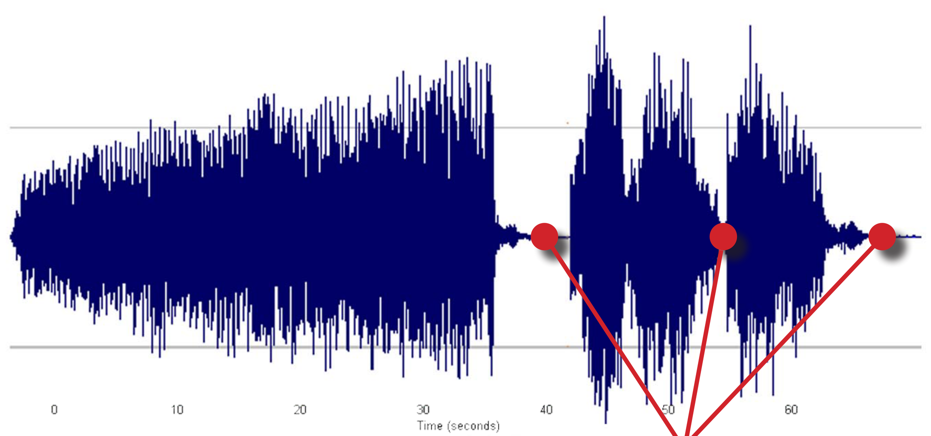
Comparing algorithms for detecting abrupt change points in data

Cody Buntain, Christopher Natoli, and Miroslav Živković

1/7 Introduction

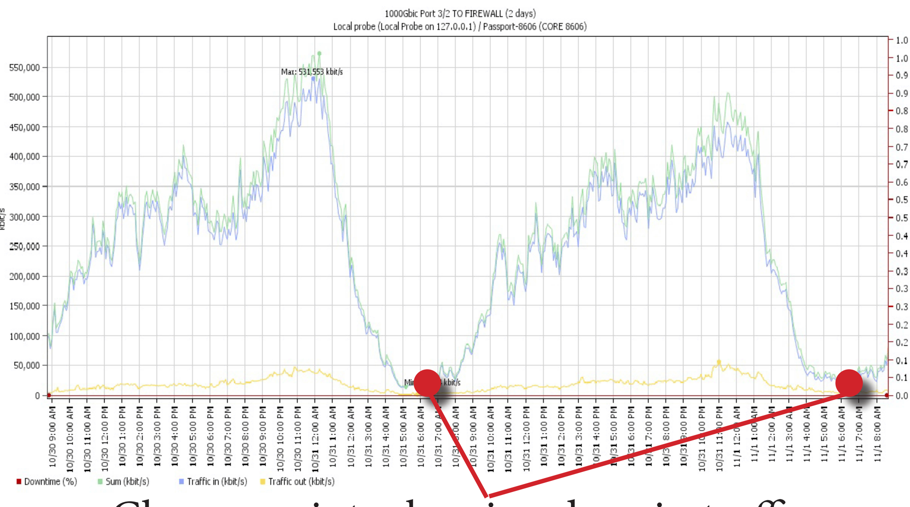
Many data-centric applications produce time-stamped streams of data ripe for analysis, and a key aspect of this data is understanding when the underlying distribution producing this data changes. These moments of change are called “change points” and have a variety of uses from fault detection to enhanced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

Musical Note Segmentation



Change points between notes

Denial-of-Service Detection



Change points showing drop in traffic

2/7 The Algorithms

Galeano and Peña’s Likelihood Ratio Test [1]

Fit a VARMA model to the data, and extract the residuals $\boldsymbol{e_i}$ from the data. For some point in time $\boldsymbol{h}$, calculate the LRT test statistic, and compare against the critical value for that dimensionality.

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

Galeano and Peña’s CUSUM Test [1]

Same as LRT but with the CUSUM test statistic.

$$C_\ell^r(h) = \frac{h}{\sqrt{2k(r-\ell+1)}} \left(\frac{\sum_{t=\ell}^{h} e_t (\hat{\Sigma}_\ell^r)^{-1} e_t'}{h} - \frac{\sum_{t=\ell}^{r-\ell+1} e_t (\hat{\Sigma}_\ell^r)^{-1} e_t'}{r-\ell+1} \right)$$

Desobry et al.’s Kernel Change Detection [2]

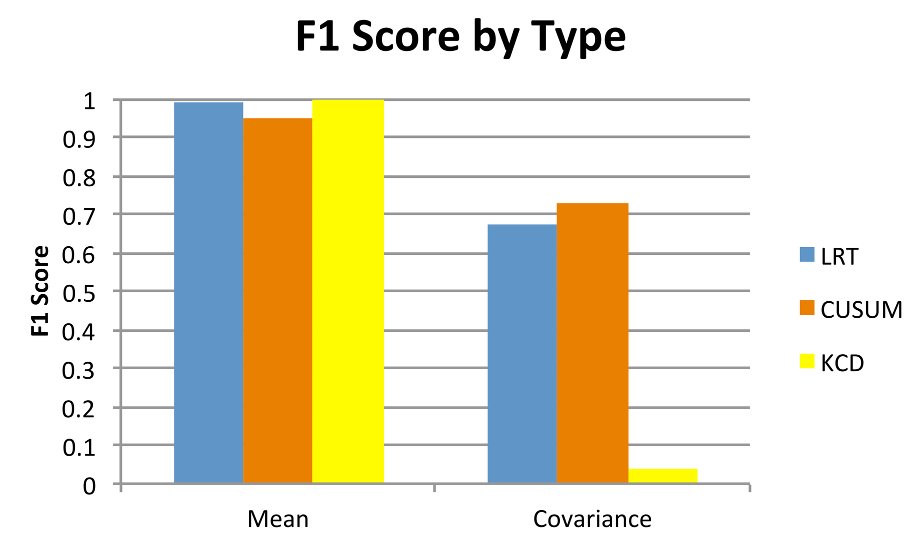
Given a data window of size $\boldsymbol{2m}$, fit two one-class SVMs to the first $\boldsymbol{m}$ points and second $\boldsymbol{m}$ points, and use the KCD statistic to calculate dissimilarity between the two data sets.

$$KCD(h) = \frac{\arccos\left(\frac{\boldsymbol{\alpha}_p^T K_{p,f} \boldsymbol{\alpha}_f}{\sqrt{\boldsymbol{\alpha}_p^T K_{p,p} \boldsymbol{\alpha}_p}\sqrt{\boldsymbol{\alpha}_f^T K_{f,f} \boldsymbol{\alpha}_f}}\right)}{\arccos\left(\frac{\rho_p}{\sqrt{\boldsymbol{\alpha}_p^T K_{p,p} \boldsymbol{\alpha}_p}}\right) + \arccos\left(\frac{\rho_f}{\sqrt{\boldsymbol{\alpha}_f^T K_{f,f} \boldsymbol{\alpha}_f}}\right)}$$

3/7 Changes in Mean vs. Covariance

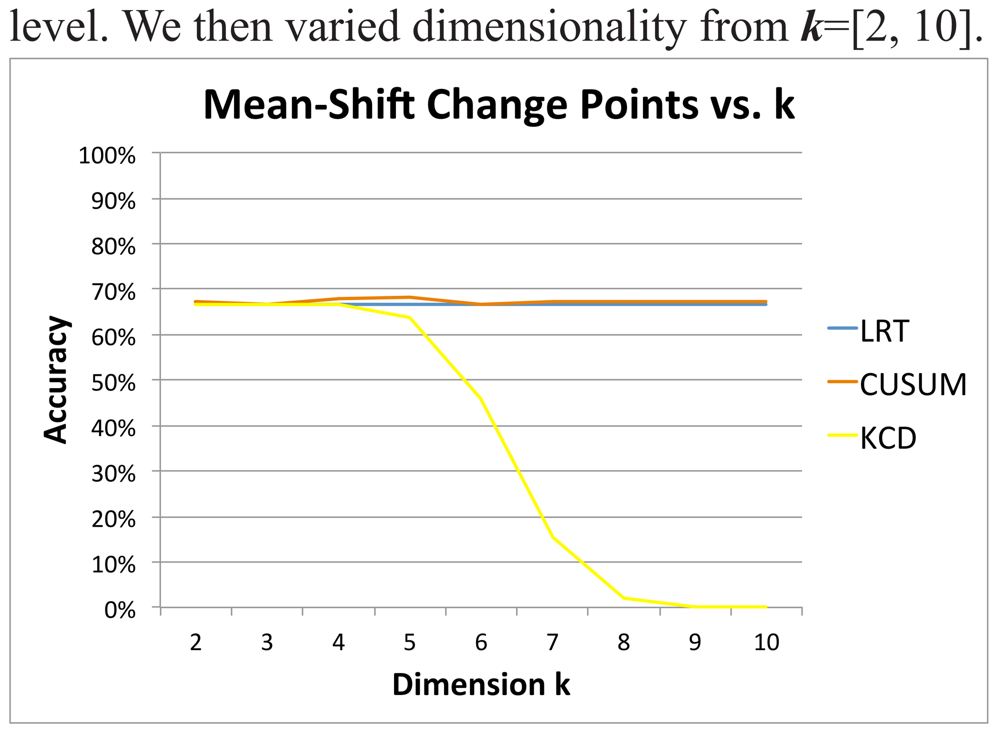
For covariance changes, we generated two regimes of data with constant mean and different covariance matrices. KCD then fit one-class SVMs to the covariance matrices within the past and future windows. Mean shifts rather used random means and constant covariance.

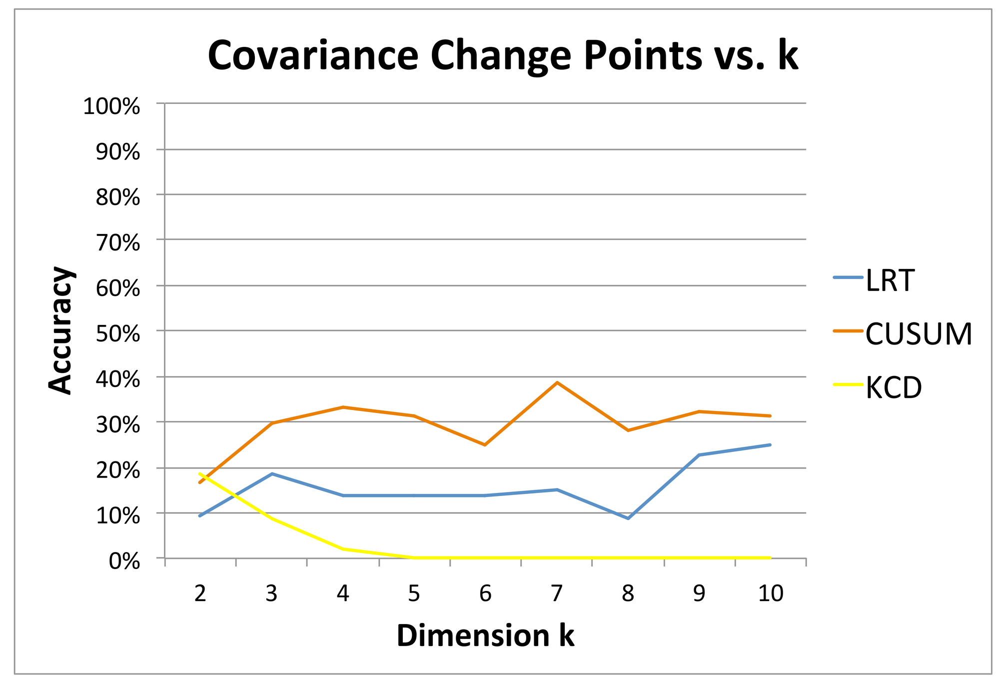
We simulated 500 bi-variate data points with a change point at $\boldsymbol{h}$=250, KCD window size of 400 ($\boldsymbol{m}$=200), and compared the LRT and CUSUM test statistics at the 95% confidence level.



4/7 Sensitivity to Dimensionality

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

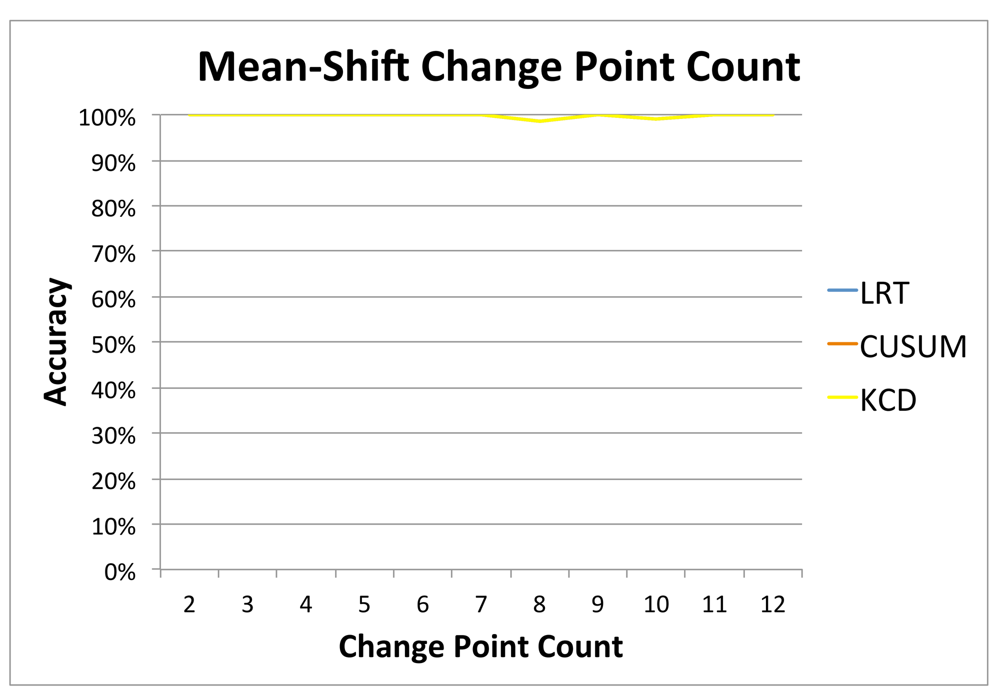


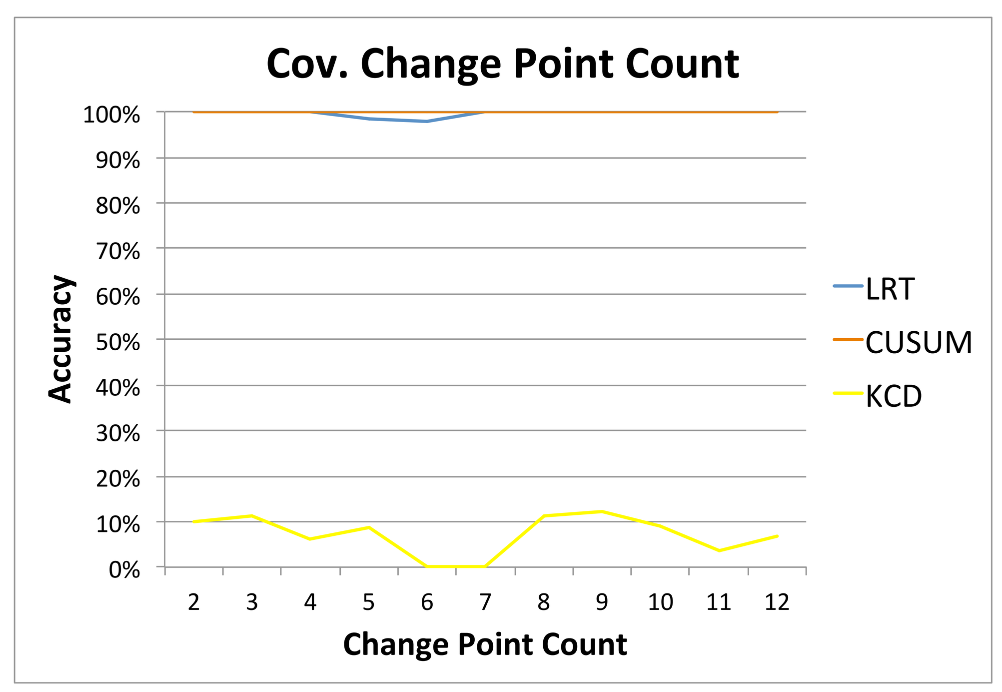


It seems the LRT and CUSUM-based algorithms are realatively insensitive to increases in dimensionality. KCD, on the other hand, seems quite sensitive with its accuracy falling to near 0% by $\boldsymbol{k}$=9.

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 ($\boldsymbol{m}$=200), and compared the LRT and CUSUM test statistics at the 95% confidence level.





All three algorithms seem robust to varying change points in the data. Only the covariance-based KCD implementation performs poorly with the large number of data points.

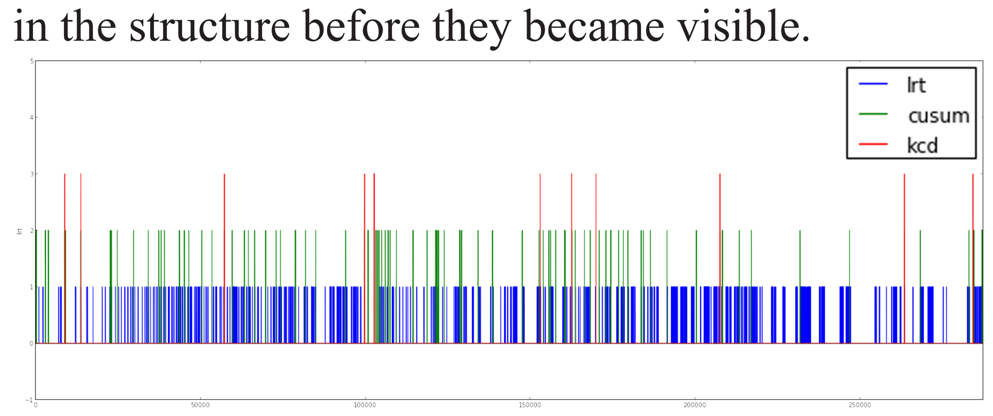
Acknowledgements This work made use of the Open Science Data Cloud (OSDC) which is an Open Cloud Consortium (OCC)-sponsored project. The OSDC is supported in part by grants from Gordon and Betty Moore Foundation and the National Science Foundation and major contributions from OCC members like the University of Chicago. This work was also supported by the National Science Foundation Partnerships for Research and Education (PIRE) Award Number 1129076. Any opinions, findings, and conclusions or recommendations expressed are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

6/7 Real Applications

We applied all three algorithms to two real data sets and sought to find change points within them:

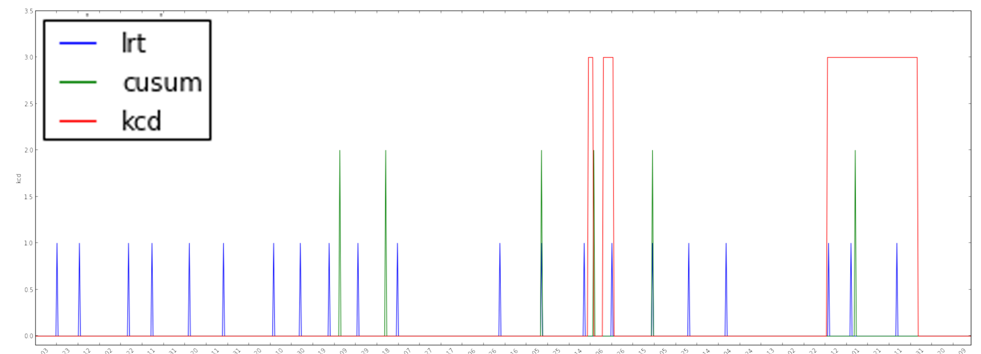
Bridge Sensor Data

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

[1] P. Galeano and D. Peña, “Covariance changes detection in multivariate time series,” J. Stat. Plan. Inference, vol. 137, no. 1, pp. 194–211, Jan. 2007.
[2] F. Desobry, M. Davy, and C. Doncarli, “An online kernel change detection algorithm,” Signal Process. IEEE Trans., vol. 53, no. 8, pp. 2961–2974, Aug. 2005.