TMA4265 Stochastic Modelling – Fall 2020

Project 2

Background information

- The deadline for the project is Sunday November 8 23:59.
- This project counts 10% of the final mark in the course. You will receive a numerical score between 0 and 10, and each of the 10 subproblems will be weighted equally.
- This project must be passed to be admitted to the final exam.
- A reasonable attempt must be made for each problem to pass this project.
- The project should be done in groups of **two** or **three** people. You must sign up as a group in Blackboard before submitting your report and code.
- The project report should be a pdf-file that includes necessary equations and explanations to justify the answer in each problem. Include the required plots in the pdf and reference them in the text. You do not need to repeat the question text in the report you submit. The computer code should be submitted as a separate file, and **not** as an appendix in the report. Make sure this code runs. We may test it.
- There is a **6 page limit** for the project report. If you submit a longer report, we may not read it. The 6 page limit does not include the computer code, which should be submitted as a separate file.
- Make your computer code readable and add comments that describe what the code is doing.
- You are free to use any programming language you want as long as the code is readable, but you can only expect to receive help with `R`, `MATLAB` and `Python`.
- We will provide physical guidance in R2 on October 27 and November 3 at 08:15–10:00 without pre-registration, and in Smia/S21 on November 2 at 08:15–10:00 **with pre-registration**. We are looking into the opportunity for an additional session with physical guidance either on Thursday or Friday (November 5 or November 6), and will provide more information when available.
- If you have questions outside the aforementioned times, please contact the teaching assistants `susan.anyosa@ntnu.no` or `mina.spremic@ntnu.no`.
- The pdf-file with the report and the files with computer code should be submitted through our Blackboard pages under "Projects". You need to sign up as a group before you can submit your answer.

Problem 1: Modelling the common cold

Assume throughout this entire problem that a year has exactly 365 days.

The common cold is an infectious disease containing more than 200 different virus strains. This means that one cannot become immune to the common cold, and that most people will get a cold several times every year. Assume that an individual only has three possible states: susceptible (S), lightly infected (I_L), and heavily infected (I_H). Further, assume that the individuals in the population are independent, and that for each susceptible individual the time until the next infection follows an exponential distribution with expected value $1/\lambda = 100$ days. When a susceptible individual becomes infected there is a probability $\alpha = 0.10$ that the individual becomes heavily infected, and a probability $1 - \alpha = 0.90$ that the individual becomes lightly infected. The durations of the light infections follow independent exponential distributions with expected values $1/\mu_L = 7$ days, and the durations of the heavy infections follow independent exponential distributions with expected values $1/\mu_H = 20$ days. When an infection ends, the individual will immediately be susceptible again.

Consider a single individual, and let $X(t)$ denote the state (S, I_L or I_H) of the individual at time $t \geq 0$ measured in days.

a) Explain why $\{X(t) : t \geq 0\}$ is a continuous-time Markov chain, calculate the jump probabilities between each pair of states, calculate the transition rates between each pair of states, and draw the transition diagram.

Note: *we here use the term* ***jump probability*** *from state i to state j to denote the probability that when the sojourn time in state i ends, the transition will be from state i to state j.*

b) Calculate (by hand) the long-run mean fractions of time in each of the states. On average, how many days per year is an individual infected (either lightly or heavily)?

c) Write code to simulate the continuous-time Markov chain over a time period of 1000 years. Assume the starting state is susceptible, and plot one realization of $\{X(t) : t \geq 0\}$ over 5 years, i.e., for $0 \leq t \leq 5 \cdot 365$.

d) Based on one realization of $\{X(t) : t \geq 0\}$ for $0 \leq t \leq 1000 \cdot 365$, estimate the long-run mean fraction of time that an individual has an infection (light or heavy). In the report, you should briefly describe how you calculated the estimate based on the realization, provide the estimated value, and compare the estimate to the values calculated in **b)**.

e) Calculate (by hand) the expected time between heavy infections. I.e., the expected time between the end of one heavy infection and the start of the next heavy infection. Verify the calculations using an estimate based on one realization of $\{X(t) : t \geq 0\}$ for $0 \leq t \leq 1000 \cdot 365$.

For the remainder of this problem, assume that the total population contains 5.26 million individuals who become infected and susceptible completely independently of each other. We simplify the individual model so that each of the 5.26 million individuals only has two states: susceptible and infected. The durations of susceptible periods are independent exponential distributions with expected value $1/\lambda = 100$ days, and the durations of infected periods are independent exponential distributions with expected value $1/\mu = 7$ days. Let $Y(t)$ denote the number of

infected individuals in the population at time $t \geq 0$ measured in days.

f) Explain why $\{Y(t) : t \geq 0\}$ is a birth and death process, specify the birth and death rates, and draw the transition diagram.

Assume the stochastic process $\{Y(t) : t \geq 0\}$ has reached its stationary distribution.

g) For each infection, there is a probability of 1% that the infection will result in serious complications that requires hospitalization. On average, the hospitals only have capacity to handle 2000 individuals with complications from a cold. Use Little's law to calculate the average treatment time required so that the average number of individuals in the hospital does not exceed the capacity.

Problem 2: Calibrating climate models

A group of climate scientists are running a climate model that outputs the temperature at every location on earth for every 6-hour period in the years 2006 and 2100[1]. The climate model is deterministic, and given the athmospheric starting conditions, external forcing, and model parameters, you will always get the same result. The challenge is that the parameters of the climate model must be selected so that the output provides as realistic evolution in time as possible. This is immensely difficult because running the model only once may require one month of computation time. For the sake of this project, assume that the only way to choose these parameters is to run the climate model for different parameter values and compare to observed temperatures.

We limit the focus to one parameter, "the albedo of sea ice", which is a measure how much sun light is reflected by sea ice. We call this parameter θ, and we decide to choose this parameter so that the temperatures observed from January 1, 2006, to October 22, 2020, matches the output of the climate model as well as possible. The fit is measured through a score $y(\theta)$ calculated based on the model output generated with parameter value θ.

The group of climate scientists have spent the last month running the model in five computing centres and provides you with five evaluation points of $(\theta, y(\theta))$: $(0.30, 0.5)$, $(0.35, 0.32)$, $(0.39, 0.40)$, $(0.41, 0.35)$, and $(0.45, 0.60)$. The observations are shown in Figure 1.

You will use a Gaussian process model $\{Y(\theta) : \theta \in [0, 1]\}$ to model the unknown relationship between the parameter value and the score. Use $\mathrm{E}[Y(\theta)] \equiv 0.5$, $\mathrm{Var}[Y(\theta)] \equiv 0.5^2$, and $\mathrm{Corr}[Y(\theta_1), Y(\theta_2)] = (1 + 15|\theta_1 - \theta_2|)\exp(-15|\theta_1 - \theta_2|)$ for $\theta_1, \theta_2 \in [0, 1]$.

a) Define a regular grid of parameter values from $\theta = 0.25$ to $\theta = 0.50$ with spacing 0.005 ($n = 51$ grid points). Construct the mean vector and the covariance matrices required to compute the conditional mean and covariance matrix of the process at the 51 grid points conditional on the five evaluation points. Display the prediction as a function of θ, along with 90% prediction intervals.

Hint: The cdf of a Gaussian distribution exits in MATLAB (normcdf), R (`pnorm`) and Python (`norm.cdf`, after adding `from scipy.stats import norm` on the top of the file)

b) The scientists' goal is to achieve $y(\theta) < 0.30$. Use the predictions from **a)** to compute the conditional probability that $Y(\theta) < 0.30$ given the 5 evalution points. Plot the probability as a function of θ.

c) The scientists decide to run the climate model again with $\theta = 0.33$ and the result is $y(\theta) = 0.40$. Add this to the set of observed values, and given the six evaluation points, compute and visualize the prediction, 90% prediction intervals, and the probabilities that $Y(\theta) < 0.30$. The scientists' budget allow for one more run of the climate model, which value of θ would you suggest them to use to have the best chance to achieve $y(\theta) < 0.30$?

[1] See, for example, `http://www.cesm.ucar.edu/projects/community-projects/LENS/`

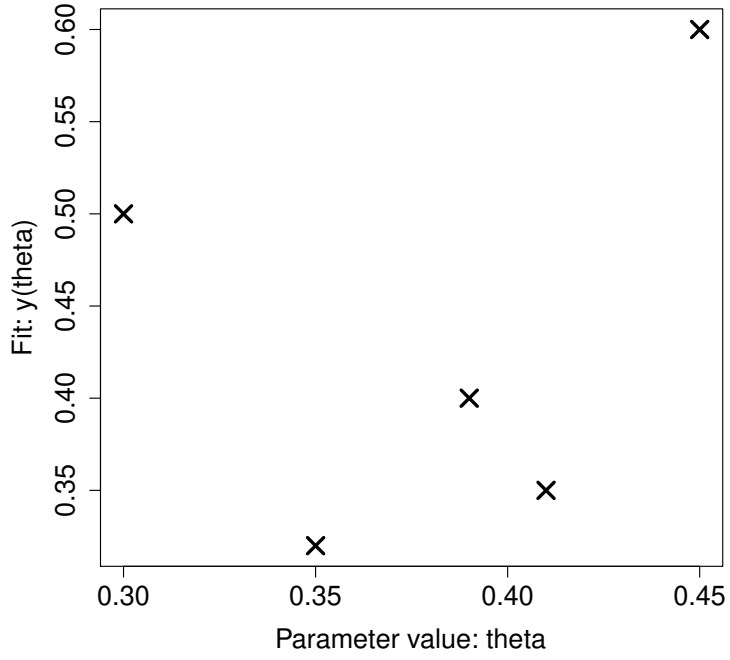

Figure 1: Observed relationship between fit and model parameter.