Comparing algorithms for detecting abrupt change points in data

Cody Buntain, Christopher Natoli, and Miroslav Živković

1/7 Introduction
Many data-centric applications produce time-stamped streams of data ripe for analysis, and a key aspect of this data is understanding when the underlying distribution producing this data changes. These moments of change are called “change points” and have a variety of uses from fault detection to enhanced forecasting to classification and many others.

2/7 The Algorithms

Galeano and Peña’s Likelihood Ratio Test [1]
Fit a VARMA model to the data, and extract the residuals $e_t$ from the data. For some point in time $h$, calculate the LRT test statistic, and compare against the critical value for that dimensionality.

$LRT(h) = n \ln \left( \frac{1}{h} \sum_{i=1}^{h} e_i \right) - \frac{1}{2} \ln \left( \frac{1}{h} \sum_{i=h+1}^{2h} e_i \right)$

Galeano and Peña’s CUSUM Test [1]
Same as LRT but with the CUSUM test statistic.

$G^2(h) = \frac{h}{2h - 1} \left( \frac{\sum_{i=1}^{h} e_i^2}{h \cdot \bar{e}^2} - \frac{\sum_{i=h+1}^{2h} e_i^2}{(2h-h) \cdot \bar{e}^2} \right)$

Desobry et al.’s Kernel Change Detection [2]
Given a data window of size $2m$, fit two one-class SVMs to the first $m$ points and second $m$ points, and use the KCD statistic to calculate dissimilarity between the two data sets.

$KCD(h) = \frac{\text{acc}_{CCD}(x, y) - \text{acc}_{CCD}(x, y)}{\text{acc}_{CCD}(x, y) - \text{acc}_{CCD}(x, y)}$

3/7 Changes in Mean vs. Covariance
For covariance changes, we generated two regimes of data with constant mean and different covariance matrices. KCD then fit one-class SVMs to the covariance matrices within the past and future windows. Mean shifts rather used random means and constant covariance.

We simulated 500 bi-variate data points with a change point at $h=250$, KCD window size of 400 ($m=200$), and compared the LRT and CUSUM test statistics at the 95% confidence level.

4/7 Sensitivity to Dimensionality
Once again, we simulated 500 multi-variate data points but included change points at $h=\{125, 250, 375\}$. We left the KCD window size at 400 ($m=200$), and compared the LRT and CUSUM test statistics at the 95% confidence level. We then varied dimensionality from $k=\{2, 10\}$.

5/7 Sensitivity to Change Point Count
Here, we simulated 3,000 bi-variate data points with 2 to 12 change points distributed evenly throughout the data set. We left the KCD window size at 400 ($m=200$), and compared the LRT and CUSUM test statistics at the 95% confidence level.

6/7 Real Applications
We applied all three algorithms to two real data sets and sought to find change points within them:

Bridge Sensor Data
Sensor data from an experiment on applying stress deformations to a bridge in a laboratory. The objective of the original data was to identify cracks in the structure before they became visible.

Mt. Gox Bitcoin Market Data
The now-defunct Mt. Gox Bitcoin exchange shared market data for Bitcoin valuations across several currencies. We analyzed two years of Bitcoin to US Dollar, Euro, GB Pound, and Polish Zloty.

7/7 Conclusions
Our performance data suggests the following results:

- The parametric LRT and CUSUM algorithms outperform the non-parametric KCD algorithm when detecting changes in covariance.
- KCD is competitive in detecting shifts in mean even with relatively small window sizes.
- LRT and CUSUM are more robust to increases in dimensionality of the data.

When applied to real data, we found the following:

- LRT detects many more change points than either CUSUM or KCD.
- KCD’s window size parameter can potentially miss change points that occur over larger periods of time.

0/0 References