STOR608 Sprint 3: Change Points and Anomaly Detection

Group 2:
Graham Burgess, Max Howell, James Neill, Wanchen Yue
Spint Leader: Idris Eckley

8 December 2022

Introduction

For data

$$Y_1, \ldots, Y_T$$

The location(time) of the n distinct change points is: $0 < t_1 < t_2 < \cdots < t_{n-1} < t_n < T$. For each segment partitioned by the change points, a parametric model with corresponding parameter vector (Θ) would be constructed.

Introduction

To determine the optimal location and number of change points for the time series, we will compare the total cost through different segmentations. The total cost is the sum of the likelihood for each segment in a particular segmentation:

$$\min_{\mathcal{T}} \sum_{i=1}^{n+1} \ell\left(\mathbf{Y}_{\tau_{i-1}:\tau_i - 1}\right) + \lambda P(n)$$

NOT (Narrowest-over-threshold) Change Point

- We first draw M sub-intervals along the total time span, i.e $(Y_{p+1}, \dots, Y_q)$.
- Next, we calculate the generalized likelihood ratio statistic for all the points (i) within one sub-intervals $(p, q]$:
$$\mathbf{R}_{i,(p,q]}(\mathbf{Y}) = 2\log\left[\frac{\sup_{\Theta^1,\Theta^2}\{l(Y_{p+1},\dots,Y_i;\Theta^1)\,l(Y_{i+1},\dots,Y_q;\Theta^2)\}}{\sup_{\Theta} l(Y_{p+1},\dots,Y_q;\Theta)}\right]$$
- Set a threshold value λ_T, compare the $\mathcal{R}_{(p,q]}(\mathbf{Y})$ with λ_T and pick out those significant maximum ratio statistics which are above the threshold value: $\mathcal{R}_{(p_s,q_s]}(\mathbf{Y})$.
- Finally, the sub-interval $(p_{s^*}, q_{s^*}]$ leading to a significant ratio statistic $\mathcal{R}_{(p_{s^*},q_{s^*}]}(\mathbf{Y})$ with narrowest length of the interval is chosen, i^* corresponding to maximum generalized likelihood ratio statistic is the (first) change point that we aim to locate. Baranowski (2019)

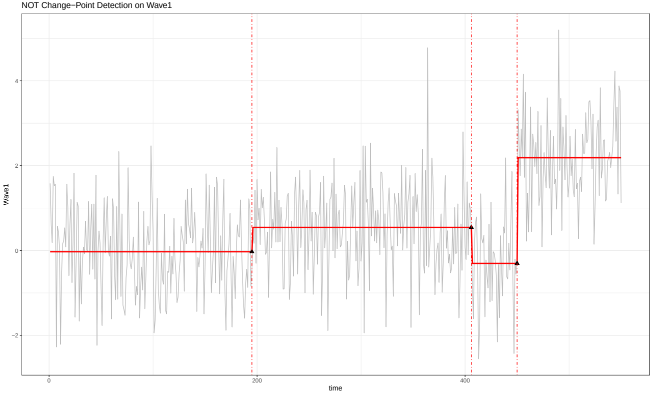
NOT Changepoint

$$Y_t = f_t + \sigma_t * \epsilon_t$$

- **Piecewise constant mean**:
  - **Constant variance**: $\sigma_t = \sigma$, $f(t) = \mu_k$
  - **Piecewise constant Variance**: $f(t) = \mu_k$, $\sigma_t = \sigma_k$
- **Constant variance**:
  - **Continuous piecewise linear mean**: $\sigma_t = \sigma$, $f(t) = \alpha_k + \beta_k t$ continuity constraint: $\alpha_k + \beta_k t_k = \alpha_{k+1} + \beta_{k+1} t_k$ with t_k being the time of the k^{th} change point with $k = 1, \ldots, n$.
  - **Piecewise linear mean (with possible discontinuity)**: $\sigma_t = \sigma$, $f(t) = \alpha_k + \beta_k t$ for all t within the k^{th} segment and $\alpha_k + \beta_k t_k \neq \alpha_{k+1} + \beta_{k+1} t_k$ for some $k = 1, \ldots, n$.

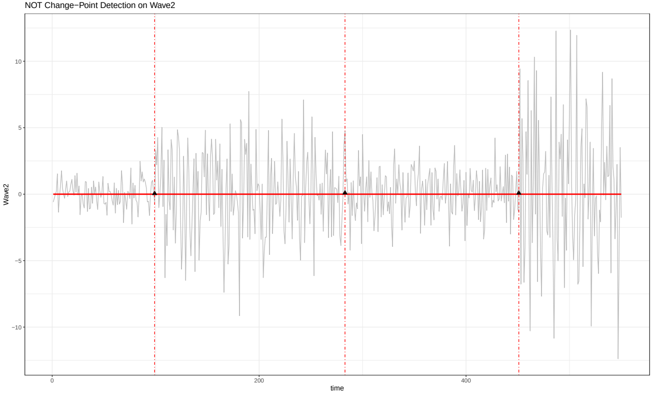
NOT Changepoint

- **Wave 1** Mean Change: length 550, Change-Points at 200, 400, 450 with mean values:0, 0.5, -0.3, 2 between the Change-Points. Standard deviation $\sigma = 1$.



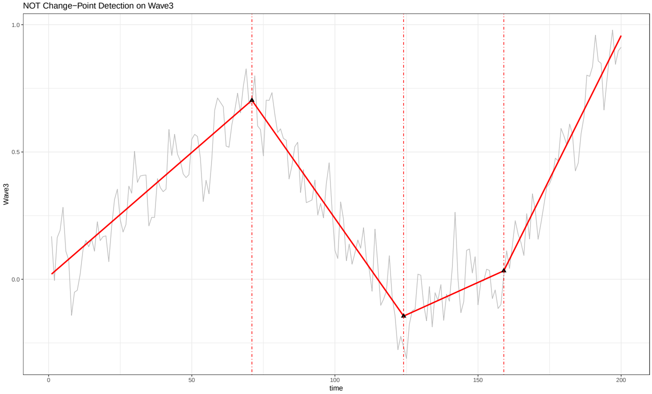
NOT Changepoint

- **Wave 2** Variance Change: length 550, Change-Points at 100, 300, 450, with standard deviation σ:1, 3, 1.5, 6 between the Change-Points.

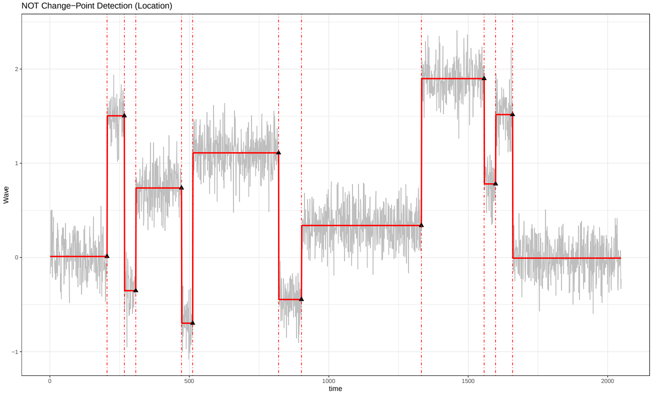


NOT Changepoint

- **Wave 3** length 200, Change-Points at 70, 120, 160 with values: 0.70, -0.05, 0.03 at the Change-Point and continuous linear trend in between. Standard deviation: $\sigma = 0.1$.

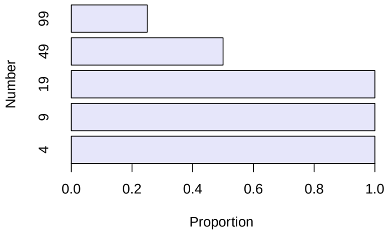


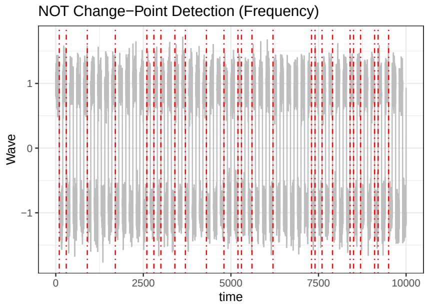
Changepoint Location



Changepoint Frequency

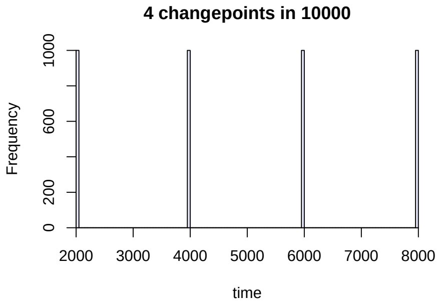
- length 10000, changepoints evenly spread, number of the changepoints:4, 9, 19, 49, 99. Standard deviation: $\sigma = 0.2$.

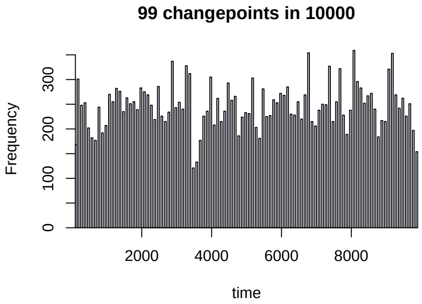




Changepoint Frequency

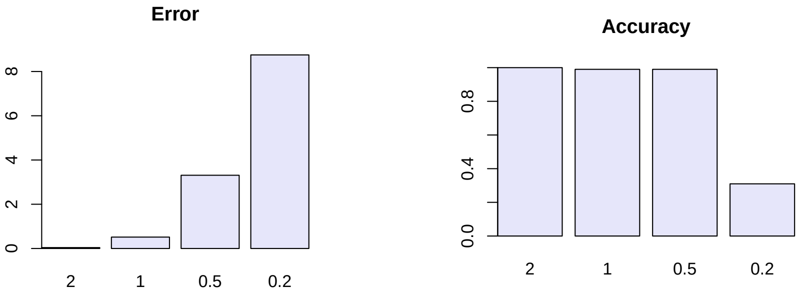
- length 10000, changepoints evenly spread, number of the changepoints:4, 9, 19, 49, 99. Standard deviation: $\sigma = 0.2$.





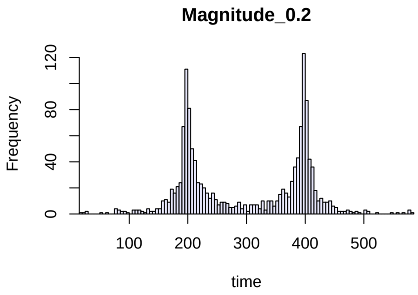
Changepoint Magnitude

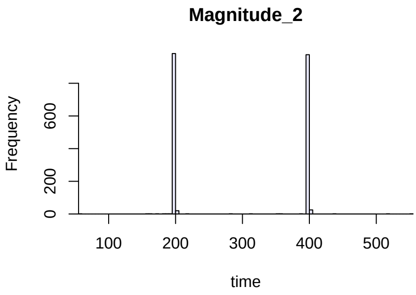
- length 600, changepoints at 200, 400 with change in mean:0.2, 0.5, 1, 2. Standard deviation: $\sigma = 1$.



Changepoint Magnitude

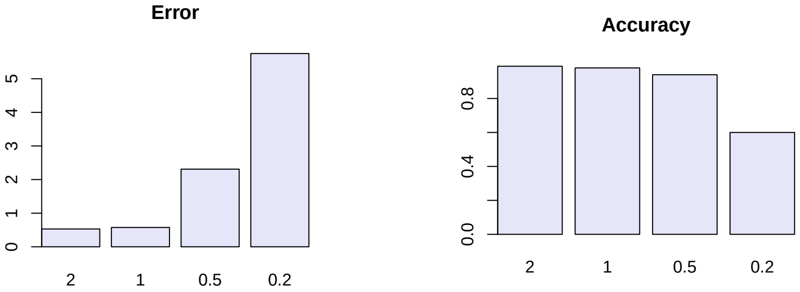
- length 600,changepoints at 200, 400 with change in mean:0.2, 0.5, 1, 2. Standard deviation: $\sigma = 1$





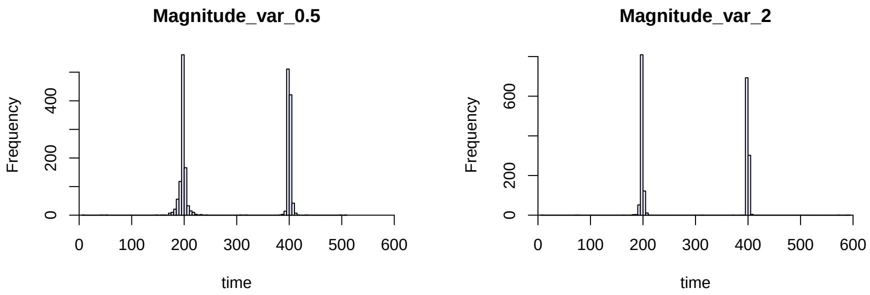
Changepoint Magnitude

- length 600, changepoints at 200, 400 with change in Variance:0.2, 0.5, 1, 2. Mean: $\mu = 0$.



Changepoint Magnitude

- length 600, changepoints at 200, 400 with change in Variance:0.2, 0.5, 1, 2. Mean: $\mu = 0$



Addition of anomalies

We start with Gaussian behaviour which changes from $N(0,1)$ to $N(2,1)$ at $t = 400$. We then add anomalies in the following manner

- A single anomaly, magnitude 10σ, **changing the location** w.r.t the changepoint
- A single anomaly, fixed location $t = 200$, **changing the magnitude**
- Fixed magnitude of of anomaly, **changing the frequency** of randomly dispersed anomalies

We evaluate the behaviour of the NOT and FPOP [Maidstone et al. (2017)] methods.

Anomalies: changing the location

Both NOT and FPOP consistently find a changepoint as the location of the anomaly changes

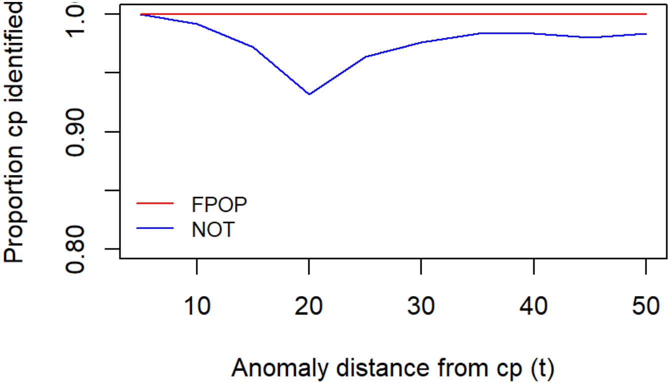


Figure: Proportion of the time NOT and FPOP identify a single changepoint

Anomalies: changing the location

When the anomaly is close to the changepoint, both NOT and FPOP can mistake the anomaly for the changepoint location

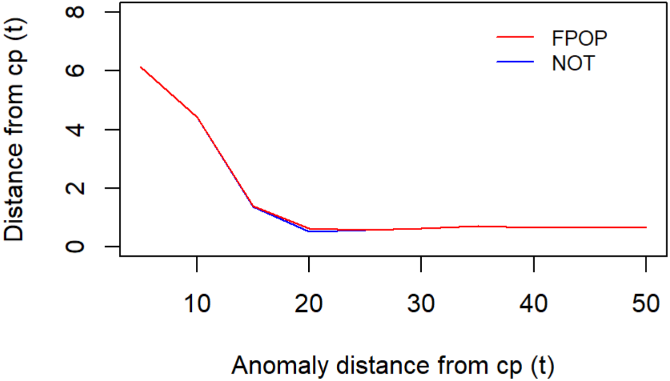


Figure: Average distance between the observed and actual changepoint

Anomalies: changing the magnitude

Both NOT and FPOP consistently find a changepoint as the magnitude of the anomaly changes

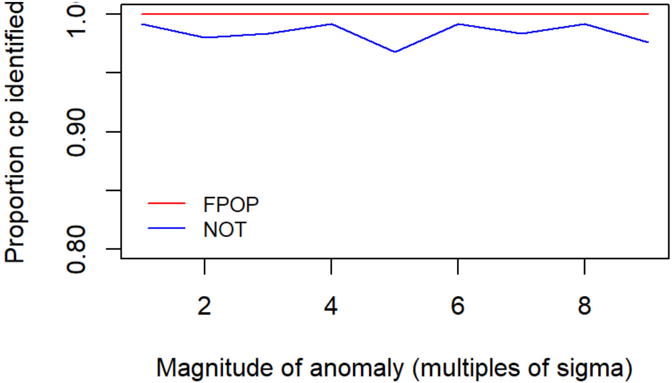


Figure: Proportion of the time NOT and FPOP identify a single changepoint

Anomalies: changing the magnitude

Both NOT and FPOP consistently find a changepoint very close to its true location as the magnitude of the anomaly changes

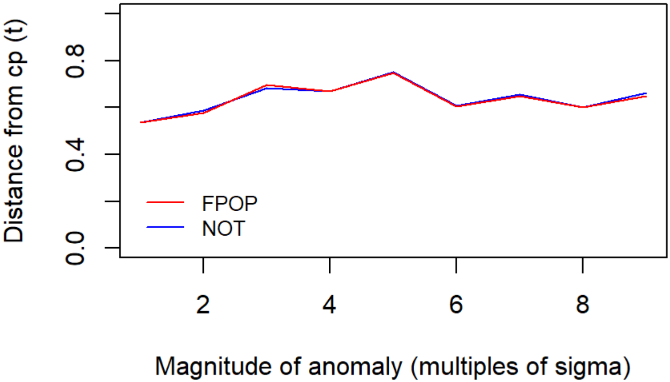


Figure: Average distance between the observed and actual changepoint

Anomalies: changing the frequency

As we increase the number of anomalies, the FPOP method consistently finds a single changepoint, whereas NOT gets worse - it tends to identify multiple changepoints.

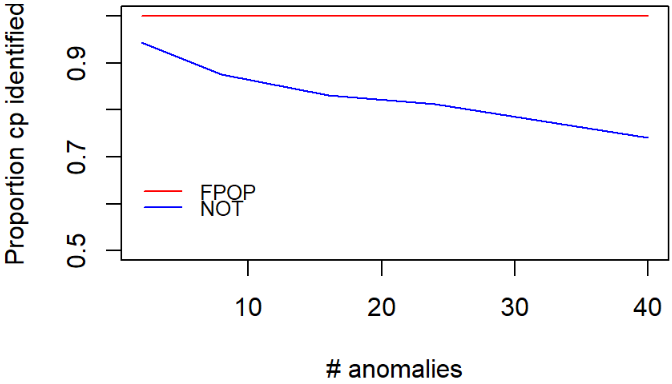


Figure: Proportion of the time NOT and FPOP identify a single changepoint

Anomalies: changing the frequency

As we increase the number of anomalies, when a single changepoint is detected, both NOT and FPOP get worse at identifying the position of the changepoint.

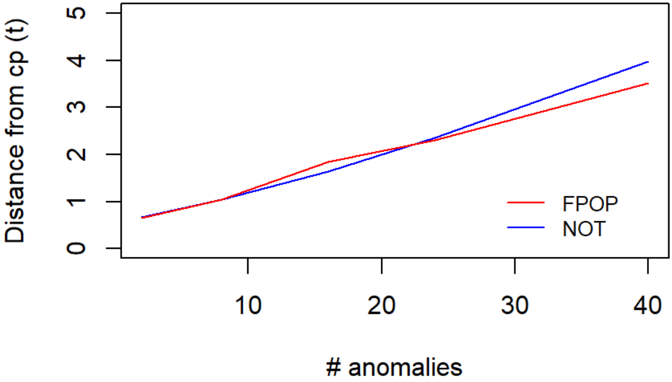


Figure: Average distance between the observed and actual changepoint

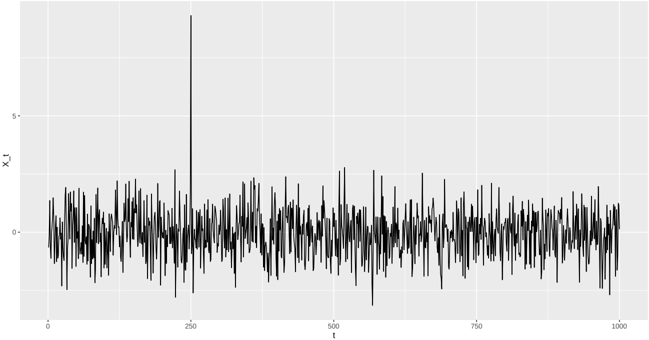
Changepoint detection summary

The NOT method works well in detecting changepoints except when the frequency of changepoints becomes large.
In the presence of anomalies, the NOT method continues to work well unless the frequency of anomalies increases.
The FPOP method outperforms the NOT method in the presence of anomalies.

Types Of Anomalies

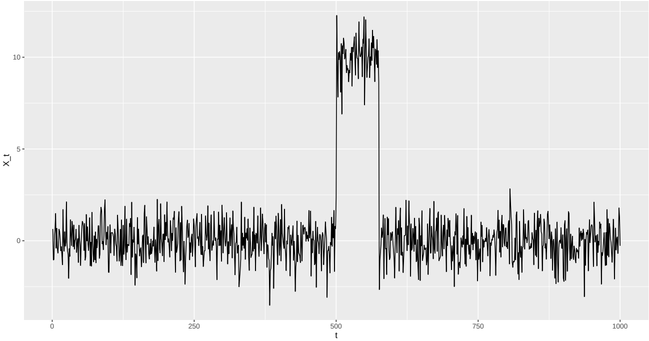
There are three kinds of anomalies:

- Global (Point) Anomalies



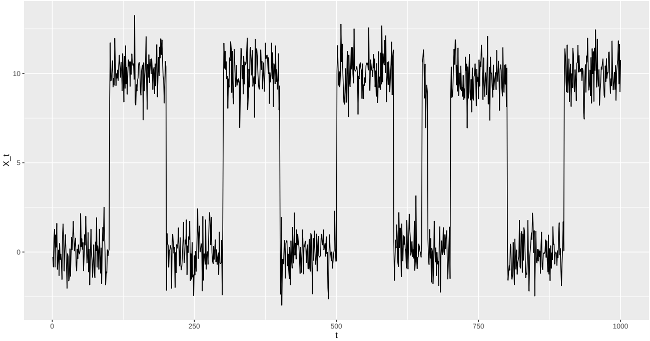
Types Of Anomalies

- Collective Anomalies



Types Of Anomalies

- Contextual Anomalies



CAPA

CAPA (Collective and Point Anomalies):

- Can detect change in mean and change in variance
- In practice, has a close to linear computational cost [Fisch et al. (2018)]
- The package we used assumes that the underlying data is normal

Minimising Cost in CAPA

The method is based on finding the points which minimise the penalised cost

$$\sum_{t \notin [\hat{s}_i+1, \hat{e}_i]} \mathcal{C}(\mathbf{x}_t, \hat{\theta}_0) + \sum_{j=1}^{\hat{K}} \left[\min_{\hat{\theta}_j} \left(\sum_{\hat{s}_j+1}^{\hat{e}_j} \mathcal{C}(\mathbf{x}_t, \hat{\theta}_j) \right) + \beta \right]$$

for each segment (s_i, e_i), where $\mathcal{C}$ is the negative log-likelihood and β is a penalty constant (often a scalar multiple of $\log(n)$, where n is the number of datapoints).

CAPA Algorithm Part 1

Input:	A set of observations of the form, $(x_1, x_2, \ldots, x_n)$ where $x_i \in \mathbb{R}$.
	Penalty constants β and β' for the introduction of a collective and a point anomaly respectively
	A minimum segment length $l \geq 2$
Initialise:	Set $C(0) = 0$, $Anom(0) = NULL$.

Figure: CAPA Algorithm from Fisch et al. (2018)

CAPA Algorithm Part 2

$\hat{\mu} \leftarrow MEDIAN(x_1, x_2, \ldots, x_n)$

$\hat{\sigma} \leftarrow IQR(x_1, x_2, \ldots, x_n)$

for $i \in \{1, \ldots, n\}$ **do**

$\quad x_i \leftarrow \frac{x_i - \hat{\mu}}{\hat{\sigma}}$

end for

Figure: CAPA Algorithm from Fisch et al. (2018)

CAPA Algorithm Part 3

for $m \in \{1, ..., n\}$ **do**

$C_1(m) \leftarrow \min_{0 \le k \le m-l} \left[C(k) + (m-k) \left[\log \left(\frac{1}{m-k} \sum_{t=k+1}^{m} \left(x_t - \bar{x}_{(k+1):m} \right)^2 \right) + 1 \right] + \beta \right]$

$s \leftarrow \arg\min_{0 \le k \le m-l} \left[C(k) + (m-k) \left[\log \left(\frac{1}{m-k} \sum_{t=k+1}^{m} \left(x_t - \bar{x}_{(k+1):m} \right)^2 \right) + 1 \right] + \beta \right]$

$C_2(m) \leftarrow C(m-1) + x_m^2$

$C_3(m) \leftarrow C(m-1) + 1 + \log\left(\gamma + x_m^2\right) + \beta' \Big]$,

Figure: CAPA Algorithm from Fisch et al. (2018)

CAPA Algorithm Part 4

$C(m) \leftarrow \min\left[C_1(m), C_2(m), C_3(m)\right]$
switch $\arg\min\left[C_1(m), C_2(m), C_3(m)\right]$ **do**
  case 1 : $Anom(m) \leftarrow [Anom(s), (s+1, m)]$
  case 2 : $Anom(m) \leftarrow Anom(m-1)$
  case 3 : $Anom(m) \leftarrow [Anom(m-1), (m)]$
end for

Output The points and segments recorded in $Anom(n)$

Figure: CAPA Algorithm from Fisch et al. (2018)

CAPA Comparison

There is also a version of the CAPA algorithm for multivariate data.

We will compare the performance of the multivariate algorithm against using the univariate algorithm on each dimension.

CAPA Application: Collective Anomalies

We first use CAPA to find collective anomalies on 5-dimensional data.

CAPA Application: Collective Anomalies

We first use CAPA to find collective anomalies on 5-dimensional data.

We generate each dimension from a standard Gaussian distribution, apart from subsections of 20 points generated from a Gaussian distribution with an increased mean μ. These subsections begin at:

- dimension 1: 100, 300, 500, 700, 900
- dimension 2: 300, 500, 700, 900
- dimension 3: 500, 700, 900
- dimension 4: 700, 900
- dimension 5: 900

Multivariate Method when $\mu = 1$

Applying CAPA to all the data, we find the following collective anomalies:

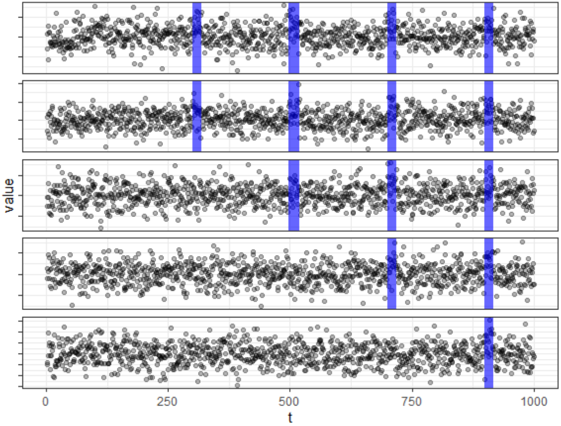


Figure: Anomaly band width: 19-23

Univariate Method when $\mu = 1$

Applying CAPA to each dimension individually, we find the following collective anomalies:

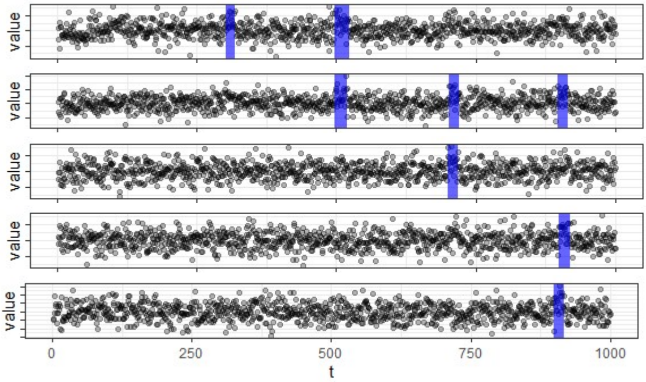


Figure: Anomaly band width: 16-27

Multivariate Method when $\mu = 2$

Applying CAPA to all the data, we find the following collective anomalies:

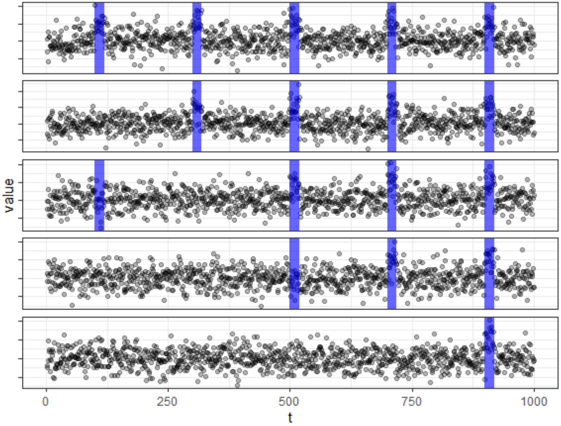


Figure: Anomaly band width: 20

Univariate Method when $\mu = 2$

Applying CAPA to each dimension individually, we find the following collective anomalies:

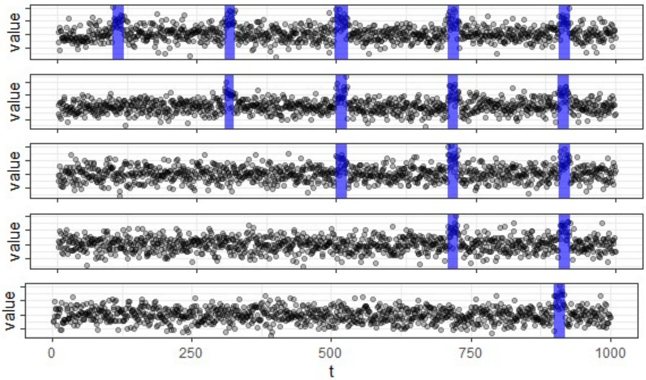


Figure: Anomaly band width: 17-25

CAPA Method Comparison

Using the multivariate version finds more anomalies than using the univariate version repeatedly – the multivariate version is more accurate when the anomalous results are less clear.

CAPA Method Comparison

Using the multivariate version finds more anomalies than using the univariate version repeatedly – the multivariate version is more accurate when the anomalous results are less clear.

However, when the anomalous results are clearer, and when the number of dimensions is increased, the multivariate method is more likely to overfit and find false positives.

CAPA Method Comparison

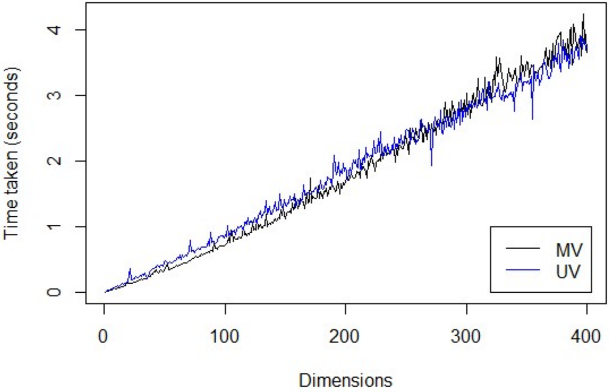
Using the multivariate version finds more anomalies than using the univariate version repeatedly – the multivariate version is more accurate when the anomalous results are less clear.

However, when the anomalous results are clearer, and when the number of dimensions is increased, the multivariate method is more likely to overfit and find false positives.

Although difficult to see on the plots, the collective anomalies selected have closer endpoints to the true anomalies by using the multivariate method than the repeated univariate method.

CAPA Method Comparison – Time Taken

Both methods take approximately the same (very small) time to run, regardless of the number of dimensions:

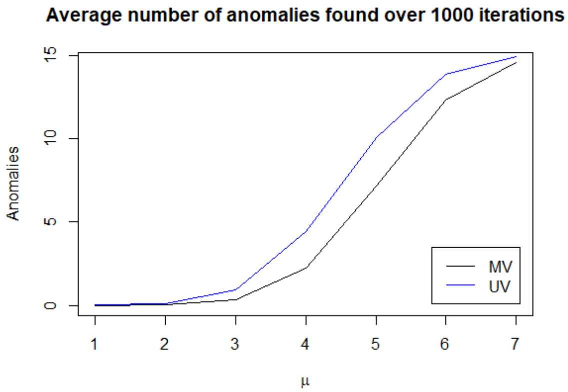


CAPA Application: Point Anomalies

We now repeat the above CAPA application on 5-dimensional data, but with single anomalous points (15 in total) instead of anomalous ranges.

Again, most of the data is from a standard Gaussian, with anomalous points from a Gaussian with mean μ.

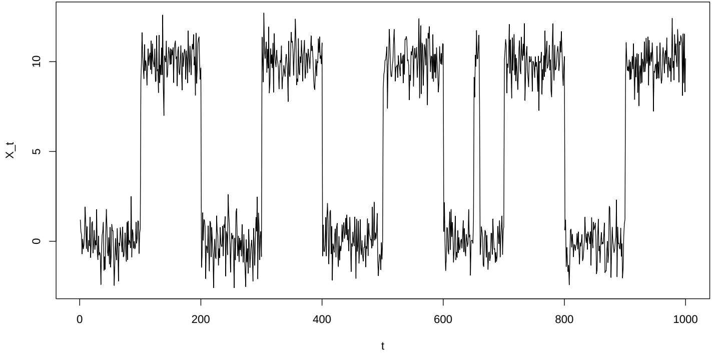
CAPA Application: Point Anomalies



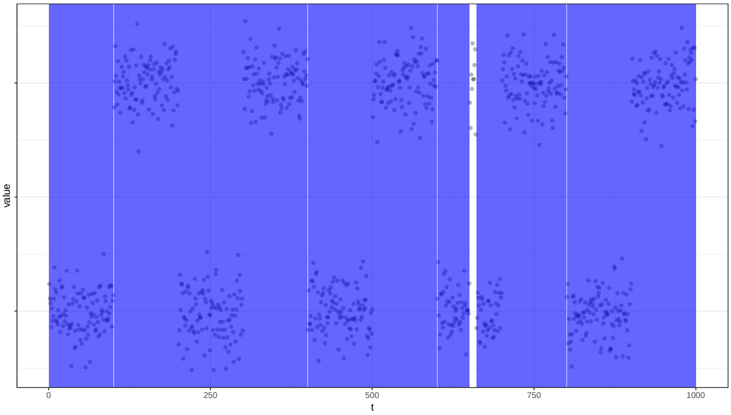
The repeated univariate method performs better for any given μ.

CAPA Application: Contextual Anomalies

Just for completeness, lets see how CAPA copes with a contextual anomaly in seasonal data (which it is not designed for).

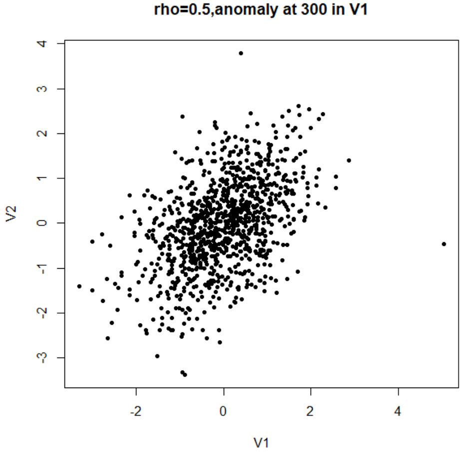


CAPA Application: Contextual Anomalies

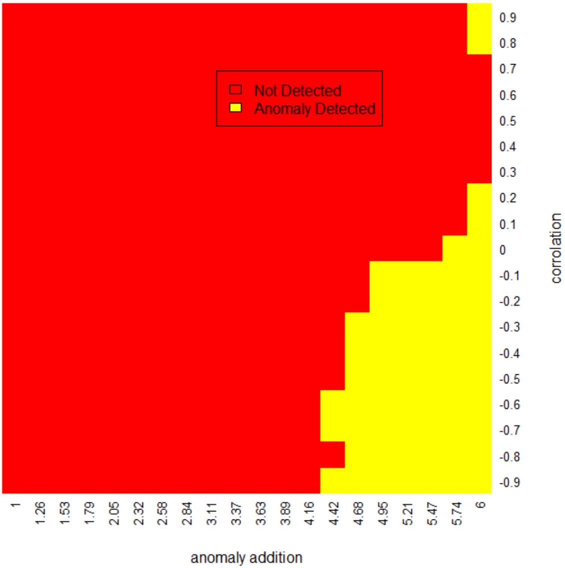


CAPA with Correlation

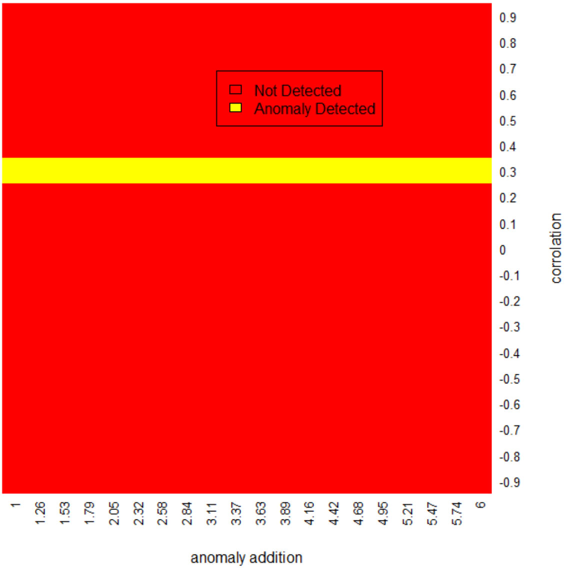
Now taking a look at how CAPA copes when there is correlation between variables...



CAPA with Correlation – Variable 1

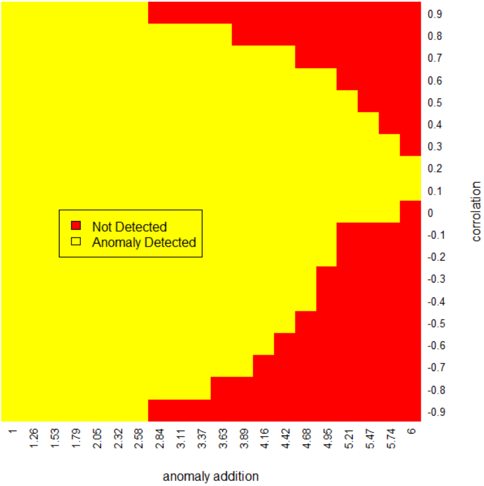


CAPA with Correlation – Variable 2



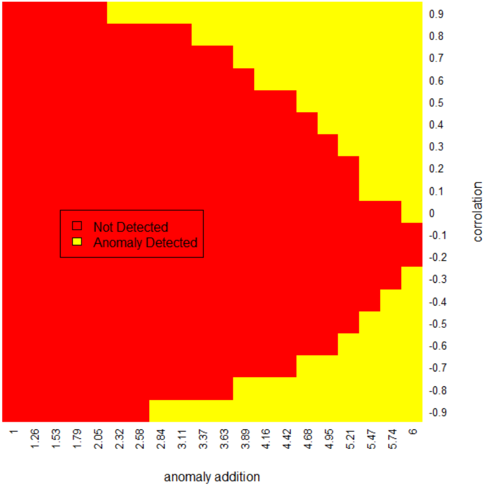
CAPACC – Variable 1

Using capa.cc we can see how it copes when there is correlation.



CAPACC – Variable 2

Changing the other variable:



Other Anomaly Detection Methods

- Brutlag's Anomaly Detection Algorithm using the Holt-Winters (triple exponential smoothing) Model [Szmit and Szmit (2012)]
- Grubbs's test (both univariate and multivariate versions)
- Regression-based analysis (both basic and robust versions)

Anomaly Detection Conclusion

- We have a method that can detect anomalies in both the univariate and multivariate case (CAPA)
- We have compared the benefits of the multivariate case against doing the univariate multiple times
- We have looked at what to do when there could be correlation between dimensions in multivariate data settings

References I

Baranowski, Yining Chen, P. F. (2019). Narrowest-over-threshold detection of multiple change points and change-point-like features. *Journal of Royal Statistical Society*, 81:649–672.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58.

Fearnhead, P. and Rigaill, G. (2018). Changepoint detection in the presence of outliers. *Journal of the American Statistical Association*, 114(525):169–183.

Fisch, A., Eckley, I. A., and Fearnhead, P. (2018). A linear time method for the detection of point and collective anomalies. *arXiv preprint arXiv:1806.01947*.

References II

Maidstone, R., Hocking, T., Rigaill, G., and Fearnhead, P. (2017). On optimal multiple changepoint algorithms for large data. *Statistics and computing*, 27(2):519–533.

Szmit, M. and Szmit, A. (2012). Usage of modified holt-winters method in the anomaly detection of network traffic: Case studies. *Journal of Computer Networks and Communications*, 2012.

Thank you for listening

Any questions?