25 Mar 2021

Summary of progress of the Master Thesis

HMM modelling for the spread of the SARS-CoV-2

Martin Benes

Supervisor: Krzysztof Bartoszek

Research question

The main goal is to model the Covid-19 pandemic and interpret the results of the simulation of the model in terms of reported vs. simulated numbers, regional comparison and imposed restrictions.

First phase is collection of the data, both demographic and Covid-19 related, the countries of focus are Czechia, Italy, Poland, and Sweden. The data are visualized, and their reliability is discussed.

Incubation period, duration of symptoms, reproduction number and infection fatality ratio, characteristics of the Covid-19 disease, relevant to the model, are essential to estimate. Their values can be determined based on relevant clinical research papers studying patients infected with the disease, information from infection spread tracing or antigen testing results.

Hidden Markov model is the main concept for the modelling of the latent states, such as true number of infected or recovered people. While emission model is based on testing strategy, transition model is driven by differential equations of SEIRD (*Susceptible-Exposed-Infectious-Recovered/Dead*). The model combines the Covid-19 statistics with the prior belief from the clinical measurements.

The simulation can be run with different date ranges and for different countries or regions. The results are compared to the confirmed cases (i.e., positively tested) and the regions to each other to seek similarities. Interpretation of the results requires knowledge of the restriction level in each of the countries or regions, demography, and events, that could change the epidemiological situation.

What has been done?

The data of Covid-19 were collected for Czechia, Sweden, and Poland separately from their official sources, such as Ministries of Health or Public Health Agencies. The data for Italy are reported regionally so it was easier to use already implemented solution, such as Covid-19 Data Hub (Guidotti and Ardia 2020) rather than to collect data per each region separately.

Demographical data of population and mortality for all four countries both country- and region-wise are acquired from Eurostat, visualized, and briefly analyzed. They are intended to normalize cases or deaths per same population unit, compare fatality with mortality and provide possible explanation of results in the discussion, for example interpret higher mortality by generally older population.

In addition, a calendar of events in all four countries has been created, intended to be an appendix of the thesis. It contains chronologically ordered events in the countries, that could potentially influence changes in epidemic, such as significant restrictions being imposed or lifted. There are also events that could have caused abnormal mobility or gatherings, such as national holidays, elections, and demonstrations. Dates of revisions and corrections of the data are included too.

Variables such as symptoms' duration or incubation period have a certain duration that is complicated to be expressed using static transition and emission matrices. Thus, model is specified in Bayesian manner

and the characteristics of the Covid-19 are represented as random variables of certain distributions, specified based on results of cited research. Distribution of symptoms' duration is fitted to a clinical data of patient cases.

SEIRD transition parameters (S-E, E-I, I-R, I-D) are related to the disease characteristics mentioned previously. To estimate the parameter priors, I draw from disease characteristics' distributions, transform the draws according to SEIRD definition and try fitting some distribution to them.

The model is implemented in Stan/RStan. The code is parametrizable by various date ranges, population sizes (SIR* assumes constant population size), parameter priors, etc. Model does not estimate parameters daily, but per timeslot called *window*, where the parameter value is constant. This way the parameters differ over time and their value can reflect epidemical curve going up and down.

For purpose of optimizer, model is designed to work with statistics normalized by population, all data have domain $[0,1]$. The latent states of the simulation reflect the assumption of SIR ($S + E + I + R + D = 1$). Results are currently being saved and presented by Matplotlib in Python.

Yet to do

Currently, the model with the parameters estimated from research yields simulation of contagion fading out. Due to this I assume that the priors might not be correct, but after discussion with supervisor we suspect that the SIR* models might not be suitable for seasonal oscillating epidemics, but rather for unimodal (single peaking) diseases with permanent immunity, such as measles.

Although I have all the data per region, I have been so far working only with country data while implementing the model. I will run the simulation per regions once the priors are tuned. Because of this no discussion over simulation results or regional comparison has been done so far, although I have a code for regional comparison using dynamic time warping and clustering using Czekanowski diagram prepared. The event calendar is currently in structured format, I plan to do a graphical presentation.

Regarding possible extensions of the model that were presented in the thesis proposal, realistic ones to implement within the thesis time frame are vaccination and non-permanent immunity.

In a violin plot of mortality in age groups I observed a small, yet surprising bubble in age group 0 – 4 in Poland (and a tiny one in Italy as well). For Poland that seems to be present throughout the past years as well. The plan is to visualize other European countries and make a discussion over this phenomenon.

References

Guidotti, Emanuele, and David Ardia. 2020. "COVID-19 Data Hub." *Journal of Open Source Software* (The Open Journal) 5 (51): 2376. doi:10.21105/joss.02376.

Appendix: Sample thesis text

See file `Master_Thesis_Midterm.pdf`.