Mathematics & Statistics | Lancaster University

Further Changepoint Analysis

Rebecca Killick(r.killick@lancs.ac.uk)
NHS Workshop 2020

Workshop Plan

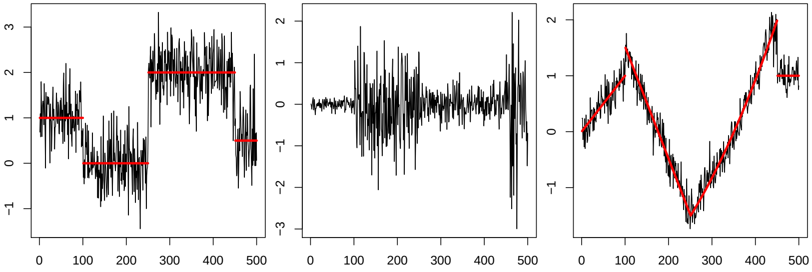
- Recap of changepoints
- Checking assumptions
- Autocorrelation
- Influence
- Multivariate changepoints

There will be tasks throughout the sections.

Recall: changepoints

For data $y_1, \ldots, y_n$, if a changepoint exists at τ, then $y_1, \ldots, y_\tau$ differ from $y_{\tau+1}, \ldots, y_n$ in some way.

There are many different types of change.



Packages

Today we will use the following packages

```
library(changepoint)
```

```
library(EnvCpt)
```

```
library(changepoint.influence)
```

```
library(changepoint.geo)
```

Other notable `R` packages are available for changepoint analysis including

- `ecp` - for univariate and multivariate energy test statistics
- `InspectChangepoint` - for multivariate Inspect projection direction mean only change
- `hdbinseg` - for multivariate double CUSUM test statistic
- `BayesProject` - for multivariate changepoints

Checking Assumptions

The main assumptions for a Normal likelihood ratio test for a change in mean are:

- Independent data points;
- Normal distributed points pre and post change;
- Constant variance across the data.

How can we check these?

How to check

- Check the residuals

```
set.seed(1)
m1=c(rnorm(100,0,1),rnorm(100,5,1))
m1.amoc=cpt.mean(m1)


means=param.est(m1.amoc)$mean
m1.resid=m1-rep(means,seg.len(m1.amoc))
shapiro.test(m1.resid)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  m1.resid
## W = 0.99228, p-value = 0.3721
```

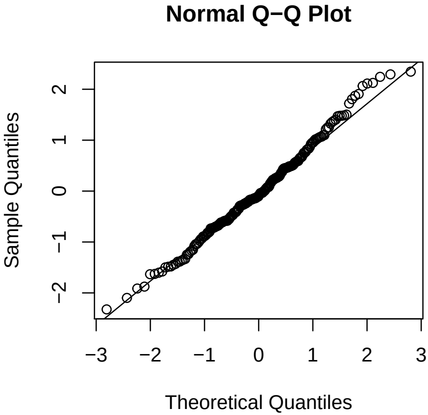
Residual Check


```
ks.test(m1.resid,pnorm,mean=mean(m1.resid),sd=sd(m1.resid))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  m1.resid
## D = 0.045812, p-value = 0.7953
## alternative hypothesis: two-sided
```

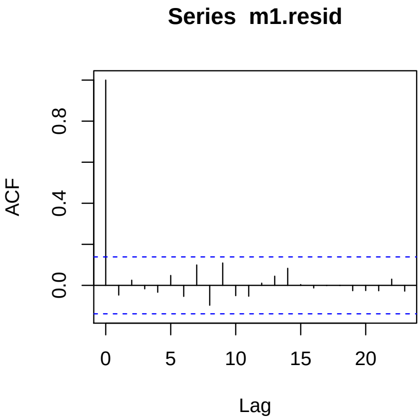
Residual Check

```r
qqnorm(m1.resid)
qqline(m1.resid)
```



Residual Check

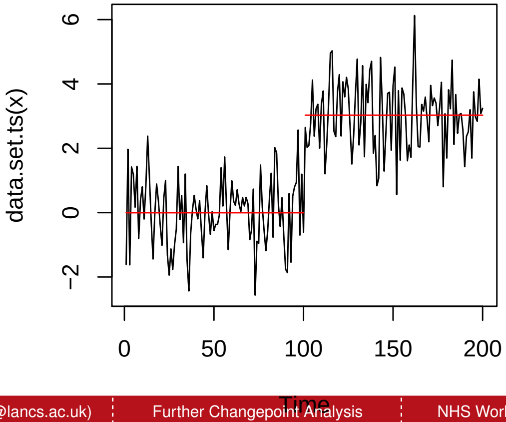

```r
acf(m1.resid)
```



Autocorrelation

What effect does autocorrelation have on our analysis?

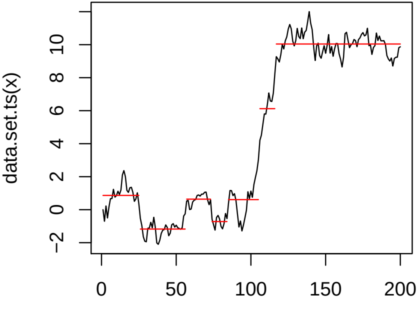
```
set.seed(879123)
x=c(rnorm(100),rnorm(100,3))
plot(cpt.meanvar(x,method='PELT'))
```



Autocorrelation +ve

What effect does autocorrelation have on our analysis?

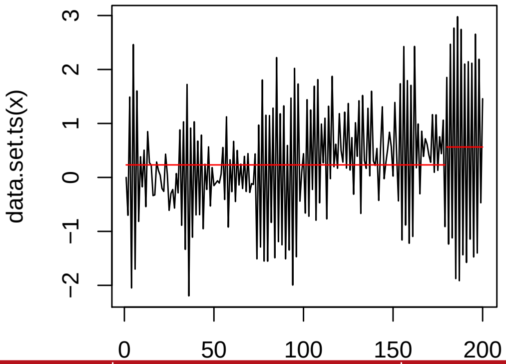
```r
source('sim.cpt.AR1.R')
set.seed(879123)
x=sim.cpt.AR1(cpts=c(0,100,200),X=cbind(rep(1,200)),init=0,
    beta=rbind(c(0,0.9),c(1,0.9)),sig2=(1-0.9^2),nsim=1)
plot(cpt.meanvar(x,method='PELT'))
```



Autocorrelation -ve

What effect does autocorrelation have on our analysis?

```
set.seed(879123)
x=sim.cpt.AR1(cpts=c(0,100,200),X=cbind(rep(1,200)),init=0,
    beta=rbind(c(0,-0.9),c(1,-0.9)),sig2=(1-0.9^2),nsim=1)
plot(cpt.meanvar(x,method='PELT'))
```



EnvCpt

Mathematics & Statistics | Lancaster University

EnvCpt automatically fits 12 different models to your data:

- Flat mean (+AR1, +AR2, +Change, +AR1+Change, +AR2+Change)
- Trend mean (+AR1, +AR2, +Change, +AR1+Change, +AR2+Change)

AR1= autoregressive of order 1 = current data point is strongly related to the last data point.

BONUS: Can see which model is best

PITFALL: Might be best to use another model which isn't checked - always look at the fit!

EnvCpt: Example


```
set.seed(879123)
x=sim.cpt.AR1(cpts=c(0,100,200),X=cbind(rep(1,200)),init=0,
    beta=rbind(c(0,0.9),c(1,0.9)),sig2=(1-0.9^2),nsim=1)
out=envcpt(x)
```

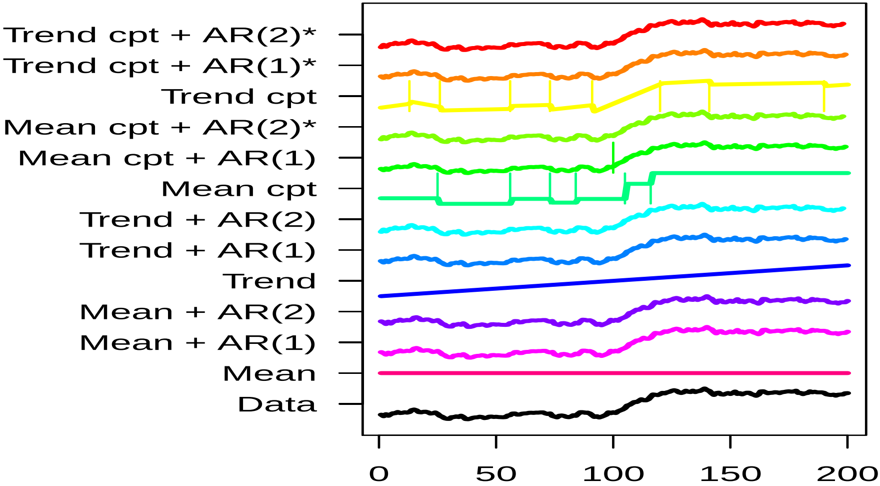
```
## Fitting 12 models

##   |
```

```
which.min(BIC(out))
```

```
## meanar1cpt
##          5
```

EnvCpt: Example


```r
plot(out)
```

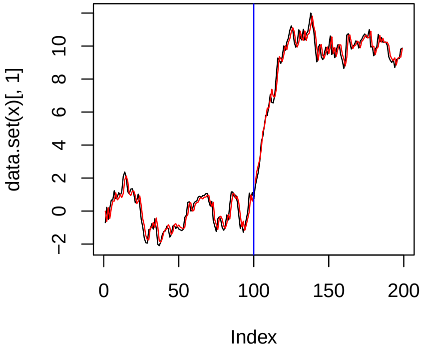


EnvCpt: Example

```
cpts(out$meanar1cpt)
```

```
## [1] 100
```

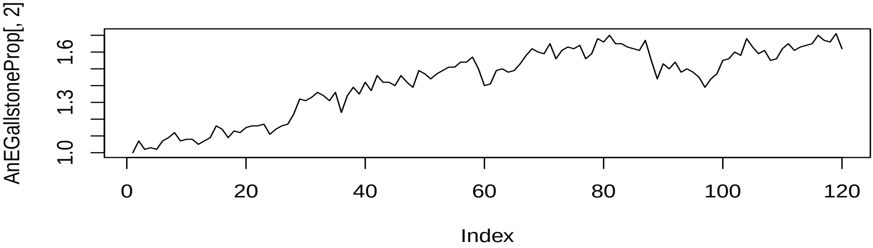
```
plot(out[[which.min(BIC(out))+1]])
abline(v=cpts(out$meanar1cpt),col='blue')
```



EnvCpt: A&E Gallstone

HES Data on monthly proportion of A&E admissions for gallstone disease from Jan 2010 - Dec 2019.

```
load('AnEGallstoneProp.Rdata')
plot(AnEGallstoneProp[,2],type='l')
```



Use EnvCpt to see if there is evidence for changes in the monthly proportion of A&E admissions for gallstone disease.

Multivariate changes

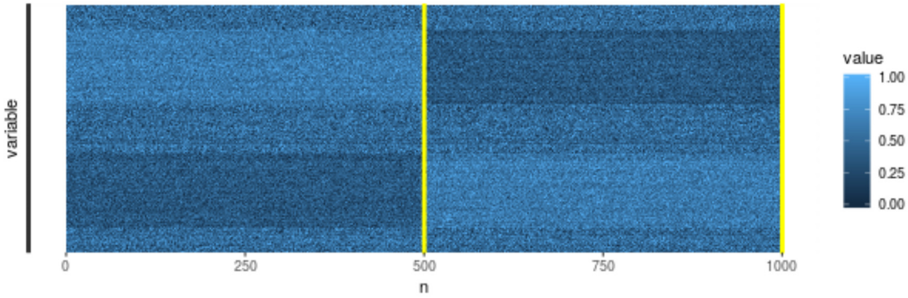
In moving to the multivariate setting a number of different scenarios could arise.

- The process in each channel could be unconnected to the rest (i.e. repeated use of univariate cpt methods might be appropriate);
- There may be some shared structure across channels. For example
  - Changes occur at the same time in all channels;
  - Changes occur in a subset of channels at the same time.
- The nature of the change could vary from one channel to another;
- ...and doubtless many more scenarios!

Multivariate changes

In the multivariate setting we encounter new challenges:

- Computational expense.
- Sparsity of changepoints.
- Incorporating multivariate power.

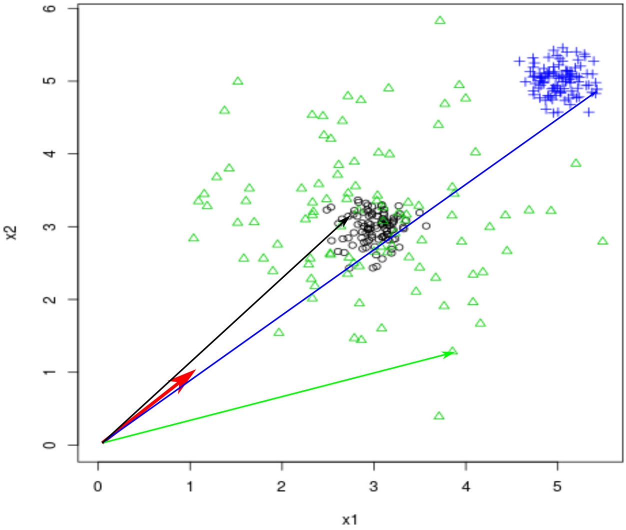


Many methods

Some well known multivariate changepoint approaches include:

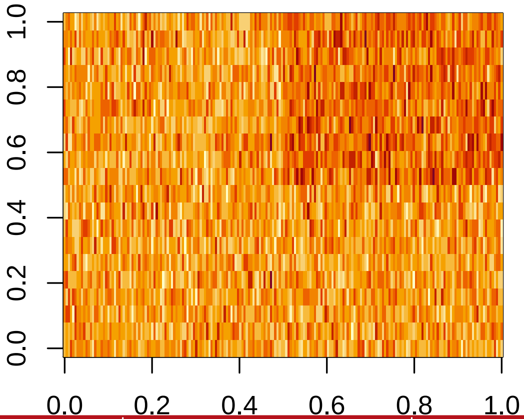
- ecp: James, Matteson (2015)
- Inspect: Wang, Samworth (2017)
- DoubleCUSUM: Cho (2016)

GeomCP Intuition



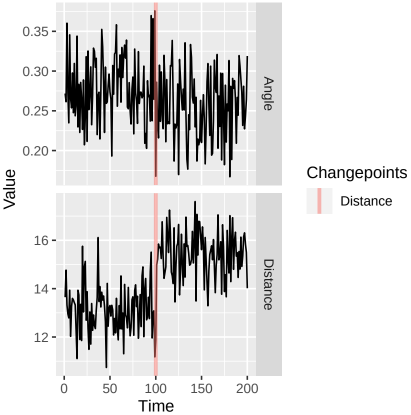
GeomCP Mean Ex

```
set.seed(1)
Y=rbind(matrix(rnorm(100*20),ncol=20),cbind(matrix(rnorm(100*
    ncol=10),matrix(rnorm(100*10,1),ncol=10)))
image(Y)
```



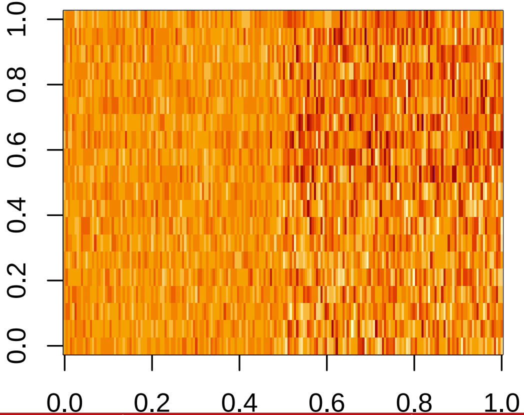
GeomCP: Mean Ex


```r
res <- geomcp(Y)
plot(res)
```



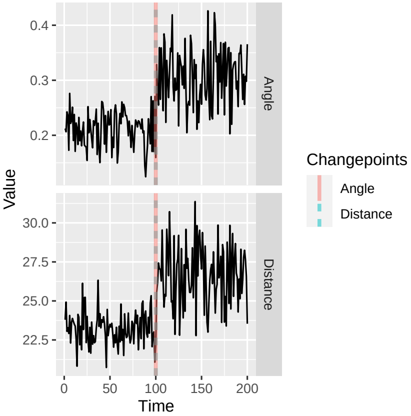
GeomCP MeanVar Ex


```
set.seed(1)
Y=rbind(matrix(rnorm(100*20),ncol=20),cbind(matrix(rnorm(100*
    ncol=10),matrix(rnorm(100*10,1,2),ncol=10)))
image(Y)
```



GeomCP: MeanVar Ex

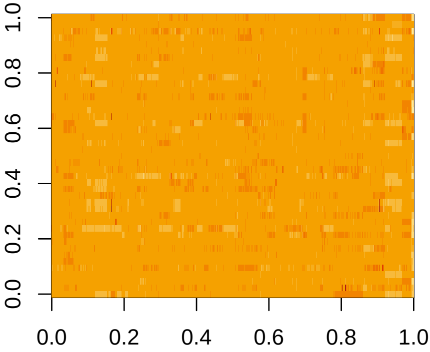
```r
res <- geomcp(Y)
plot(res)
```



Task: Genetics

Analyse the ACGH Bladder Tumor data from the ecp package. It is 2215x43 with 43 patients. How many changes do you find?

```
data(ACGH)
image(ACGH$data)
```

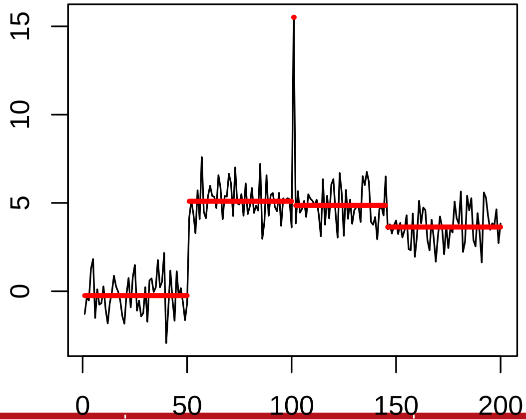


Influence

- Which data points are *influential* for obtaining the segmentation?
  - Changepoints versus Outliers
  - How to measure influence?
- How *stable* is the obtained segmentation?

Influence: Example


```
set.seed(30)
x=c(rnorm(50),rnorm(50,mean=5),rnorm(1,mean=15),rnorm(49,mear
xcpt=cpt.mean(x,method='PELT')
plot(xcpt,cpt.width=3,ylab='')
```



How to measure?

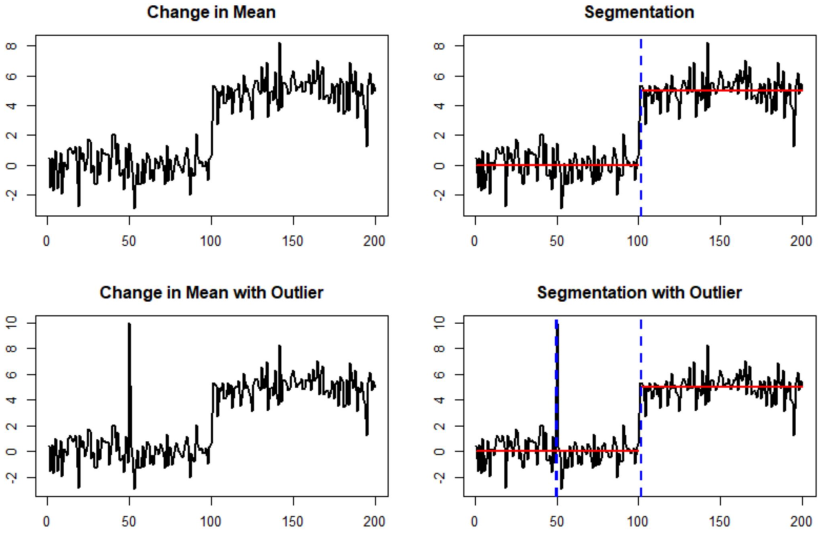
Sources of Inspiration:

- Regression Analysis: Measures of Influence (e.g., Cook's distance)
- Robust Statistics: Influence Functions

Two routes:

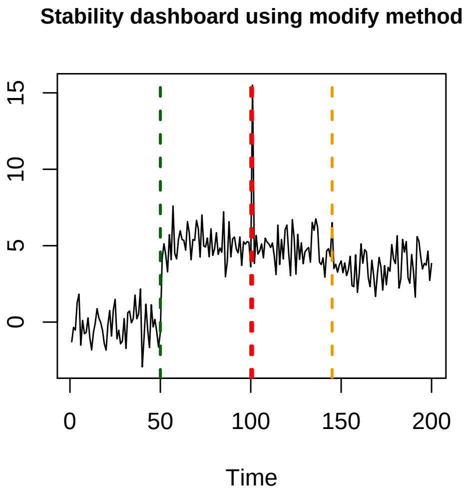
- Modifying an observation
- Leaving out an observation

Modify


Change in Mean
Segmentation
Change in Mean with Outlier
Segmentation with Outlier

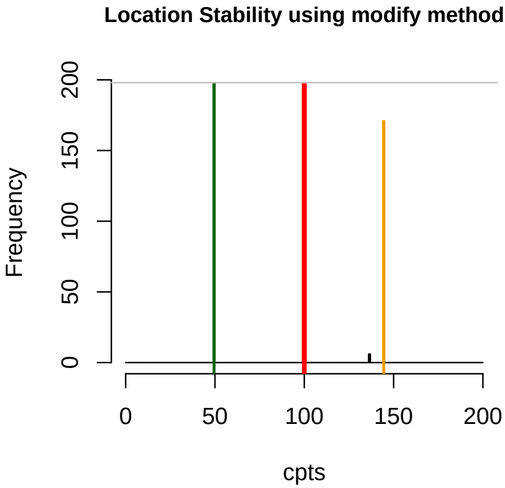
Stability Dashboard: Mod

```r
x.inf=influence.generate.PELT(xcpt,method='modify')
stability.overview(x,cpts(xcpt),x.inf,cex.main=0.9)
```



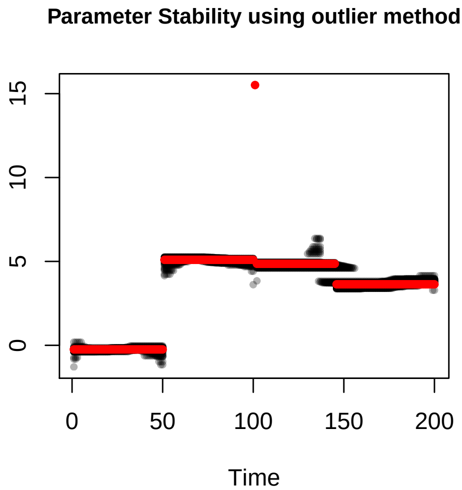
Location Stability: Mod


```
location.stability(cpts(xcpt),x.inf,cex.main=0.9)
```

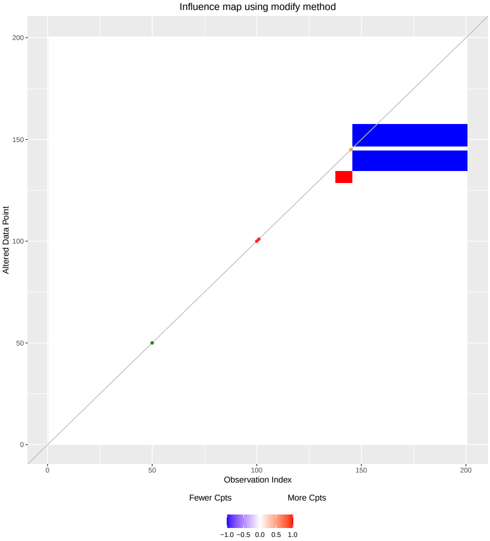


Parameter Stability: Mod


```
parameter.stability(x.inf,original.mean=rep(param.est(xcpt)$m
    times=diff(c(0,xcpt@cpts))),cex.main=0.9)
```

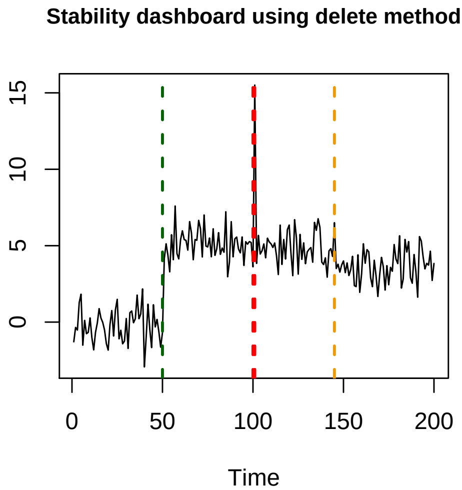


Influence Map: Mod



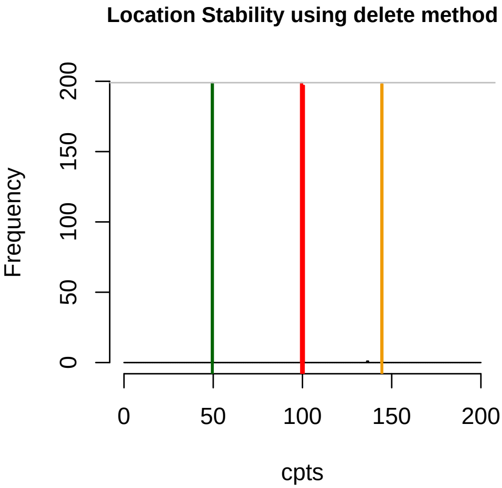
Stability Dashboard: Del


```
x.inf=influence.generate.PELT(xcpt,method='delete')
stability.overview(x,cpts(xcpt),x.inf,cex.main=0.9)
```



Location Stability: Del

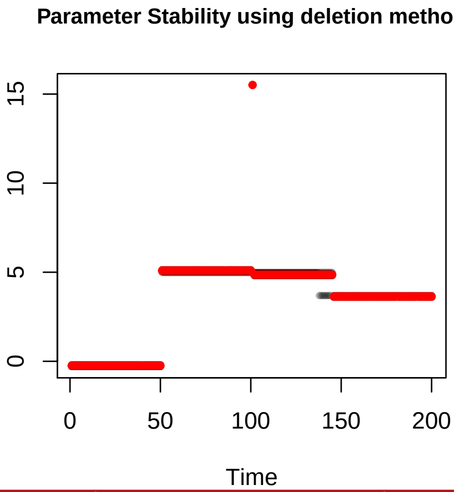

```
location.stability(cpts(xcpt),x.inf,cex.main=0.9)
```



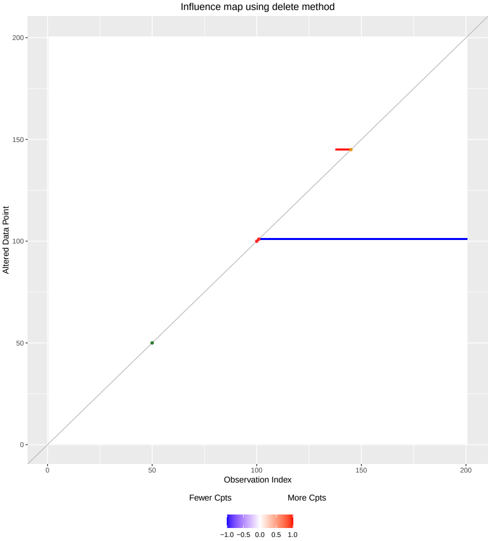
Parameter Stability: Del

Mathematics & Statistics | Lancaster University

```
parameter.stability(x.inf,original.mean=rep(param.est(xcpt)$m
    times=diff(c(0,xcpt@cpts))),cex.main=0.9)
```



Influence Map: Del



Summary

- Multivariate is interesting but still lots of challenges in the univariate space
- Lots of interesting research in the changepoint space
- Always looking for interesting problems to work on
- Reach out if you want help / guidance

References

PELT: Killick, Fearnhead, Eckley (2012)

EnvCpt: Beaulieu, Killick (2018)

geomCP: Grundy, Killick (2020)

Influence: Wilms, Killick, Matteson (2020+)