9994,] "I"
```

```
## [9995,] "I"
## [9996,] "I"
## [9997,] "I"
## [9998,] "I"
## [9999,] "R"
## [10000,] "R"
```

```
table(x)/nsteps
```

```
## I      R      S
## 0.0873 0.5756 0.3371
```

We now want to find the Stationary distribution pi. We know that the chain is aperiodic and irreducible, thus it will converge to a unique stationary distribution after a high number of steps.

We will first use the "Eigenvector method", exploiting the definition of the stationary distribution and then try a "Naive approach", checking that taking our transition matrix P to a high number, say m = 100000, and multiplying it by pi we still obtain pi.

```
# EIGENVECTOR METHOD
# the stationary distribution (pi) of the chain is the normalized eigenvector corresponding
to eigenvalue 1 of the matrix P transposed
# we start writing the transposed matrix of the transition matrix P
Ptrans<-matrix(c(0.93,0,0.04,0.07,0.75,0,0,0.25,0.96),nrow = 3, ncol = 3, byrow=TRUE)
print(Ptrans)
```

```
##      [,1] [,2] [,3]
## [1,] 0.93 0.00 0.04
## [2,] 0.07 0.75 0.00
## [3,] 0.00 0.25 0.96
```

```
# we find the eigenvalues of the matrix, and check that 1 is an eigenvalue of P trans
e <- eigen(Ptrans)
e$values
## [1] 1.0000000 0.8658258 0.7741742

# now we find the eigenvectors corresponding to the different eigenvalues. We are interested
in the first column, related to the eigenvalue 1
e$vectors
```

```
##      [,1]      [,2]       [,3]
## [1,] 0.4914198 -0.5038436 0.2017803
## [2,] 0.1375975 -0.3045009 0.5842839
## [3,] 0.8599846 0.8083445 -0.7860642
```

```
e$vectors[,1]

## [1] 0.4914198 0.1375975 0.8599846

v <- e$vectors[,1]
```

To find the stationary distribution we need to normalize this vector v. We can multiply v by any scalar, and still have an eigenvector of P trans corresponding to eigenvalue 1. To obtain the normalized vector, we must multiply v for a scalar such that the product of v times that scalar is a probability distribution, i.e. each of the values of that product pertain to the interval [0,1], and the sum of all values v(i) equals to 1.

```
# to find the scalar by which we must multiply the non-normalized vector, we write a linear
equation and solve it.
```

```
S<-v[1]
I<-v[2]
R<-v[3]

solve(S+I+R,1)


## [1] 0.6715908


# finally, we multiply v for the scalar we just found and get the stationary distribution pi
of the (transposed) transition matrix P

pi <- v*(solve(S+I+R,1))
pi
## [1] 0.33003301 0.09240924 0.57755776
```

A more naive way to look at the stationary distribution is simply to take any row of the transition matrix at a high power. We know in fact that the stationary distribution corresponds to the asymptotic behaviour of the chain, and once it enters it, it never changes.

```
# NAIVE APPROACH

 pi_naive <- P^100000
 pi_naive <- pi_naive[1,]
 pi_naive-pi_naive%*%P

     [,1] [,2] [,3]
[1,]    0    0    0
```

We see that the error made is 0,and this is consistent with the existence of the stationary distribution of the chain. Now, we can find the exact number of individuals in the 3 states by multiplying the vector pi by the number of individuals (nindv), equal to 10000, and print the result as IndividualsPerState (IXS), i.e. the number of individuals in the states, respectively, S, I, and R.

```
# counting individuals in each state
nindv <- 10000
IXS <- pi*nindv
round(IXS)

[1] 3300  924  5776
```

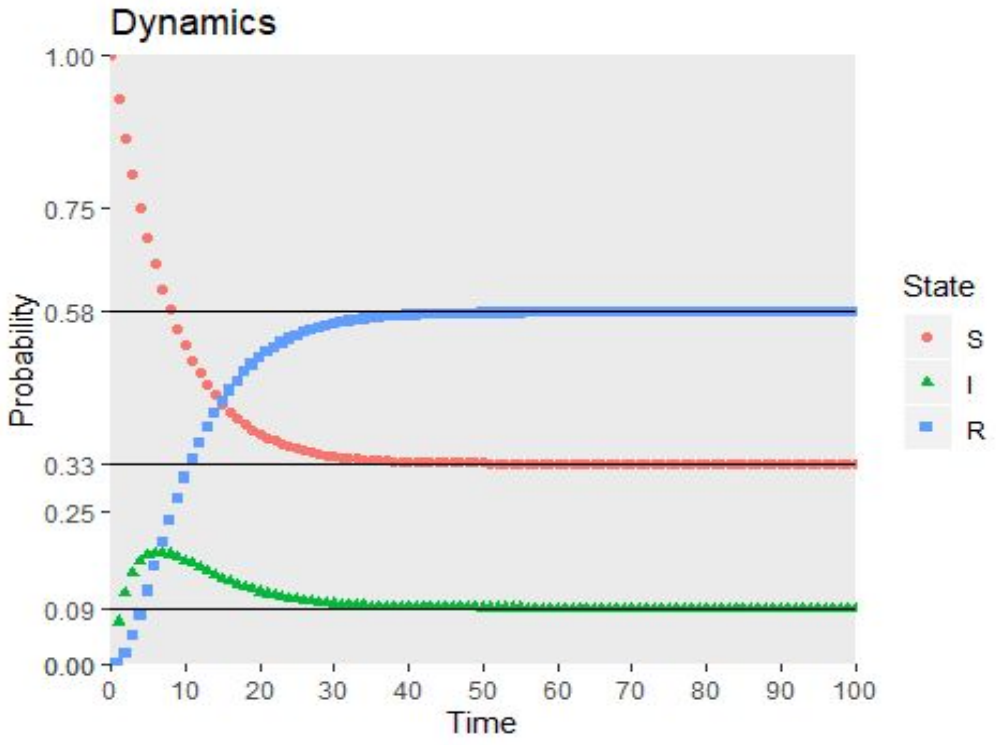
To have a visual representation of the asymptotic convergence of the MC to its stationary distribution, we plot a graph displaying the probability of being in each state across time. By plotting the dynamic graph of the distribution, we observe that the probability of individuals to be in State S drops fast, while the probability of individuals to be in State R keeps increasing, and then converges. Eventually, more than 50% of the individuals will get immune or die. Noticeably, the probability of individuals to be infected (to be in State I) first increases fast until it reaches a peak, and then drops to its stationary value.

```
library(Matrix)
library(expm)
```

```
# initial distribution in matrix form
u0 <- matrix(c(1,0,0),nrow=1,byrow = TRUE)

##      [,1] [,2] [,3]
## [1,]    1    0    0

list_S <- numeric()
list_I <- numeric()
list_R <- numeric()
for (i in 1:100){
  u <- u0 %*% (P %^% i)
  S <- u[1]
  I <- u[2]
  R <- u[3]
  list_S[i] <-S
  list_I[i] <-I
  list_R[i] <-R
}
df <- data.frame(Time= 1:100, S =list_S, I =list_I, R=list_R)
initial <- c(0,1,0,0)
df <- rbind(initial,df)
library(ggplot2)
library(reshape2)
Dynamics <- melt(df, id.vars = "Time")
colnames(Dynamics) <- c("Time", "State", "Probability")
ggplot(data=Dynamics, aes(x=Time, y=Probability, group=State)) + geom_point(aes(color=State, shape=State)) + ylab('Probability') + scale_x_continuous(expand = c(0,0), breaks =seq(0,100,10)) + scale_y_continuous(expand = c(0,0), breaks = c(0, 0.09, 0.25, 0.33, 0,50, 0.58, 0.75, 1.00)) + ggtitle("Dynamics") + geom_hline(yintercept = c(0.09, 0.33, 0.58)) + theme(panel.grid =element_blank())
```



Conclusions

Considering the evolution of the epidemics as a Markov Chain, this random process $(X_0, X_1, X_2, \dots)$ (X_n is the state at which the individual stays at time n) has a finite state space $S = \{ S, I, R \}$ with transition matrix P. After drawing the transition graph, we get the transition matrix P: $[\, 0.93 \;\; 0.07 \;\; 0 \;\; 0 \;\; 0.75 \;\; 0.25 \;\; 0.04 \;\; 0 \;\; 0.96 \,]$. After the analysis, we found this Markov Chain to have the following properties:

- *Memoryless*: the conditional distribution of X_{n+1} given $(X_0, X_1, X_2, \dots)$ depends only on X_n, so the present state of the individual at present step depends only on the last step
- *Homogeneous*: at each step, the transition matrix never changes
- *Irreducible*: all three states of the Markov Chain can be reached from all the three states
- *Aperiodic*: at each state, one step is enough to get back to that state

According to the properties of this Markov Chain, there exists a unique stationary distribution to which the distribution will converge in the long run. Assuming that the initial distribution of this random process is (1,0,0) , which means that all the individuals are susceptible to the virus in the beginning, we manage to find the stationary distribution which is $(0.33003301,\ 0.09240924,\ 0.57755776)$. This result means that in the long run, it will be reached a situation where each individual has probability **33.00%** to be in state **“S” (**i.e. to be susceptible), a probability of **9.24%** to be in state **“I”** (i.e. to be infected), and probability **57.76%** to be in state **“R”** (i.e. to be recovered and immune, or dead). Thus when the number of individuals N in the population is 10,000, the distribution will be the same of that observed after many time steps, the number of individuals in state **“S”** is **3,300**, the number of individuals in state **“I”** is **924** and the number of individuals in state **“R”** is **5,776**.

Finally, it is worth noticing that while the settings of the problem have led us to model the epidemic cycle with a homogenous Markov Chain, a real world scenario would require different assumptions. In fact, the Covid-19 crisis has shown that several factors can alter the initial transition matrix of the chain. First, governments have imposed restrictions on the movement of people and provided guidelines to help mitigate the spread of the epidemic (social distancing, face masks, frequent use of sanitizers etc). Secondly, the research and development carried out in the medical field has allowed hospitals to improve the efficiency of the medical treatments for covid-patients. Therefore, we would expect these two factors to contribute over time to a reduction of the probability of being infected and to an increase in the probability of being “removed”. Lastly, most of the countries hit by the epidemic have faced more than a wave of Covid, thus the conditional probabilities of the transition matrix could be assumed to follow an undulatory pattern over time. Having said so, a inhomogeneous Markov Chain would help modelling the reality in a more precise and comprehensive manner.