Change Point Detection with Dynamic Linear Modeling on COVID-19 Time Series Data

Denison University
Riley Coburn
Research Advisor: Dr. Zhe Wang

DENISON

Introduction

COVID-19 has had a massive impact on society over the last two and a half years. Lives have been and continue to be permanently affected by this virus.

There has been many different phases to the COVID virus. Initially, most governments tried to contain it until a suitable vaccine could be developed and distributed to its citizens. Although this was attempted, it was rarely successful. Different variants of COVID-19 ran rampant worldwide causing the number of new cases in many countries to fluctuate with each new variant. The these fluctuations are marked as change points, or points where the transmission and contraction of the virus structurally changes. The following presents an offline change point detection method for COVID-19 time series.

Background & Methods

As aforementioned, the goal of this research is to present an offline change point detection method that could be implemented in an online setting. The offline method is one in which all of our data is known and we're able to optimize model parameters using maximum likelihood estimation. With optimized parameters, the model could then be used on new observations and implemented as we receive new data.

There are different ways of determining change points given some data. One such method is through a dynamic linear model. The premise of this model is that each observation, y_t is dependent on the underlying state at time t, θ_t.

The first step in specifying a dynamic linear model is choice of prior distribution of θ_t. This initial parameter is

$$\theta_0 \sim N(m_0, C_0)$$

where m_0 is our initial guess at the state mean and C_0 is the initial guess of state variance. C_0 can easily be adjusted to reflect uncertainty in the initial state value. A higher value of C_0 represents a more uncertain guess in m_0.

From here, we can obtain a value for θ_1 and subsequent state values using

$$\theta_t = G_t\theta_{t-1} + w_t$$

where G_t is the known state transition matrix and w_t is some Gaussian distributed error. The state updated equation is an autoregressive equation, with future state values depending only on the previous state.

To find the value of an observation at time t we use the observation equation:

$$Y_t = F_t\theta_t + v_t$$

where F_t is the known observation matrix and v_t is some Gaussian error. As can be seen, the observation at time t depends on the state at time t, θ_t.

We use the conditional probability $\theta_t|y_{1:t} \sim N(m_t, C_t)$ to find the updating equation necessary for predictions. Our prediction of the state at time t given the data up to that point has a gaussian distribution with mean m_t and variance C_t. This is reminiscent of the a traditional Bayesian filtering algorithm which estimates state values up to and including the data up to that time t. This is not to be confused with a Bayesian smoothing algorithm which makes state estimations given the entire set of data all at once.

We use the conditional probability $\theta_n|y_{1:n} \sim N(m_n, C_n)$ to find the updating equation necessary for predictions. Our prediction of the state at time t given the data up to that point has a gaussian distribution with mean m_t and variance C_t.

The updated procedure is simple from here. We use Bayes formula where we have

$$\pi(\theta_n|y_{1:n}) \propto \prod_{t=1}^{n} \pi(y_t|\theta_t)\pi(\theta_t)$$

Evaluating this gives the kernel of a normal distribution and a way of updating the distribution of $\theta_t|y_{1:t}$. Now, given $\pi(y_n|\theta_n) = N(y_n; \theta_n, \sigma^2)$ we can update the prior distribution for $\theta_{n-1}|y_{1:n-1} \sim N(m_{n-1}, C_{n-1})$ based on y_n. The equations that are used for updating the state distribution are

$$m_n = m_{n-1} + \frac{C_{n-1}}{C_{n-1} + \sigma^2}(y_n - m_{n-1})$$

and

$$C_n = \frac{\sigma^2 C_{n-1}}{C_{n-1} + \sigma^2}.$$

The predictive distribution of $Y_{n+1}|y_{1:n}$ is normal with mean m_n and variance $C_n + \sigma^2$. Thus, m_n is the posterior expected value of θ_n and the forecasting expected value.

We now cover the predictive distribution $\theta_n|y_{1:n-1} \sim N(a_n, R_n)$. This distribution follows the formula $\theta_t = G_t\theta_{t-1} + w_t|y_{1:t-1}$ where the error w_t has yet to be updated.

By this point, change points can be calculated since we have a conditional distribution and predictive distribution for θ_n in $\theta_n|y_{1:n} \sim N(m_n, C_n)$ and $\theta_n|y_{1:n-1} \sim N(a_n, R_n)$, respectively. A simple and intuitive way of considering if $t = n$ is a change point would be to see if m_n is an outlier in $N(a_n, R_n)$. In other words, we want to see whether what we expect for θ_n aligns with what was predicted for θ_n. We can calculate this z-score by calculating $z = (m_n - a_n)/R_n$. These z-scores can then be converted to p-values in the normal distribution using a z table or equivalent programming command.

Another way of calculating the extremeness of a z-score is through extreme value theory. We want to calculate $P_{EV}(z|Z_m) = P(\max(Z_m) \leq z)$ where $Z_m = \{z_i\}_{i=1}^{m}$. We are calculating the probability that, given some sequence of m z's, the z that we are observing is a maximum. I will show both approaches in the below analysis.

Background and Methods (cont.)

Analysis

Both of the following analyses were conducted using R. The *dlm* package was used to implement the dynamic linear model formulae. This analysis was implemented on the smoothed number of new COVID-19 cases in Sierra Leone because of its relatively smooth COVID-19 numbers. This made it easy to empirically determine change points and see whether the model was correctly characterizing change points.

The first analysis conducted focused on the z-scores calculated by taking $\frac{|a_n - m_n|}{R_n} \sim |N(0,1)|$. These z-scores were converted to p-values in the one-sided standard Gaussian distribution. A significance level of 0.05 was used to observe the locations of change points. Even then, the model seems to identify large sequences of change points. Moreover, it fails to identify obvious changepoints, especially those at earlier times. In a lot of cases, change points are identified at times just past where change points should be appearing.

The extreme value method doesn't perform too much better, if at all. Again, large sequences of change points are being found. In addition, it also fails to identify apparent change points. The one advantage that this model takes is that, when change points are detected, this method much better characterizes them at or near the change point, not after the change point has appeared.

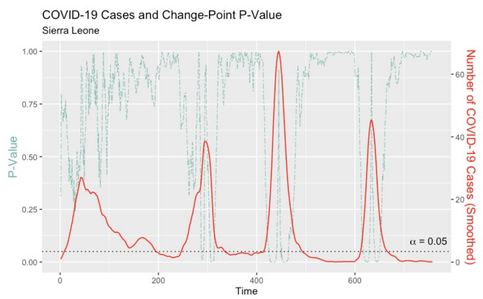

Figure 1: Predicted change points using p-values of m_n in $N(a_n, R_n)$.

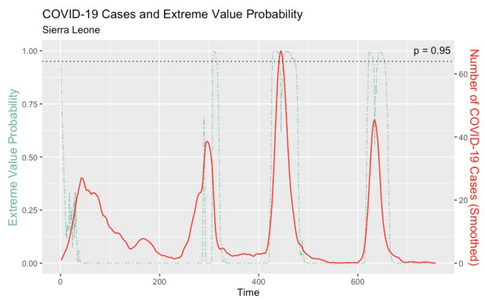

Figure 2: Predicted change points using the extreme value probability of m_n in $N(a_n, R_n)$.

Future Work

The most pressing work would be to determine why some change points are being detected and others aren't, especially in the case of the extreme value method. This method was one which showed the ability to correctly identify change points, albeit not all that were present. More work could be done with optimizing a cutoff point for change points as well as removing the ambiguity of having a long sequence of change points. Once the method has been improved, it can then be implemented on other countries.

References

Daumer M., and Falk M. (1998), "On-line change-point detection (for state space models) using multi-process Kalman filters," *Linear Algebra and its Applications*, 284, 1–3.

Lee H., and, Roberts S.J. (2008), "On-line novelty detection using the Kalman filter and extreme value theory," *19th International Conference on Pattern Recognition*, pp. 1-4.

Petris G., Petrone S., and Campagnoli P. (2009), "Dynamic Linear Models with R," New York, NY: Springer.

Acknowledgements

A very special thank you to the Laurie and David Hodgson Faculty Support Endowment, without whom the funding for this project would not be possible. Another thank you goes out to Dr. Zhe Wang, whose guidance and reassurance were essential in my completion of this research project.