Bayesian change point detection: a case study on COVID-19 confirmed cases in UK

Zeliang Wang

1. Problem definition

As COVID-19 is rapidly spreading across the globe, I have been trying for some time to come up with an interesting COVID-19 problem to address with statistics, in particular with Bayesian models. After examining some data, it was clear that at certain date, the growth of new cases stopped being exponential, which implies the distribution of new cases changes. Also, that date varies among different countries. The aim of the study is to identify the date when the distribution of new COVID-19 cases in a particular country changes. This can help policy maker evaluate the effectiveness of interventions, such as social distance measure.

By looking at the number of daily cases as shown in Fig. 1, it is very difficult to visually determine a change point, and it is even more difficult by looking at the accumulated number of cases.

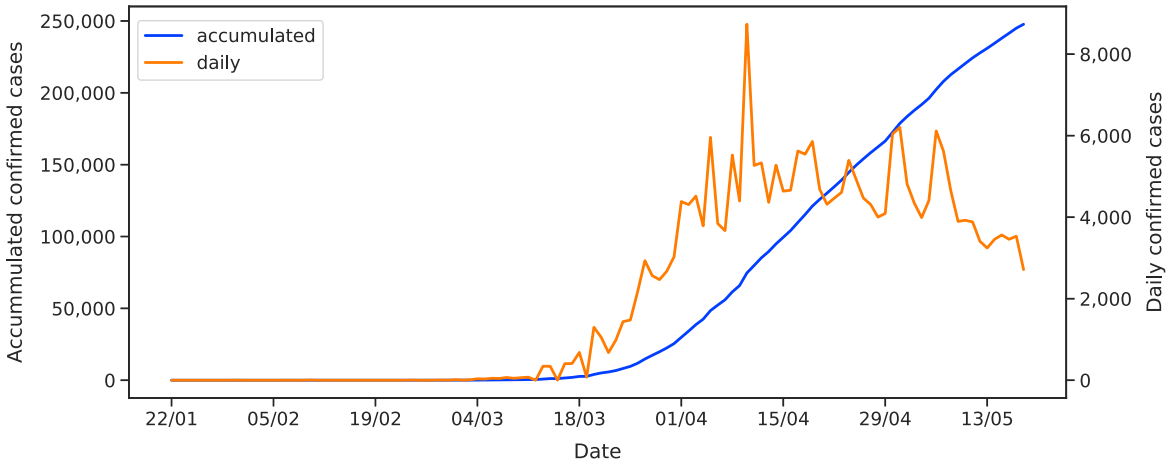

Fig. 1. Confirmed COVID-19 cases in UK.

All the code used to produce the results in this study can be found here.

Extend to Eniscope Recorded Data

Potentially, Bayesian change point analysis can also be applied our Eniscope recorded data to inform the change points of energy consumption for a particular meter channel. If the energy usage of a monitored device changed in a characteristic way at some date, for example, an air conditioning unit with old, clogged filters will gradually use more energy, we would like to identify that date on which changes to the normal usage occurred by using Bayesian change point analysis. This will lead to the development of the predictive maintenance software.

2. Proof of concept

To check if there exists a real change point (flattened curve) of the daily confirmed cases in UK, we firstly select the UK's COVID-19 data from 3rd March till 30th April, and then split the data into two parts based on a randomly chosen date, in this case, 42 days since 3rd March. We then fit linear regression models to both parts of data (See Fig. 1 for the piecewise linear regression models).

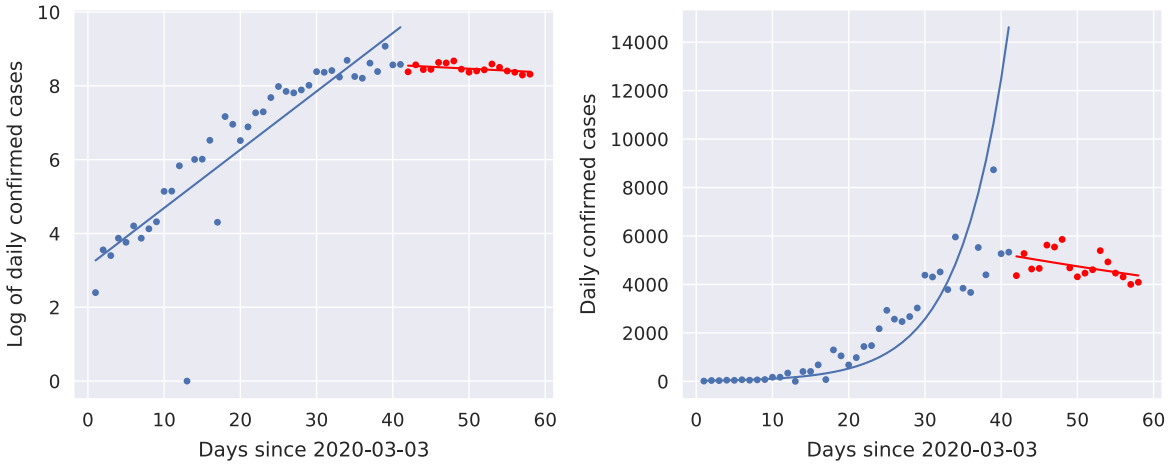

Fig. 2. Piecewise regression model for COVID-19 cases in UK.

As shown in Fig. 2, it is clear that these two linear regressions are different both in slope and intercept. This means that there exists a real change point during this period when the number of daily confirmed cases start flattening. Now the problem becomes **how we can estimate that change point using Bayesian inference**.

3. Modelling

To model the relationship between y, log of the number of new COVID-19 cases in UK each day, and t, the number of days since the "start" date, we will use a segmented regression model. The point at which we segment will be determined by a learned parameter, τ.

We firstly need to find the likelihood function of data represented by a line in the form of $y = wt + b$, where any reason for the data to deviate from a linear relation is an added offset in the y direction. The error y_i was drawn from a Gaussian distribution with a *zero mean* and *variance* σ^2. In this model, given an independent position of t_i, an uncertainty σ, a weight (or slope) w, a bias (or intercept) b, the probability density function p is

$$p(y_i|t_i, \sigma, w, b) \sim N(y_i - wt_i - b, \sigma^2)$$

where

$$w = \begin{cases} w_1 & \text{if} \quad \tau \leq t \\ w_2 & \text{if} \quad \tau > t \end{cases}, \qquad b = \begin{cases} b_1 & \text{if} \quad \tau \leq t \\ b_2 & \text{if} \quad \tau > t \end{cases}.$$

Therefore, the likelihood can be expressed as

$$\mathcal{L} = \prod_{i=1}^{N} p(y_i | t_i, \sigma, w, b)$$

As the amount of data available is very limited, and the growth rate of the virus is very sensitive to population dynamics of different countries, it is important to supplement the model with appropriate priors. In our case, the priors are given by

$$w_1 \sim N(\mu_{w_1}, \sigma^2_{w_1}) \ , \ w_2 \sim N(\mu_{w_2}, \sigma^2_{w_2}) \ ,$$
$$b_1 \sim N(\mu_{b_1}, \sigma^2_{b_1}) \ , \ b_2 \sim N(\mu_{b_2}, \sigma^2_{b_2}) \ ,$$
$$\tau \sim Beta(\alpha, \beta) \ , \ \sigma \sim U(0, 3)$$

In other words, y will be modelled as $w_1 t + b_1$ for days up until day τ, and after that it will be modelled as $w_2 t + b_2$. In this context, w_1 and w_2 can be seen as the growth rate of the virus before and after the change point τ. b_1 and b_2 are the bias terms for the first and second regression respectively, and they are sensitive to country characteristics. Countries that are more exposed to COVID-19 will have more confirmed cases than those less exposed. Our initial guess of these priors are

$$w_1 \sim N(0.5, 0.25) \ , \ w_2 \sim N(0, 0.25) \ ,$$
$$b_1 \sim N(\mu_{q_1}, 1) \ , \ b_2 \sim N(\mu_{q_4}, \frac{\mu_{q_4}}{4}) \ ,$$
$$\tau \sim Beta(4, 3) \ , \ \sigma \sim U(0, 3)$$

where μ_{q_1} and μ_{q_4} represents the mean of the first and forth quartiles of y.

4. Results

To compute the distribution over the parameters, we use Markov Chain Monte Carlo (MCMC) to approximate inference. Specifically, Hamiltonian Monte Carlo algorithm with No-U-Turn Sampler (NUTS) is used for sampling the posteriors. For the implementation of the model, we use a python package called Pyro, a probabilistic programming language built on PyTorch.

The posterior distributions of w_1, w_2, b_1, b_2, τ and σ are illustrated in Fig. 3. **Clearly, the posteriors for w_1 and w_2, along with b_1 and b_2 do not overlap with each other, which proves that the change point estimated by our model is true**.

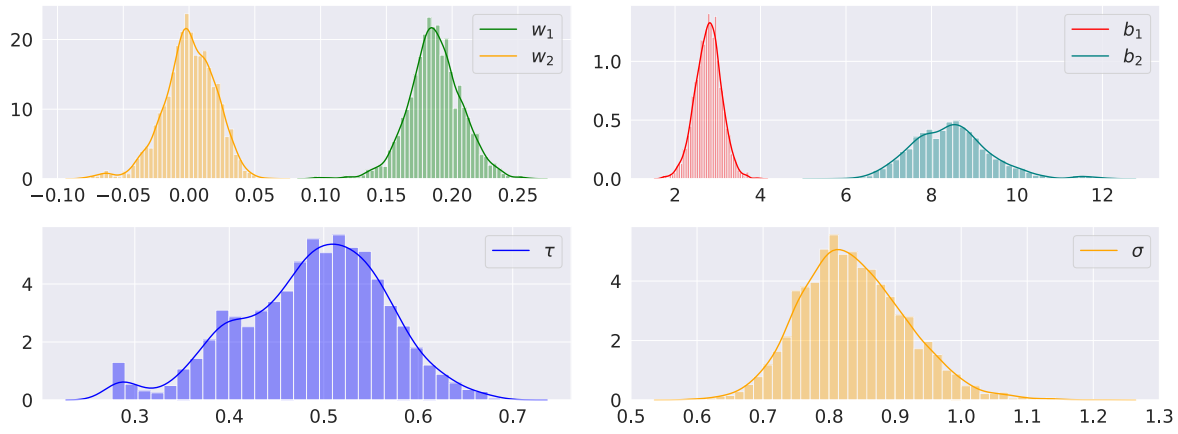

Fig. 3 Posterior distributions of model parameters for COVID-19 confirmed cases in UK.

Therefore, the best estimate of the parameter τ is given 0.487, and since we only selected 58 days (from 3rd March to 30th April) data for the model, the change point is estimated as **1st April 2020**.

If we recall the timeline of COVID-19 in UK, on 16th March, the UK government advised everyone in the UK against "non-essential" travel and contact with each other, as well as suggesting people working from home if possible. In addition, assuming the incubation period of the virus is up to 14 days, this estimated change point does make sense as it is 15 days after widespread social distancing measures began.

Predictions

Based on the posterior distributions of these parameters, the predicted daily news cases on day $t_i = \{1, 2, \ldots, N\}$ can be formulated as

$$\mathbf{y}_i = \mathbf{w}t_i + \mathbf{b} \ , \qquad \mathbf{y}_i, \mathbf{w} \text{ and } \mathbf{b} \in \mathbb{R}^{M \times 1} \ ,$$

where M denotes the number of accepted samples after using MCMC.

The left plot of Fig 4 shows the log of number of daily cases which is what we used to train the model, along with the predicted mean and 90% credible interval. The dotted vertical line indicates the change point along with its 90% credible interval estimated by the model. Also the right plot shows the real number of daily cases.

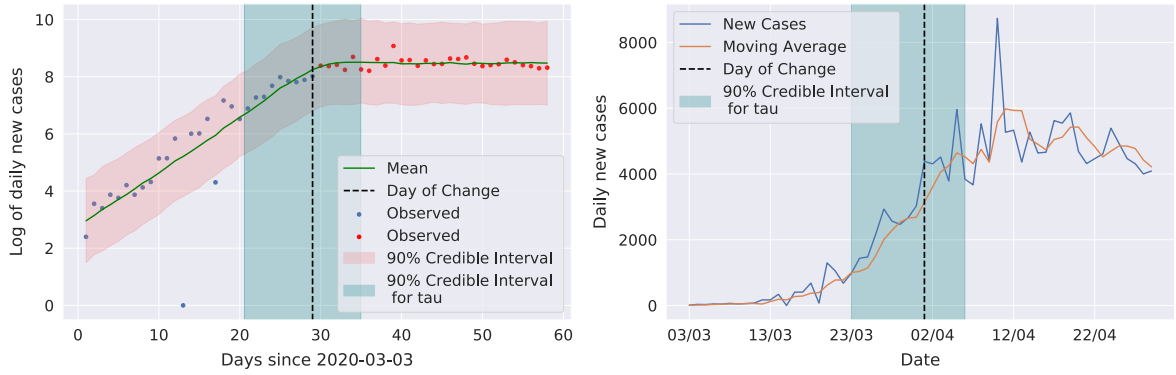

Fig. 4 Estimated change point for COVID-19 cases in UK.

References

- Wikipedia: COVID-19 pandemic in the United Kingdom
- Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions
- Probabilistic programming in Python: Pyro versus PyMC3