Change-point Analysis: Looking at Forced Marriage Call Data

Nicholas Geere, Fabian Hernandez, Dr Rowland Seymour

Rights Lab, University of Nottingham

Forced Marriage Helpline Data used in Change-point Analysis

Firstly, a change-point is a time where the response variable starts to follow a different distribution. Below is an example of a Change-point algorithm being applied to some time-series data, data indexed in time order.

$$X_i \overset{ind}{\sim} Poi[\lambda_i] \qquad \lambda = \begin{cases} \lambda_1 & \text{for i} < t^1 \\ \lambda_2 & \text{for i} > t^1 \end{cases}$$

The type of data used is the call data received from two call centres which are set up to help victims of forced marriage and other honour-based human rights abuses. The first is the Forced Marriage Unit (FMU) based in London and the second is Karma Nirvana (KN) based in Leeds.

The data contains the call volumes received by both organisations and details about the calls such as age of the victim, the person who got in contact with the helpline, the region of the world a particular call might be related to such as the victims birth county and region of the UK the call is from. However, these are all vague enough to keep the victims identity hidden and safe.

Models and Methods

To uncover information about the data, specifically with respect to Change-point, the analysis was done within R using 2 packages called changepoint and mcp. The benefit of using the changepoint packages:

- **The speed** is faster allowing dynamic/reactive presentation tools e.g flexdashboard
- **It's simple:** and can be learnt and used to present easily
- **It's applicable at scale** and can easily be used to analyse a lot of data in little time

This other package used was mcp, which goes into more detail than changepoint at the cost of run-time. Some of the benifits of using mcp are below:

- **Prior distributions** allow you to account for and test any prior belief you might have about the data
- **More model choice** in the model selection allows you to choose sloping patterns between change-points
- **auto-regression** in mcp allows you to account for any lag effects from previous data points.

The approach has been to look at the changepoint package first, to analyse the data to find when the optimum location for a certain number of change-points would be and to test how certain penalties such as the Bayesian Information Criterion (BIC) affected the number of change-points.

Then take a few models from that to work out the models whose parameters have distributions that fit well Monte Carlo Markov Chain (MCMC) sampling. This should produce data which follow similar distributions, indicating that the parameters were chosen well and gives some evidence that the model fit is accurate.

However, the graphs produced from MCMC sampling may produce similar outcomes and you might have multiple models which you want to compare. This is where a method of comparison helps, called Leave-One-Out (LOO) cross validation for determining which is a better fit to the data. This gives a comparison of the models and shows you how many time better the best model is to the one you're judging. It uses the LOO Estimated Log Predictive Density (ELPD) and Standard error $elpd/se$ ratio as the comparison.

Bayes' Rule

Bayesian Statistics is heavily used to judge the details about the change-point models in the packages mentioned above. Fundamental to Bayesian statistics is Bayes' Rule:

$$P(A \mid B) \propto P(B \mid A)P(A)$$

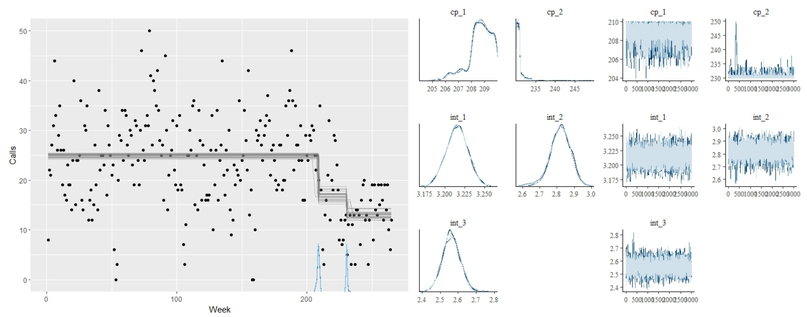
On the Left is the posterior distribution or the distribution the data follows accounting for the data. The probability B given A is the likelihood and the P(A) also represents prior beliefs about the data, such as how the population of A is distributed. It's better to leave this vague at the beginning when you're just starting with a data set to avoid over-fitting. Understanding these concepts is what allows you to understand how the models in mcp work.

Fitting Models

When fitting models, testing the parameters helps get an idea of how well the model fits to the data. In the changepoint package, the cpt.meanvar function allows you to change the maximum number of change-points, minimum segments length and the type of penalty applied to finding change-points, allowing you to good initial models to fit the data. The penalties are:

1. **Binary Segmentation** runs through the data to find the optimum location of one change-point and then repeats the algorithm on either side of that change-point until the algorithm can't find more in each sub-segment. An approximate method.
2. **PELT** (Pruned Exact Linear Time) algorithm works its way through the time series data, only considering the last change-point and where there would be any new change-points from then on.

Below the model fits 2 change-points to the data, now using **mcp**, shown by the lines at the end of the graph on the left. On the right shows the convergence of the MCMCs (Monte Carlo Markov Chains), which if the parameters fit the data well, should all have overlapping distributions if they all produce similar models. Below is all the data from the FMU since 2015 but not including 2018 (missing):



As the graph shows the data seems to fit two change-points but another model which also is a good fit to the data is a sloping down model at the right end of the time series. This choice of multiple models allows for different conclusions to presented, which can be checked against the call handlers and then get a good overall picture of what underlying trends exist. Before making any conclusions, it's worth mentioning the data can also be affected by change in definitions of forced marriage, something which affects both data sets to varying degrees.

Estimation of Missing Data

The FMU has missing weekly data for 2018 and Bayesian Inference was used to estimate values and create a 95% CI. First, we assume the call data is Poisson distributed and our prior belief is the lambda of the Poisson data is Gamma distributed:

$$X_i \overset{ind}{\sim} Poi[\lambda]$$
$$\lambda_i \overset{ind}{\sim} Gamma[\alpha, \beta]$$

Firstly, we find out the probability of a certain λ value given the data.

$$P(\lambda \mid \underline{\mathbf{x}}) \propto P(\underline{\mathbf{x}} \mid \lambda)P(\lambda)$$
$$\propto \lambda^{\Sigma x_i+\alpha-1}e^{-(n+\beta)\lambda}$$
$$\overset{iid}{\sim} Gamma(\Sigma x_i + \alpha, n + \beta)$$

Then we want to find the probability of a new data point, given the data we already know.

$$P(x_{new} \mid \underline{\mathbf{x}}) = \int_0^\infty P(x_{new} \mid \lambda)P(\lambda \mid \underline{\mathbf{x}})\, dx$$

$$= \int_0^\infty \frac{e^{-\lambda}\lambda^k}{k!} \cdot \frac{(n+\beta)^{(\Sigma x_i+\alpha)}}{\Gamma(\Sigma x_i+\alpha)} \lambda^{\Sigma x_i+\alpha-1}e^{-(n+\beta)\lambda}\, dx$$

$$= \frac{\Gamma(\Sigma x_i+\alpha+x_{new})}{x_{new}!\Gamma(\Sigma x_i+\alpha)} \cdot \int_0^\infty \frac{(1+n+\beta)^{(\Sigma x_i+\alpha+x_{new})}}{\Gamma(\Sigma x_i+\alpha+x_{new})} \lambda^{\Sigma x_i+\alpha+x_{new}-1}e^{-(1+n+\beta)\lambda}\, dx$$

$$= \frac{\Gamma(\Sigma x_i+\alpha+x_{new})}{x_{new}!\Gamma(\Sigma x_i+\alpha)} \left(\frac{n+\beta}{1+n\beta}\right)^{\Sigma x_i+\alpha} \cdot \left(\frac{1}{1+n+\beta}\right)^{x_{new}}$$

$$\overset{iid}{\sim} NegBin(r = \Sigma x_i + \alpha, p = \frac{1}{1+n+\beta})$$

Results

In 2018, KN received more calls than the other years before and after, resulting in a change-point for that year with a larger λ. Reasons for is could be that there were two court cases, one on 23rd may and one on 30th July where parents were trying to force their daughters into marriage. This received media attention and coverage, possibly encouraging other victims to come forward. However, the FMU had a larger amount of cases with a higher change-point but with a less profound difference to the KN data.

Over the first lockdown, The FMU data had dropped by 56% to an average of 10 calls per week. But as calls decreased by 56% again from 2020 to 2021. It's worth noting that the FMU (and KN) both increased their outreach to the networks available to them during lockdown, but part of the FMUs decrease in calls is that some of the calls related to forced marriage in some ways weren't counted as forced marriage calls.

The KN received over 25% more calls over the first lockdown, increasing from an average of 60 to averages of 75 calls per week. This is likely due to their increased outreach by using the networks they're connected to such as departments of the NHS and police services.

The results show that while KN's data is more sensitive to external or global factors, there is an effect in both data set with change-points being detected at higher values when outreach to effected communities is improved and increased.

enyng1@nottingham.ac.uk