Data analysis on Covid-19 in Thailand by Machine Learning

Menghorng Bun (6222040096)
Ming Hsien Chuang (6214552743)
Nyan Lin Htet (6222040351)

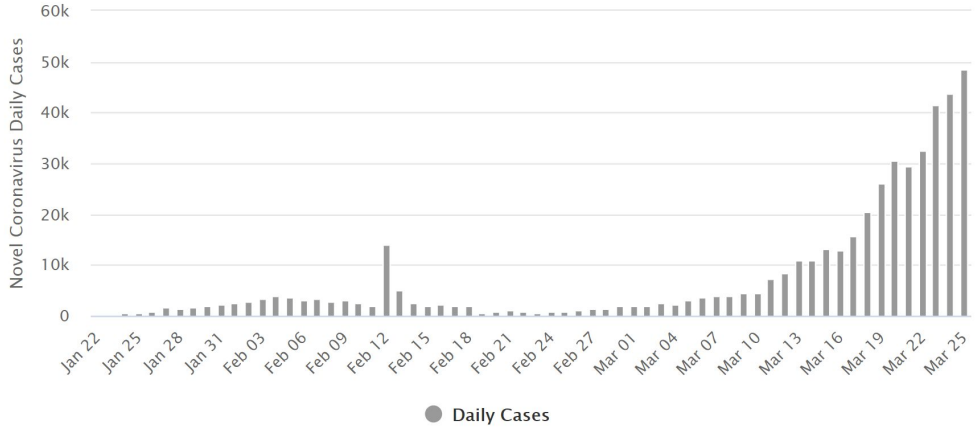
Outline

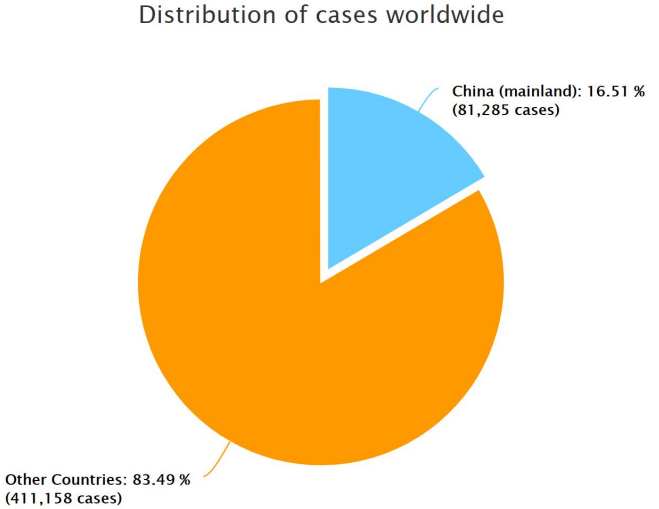
1. Introduction and background
2. Problem and understanding
3. Data
4. Model and Methods
5. Estimating model parameters
6. Scenarios
7. Results and discussion
8. Conclusion
9. Reference

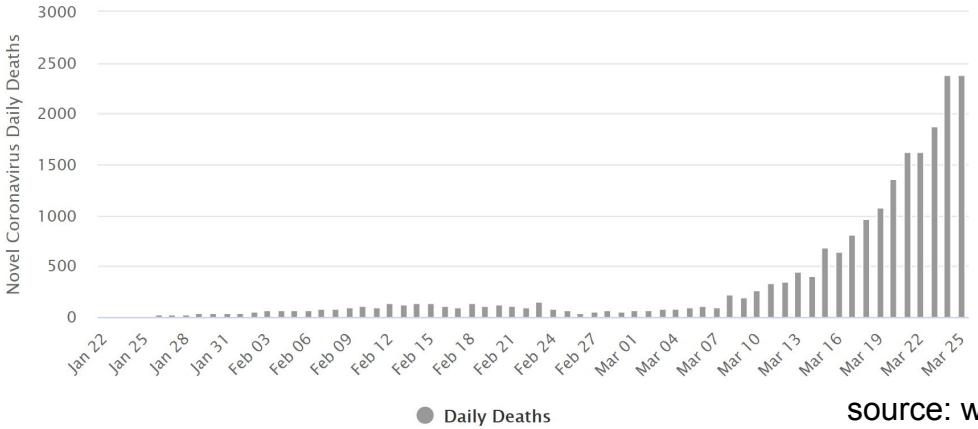
Introduction and Background

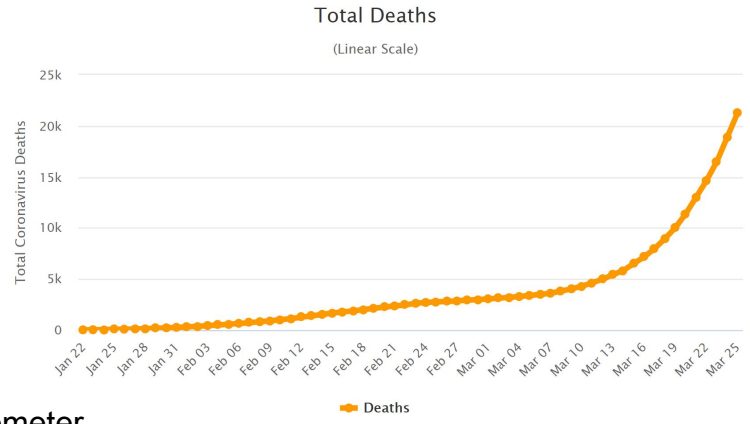
- COVID-19 is an emergent viral infection which rose in December 2019 in the Chinese province of Hubei, Wuhan.
- Confirmed as an ongoing pandemic by WHO on the 11th March 2020
- Affecting many lives and activities severely across the globe
- Our interest to capture the trend of evolving situation and provide prognosis of the disease with a robust measure with mathematical model aided analysis.

Cases all around the world



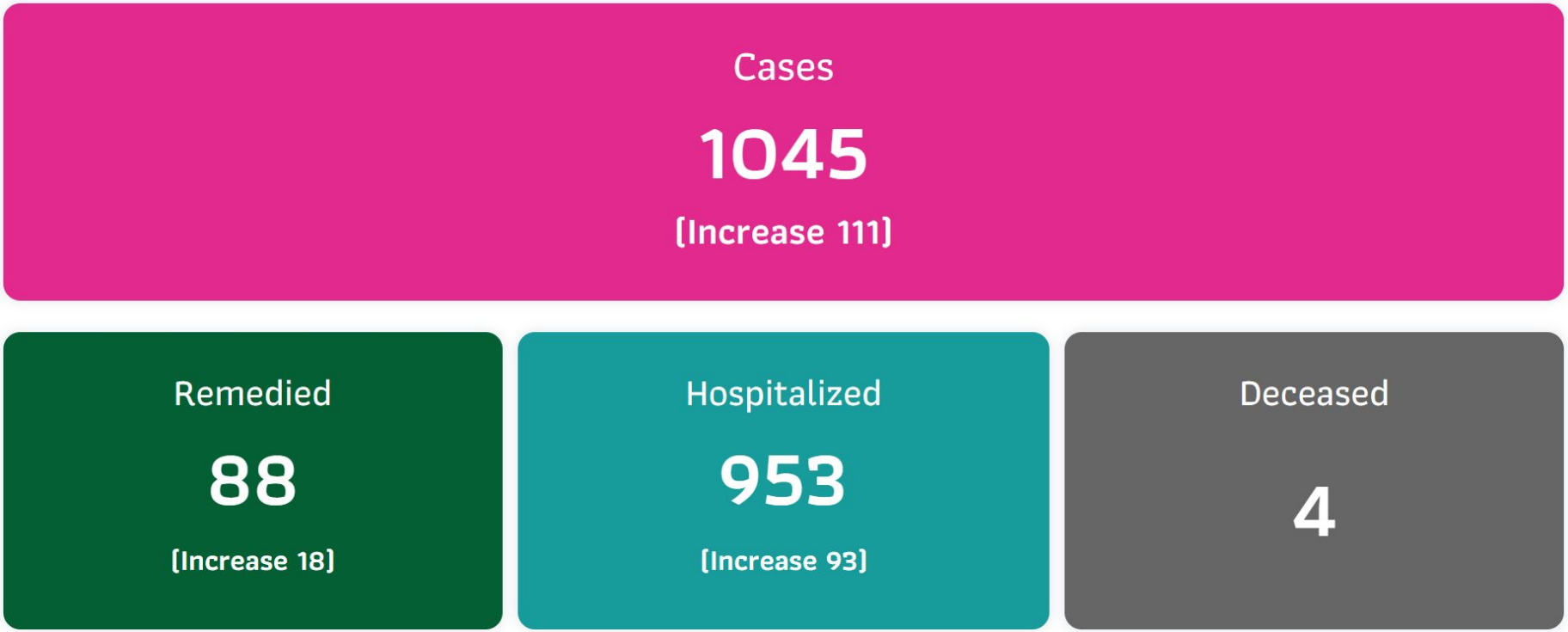






source: worldometer

Focused Area of Study (Thailand)



source: http://covid19.th-stat.com/en
Time: 26th March

Problem and understanding

- Forecasting different scenarios on the spread of COVID-19 in Thailand. (aim)
- Main Challenge - estimating model parameters
- Differential equations of the well-established SIR (Susceptible-Infected-Recovered) model.
- Parameter estimation - Monte Carlo importance sampling on the model parameters - distribution of parameters that well describes the observed data
- Forecast - evolving the model equations with parameter samples from the distribution

Data

- The underlying data used in this analysis is based on the daily reported Onset, Recover and Death toll from the World Health Organization (WHO) situation reports available publicly
- The time series recorded of confirmed cases and deaths for this study starts from 1st March until 24th March, a total observation of 24 days.
    - https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv
    - https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_deaths_global.csv

Model for Disease Spreading

- Population Classification
  - Susceptible - have not been infected, but can be infected
  - Exposure - have been infected with no symptoms and can spread the disease to others
  - Infection - have been infected
  - Recover - recovered and can’t be infected
- Common models for Disease Spreading
  - SI - HIV
  - SIS - Fluent
  - SIR - Chickenpox
  - SIER - SARS, COVID-19

Model of SIR Method

- Model Equations: $S + I \xrightarrow{\lambda} I + I$

$$I \xrightarrow{\mu} R\,.$$

- Model Equations with Population:

Assume that the disease spreads at rate λ from an infected person (I) to a susceptible person (S) and that an infected person becomes a recovered person (R) at rate μ. N is total population.

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\lambda\frac{SI}{N} \qquad S_t - S_{t-1} = -\lambda\Delta t\frac{S_{t-1}}{N}I_{t-1} =: -I_t^{\text{new}}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \lambda\frac{SI}{N} - \mu I \qquad R_t - R_{t-1} = \mu\Delta t I_{t-1} =: R_t^{\text{new}}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \mu I\,. \qquad I_t - I_{t-1} = \left(\lambda\frac{S_{t-1}}{N} - \mu\right)\Delta t I_{t-1} = I_t^{\text{new}} - R_t^{\text{new}}$$

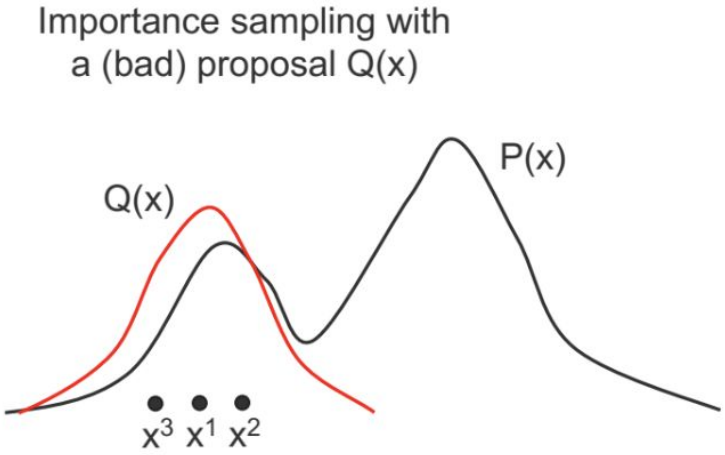
Exponential growth during outbreak onset

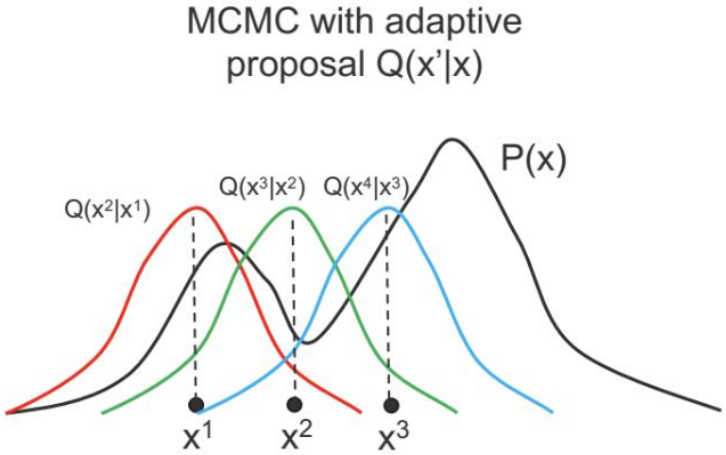
- Only a tiny fraction of the population is infected (I) or recovered (R)
    - $S \approx N \gg I$ such that $S/N \approx 1$
- Reduces to a simple linear equation

$$\frac{dI}{dt} = (\lambda - \mu)I \quad \text{solved by} \quad I(t) = I(0)\, e^{(\lambda-\mu)t}$$

Markov-chain Monte-Carlo (MCMC)

- To know high-level and complexity posterior distribution
  - Sampling the distribution
  - Next sampling point only related to current point - Markov-chain property





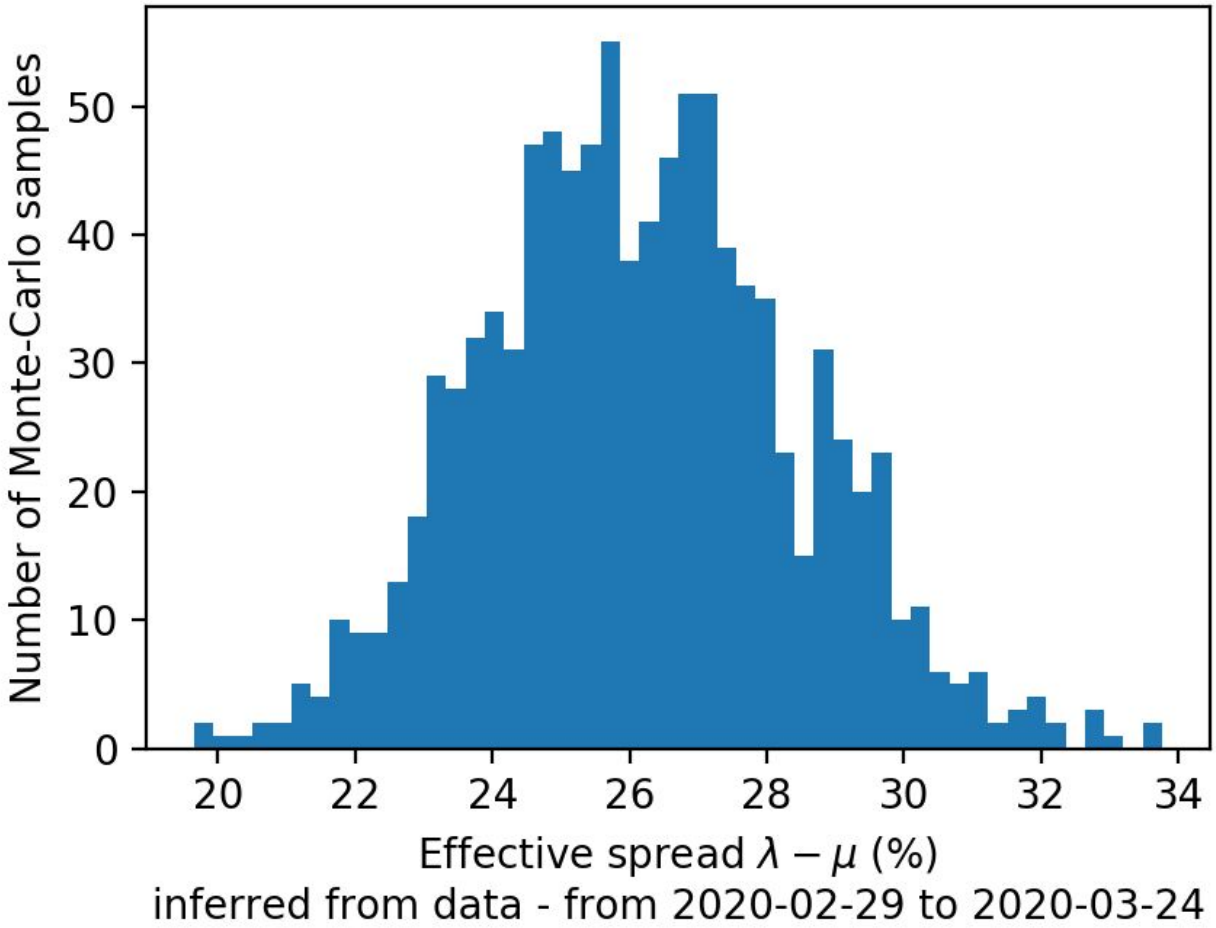
Estimating model parameters

- We estimate the set of model parameters $\theta=\{\lambda,\mu,\sigma,I\}$ using Bayesian inference with Markov-chain Monte-Carlo (MCMC)
- The structure of our approach is the following:
  - **Choose random initial parameters and evolve according to model equations:**

$$\hat{I}^{\text{new}} = \left\{ \hat{I}_t^{\text{new}} \right\}$$

  - **Recursively update the parameters using MCMC:**

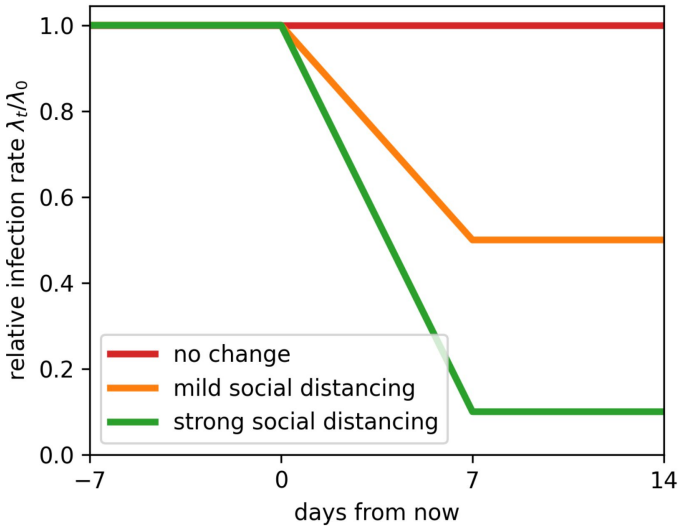
$$p\left(\hat{I}_t^{\text{new}} \middle| \theta\right) \sim \text{StudentT}_{\nu=4}\left(\text{mean} = I_t^{\text{new}}(\theta),\ \text{width} = \sigma\sqrt{I_t^{\text{new}}(\theta)}\right)$$


Number of Monte-Carlo samples
0
10
20
30
40
50
20
22
24
26
28
30
32
34
Effective spread $\lambda - \mu$ (%)
inferred from data - from 2020-02-29 to 2020-03-24

Scenarios

- Everything stays the same and the spread continues with the inferred rate $\lambda_t=\lambda$
- Mild social distancing: Assume λ_t gets reduced down to 50%, linearly within 7 days.
- Strong soial distancing: Assume λ_t gets reduced down to 10% , lineraly within 7 days.

Three potential future scenarios and they are implemented by introducing a time-dependent spreading rate λ_t

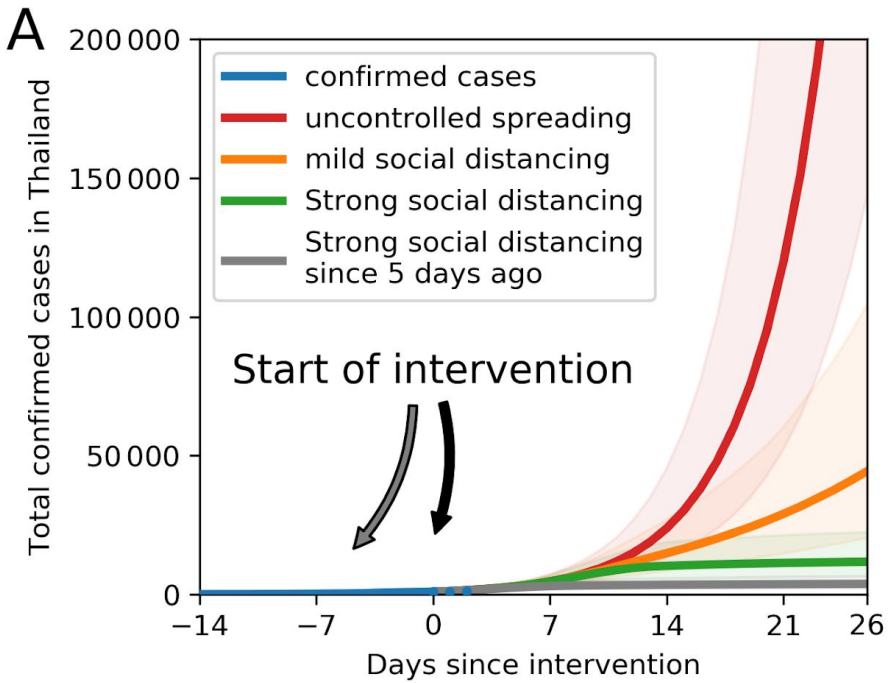


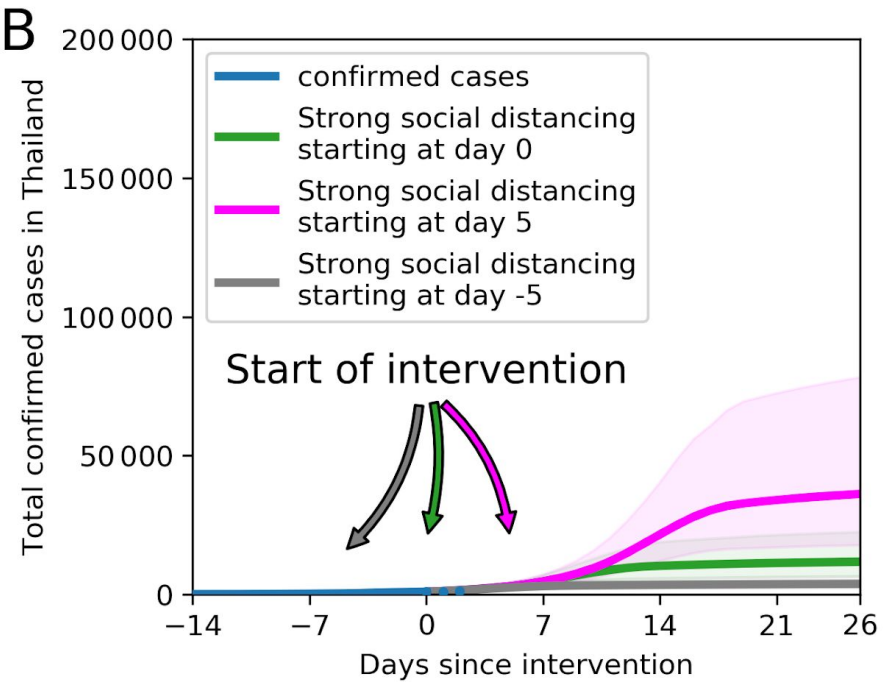
Overview model parameters

Variable	Parameter
$\theta = \{\lambda, \mu, \sigma, I_0\}$	Set of model parameters that are optimized
λ	Spreading rate
μ	Recovery rate
σ	Scale factor of the width of Student's t-distribution
N	Population size
S_t	Susceptible at time t
I_t	Infected at time t
R_t	Recovered at time t
Δt	Time step
$I_t^{\text{new}} = \lambda \Delta t \frac{S_{t-1}}{N} I_{t-1}$	New infections at time t
$R_t^{\text{new}} = \mu \Delta t I_{t-1}$	New recoveries at time t

Variable	Parameter
$A_t = \sum_{t'=0}^{t} I_{t'}^{\text{new}}$	Cumulative active cases until time t
$a_t = \alpha A_t$	Subsampled cum. active cases until time t
α	Subsampling fraction
D	Delay of case detection

Results and discussion





Noted: Discussion on page 26

Conclusion

Summury what we have learned:

- Model equation of discrete SIR model
- Estimating model parameters usign Bayesian inference with Markov-chain Monte-Carlo (MCMC)
- Choice of future scenarios (three future scenarios)
- Parameterizing future scenarios with changes in λ
- Effectiveness of strong social distancing on COVID-19

Reference

Acknowledgement: The model we used is not our own model. The model is wholly based on the model that was used for COVID-19 in germany.

- Zhang, Y., You, C., Cai, Z., Sun, J., Hu, W., & Zhou, X.-H. (2020). *Prediction of the COVID-19 outbreak based on a realistic stochastic model* [Preprint]. Infectious Diseases (except HIV/AIDS). https://doi.org/10.1101/2020.03.10.20033803
- Liu, Y. (n.d.). *Estimating the Case Fatality Rate for the COVID-19 virus: A Markov Model Application*. 17.
- Song, P. X., Wang, L., Zhou, Y., He, J., Zhu, B., Wang, F., Tang, L., & Eisenberg, M. (2020). *An epidemiological forecast model and software assessing interventions on COVID-19 epidemic in China* [Preprint]. Infectious Diseases (except HIV/AIDS). https://doi.org/10.1101/2020.02.29.20029421

Reference (cont.)

- Maier, B. F., & Brockmann, D. (2020). Effective containment explains sub-exponential growth in confirmed cases of recent COVID-19 outbreak in Mainland China. *ArXiv:2002.07572 [Physics, q-Bio]*. http://arxiv.org/abs/2002.07572
- Allen, L. J. S. (2008). An Introduction to Stochastic Epidemic Models. In F. Brauer, P. van den Driessche, & J. Wu (Eds.), *Mathematical Epidemiology* (Vol. 1945, pp. 81–130). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-78911-6_3
- Irene Li, Ziheng Cai, Zhuoran Zhang, Jiacheng Zhu. Lecture 14: Approximate Inference: Markov Chain Monte Carlo. https://sailinglab.github.io/pgm-spring-2019/notes/lecture-14/

Answers to Questions

$$S + I \xrightarrow{\lambda} I + I$$
$$I \xrightarrow{\mu} R.$$

Q: Explain the differential equations of SIR model?

Ans: the disease spreads at rate λ from an infected person (I) to a susceptible person (S) and that an infected person becomes a recovered person (R) at rate μ. N is total population. In this study, we assumed (1)The spreading rate is set to 0.4 (an initial guess that corresponds to 40% new infections day over day. (2) The recovery rate is set to 0.2 which corresponds to median recovery time of 8 days

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\lambda\frac{SI}{N}$$
$$\frac{\mathrm{d}I}{\mathrm{d}t} = \lambda\frac{SI}{N} - \mu I$$
$$\frac{\mathrm{d}R}{\mathrm{d}t} = \mu I\,.$$

The SIR system without so-called vital dynamics (birth and death, sometimes called demography) described above can be expressed by the following set of ordinary differential equations (see more on wikipedia where S is the stock of susceptible population, I is the stock of infected, R is the stock of recovered population, and N is the sum of these three.

Answers to Questions (cont.)

Q: Why did we use student t-distribution ?

Ans: because it approaches a Gaussian distribution but features heavy tails, which make the MCMC more robust with respect to outliers. The square root width models the demographic noise of typical mean-field solutions for epidemic spreading. sigma = the scale factor of the width of the Student's t-distribution of new cases.

$$p\left(\hat{I}_t^{\text{new}} \middle| \theta\right) \sim \text{StudentT}_{\nu=4}\left(\text{mean} = I_t^{\text{new}}(\theta),\ \text{width} = \sigma\sqrt{I_t^{\text{new}}(\theta)}\right)$$

Answers to Questions (cont.)

Q: what is the new notation {} ? in page 12 $\hat{I}^{new} = \left\{ \hat{I}_t^{new} \right\}$

Ans:{} is the sign of set, so $\left\{ \hat{I}_t^{new} \right\}$ is the set of time series of new infected cases of $\hat{I}^{new}$

Q: what is theta?

Ans: θ={λ,μ,σ,I} is the set of estimation of variable λ,μ,σ,I,

e.g: $p(I_t^{new}|\{\lambda, \mu, \sigma, I\}) = p(I_t^{new}|\theta)$

The MCMC sampler finds the posterior distribution p of model parameters θ such that the model equations match the real-world data. In the figure page 13, we plot the effective spread (λ−μ, which corresponds to the daily rate of case increase) inferred from the data of the past.

Answers to Questions (cont.)

Our data set is discrete in time ($\Delta t = 1$ day), we solved the previous differential equations with a discrete time step ($dI/dt \approx \Delta I/\Delta t$), by the following equations;

$$S_t - S_{t-1} = -\lambda \Delta t \frac{S_{t-1}}{N} I_{t-1} =: -I_t^{\text{new}}$$

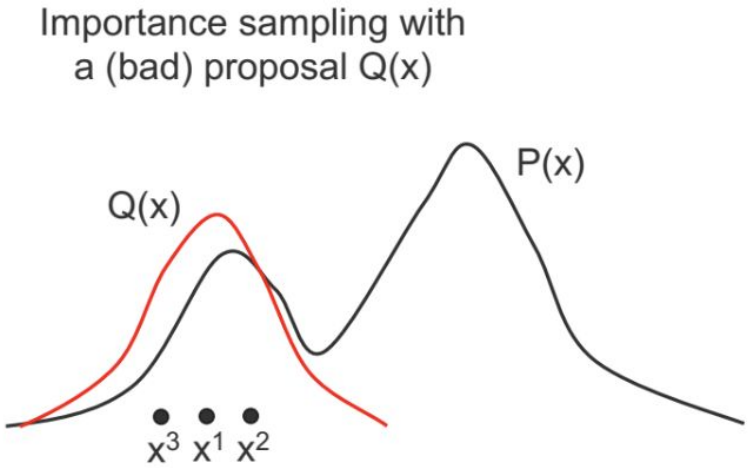
$$R_t - R_{t-1} = \mu \Delta t I_{t-1} =: R_t^{\text{new}}$$

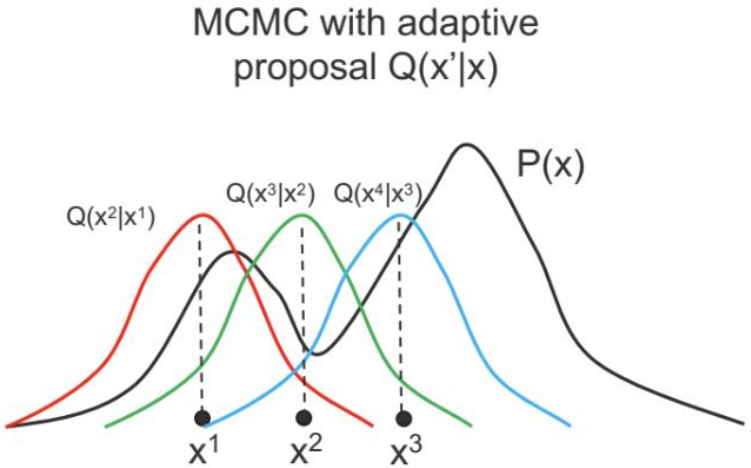
$$I_t - I_{t-1} = \left(\lambda \frac{S_{t-1}}{N} - \mu\right) \Delta t I_{t-1} = I_t^{\text{new}} - R_t^{\text{new}}$$

I_t models the number of all active, (currently) infected people, while I_t^{new} is the number of new infections that is reported according to standard WHO convention. Furthermore, we explicitly include a reporting delay D between new infections I_t^{new} and reported cases when generating the forecast.

Answers to Questions (cont.)

Q:What is P(x) and Q(x)? Why one image is bad and the other one is good? How MCMC helps?





Answers to Questions (cont.)

P(x) is the posterior distribution we want to know and sample it. Q(x) is the distribution we use to sample.

The accuracy of sampling result relays on choosing a proper sampling distribution. The sampling distribution is better when it is similar to the P(x). In the left graph, it is just Q(x) distribution. It is hard to be similar to the posterior distribution. On the right, graph , it uses Q(x'|x) as a sampling distribution. x' is the new sampling state. x changes, Q(x' | x) can also changes. It will be more possible similar to the posterior distribution and cover it. The sample result will be better.

Answers to Questions (cont.)

Discussion on figure page 16

Figure A: We modeled three different future scenarios and a past model for the development of confirmed COVID-19 cases in Thailand: unchanged spread, mild restriction of contacts, or strict restriction of contacts.

Figure B: We also analyzed how a delayed restriction impacts case numbers: Strict restrictions starting on day zero, or five days later can make a substantial difference in case numbers. A delay of five days in implementing restrictions has a major impact on the expected case numbers.

Model assumptions: The forecasts are based on the SIR model. **Figure A, red:** The infection rate is unaltered. **Figure A, orange:** The infection rate (which scales as a function of the number of contacts per person) is cut in half. **Figure A, green:** The infection rate is reduced to 1/10th. Shaded areas indicate the variability in the prediction (95 % confidence interval).

The evolution of the total number of cases, it can be seen very clearly that a change in behavior will only be visible in 7-10 days. Only on day 7, the curves start to show a difference from each other. This is because many people are already infected, but they have not yet been identified and tested. Due to this uncertainty, it is especially important to reduce the number of contacts for everyone. If we wait another 5 days, the number of cases will increase threefold, even in the best scenario **(Figure B).**