First Assignment - Discrete-time Markov chains

Stochastic Processes

Master in Statistics for Data Science, 2020 - Univ. Carlos III de Madrid

Ana Arribas Gil

Instructions:

- This assignment is due Dec. 3rd, by 10 am.
- The assignment must be submitted to the specific link in Aula Global. You can hand write all the parts requiring calculations and math notation. You may also include, as an appendix, scripts for the R code used and refer to them within the text. Please do not directly include R outputs unless properly formated as text, table, equations, ... All of the material must be put together an converted into a single pdf file that will be uploaded to Aula Global.
- The data for exercice 1 will vary according to your **group number**, as explained below. Your group number will be provided by e-mail after one member of each group sends me a message with the group composition (`ana.arribas@uc3m.es`).
- Provide (short and clear) explanations for every part. Final answers without any comment won't be taken into account.
- Give your results with 4 decimal places.

1. [**6 points**] The antipollution protocole of the city of Madrid defines five scenarios (from 1 to 4, plus scenario Alert) to warn citizenship and endforce traffic restrictions when the level of pollutants are higher that some given threshold. Scenario 1 is declared the first day of any pollution episode, and then, according to the number of days in a row in which pollution levels remain high, Scenario 2, 3, or 4 are declared (Scenario 4 is used after 4 or more days in a row with high pollution levels). If the pollution levels go back to normal, the protocole is deactivated. The Alert scenario is declared at any moment in which an alert threshold is exceeded (this threshold is above the one used for Scenarios 1 to 4). Once the Alert scenario is declared, even if pollution levels decrease to values below the alert threshold in the following days, the Alert scenario remains in place until the end of that pollution episode (that is, until pollution decrease under the threshold used for Scenarios 1 to 4).

   a) [1.25 point] Explain why the pollution scenario sequence over days can be modelled as a Markov chain (considering also a "No pollution episode" scenario). Define its states and all possible transitions between states. Has this Markov chain a unique stationary distribution? Argue it without using any of the data that follows.

   b) [1 point] The file `PollutionMadrid.RData` contains data of (simulated) pollution episodes in Madrid in the last four years. The file contains 20 different data sets, and you should use the one that is assigned to your according to your group number. That is, you must load the data and **only use row** i**, where** i **is your group number.**
   ```
   load('PollutionMadrid.RData')
   x=X[i,] # i is your position in the class list
   ```
   The states are labelled as `NR, Sc1, Sc2, Sc3, Sc4` and `Alert`. Estimate by maximum likelihood the transition probabilities of the chain. What can you say of the comparison of your estimates and the possible transitions between states that you had argued in part a)?

   c) [0.75 points] From the estimated transition matrix, compute the joint probability of the observed sequence of scenarios for the first week of data, assuming the initial state of the chain is the "No pollution" scenario.

   d) [1.75 points] Obtain the stationary distribution of the chain from the transition matrix estimated in part b). How would you estimate it directly from the data? Implement this second approach and provide another estimate for the stationary distribution. Discuss the differences between the two methods.

e) [0.75 points] Does this chain has a limiting distribution? Argue about it without using any of the data. What does this mean, in terms of pollution episodes?

f) [0.5 points] If driving is forbiden whenever scenarios 3, 4 or Alert are declared, what is the expected number of days over a whole year in which you won't be allowed to use your car?

2. [**4 points**] Consider the following Markov chain on $\mathbb{N} \cup \{0\}$ with transition probabilities

$$p_{i0} = p, \quad p_{i,i+1} = 1 - p, \quad p_{ij} = 0 \text{ if } j \neq 0, i+1, \quad \forall i \geq 0.$$

This chain has a stationary distribution if $p \neq 0, 1$.

a) [1.25 point] Obtain the stationary distribution of the chain (for $p \neq 0, 1$). What is the expected return time to 0? *Hint:* Use that for any $|r| < 1$, $\sum_{k \geq 0} r^k = \dfrac{1}{1-r}$.

b) [0.75 points] Analyse the communication classes of the chain, and the classification of the states (null/positive recurrence, transience, periodicity). Does this chain has a limiting distribution?

c) [1 point] Write an R function to simulate sequences from this chain, with arguments p, the initial value of the chain, and the number of steps to generate.

d) [1 point] **According to the class list, set** $p = \dfrac{1}{1+k}$**, where** k **is your position on that list** and generate 8 trajectories of length 1000 of this chain with different initial values on $[100, 1000]$. Represent the trajectories graphically as a function of time. What do you observe? Comment your results in relation to part a) and part b).