TMA4265 Stochastic Modelling – Fall 2019

Project 2

Background information

- The deadline for the project is Sunday November 10 23:59.
- This project counts 10% of the final mark in the course.
- 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** people. Please sign up as a group in Blackboard before submitting your report and code.
- The project report should include equations with calculations, plots and interpretation as text. The computer code should be submitted as a separate file. 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 will 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`.
- Lectures on November 5 and November 6 will be used for work on the project, and you can get assistance in the lecture room these days. The exercise classes on October 31 and November 7 takes place as normal and you can receive help with the project.
- If you have questions outside the aforementioned times, please contact `susan.anyosa@ntnu.no` or `rasmus.erlemann@ntnu.no`.

Problem 1: Modelling the common cold

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 two possible states: susceptible (S) and infected (I). 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 and that the durations of the infections follow independent exponential distributions with expected values $1/\mu = 7$ days.

a) Consider a single individual, and let $X(t)$ be the state (S or I) of the individual at time t measured in days. Explain why this is a continuous-time Markov chain, specify the transition rates, and draw the transition diagram.

b) Calculate (by hand) the long-run mean fraction of time per year that an individual has a cold.

c) Write code to simulate the continuous-time Markov chain over a time period of 1000 years. Use this code to:

c.1) plot one realization of $X(\cdot)$ over 5 years, i.e., for $0 \leq t \leq 5 \cdot 365$.

c.2) based on one realization of $X(\cdot)$ for $0 \leq t \leq 1000 \cdot 365$, estimate the long-run mean fraction of time per year that an individual has a cold. In the report, you should briefly describe how you calculated the estimate and provide the estimated value.

The total population contains 5.26 million individuals.

d) Let $Y(t)$ denote the number of infected individuals in the population at time t measured in days. Explain why this process can be modelled with a birth and death process, specify the birth and death rates, and draw the transition diagram. Did you have to make any simplifying assumptions?

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

e) 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 to not exceed the capacity.

*Hint: You do not need to calculate the stationary distribution of $Y(t)$, but you can instead take advantage of the result from **b)**.*

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 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 are 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 temperature observed on October 18, 2019, at 12:00–18:00 best matches the output of the climate model. 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 ($n = 51$ points). Construct the mean vector and the covariance matrix required to compute the conditional means and covariances of the process at the 51 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 exit 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/`

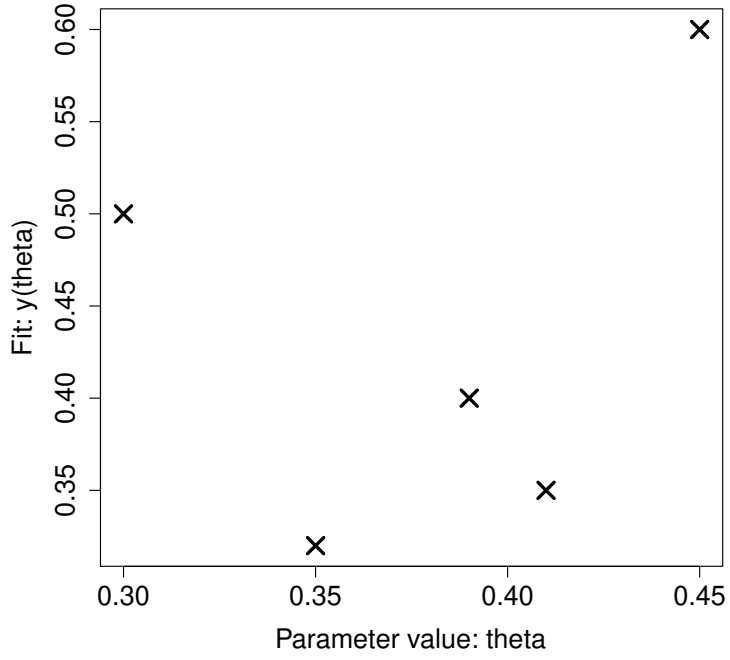

Figure 1: Observed relationship between fit and model parameter.