Bayesian Online Changepoint Detection on COVID-19 Data and Policy Analysis

Yuhan Fu
Data Analytics Program, Denison University
Mentor: Dr. Zhe Wang

INTRODUCTION

Methods for analyzing COVID-19 data and policies have become necessary during recent pandemics. The combination of the changepoint detection method and the policy analysis effectively illustrates the COVID-19 pandemic pattern and the potential time lag effects of policies in a country. A Bayesian online changepoint detection method models the sequential COVID-19 data and infers the most recent changepoints using a recursive online algorithm. The corresponding dates on which the changepoints occur often mark critical time points. Successful detection of changepoints allows us to investigate how scientific methods and government policies infected the COVID-19 situation in various countries.

METHODS

- **Data sequence and product partitions** [1]
- $\mathbf{x}_{s:t}$ denotes a sequence of observations $x_s, x_{s+1}, \ldots, x_{t-1}, x_t$ $(s \le t)$.
- Assume T data points $\mathbf{x}_{1:T}$ can be divided into product partitions (each partition are i.i.d.)
- Changepoints occur between partitions.

- **Run length**
- Two cases of run length (the time since the last changepoint):

$$r_t = \begin{cases} 0 \text{ (changepoint at } t) \\ r_{t-1} + 1 \text{ (otherwise).} \end{cases}$$

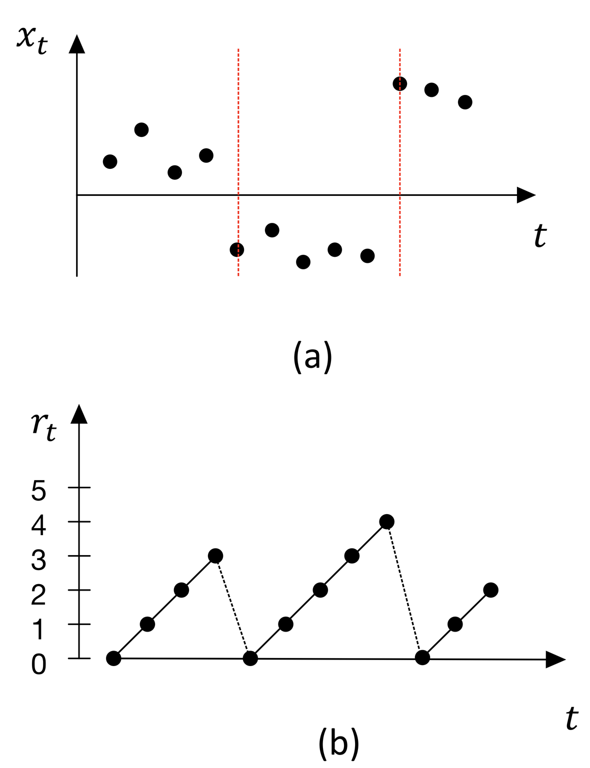

Figure 1. Conceptual diagrams of **(a)** data partitioned by two changepoints, and **(b)** the run length with respect to time.

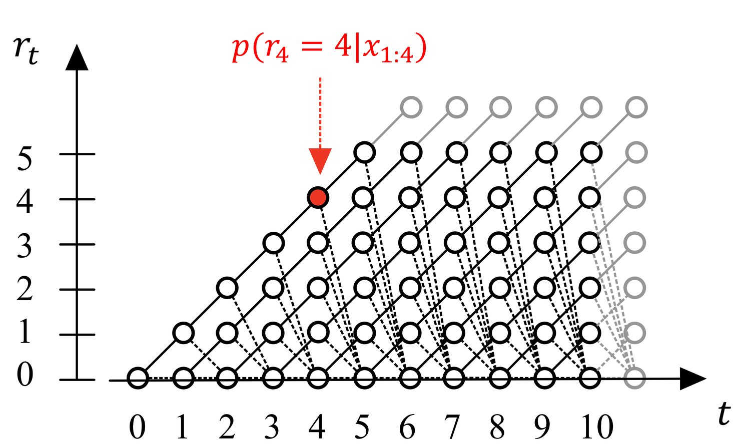

Figure 2. The trellis of message-passing algorithm. Each node has associated mass. For example, the red node has probability $p(r_4 = 4|x_{1:4})$.

METHODS

- **Recursive run length posterior estimation**
- A sequential inference algorithm that calculates the run length posterior recursively, i.e. the probability of each node in Figure 2.
  - Boundary conditions, Changepoint prior, Posterior predictive, Message-passing parameters

- **Changepoint detection method**
- To detect the changepoint more accurately, we plot the probability that a changepoint occurs at each time point t and find all local maxima that meet a certain threshold probability, the lowest probability we accept to suggest a changepoint.

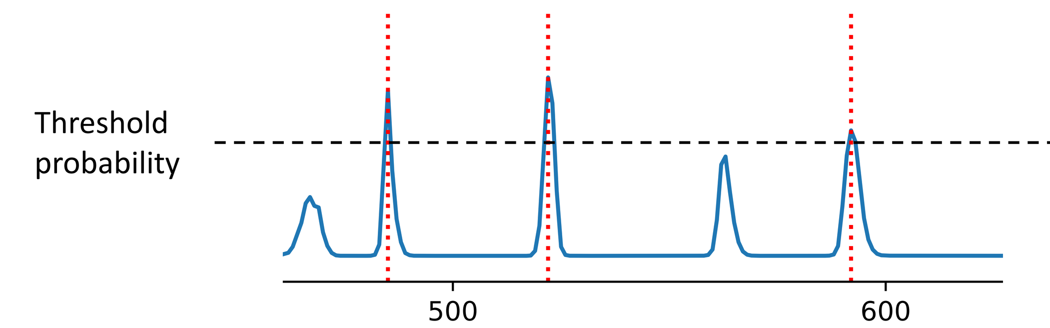

Figure 3. The diagram of probabilities of changepoints. The peaks above the threshold probability are treated as changepoints.

RESULTS

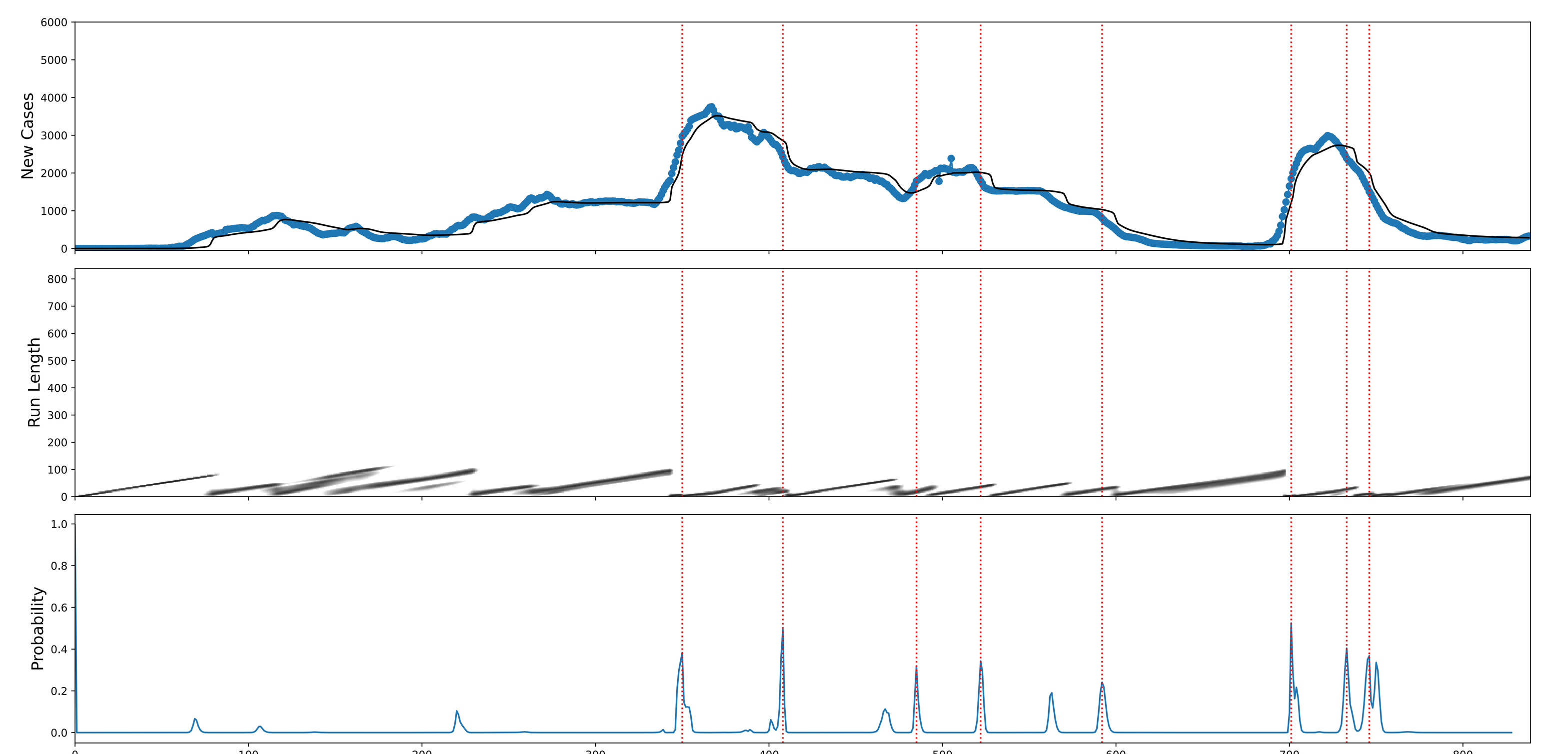

Figure 4. (Top) The smoothed daily new cases of COVID-19 (blue line) in United Arab Emirates and the predictive mean (black line). **(Middle)** The run length posterior at each time step using a logarithmic color scale. Darker pixels indicate higher probability. **(Bottom)** The probability that a changepoint occurs at each time step. Red dashed lines denote changepoints.

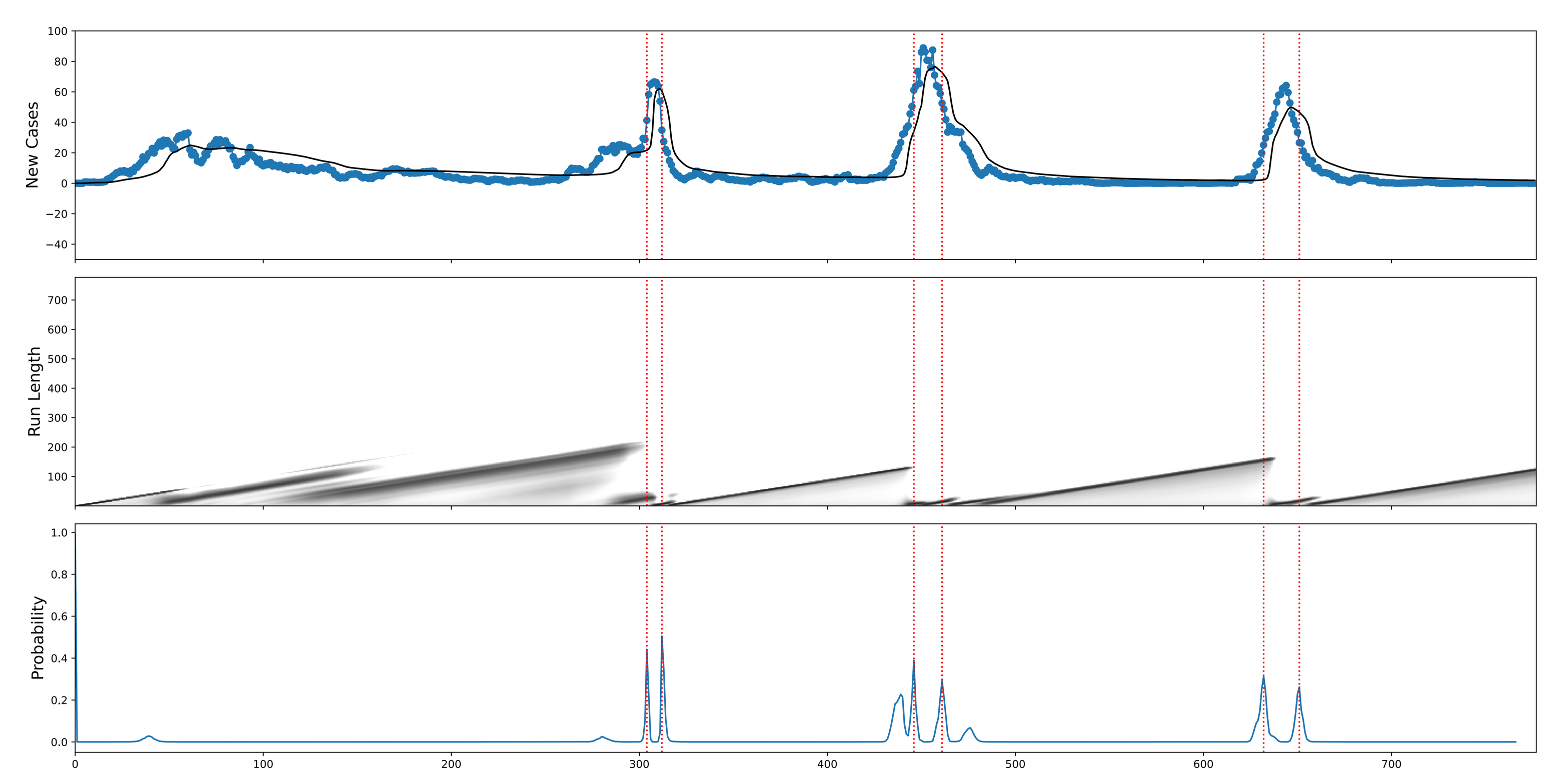

Figure 5. (Top) The smoothed daily new cases of COVID-19 (blue line) in Sierra Leone and the predictive mean (black line). See Figure 4 for interpretations of Middle and Bottom plots.

- We apply the model to seven countries, but only two are shown here as examples: United Arab Emirates and Sierra Leone.

- Three subplots: New Cases (smoothed), Run length, and Probability. Red dashed lines mark the locations of changepoints.

United Arab Emirates
- Starting date: 01/29/2020
- Changepoints: [350, 408, 485, 522, 592, 701, 733, 746]

Examples of changepoints and relevant events:
- 01/31/2022 (Day 733) – new cases dropped
  - 01/17/2022 – boosters available for all types of vaccinations

- 02/13/2022 (Day 746) – new cases dropped
  - 02/25/2022 – Abu Dhabi Emergency, Crisis and Disasters Committee approved reduced precautionary measures

Sierra Leone
- Starting date: 03/31/2020
- Changepoints: [304, 312, 446, 461, 632, 651]

Examples of changepoints and relevant events:
- 06/20/2021 (Day 446) – new cases increased
  - 06/17/2021 – required face covering outdoor at all times

- 01/11/2022 (Day 651) – new cases dropped
  - 12/10/2021 – WHO offered a consignment of COVID-19 laboratory commodities worth over $600,000 to Sierra Leone

DISCUSSION

- This Bayesian Online Changepoint Detection method is based on a Gaussian univariate model. Depending on the type and our prior knowledge of the data, we can also parameterize various distributions differently, such as using an inverse gamma or an inverse chi-squared [2]. There are also methods using "Changepoint Hierarchy Generative Model" [3] to do a similar Bayesian inference but assuming the hazard rate is unknown.

- Countries are chosen based on the cluster analysis but not by regions. Some countries are not fully analyzed mostly because there is already a representative of the cluster. Similarly, some regions have more than one countries analyzed because they fall into different clusters.

- Higher probability often results from a larger number of changepoints. We sometimes sacrifice the probability to achieve a decent amount of changepoints. The threshold probability also varies for different data.

- The probability shown is the probability we predict backward after waiting for a few data points. Thus, a relatively higher probability compared to its neighbors is enough to indicate a changepoint.

- The issue and implementation of a policy can depend on various factors other than the rise and drop in new daily cases. The connection or disconnection between the COVID-19 cases and the policies is not an absolute standard for the appropriateness of the response.

ACKNOWLEDGMENTS

This research project is supported by Denison University Research Foundation. Thanks Dr. Zhe Wang for giving me useful instructions during the research and paper writing process. I also appreciate the work done by Mark Raney and Riley Coburn, who contribute to other parts of the entire project, including the data mining, the cluster analysis and the dynamic linear modeling.

REFERENCES

[1] Ryan Prescott Adams and David JC MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.
[2] Kevin P Murphy. Conjugate bayesian analysis of the gaussian distribution. *def*, 1(2σ2):16, 2007.
[3] Robert C Wilson, Matthew R Nassar, and Joshua I Gold. Bayesian online learning of the hazard rate in change-point problems. *Neural computation*, 22(9):2452–2476, 2010.