

Further Changepoint Analysis

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



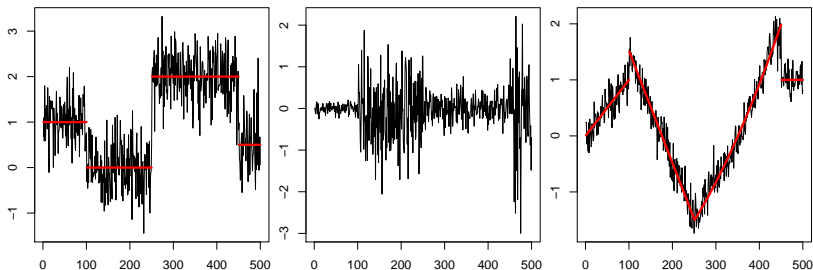
- Recap of changepoints
- Checking assumptions
- Autocorrelation
- Influence
- Multivariate changepoints

There will be tasks throughout the sections.

Recall: changepoints

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

There are many different types of change.



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

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?

- 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)

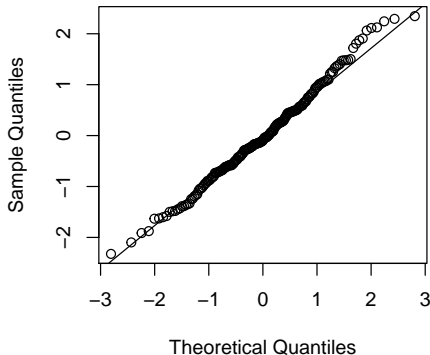
##
##  Shapiro-Wilk normality test
##
## data:  m1.resid
## W = 0.99228, p-value = 0.3721
```

```
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
```

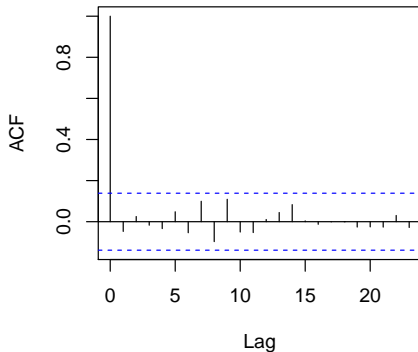
```
qqnorm(m1.resid)  
qqline(m1.resid)
```

Normal Q-Q Plot



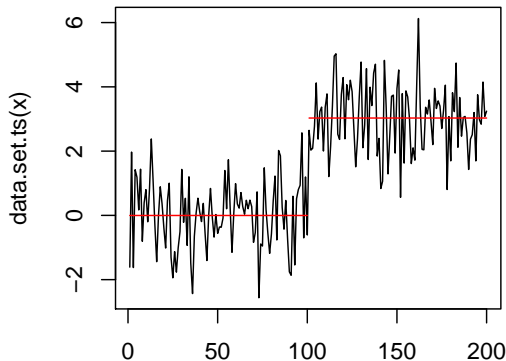

```
acf(m1.resid)
```

Series m1.resid



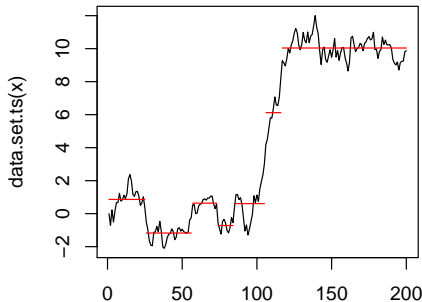
What effect does autocorrelation have on our analysis?

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



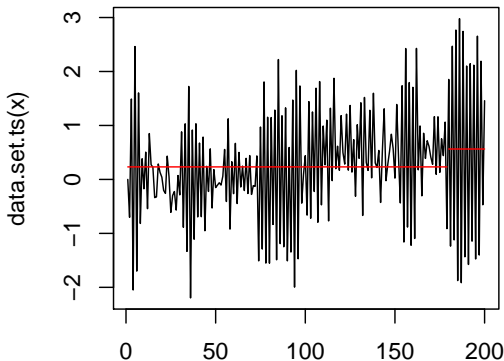
What effect does autocorrelation have on our analysis?

```
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'))
```



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 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!

```
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
```

```
## |
```

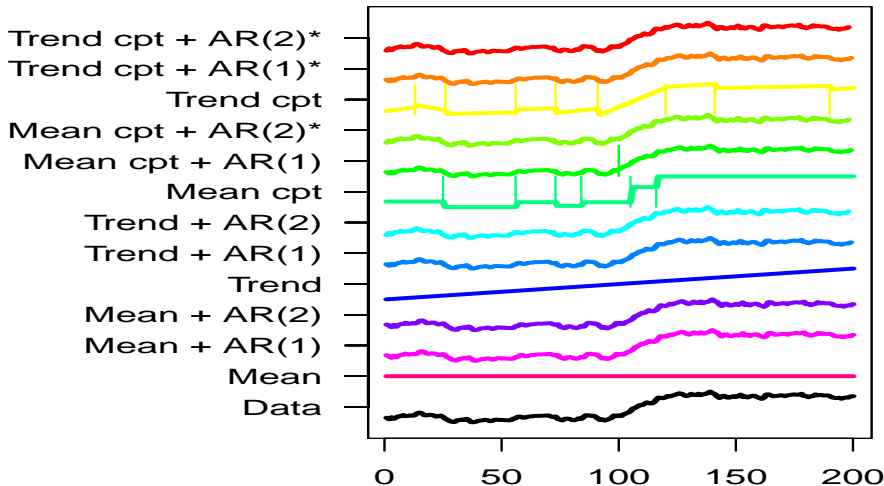
```
which.min(BIC(out))
```

```
## meanar1cpt
```

```
## 5
```

EnvCpt: Example

`plot(out)`

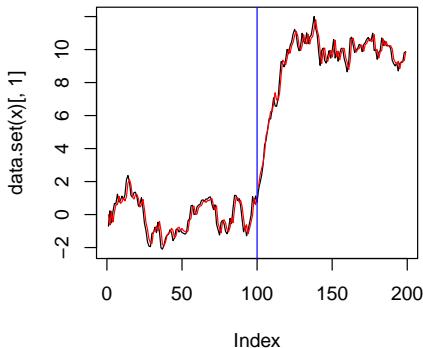


```
cpts(out$meanar1cpt)
```

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

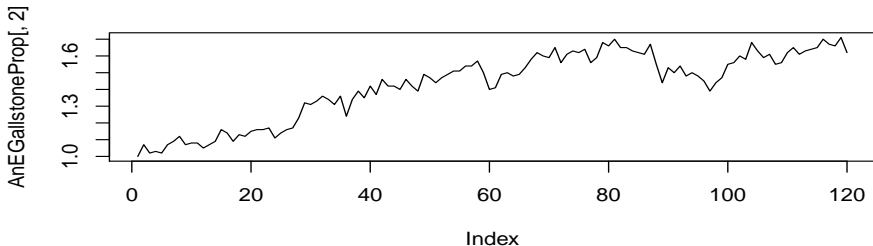
```
plot(out[[which.min(BIC(out))+1]])
```

```
abline(v=cpts(out$meanar1cpt),col='blue')
```



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.

```
out=envcpt(AnEGallstoneProp[,2])
```

```
## Fitting 12 models
```

```
## |
```

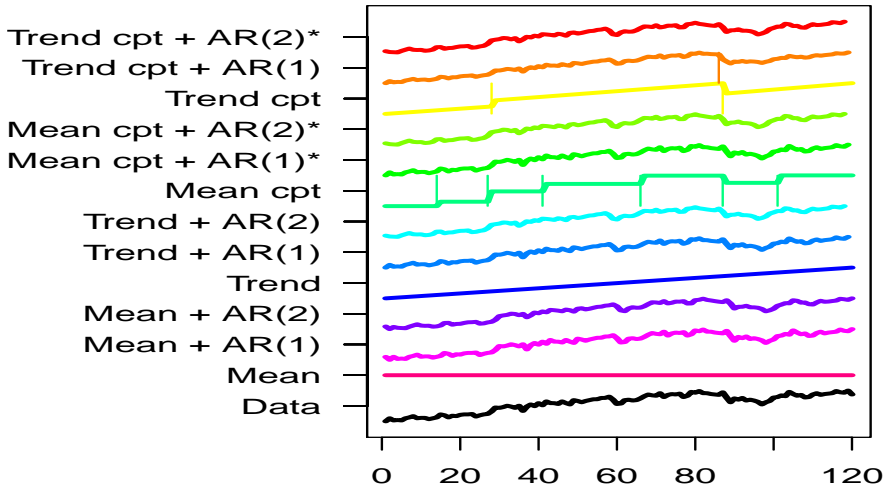
```
which.min(BIC(out))
```

```
## trendar1cpt
```

```
## 11
```

Gallstone Solution

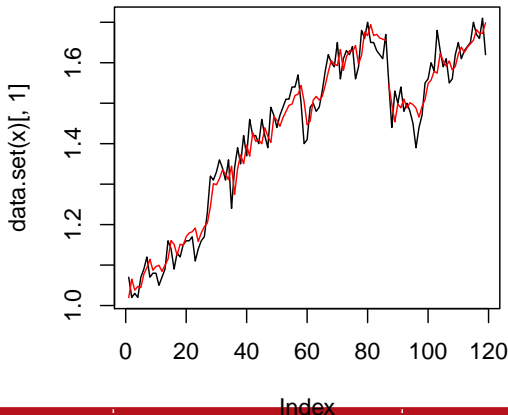
`plot(out)`



```
AnEGallstoneProp$date[cpts(out$trendar1cpt)]
```

```
## [1] "1-02-17"
```

```
plot(out$trendar1cpt)
```

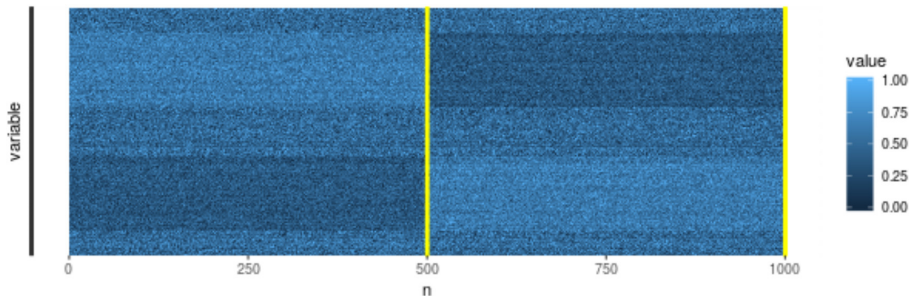


In moving to the multivariate setting a number of different scenarios could arise.

- 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!

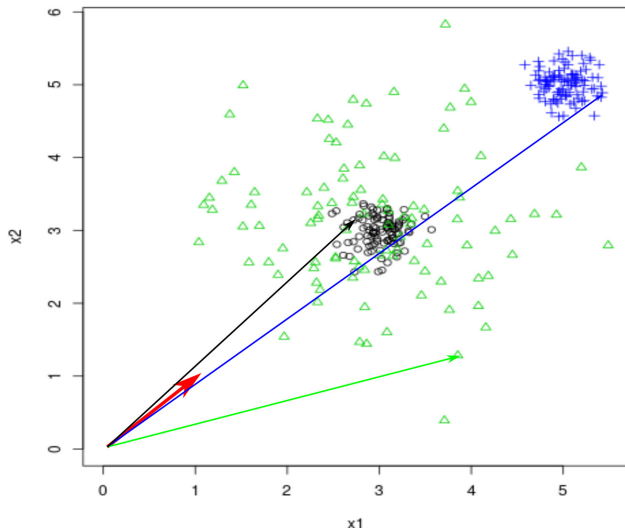
In the multivariate setting we encounter new challenges:

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

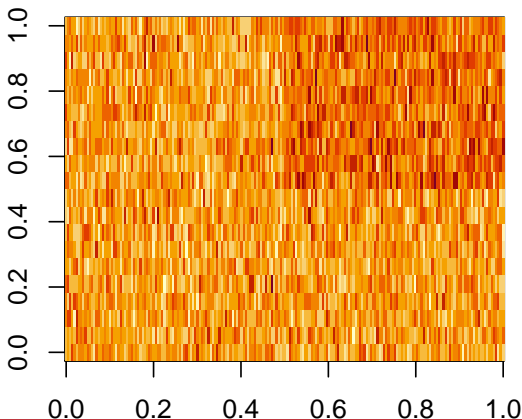


Some well known multivariate changepoint approaches include:

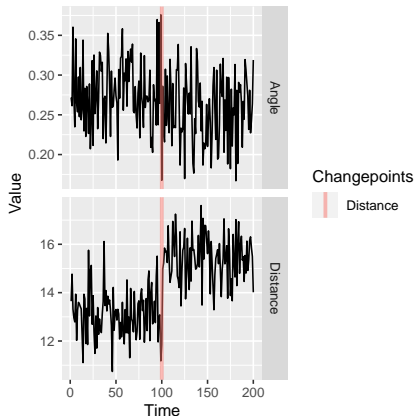
- ecp: James, Matteson (2015)
- Inspect: Wang, Samworth (2017)
- DoubleCUSUM: Cho (2016)



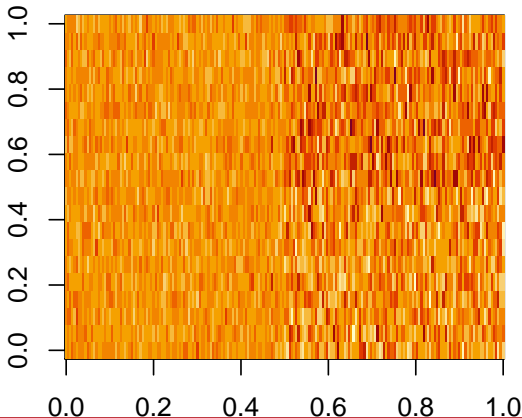

```
set.seed(1)  
Y=rbind(matrix(rnorm(100*20),ncol=20),cbind(matrix(rnorm(100*10),  
ncol=10),matrix(rnorm(100*10,1),ncol=10)))  
image(Y)
```



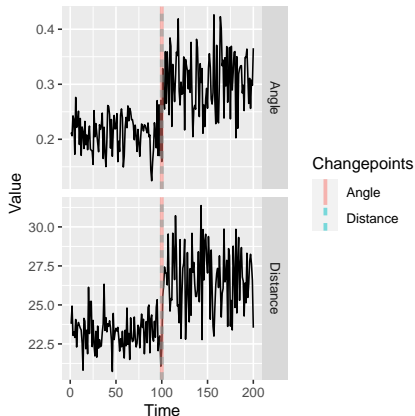
```
res <- geomcp(Y)  
plot(res)
```



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



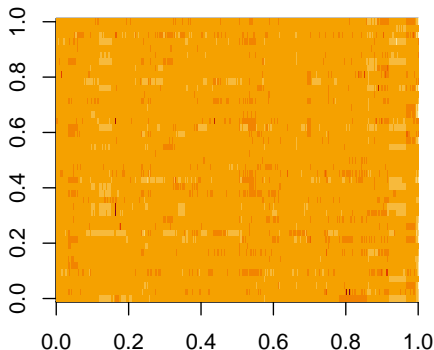
```
res <- geomcp(Y)  
plot(res)
```



Task: Genetics

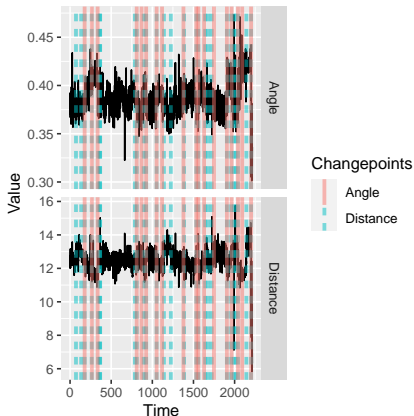
Analyse the ACGH Bladder Tumor data from the ecp package. It is 2215x43 with 43 patients. How many changes do you find?

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



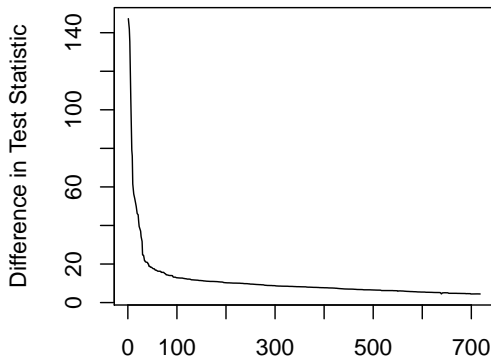
Solution: Genetics

```
outACGH <- geomcp(ACGH$data)  
plot(outACGH)
```



Can also take the distance and angle vectors and analyse using CROPS

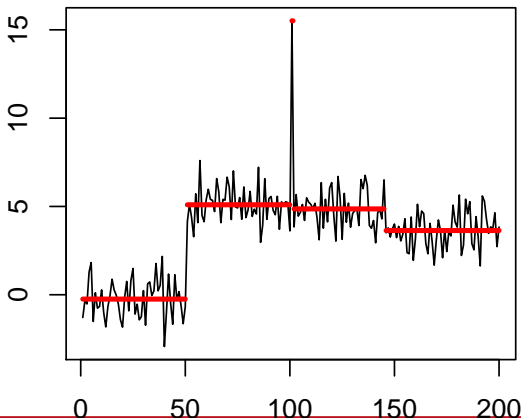
```
outACGH.dist=cpt.meanvar(distance(outACGH),method='PELT',  
    penalty='CROPS',pen.value=c(5,500))  
plot(outACGH.dist,diagnostic=TRUE)
```



- 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,mean=5))
xcpt=cpt.mean(x,method='PELT')
plot(xcpt,cpt.width=3,ylab='')
```



Sources of Inspiration:

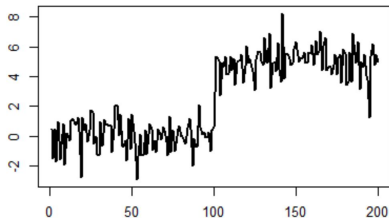
- Regression Analysis: Measures of Influence (e.g., Cook's distance)
- Robust Statistics: Influence Functions

Two routes:

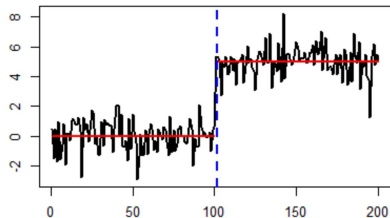
- Modifying an observation
- Leaving out an observation



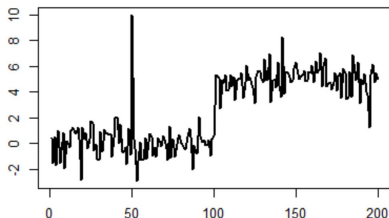
Change in Mean



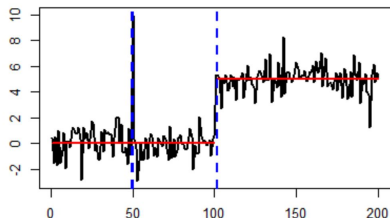
Segmentation



Change in Mean with Outlier



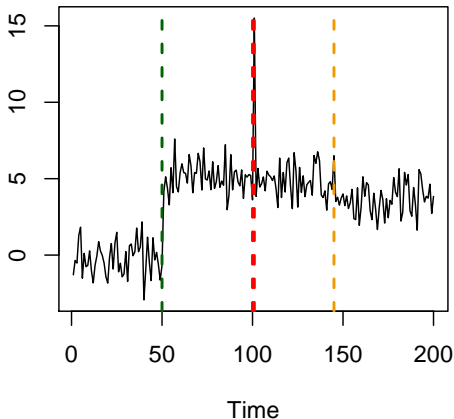
Segmentation with Outlier





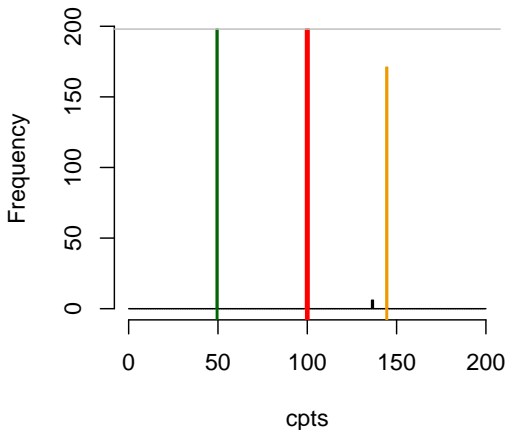
```
x.inf=influence.generate.PELT(xcpt,method='modify')  
stability.overview(x,cpts(xcpt),x.inf,cex.main=0.9)
```

Stability dashboard using modify method



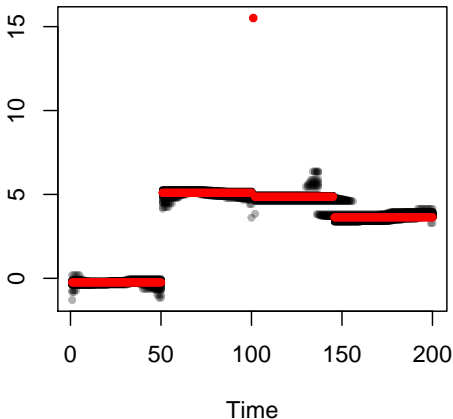
```
location.stability(cpts(xcpt),x.inf,cex.main=0.9)
```

Location Stability using modify method

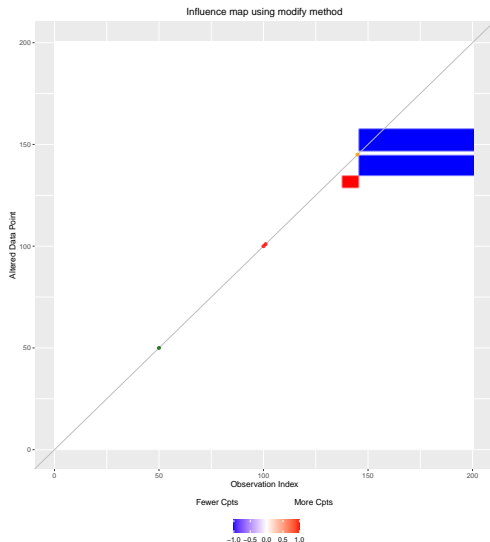


```
parameter.stability(x.inf,original.mean=rep(param.est(xcpt)$  
times=diff(c(0,xcpt@cpts))),cex.main=0.9)
```

Parameter Stability using outlier method



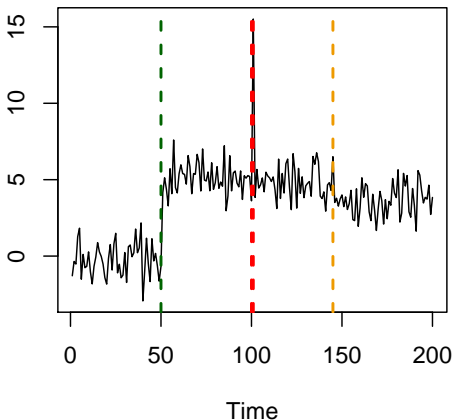
Influence Map: Mod





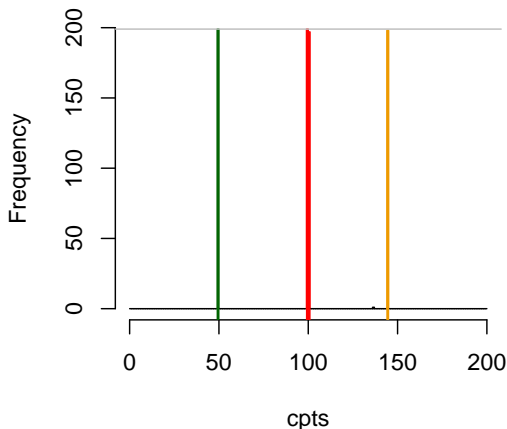
```
x.inf=influence.generate.PELT(xcpt,method='delete')  
stability.overview(x,cpts(xcpt),x.inf,cex.main=0.9)
```

Stability dashboard using delete method



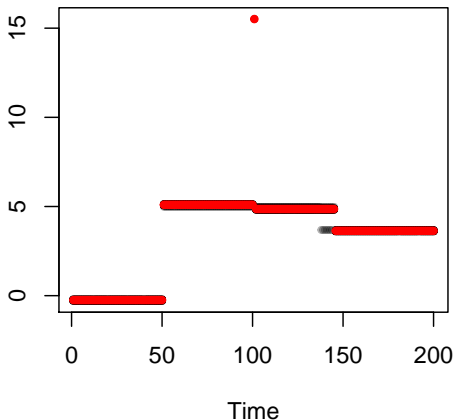

```
location.stability(cpts(xcpt),x.inf,cex.main=0.9)
```

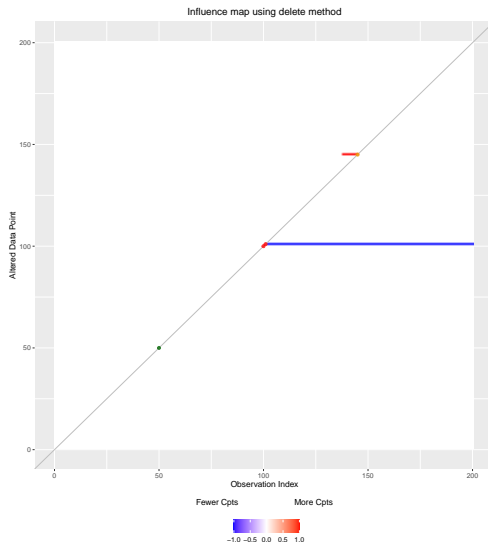
Location Stability using delete method



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

Parameter Stability using deletion method





- 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

PELT: Killick, Fearnhead, Eckley (2012)

EnvCpt: Beaulieu, Killick (2018)

geomCP: Grundy, Killick (2020)

Influence: Wilms, Killick, Matteson (2020+)