An Analysis of a Circular Changepoint Model – A Covid-19 Case Study –

Matthew Scholes with Owen Li

University of Lancaster and STOR-i

Introducing Changepoint Detection

- Changepoint detection aims to find the location in a time series where there is a sudden **change in parameter**. By locating this changepoint (cpt) one can split the data into a collection of segments which all have unique parameters to describe them.
- This allows one to build a model, based off these segments, that much better defines the data, rather than having a singular set of parameters for the whole time series.

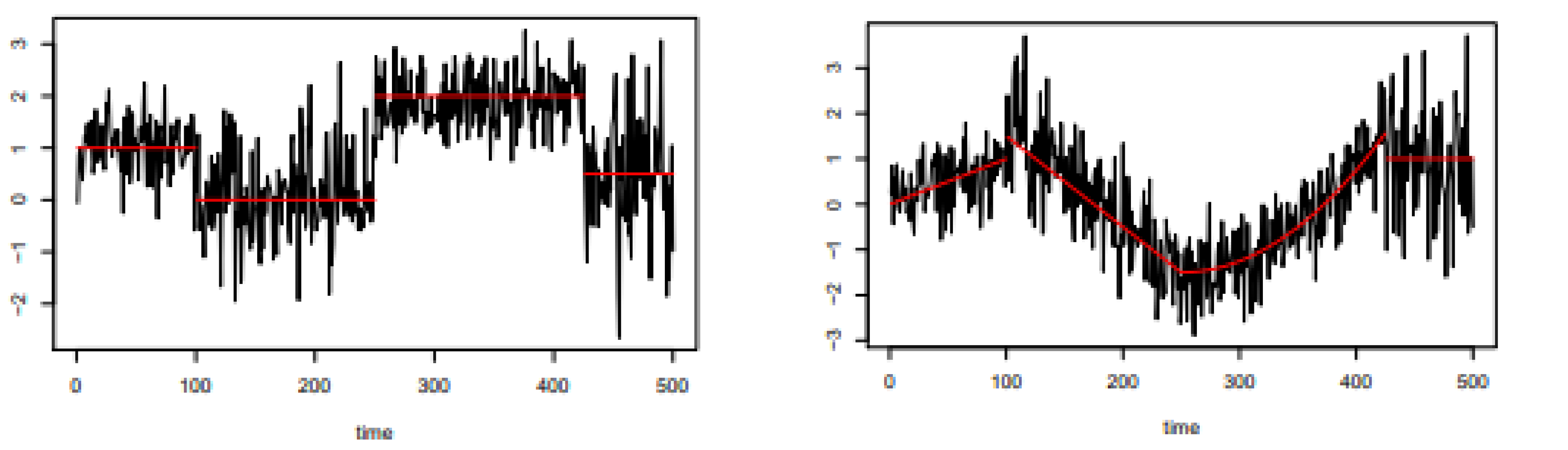

Figure 1: Graphs show examples of how parameters change given at changepoint locations: LHS - mean; RHS - trend [1].

- One approach to locate changepoints is to **maximise the likelihood** of the model to the data, while introducing a penalty to limit the number of cpts chosen, shown by the **Penalised Likelihood Function** [2]:

$$L(\mathcal{M}_\kappa) = -2\log\max L(\theta_\kappa) + p_\kappa \phi(n) \tag{1}$$

Where $L(\mathcal{M}_\kappa)$ is the likelihood of a model with p_κ parameters fitting the data, this is also called the cost function ($\mathcal{C}$). The parameters themselves are denoted as θ_κ and the penalty term is introduced as $\phi(n)$.

Defining The Circular Model

- The examples in figure 1 finds changepoints in a linear time series; however, if there are periodic patterns in the data then it can be beneficial to think of the time axis as circular, repeating on that period.
- This can helpful as a deviated point may be considered as an outlier when it is considered in isolation, however if repeated periodically then it is likely significant.
- A method developed by Li and Killick (2022+) is based off the **Segment Neighbourhood** approach and is detailed below.

Segment Neighbourhood Search (SegN):

- This linear approach takes in a maximum number of changepoints (m) giving a maximum number of segments ($M = m + 1$).
- The algorithm then minimises the cost for $0, 1, ..., m$ changepoints using dynamic programming, until an optimum solution is found.
- This finds the exact solution as once a cpt is found it is not restricted to that location, meaning the solution converges on the global minimum.
- This can be implemented by minimizing the following function where τ_i is the location of a changepoint [3]:

$$\sum_{i=0}^{m} \mathcal{C}(y_{(\tau_i+1):\tau_{i+1}}) \tag{2}$$

The Circular Model

- This method is structured by wrapping the time axis on itself; so if the time series is n points long with a period of T, then there will be $\frac{n}{T}$ data points at each time location.
- By doing this even small deviations can be considered segments as there is now more data to make them significant.
- The algorithm is based off SegN (equation 2) and simply takes $\mathbf{x}$ as a vector instead of a singular point:

$$\sum_{i=0}^{m} \mathcal{C}(\mathbf{x}_{(\tau_i+1):\tau_{i+1}}) \tag{3}$$

Success of The Circular Model

- To quantify how successful this method is, a simulation was run by taking normally distributed, periodic data with a constant variance of $\sigma = 1$ and a varying mean (μ).
- Success is defined only if both changepoints are located at the correct place.

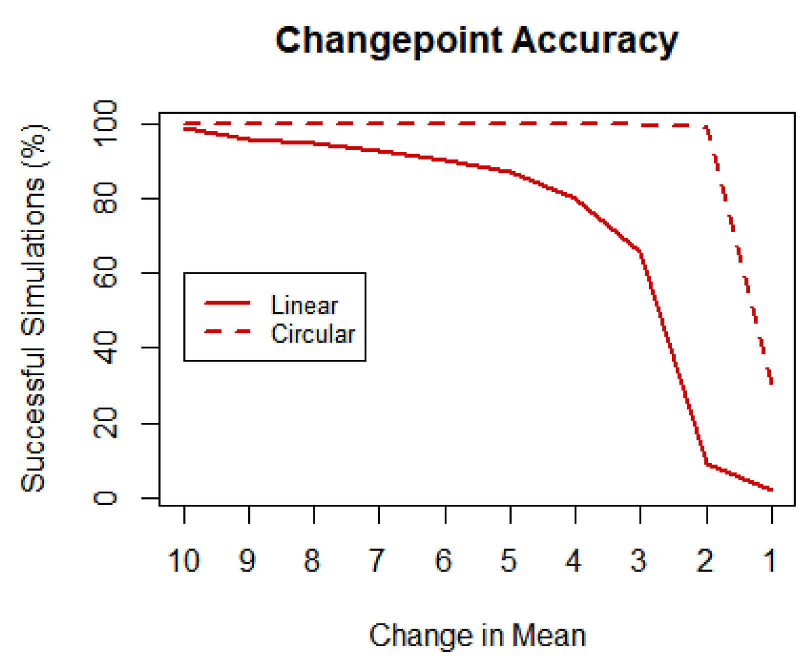

Figure 2: 100 simulations were run for normally distributed data for varying $\Delta\mu$, comparing linear and circular changepoint detection methods, with the results displayed.

- The **linear method struggles** to consistently find the correct cpt location, even when means are clearly distinct and completely fails when the $\Delta\mu = 1$.
- However the **circular method succeeds** in finding the correct changepoints up to and including $\Delta\mu = 2$ and still performs better than SegN for $\Delta\mu = 1$.

A Covid-19 Case Study

- Evidence suggests that there is a weekly periodicity in positive cases of Covid-19 in the UK, this is likely because testing reduces over the weekend.
- For this reason we expect to find changepoints on Fridays and Sundays [4].

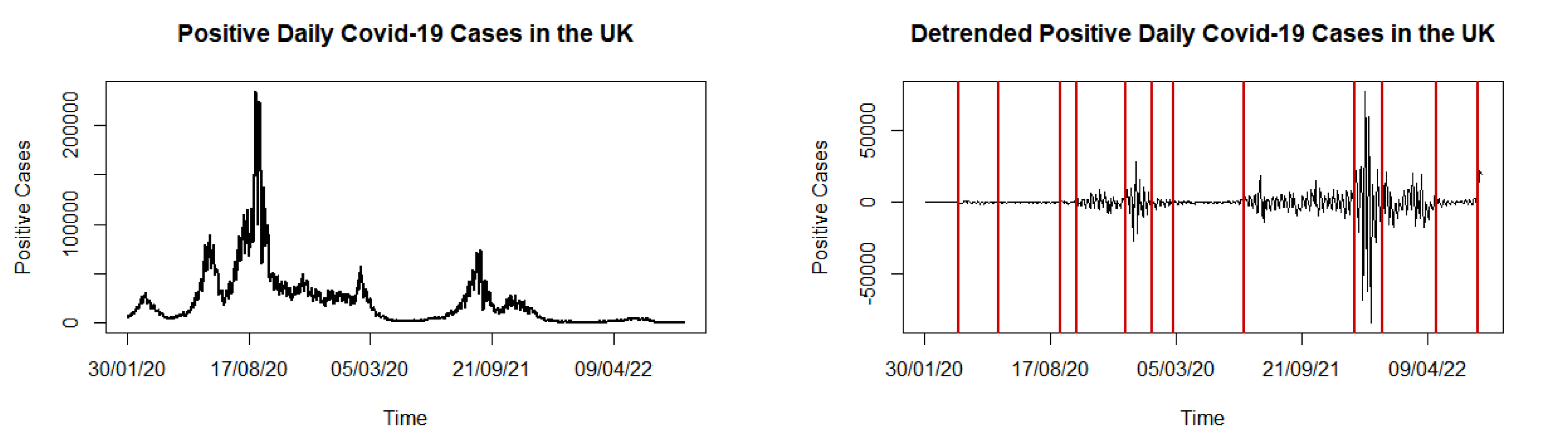

Figure 3: LHS - Daily positive cases in England, RHS - Same data cleaned, with changepoints found by SegN

- The RHS of Figure 3 shows the changepoints found by SegN on the cleaned data, clearly the **linear method fails** to locate the weekly changes.
- By conducting a Fast Fourier Transform on cleaned data, a periodicity of 7 days was found, allowing the circular method to be used.
- Due to high variance in the data the method was applied to the segments found using SegN with the results shown below.

Mon	Tue	Wed	Thu	Fri	Sat	Sun
2	2	1	3	5	3	7

Table 1: Location of changepoints found by Circular Model

- Despite not consistently finding the correct cpt location, Friday and Sunday are the two most popular, found 43% of the time.
- This is a **big improvement** on SegN which finds no periodicity at all.

References

[1] I. A. Eckley *et al.*, "Analysis of changepoint models," *Bayesian time series models*, vol. 205, p. 224, 2011.

[2] Hinkley *et al.*, "Inference about the change-point in a sequence of random variables," 1970.

[3] J. V. Braun and H.-G. Muller, "Statistical methods for dna sequence segmentation," *Statistical Science*, pp. 142–162, 1998.

[4] Q. Bukhari *et al.*, "Periodic oscillations in daily reported infections and deaths for coronavirus disease 2019," *JAMA network open*, vol. 3, no. 8, pp. e2 017 521–e2 017 521, 2020.

Contact: s1748379@ed.ac.uk