Avd. Matematisk statistik

KTH Matematik

HOME ASSIGNMENT 2, SF2955 COMPUTER INTENSIVE METHODS IN MATHEMATICAL STATISTICS

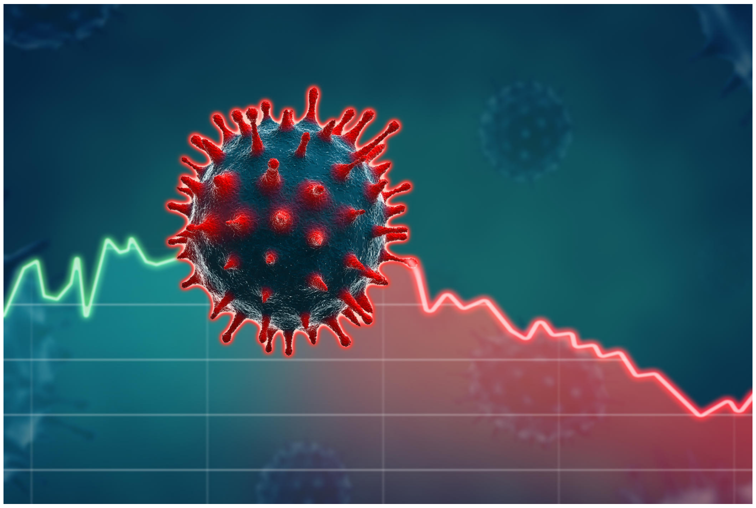
Examiner: Jimmy Olsson
All MATLAB-files needed are available through the course home page.
The following is to be submitted in Canvas by Thursday 19 May, 12:00:

- A report, named `group number-HA2-report.pdf`, of **maximum 7 pages** in `pdf` format. The report should provide detailed solutions to all problems. The presentation should be self-contained and understandable without access to the code.
- *All* your `m`-files (or similar depending on your language of choice) along with a file named `group number-HA2-matlab.m` that runs your analysis.

Discussion between groups is permitted, as long as your report reflects your own work.

Bayesian inference from COVID-19 pandemic data using Markov chain Monte Carlo

29 April 2022



Bayesian analysis of COVID-19 data—constructing a complex MCMC algorithm

In this assignment you will work with a basic epidemic model, whose unknown model parameters will be estimated by means of Bayesian inference. Since the posterior will be a complex probability distribution, you will sample from the same using a Markov chain Monte Carlo (MCMC) algorithm.

Model description

The SIR model

Consider an epidemic spreading in a population of P individuals. In the so-called *SIR model*, a discrete-time stochastic process $(S_t, I_t, R_t)_{t \in \mathbb{N}}$ keeps track of the numbers

– S_t of *susceptible individuals* at the beginning of day t, *i.e.*, the individuals who have not got the infection before day t,

– I_t of *infected individuals* at the beginning of day t,

– R_t of *removed individuals* at the beginning of day t, *i.e.*, the individuals who have either recovered—and, consequently, have become immune to the disease—or died before day t, and therefore cannot become infected again.

We assume that the population is isolated, in the sense that at each day t, $P = S_t + I_t + R_t$; thus, knowing P, the process may be described by two states only, say S_t and I_t. The epidemic evolves as follows. The outbreak starts with an initial number I_0 of cases, yielding $R_0 = 0$ and $S_0 = P - I_0$. Then, during day t, an additional number Δ_t^I of individuals become infected, which means that the number S_t of susceptible ones decreases by the same amount. At the same time, an additional number Δ_t^R of infected individuals either recover or die, which means that the number R_t of removed individuals increases by the same amount. Consequently, the group of infected individuals at time t is subjected to an inflow of Δ_t^I individuals and an outflow of Δ_t^R individuals. This dynamics is summarised by the state equations

$$\begin{cases} S_{t+1} = S_t - \Delta_t^I, \\ I_{t+1} = I_t + \Delta_t^I - \Delta_t^R, \\ R_{t+1} = R_t + \Delta_t^R, \end{cases} \qquad t \in \mathbb{N}.$$

Distribution of Δ_t^R

As mentioned, the process is stochastic and we need to describe the distributions of Δ_t^I and Δ_t^R. Regarding Δ_t^R, we assume that each of the I_t infected individuals at time t either recovers or dies the until next day with a constant probability $p^{i \to r} \in (0, 1)$ (since we assume that a recovered individual cannot become susceptible again, we do not need to make any distinction between the cases of recovery and death). We may hence assign Δ_t^R a binomial distribution with parameters I_t and $p^{i\to r}$, *i.e.*

$$\Delta_t^R \sim \mathrm{Bin}(I_t, p^{i\to r}).$$

Distribution of Δ_t^I

Since the epidemic is transmitted between individuals, we may assume that Δ_t^I, the number of individuals that become infected during day t, could, conditionally on the states (S_t, I_t, R_t), be assumed to follow another binomial distribution with parameters S_t and $p_t^{s\to i} \in (0,1)$, where

$$p_t^{s\to i} = 1 - \exp\left(-\lambda(t)\frac{I_t}{P}\right),$$

is the probability that a susceptible individual becomes infected during day t. Here $\lambda(t) > 0$ is a time-dependent parameter that reflects the average number of interactions per individual on day t. Still, there is a problem of using a binomial distribution in this modeling step, since for large populations, such as the inhabitants of an entire country, the model becomes close to deterministic, while real data usually exhibit significant noise. Thus, a better choice is to use a (generalised) *negative binomial distribution* (see Section A)

$$\Delta_t^I \sim \mathrm{NegBin}(\kappa, \varphi) \tag{1}$$

with parameters $\varphi \in (0,1)$ and

$$\kappa = \left(\frac{1}{\varphi} - 1\right) S_t p_t^{s\to i}.$$

Under (1), the mean of Δ_t^I is, for every φ, the same, $S_t p_t^{s\to i}$, as in the binomial case, while the variance $S_t p_t^{s\to i}/(1-\varphi)$ is always larger than in the binomial case. Furthermore, the closer φ is to one, the larger the variance, which gives us a way to obtain a sufficiently dispersed distribution of Δ_t^I. We suggest using $\varphi = 0.995$ throughout this assignment.

The parameter $\lambda(t)$ might be considered constant as long as the individuals of the population do not change their social habits; however, in the event of a serious disease with extensive spread, the government will typically enforce restrictions of various kinds, which will most certainly effect the interaction patterns in the population. Thus, we assume that there are $d-1$ breakpoints that divide the time frame into d intervals during which $\lambda(t)$ is constant. More precisely, let T denote the last day of the modeling period; then, given integer breakpoints $\mathbf{t} = (t_i)_{i=1}^{d-1}$ such that $0 = t_0 < t_1 < \ldots < t_{d-1} < t_d = T$ and positive parameters $\boldsymbol{\lambda} = (\lambda_i)_{i=1}^{d}$, we let

$$\lambda(t) = \sum_{i=1}^{d-1} \lambda_i \mathbb{1}_{[t_{i-1}, t_i)}(t) + \lambda_d \mathbb{1}_{[t_{d-1}, t_d]}(t), \quad t \in \{0, \ldots, T\}.$$

In the model described in this section, the parameters $\theta = (\boldsymbol{\lambda}, \mathbf{t}, p^{i\to r})$ are unknown and need to be estimated on the basis of a given record of observed epidemic states up to T. This will be discussed below.

Problem 1

Convince yourself that $(S_t, I_t)_{t\in\mathbb{N}}$ is a Markov chain and determine, for a given parameter vector θ, its transition probabilities

$$q_\theta(s_t, i_t; s_{t+1}, i_{t+1}) = \mathbb{P}_\theta\left(S_{t+1} = s_{t+1}, I_{t+1} = i_{t+1} \mid S_t = s_t, I_t = i_t\right).$$

Prior distributions

The main goal of this assignment is to provide a Bayesian solution to the problem of inferring θ by calculating the posterior distribution of θ given the observed data. In order to define a Bayesian model, we need to specify a prior distribution $\pi(\theta)$ for θ. We will assume that the parameters are *a priori* independent, *i.e.* $\pi(\theta) = \pi(\boldsymbol{\lambda})\pi(\mathbf{t})\pi(p^{i\to r})$, with the following marginals. First, we assign $\mathbf{t}$ a flat prior over the set of all ordered breakpoints, *i.e.*

$$\pi(\mathbf{t}) \propto \mathbb{1}_{\{0<t_1<t_2<\ldots<t_{d-1}<T\}}(\mathbf{t}), \quad \mathbf{t} \in \mathbb{R}^{d-1}.$$

Furthermore, we let the components of $\boldsymbol{\lambda}$ be a priori independent with $\Gamma(\alpha_i, \beta_i)$ marginal distributions, *i.e.*

$$\pi(\boldsymbol{\lambda}) = \prod_{i=1}^{d} \pi(\lambda_i) = \prod_{i=1}^{d} \frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)} \lambda_i^{\alpha_i - 1} e^{-\beta_i \lambda_i}, \quad \boldsymbol{\lambda} \in (0, \infty)^d,$$

where $(\alpha_i, \beta_i)_{i=1}^d$ are positive hyperparameters. We suggest to use $\alpha_i = 2$ for all $i \in \{1, \ldots, d\}$ throughout this assignment. Finally, we assign a Beta(a, b) prior to $p^{i\to r}$, *i.e.*

$$\pi(p^{i\to r}) = \frac{1}{B(a, b)} (p^{i\to r})^{a-1} (1 - p^{i\to r})^{b-1}, \quad p^{i\to r} \in (0, 1),$$

where $a > 0$ and $b > 0$ are hyperparameters.

Data

In the present assignment you are supposed to calibrate the model using COVID-19 time-series data from two given countries[1]. More specifically, you are going to estimate the parameters θ on the basis of two given realisations $(i_t, r_t)_{t=0}^T$ of the process $(I_t, R_t)_{t=0}^T$. Knowing the population size P, the realisation $(s_t)_{t=0}^T$ of $(S_t)_{t=0}^T$ can be obtained by letting $s_t = P - i_t - r_t$. For every t, let $y_t = (s_t, i_t)$ and collect these data in $\mathbf{y} = (y_t)_{t=0}^T$. For robustness we provide data starting a few days after the first cases, for which there are already a certain number of removed individuals at the very beginning (*i.e.*, $r_0 > 0$). The data sets you are going to use are the following.

- German data, ranging from 1 March 2020 ($t = 0$) to 15 June 2020 ($t = T$), named `germany_infected.csv` and `germany_removed.csv`. Population size P is 83 millions.
- Iranian data, ranging from 1 March 2020 ($t = 0$) to 1 June 2020 ($t = T$), named `iran_infected.csv` and `iran_removed.csv`. Population size P is 84 millions.

[1] Source: `https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases`

Metropolis-within-Gibbs sampling of posterior distributions

We proceed stepwise.

Problem 2

Determine the likelihood

$$f(\mathbf{y} \mid \theta) = \mathbb{P}\left(S_0 = s_0, I_0 = i_0, S_1 = s_1, I_1 = i_1, \ldots, S_T = s_T, I_T = i_T \mid \theta\right).$$

Here you may assume that the initial state (S_0, I_0) is set deterministically to (s_0, i_0), implying that $\mathbb{P}(S_0 = s_0, I_0 = i_0 \mid \theta) = 1$.

Problem 3

Compute, up to normalising constants, the full conditionals $\pi(\boldsymbol{\lambda} \mid \mathbf{y}, \mathbf{t}, p^{i\to r})$, $\pi(\mathbf{t} \mid \mathbf{y}, p^{i\to r}, \boldsymbol{\lambda})$, and $\pi(p^{i\to r} \mid \mathbf{y}, \boldsymbol{\lambda}, \mathbf{t})$. When possible, try to identify the distributions.

Problem 4

Implement a hybrid sampler simulating from the joint posterior $\pi(\theta \mid \mathbf{y})$ for each available data set $\mathbf{y}$. Use a standard Gibbs step for $p^{i\to r}$, while each of the components of $\boldsymbol{\lambda}$ and $\mathbf{t}$ is updated using local Metropolis–Hastings moves. Explore, for both data sets, the behavior of the MCMC chain for *one, two, or three breakpoints.* We suggest using a Gaussian random-walk proposal for each λ_i, where a candidate λ_i^* is generated according to

$$\lambda_i^* = \lambda_i + \sigma\epsilon,$$

with ϵ being standard normally distributed and $\sigma > 0$ is an algorithmic parameter. If you prefer you may also use an independent proposal, *e.g.* the prior, or something else. For each breakpoint we recommend generating candidates using a random-walk proposal, *i.e.*

$$t_i^* = t_i + \varepsilon,$$

where ε has discrete uniform distribution on $\{\pm 1, \ldots, \pm(M-1), \pm M\}$, with $M > 0$ being an integer-valued algorithmic parameter to be tuned. Be careful when calculating the acceptance probabilities!

Problem 5

Investigate the sensitivity of the posteriors and the mixing with regard to the hyperparameters $(\beta_i)_{i=1}^d$, a, and b as well as the algorithmic parameters σ and M.

Problem 6

Comment on your results. Are the posteriors of $\boldsymbol{\lambda}$ and $\mathbf{t}$ consistent with the data? It might be helpful for your empirical analysis to compute, from the given data, the time series $(\Delta_t^I)_{t=1}^T$. Regarding the German data, we know that "*interventions to contain the COVID-19 outbreak were implemented in three steps over 3 weeks: (i) Around 9 March 2020, large public events* *such as soccer matches were canceled; (ii) around 16 March 2020, schools, childcare facilities, and many stores were closed; and (iii) on 23 March 2020, a far-reaching contact ban (Kontaktsperre) was imposed by government authorities; this included the prohibition of even small public gatherings as well as the closing of restaurants and all nonessential stores."*[2]. In the same paper, in which a more complex approach is developed, the authors estimated the reporting delay (including both incubation period and test delay) to be within 9 and 14 days. Given this information, are the posterior distributions of the breakpoints reasonable?

Problem 7

Assume that we were interested only in the parameter $p^{i\to r}$, and wanted to infer the same by computing the marginal posterior $\pi(p^{i\to r} \mid \mathbf{y})$. Would it be necessary to use an MCMC algorithm in that case? Motivate your answer!

A The negative binomial distribution

The *negative binomial distribution*, denoted by $\mathrm{NegBin}(k, p)$, models the number of successes in a sequence of independent and identically distributed Bernoulli trials (with probability p) before k failures occur. If $N \sim \mathrm{NegBin}(k, p)$, then for every $n \in \{0, 1, \dots\}$ it holds that $\mathbb{P}(N = n) = \binom{n+k-1}{n}(1-p)^k p^n$. Moreover, $\mathbb{E}[N] = kp/(1-p)$ and $\mathrm{Var}(N) = kp/(1-p)^2$. For a general $k \in (0, \infty)$, the *generalised negative binomial distribution* has probability function

$$\mathbb{P}(N = n) = \frac{\Gamma(n+k)}{n!\Gamma(k)}(1-p)^k p^n, \quad n \in \{0, 1, \dots\},$$

with the same mean and variance.

[2] https://www.science.org/doi/abs/10.1126/science.abb9789