Collective & Point Anomaly Detection in R

useR! 2019. Toulouse, France.

Daniel Grose

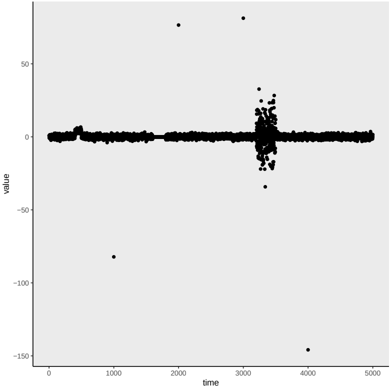
July 10, 2019

What is an Anomaly ?

Simulated Data

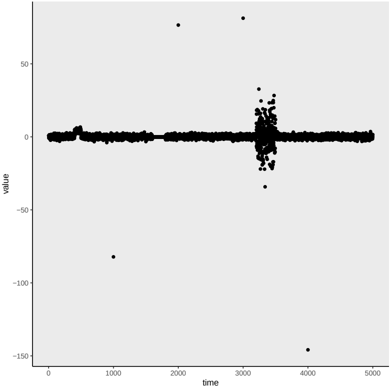
What is an Anomaly ?

Simulated Data



What is an Anomaly ?

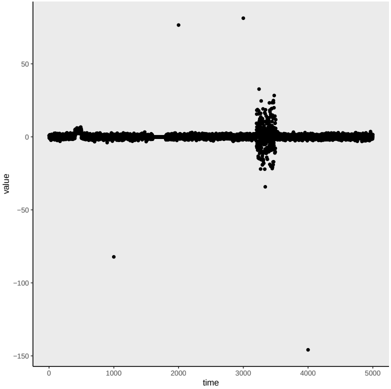
Simulated Data



Anomalies

What is an Anomaly ?

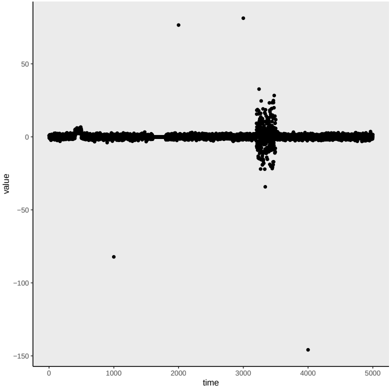
Simulated Data



Anomalies - observations that do not conform with the pattern of the data.

What is an Anomaly ?

Simulated Data

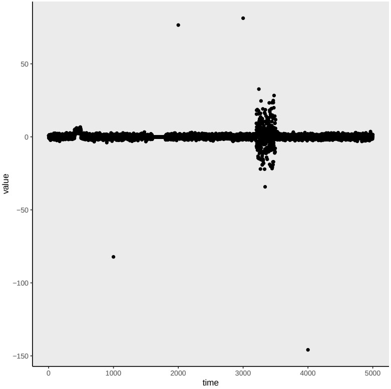


Anomalies - observations that do not conform with the pattern of the data.

- Point Anomaly

What is an Anomaly ?

Simulated Data

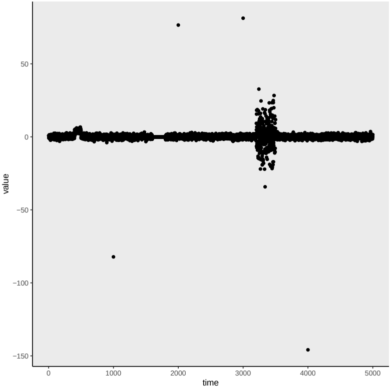


Anomalies - observations that do not conform with the pattern of the data.

- Point Anomaly - a *global* or *contextual* outlier.

What is an Anomaly ?

Simulated Data

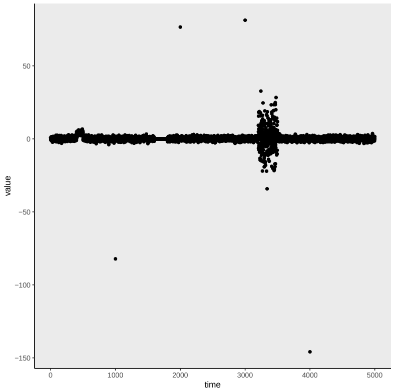


Anomalies - observations that do not conform with the pattern of the data.

- Point Anomaly - a *global* or *contextual* outlier.
- Collective Anomaly

What is an Anomaly ?

Simulated Data



Anomalies - observations that do not conform with the pattern of the data.

- Point Anomaly - a *global* or *contextual* outlier.
- Collective Anomaly - observations that are credible in their context but unusual as a group.

Modelling Anomalies

Overview

Assume a parametric model for x_t with PDF $f(x, \theta(t))$

Modelling Anomalies
Overview

Assume a parametric model for x_t with PDF $f(x, \theta(t))$

Modelling Anomalies
Overview

Assume a parametric model for x_t with PDF $f(x, \theta(t))$

- Infer the parameter θ_0 for the typical distribution using robust statistics.

Modelling Anomalies
Overview

Assume a parametric model for x_t with PDF $f(x, \theta(t))$

- Infer the parameter θ_0 for the typical distribution using robust statistics.
- Model collective anomalies as epidemics in $\theta(t)$ that deviate from θ_0.

Modelling Anomalies

Overview

Assume a parametric model for x_t with PDF $f(x, \theta(t))$

- Infer the parameter θ_0 for the typical distribution using robust statistics.
- Model collective anomalies as epidemics in $\theta(t)$ that deviate from θ_0.
- Model point anomalies as epidemic anomalies of length 1.

Modelling Anomalies
Overview

Assume a parametric model for x_t with PDF $f(x, \theta(t))$

- Infer the parameter θ_0 for the typical distribution using robust statistics.
- Model collective anomalies as epidemics in $\theta(t)$ that deviate from θ_0.
- Model point anomalies as epidemic anomalies of length 1.

The number and location of changepoints are determined by dynamically minimising a penalised cost problem [Fisch et al., 2018a].

Modelling Anomalies

Overview

Assume a parametric model for x_t with PDF $f(x, \theta(t))$

- Infer the parameter θ_0 for the typical distribution using robust statistics.
- Model collective anomalies as epidemics in $\theta(t)$ that deviate from θ_0.
- Model point anomalies as epidemic anomalies of length 1.

The number and location of changepoints are determined by dynamically minimising a penalised cost problem [Fisch et al., 2018a]. Dynamic programming combined with pruning results in a worst case $\mathcal{O}(n^2)$ method. Typically it is $\mathcal{O}(n)$.

The **anomaly** package

Example 1 - simulated data

The **anomaly** package

Example 1 - simulated data

```
# generate simulated data
R> x = rnorm(5000)
R> x[401:500] = rnorm(100,4,1)
R> x[1601:1800] = rnorm(200,0,0.01)
R> x[3201:3500] = rnorm(300,0,10)
R> x[c(1000,2000,3000,4000)] = rnorm(4,0,100)
```

The **anomaly** package

Example 1 - simulated data

```
# generate simulated data
R> x = rnorm(5000)
R> x[401:500] = rnorm(100,4,1)
R> x[1601:1800] = rnorm(200,0,0.01)
R> x[3201:3500] = rnorm(300,0,10)
R> x[c(1000,2000,3000,4000)] = rnorm(4,0,100)
```

```
R> res<-uvcapa(x)
R> plot(res)
R> res
Univariate CAPA detecting changes in mean and variance.
observations = 5000
minimum segment length = 10
maximum segment length = 5000
Point anomalies detected : 4
  location  strength
1     1000  67.19607
2     2000  41.68110
3     3000 136.37800
4     4000  45.35898
Collective anomalies detected : 3
  start  end  mean.change variance.change
1   401  500 1.553523e+01      0.02204117
2  1601 1800 1.698052e-04    106.40481634
3  3202 3500 2.562775e-02      7.33503544
```

The **anomaly** package
Example 1 - simulated data

```
# generate simulated data
R> x = rnorm(5000)
R> x[401:500] = rnorm(100,4,1)
R> x[1601:1800] = rnorm(200,0,0.01)
R> x[3201:3500] = rnorm(300,0,10)
R> x[c(1000,2000,3000,4000)] = rnorm(4,0,100)
```

```
R> res<-uvcapa(x)
R> plot(res)
R> res
Univariate CAPA detecting changes in mean and variance.
observations = 5000
minimum segment length = 10
maximum segment length = 5000
Point anomalies detected : 4
  location  strength
1     1000  67.19607
2     2000  41.68110
3     3000 136.37800
4     4000  45.35898
Collective anomalies detected : 3
  start  end  mean.change variance.change
1   401  500 1.553523e+01      0.02204117
2  1601 1800 1.698052e-04    106.40481634
3  3202 3500 2.562775e-02      7.33503544
```

Package available from CRAN

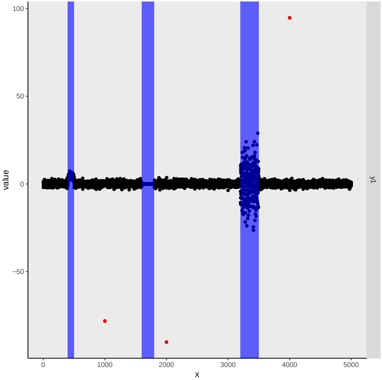
[Fisch et al., 2018b].

The **anomaly** package

Example 1 - simulated data

```
# generate simulated data
R> x = rnorm(5000)
R> x[401:500] = rnorm(100,4,1)
R> x[1601:1800] = rnorm(200,0,0.01)
R> x[3201:3500] = rnorm(300,0,10)
R> x[c(1000,2000,3000,4000)] = rnorm(4,0,100)


R> res<-uvcapa(x)
R> plot(res)
R> res
Univariate CAPA detecting changes in mean and variance.
observations = 5000
minimum segment length = 10
maximum segment length = 5000
Point anomalies detected : 4
  location  strength
1     1000  67.19607
2     2000  41.68110
3     3000 136.37800
4     4000  45.35898
Collective anomalies detected : 3
  start  end  mean.change variance.change
1   401  500 1.553523e+01      0.02204117
2  1601 1800 1.698052e-04    106.40481634
3  3202 3500 2.562775e-02      7.33503544
```



Package available from CRAN

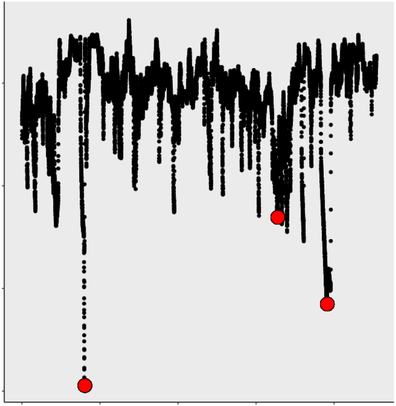
[Fisch et al., 2018b].

The **anomaly** package

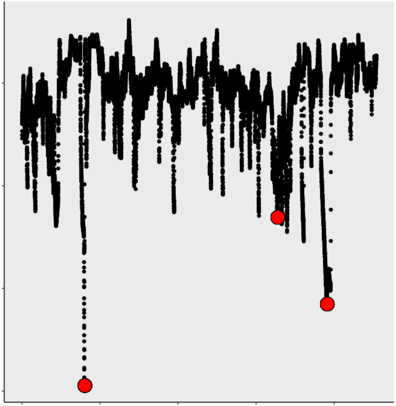
Example 2 - real data

The **anomaly** package

Example 2 - real data

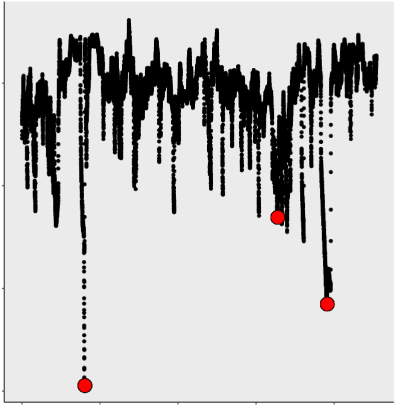


The **anomaly** package
Example 2 - real data



The **anomaly** package

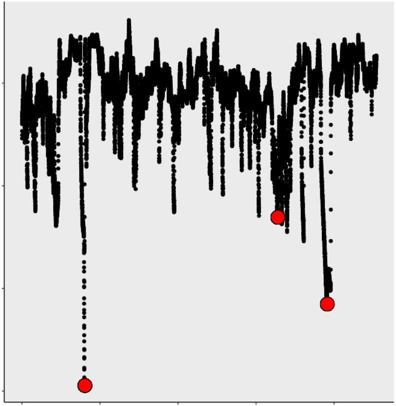
Example 2 - real data



- Temperature sensor in a large industrial machine.

The **anomaly** package

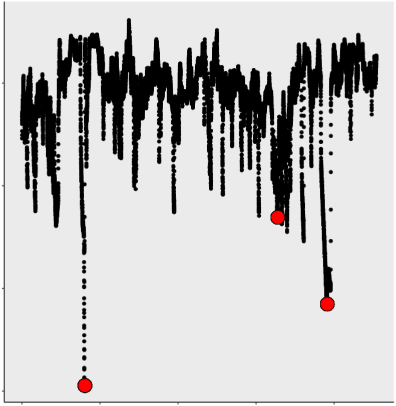
Example 2 - real data



- Temperature sensor in a large industrial machine.
- Taken from the Numenta Anomaly Benchmark (NAB).

The **anomaly** package

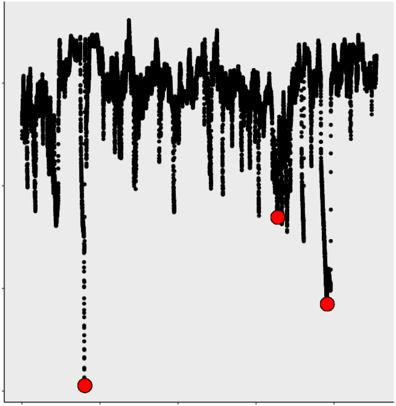
Example 2 - real data



- Temperature sensor in a large industrial machine.
- Taken from the Numenta Anomaly Benchmark (NAB).
- 22695 observations at 5 minute intervals.

The **anomaly** package

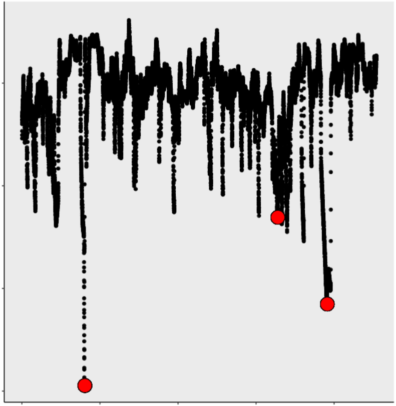
Example 2 - real data



- Temperature sensor in a large industrial machine.
- Taken from the Numenta Anomaly Benchmark (NAB).
- 22695 observations at 5 minute intervals.
- Three known anomalies as identified by an engineer.

The **anomaly** package

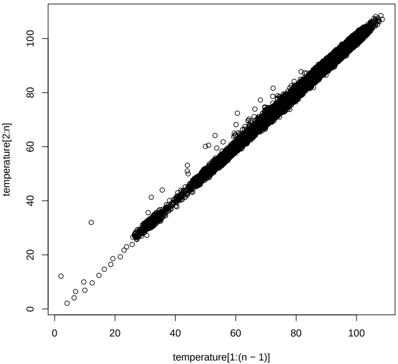
Example 2 - real data

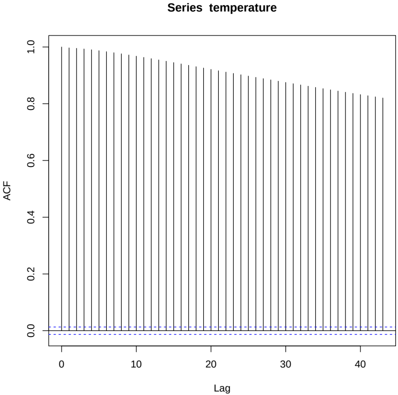


- Temperature sensor in a large industrial machine.
- Taken from the Numenta Anomaly Benchmark (NAB).
- 22695 observations at 5 minute intervals.
- Three known anomalies as identified by an engineer.
- Data is highly autocorrelated.

The **anomaly** package

Example 2 - real data - accounting for autocorrelation





The **anomaly** package

Example 2 - real data - accounting for autocorrelation

The **anomaly** package

Example 2 - real data - accounting for autocorrelation

```
R> n<-length(temperature)
R> library(robust)
R> rcov<-covRob(matrix(c(temperature[2:n],temperature[1:(n-1)]),
                       ncol=2),
             corr=TRUE,estim="M")
R> psi<-rcov$cov[1,2]
R> psi
0.986951
```

The **anomaly** package

Example 2 - real data - accounting for autocorrelation

```
R> n<-length(temperature)
R> library(robust)
R> rcov<-covRob(matrix(c(temperature[2:n],temperature[1:(n-1)]),
                       ncol=2),
             corr=TRUE,estim="M")
R> psi<-rcov$cov[1,2]
R> psi
0.986951


# NB - default value is 3*log(n)
R> inflated_penalty<-3*(1+psi)/(1-psi)*log(n)
R> res<-uvcapa(temperature,type="mean",beta=inflated_penalty,
               beta_tilde=inflated_penalty)
R> res # summary of results
Univariate CAPA detecting changes in mean.
observations = 22695
minimum segment length = 10
maximum segment length = 22695
Collective anomalies detected : 4
  start   end mean.change test.statistic
1  1612  2327    9.148952       6550.650
2  3773  4002   25.648888       5899.244
3 16023 17204    8.191733       9682.628
4 19166 19775   39.426847      24050.377
R> plot(res)
```

The **anomaly** package

Example 2 - real data - accounting for autocorrelation

```
R> n<-length(temperature)
R> library(robust)
R> rcov<-covRob(matrix(c(temperature[2:n],temperature[1:(n-1)]),
                    ncol=2),
             corr=TRUE,estim="M")
R> psi<-rcov$cov[1,2]
R> psi
0.986951
```

```
# NB - default value is 3*log(n)
R> inflated_penalty<-3*(1+psi)/(1-psi)*log(n)
R> res<-uvcapa(temperature,type="mean",beta=inflated_penalty,
               beta_tilde=inflated_penalty)
R> res # summary of results
Univariate CAPA detecting changes in mean.
observations = 22695
minimum segment length = 10
maximum segment length = 22695
Collective anomalies detected : 4
  start   end mean.change test.statistic
1  1612  2327    9.148952       6550.650
2  3773  4002   25.648888       5899.244
3 16023 17204    8.191733       9682.628
4 19166 19775   39.426847      24050.377
R> plot(res)
```

For details see

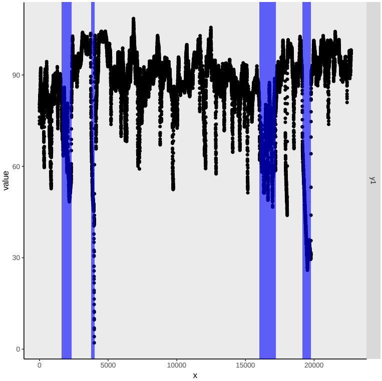
[Lavielle and Moulines, 2000]

The **anomaly** package

Example 2 - real data - accounting for autocorrelation

```
R> n<-length(temperature)
R> library(robust)
R> rcov<-covRob(matrix(c(temperature[2:n],temperature[1:(n-1)]),
                      ncol=2),
             corr=TRUE,estim="M")
R> psi<-rcov$cov[1,2]
R> psi
0.986951


# NB - default value is 3*log(n)
R> inflated_penalty<-3*(1+psi)/(1-psi)*log(n)
R> res<-uvcapa(temperature,type="mean",beta=inflated_penalty,
              beta_tilde=inflated_penalty)
R> res # summary of results
Univariate CAPA detecting changes in mean.
observations = 22695
minimum segment length = 10
maximum segment length = 22695
Collective anomalies detected : 4
  start   end mean.change test.statistic
1  1612  2327    9.148952       6550.650
2  3773  4002   25.648888       5899.244
3 16023 17204    8.191733       9682.628
4 19166 19775   39.426847      24050.377
R> plot(res)
```

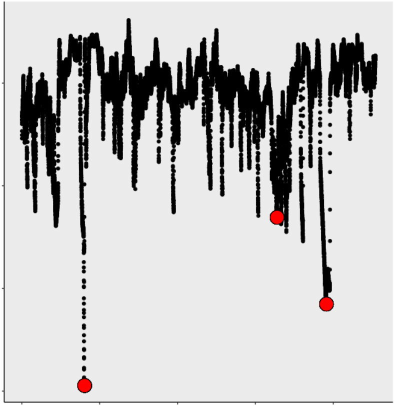


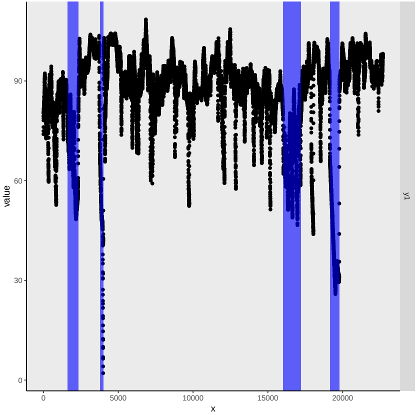
For details see

[Lavielle and Moulines, 2000]

The **anomaly** package

Example 2 - real data - results





The **anomaly** package

Features

The **anomaly** package

Features

- Detects either change in mean or change in mean and variance

The **anomaly** package

Features

- Detects either change in mean or change in mean and variance
- Penalty value can be modified

The **anomaly** package

Features

- Detects either change in mean or change in mean and variance
- Penalty value can be modified
- Maximum and minimum segment lengths

The **anomaly** package

Features

- Detects either change in mean or change in mean and variance
- Penalty value can be modified
- Maximum and minimum segment lengths
- Returns S4 class containing collective and point anomaly information

The **anomaly** package

Features

- Detects either change in mean or change in mean and variance
- Penalty value can be modified
- Maximum and minimum segment lengths
- Returns S4 class containing collective and point anomaly information
- Generic methods for **plot** and **summary**

The **anomaly** package

Features

- Detects either change in mean or change in mean and variance
- Penalty value can be modified
- Maximum and minimum segment lengths
- Returns S4 class containing collective and point anomaly information
- Generic methods for **plot** and **summary**
- Several methods for post processing anomaly information

The **anomaly** package

Features

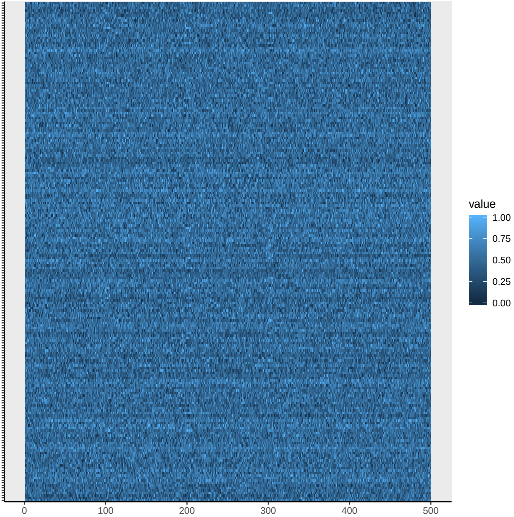
- Detects either change in mean or change in mean and variance
- Penalty value can be modified
- Maximum and minimum segment lengths
- Returns S4 class containing collective and point anomaly information
- Generic methods for **plot** and **summary**
- Several methods for post processing anomaly information

The **anomaly** package

Recent Developments - multivariate CAPA

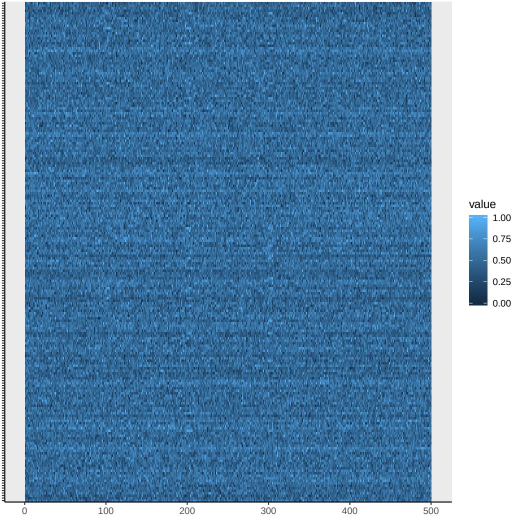
The **anomaly** package

Recent Developments - multivariate CAPA



The **anomaly** package

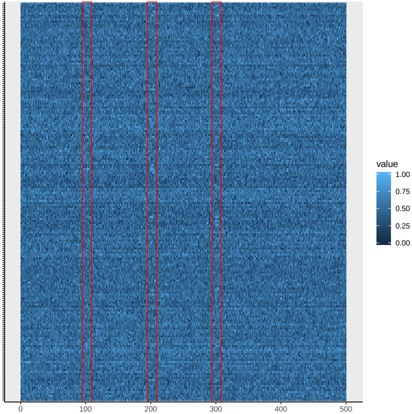
Recent Developments - multivariate CAPA



```
R> data(simulated)
# randomly shuffle the columns
R> set.seed(0)
R> m<-ncol(sim.data)
R> res<-capa(sim.data[,sample(1:m)],
           type="mean",max_lag=5)
R> plot(res)
```

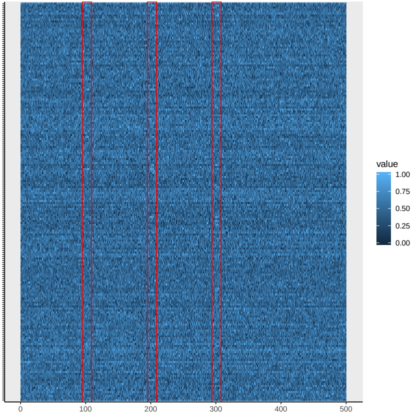
The **anomaly** package

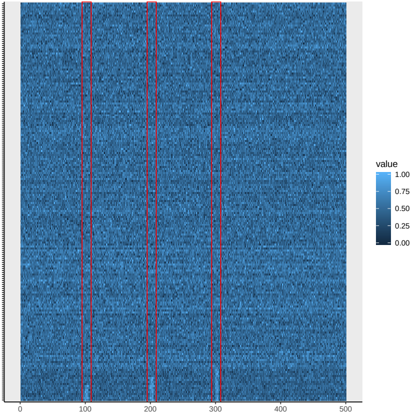
Recent Developments - multivariate CAPA



The **anomaly** package

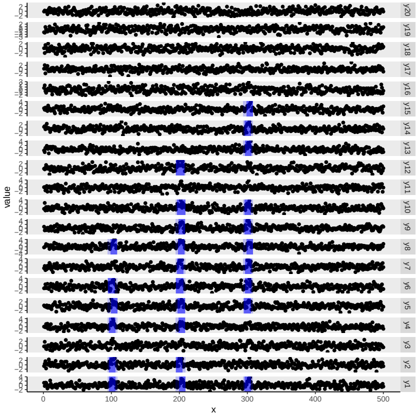
Recent Developments - multivariate CAPA





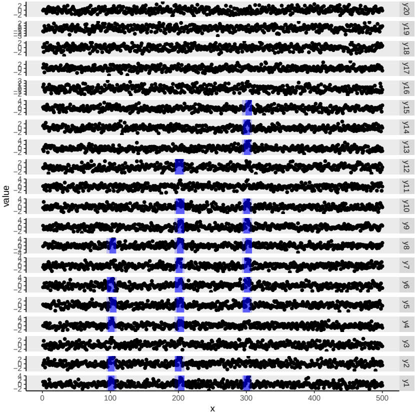
The **anomaly** package

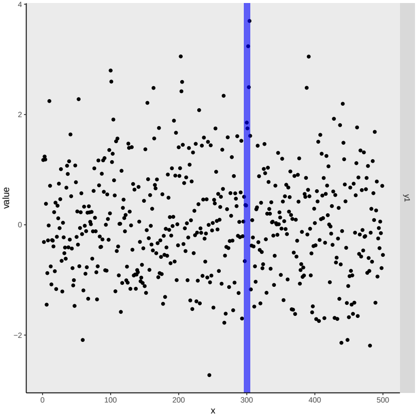
Recent Developments - multivariate CAPA



The **anomaly** package

Recent Developments - multivariate CAPA





The **anomaly** package

References

Fisch, A. T. M., Eckley, I. A., and Fearnhead, P. (2018a).
A linear time method for the detection of point and collective anomalies.
arXiv e-prints, page arXiv:1806.01947.
`https://ui.adsabs.harvard.edu/abs/2018arXiv180601947F`.

Fisch, A. T. M., Grose, D. J., Eckley, I. A., and Fearnhead, P. (2018b).
anomaly: An R package for detecting anomalies in data.
R package version 1.2.0.

Lavielle, M. and Moulines, E. (2000).
Least-squares estimation of an unknown number of shifts in a time series.
Journal of Time Series Analysis, 21(1):33–59.
`https://doi.org/10.1111/1467-9892.00172`.