Pairwise SNHT for changepoint detection.

Carina Schneider

Abstract

This vignette provides an example on how to use the function **pairwiseSNHT** of the package **snht** for changepoint detection.

Keywords: pairwiseSNHT, time series, climate data, temperature data.

1. Methodology

1.1. Main idea

The pairwiseSNHT algorithm performs a relative homogeneity test, i.e., reference series are used to detect change points at a certain location. For this purpose it calculates pairwise difference series from highly correlated neighbor time series as described in Menne and Williams (2009). This has the advantage that overall periodic or linear trends are no longer present in the difference series which then can be investigated through the Standard Normal Homogeneity Test (snht). This of course only works under the assumption that "similar variations in climate occur at nearby locations", Menne and Williams (2009).

Relative homogeneity tests for analyzing a time series at one location with the help of reference series often have the disadvantage that reference series must first be investigated for homogeneity to guarantee that the inhomogeneity is found at the right location, Menne and Williams (2009). The pairwiseSNHT therefore investigates for each location more than one difference series. In concrete terms, it investigates for each location k difference series. These k difference series come about through subtracting the k closest neighbor series from the investigated location series. It then counts the number of inhomogeneities that occurred at a certain time involving a specific location. The locations being involved with the highest counts of inhomogeneities are then assumed to be the locations where the change points occurred.

If, however, the same maximal count of inhomogeneities are measured for two neighbor locations, pairwiseSNHT just assumes that the inhomogeneity occurred at the location with the lower enumeration number. This is, so to speak, arbitrary since tie-breaking is non-trivial.

1.2. Code review

1. Calculate a distance matrix with as (i,j)-th entry the geographical distance between locations i and j.
2. Choose k (number of neighbors for each location), period (see description of the snht), critical value (threshold for the SNHT statistic, if it exceeds this critical value then a change point is assumed to have occurred).
3. Pass the above arguments into the pairwiseSNHT function, i.e.:

```
pairwiseSNHT(data, dist, k, period, crit, returnStat=TRUE/FALSE)
```

4. pairwiseSNHT calculates unique pairs of neighbor locations using the distance matrix and *k*.

5. The snht function is applied to each pair, i.e., its difference series is applied to snht with scaled=TRUE (snht statistic has a chi-squared distribution) and robust=FALSE (non-robust estimator is used) and the period that was set as an input parameter in pairwiseSNHT.

6. If returnStat=TRUE, a matrix with dimension time x (number of pairs) containing the snht statistic is returned.

7. If returnStat=FALSE, a candidate matrix is created with dimension (time x number of locations) containing the number of change points at each location and time.

8. This candidate matrix is then given to the function *unconfoundCandidateMatrix()* which looks for the largest counts in the candidate matrix and in that way assigns the change points to a location and time. It also returns the magnitude of the change point.

9. In the end a data.frame is returned containing the corrected data and the location, time of all the breaks and their magnitude. These can be accessed through *output$breaks*, *output$data*.

2. Usage of the PairwiseSNHT

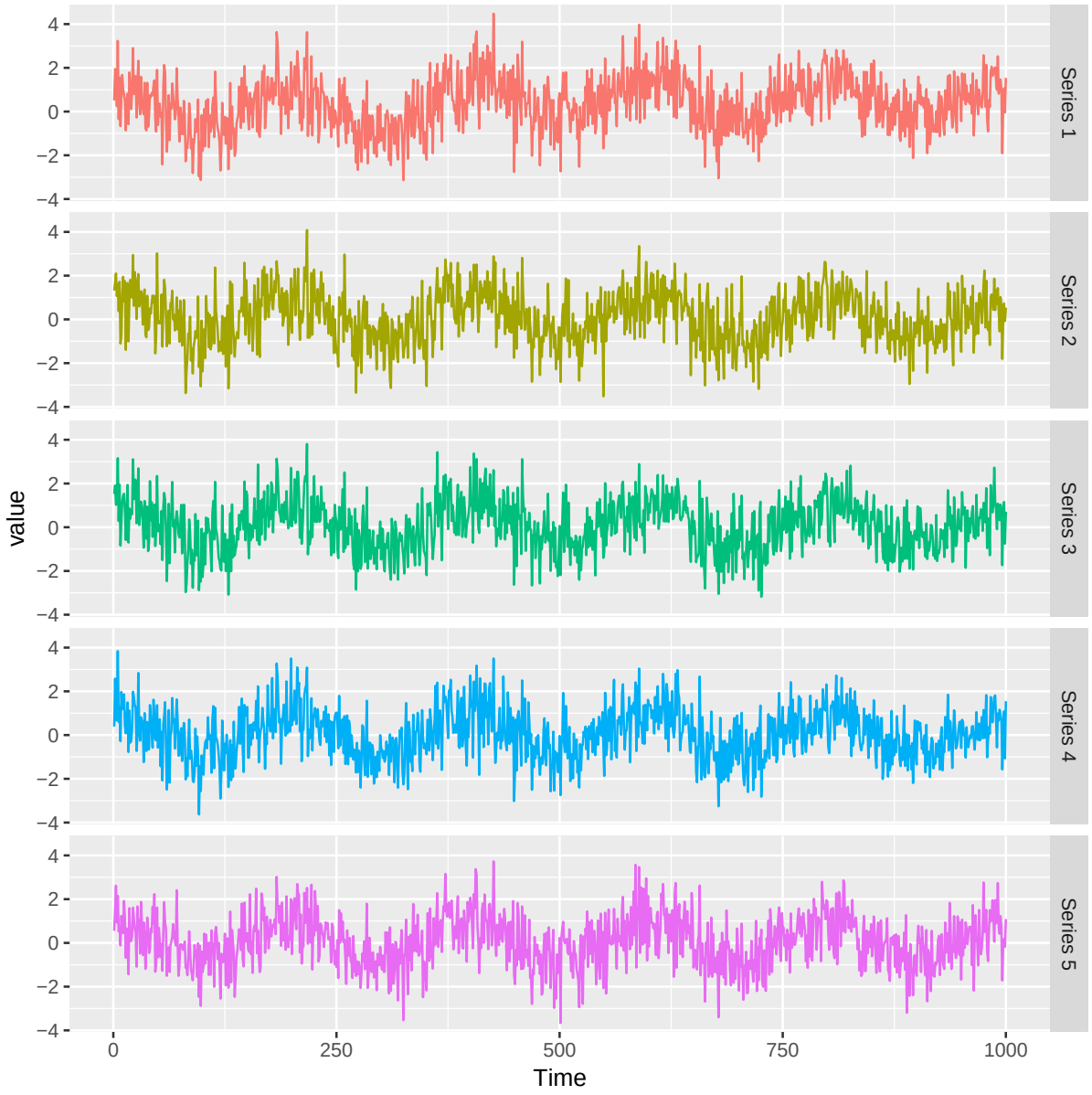
This section is intended to show how the function pairwiseSNHT is used in R.

2.1. Example 1: Seasonal data, no linear trend, shift at center location

Let us consider 5 stations where station 1 is surrounded by station 2,...,5, all having the distance 1 to station 1. The data these stations measured is assumed to be periodic, see below. First, though, let's load the required libraries:

```
library(snht)
library(reshape2)
library(ggplot2)
library(mvtnorm)
```

```
set.seed(2)
Cor<-rbind(c(0.5,0.8,0.8,0.8,0.8),c(0,0.5,0.8,0.5,0.6),
c(0,0,0.5,0.8,0.5),c(0,0,0,0.5,0.6),c(0,0,0,0,0.5))
Cor<-t(Cor)+Cor
baseData<-rmvnorm(mean=rep(0,5),sig=Cor,n=1000)+cos(1:1000*2*pi/200)
baseData[401:1000,1]<-baseData[401:1000,1]+0.5
```




```
dist<-matrix(0,5,5)
dist<-dist(rbind(c(1,0),c(1,1),c(0,0),c(1,-1),c(2,0)))
dist<-as.matrix(dist)
colnames(dist)<-rownames(dist)<-1:5
dist

##   1        2        3        4        5
## 1 0 1.000000 1.000000 1.000000 1.000000
## 2 1 0.000000 1.414214 2.000000 1.414214
## 3 1 1.414214 0.000000 1.414214 2.000000
## 4 1 2.000000 1.414214 0.000000 1.414214
## 5 1 1.414214 2.000000 1.414214 0.000000
```

This is the distance matrix which is needed as an input parameter for pairwiseSNHT. It contains as (i,j)-th entry the distance between location i and j. It is therefore always a symmetric matrix with diagonal 0.

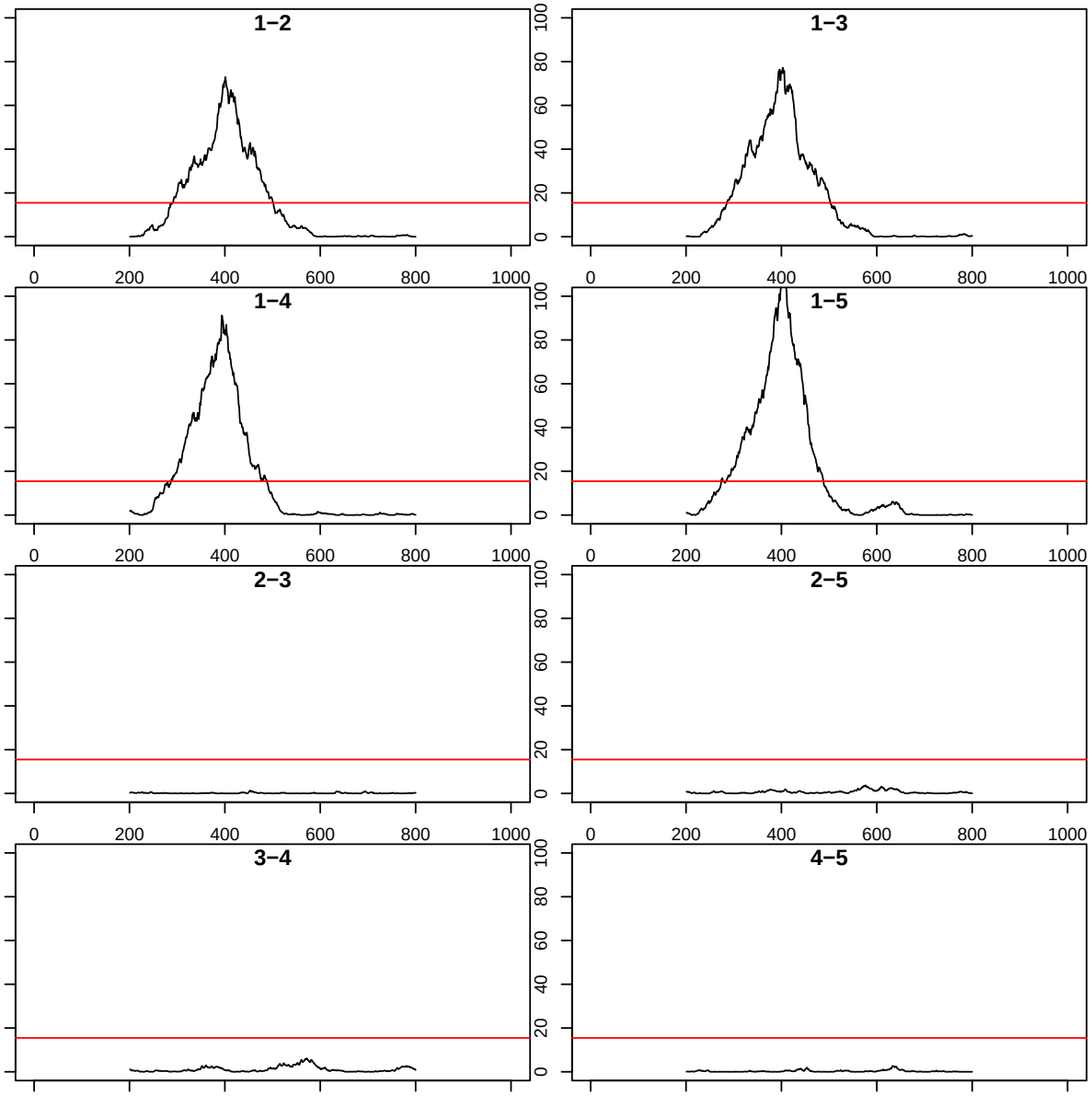
```
colnames(baseData) <-"1":"5"
baseData <- data.frame(time = 1:1000, baseData)
baseData <- melt(baseData, id.vars = "time", variable.name = "location",
                 value.name = "data")
baseData$location<-gsub("X","",baseData$location)

out1 <- pairwiseSNHT(baseData, dist, k=3, period=200,
crit=qchisq(1-0.05/600,df=1), returnStat=T)

pairs <- colnames(out1)
pairs

## [1] "1-2" "1-3" "1-4" "1-5" "2-3" "2-5" "3-4" "4-5"

out1[300:302, ]

##           1-2      1-3      1-4      1-5        2-3         2-5        3-4
## [1,] 20.25952 21.63336 20.78306 21.53806 0.02320501 0.009053666 0.10284340
## [2,] 21.07620 22.44238 21.44807 22.08905 0.02901951 0.007040604 0.11129682
## [3,] 23.07201 24.42300 22.09088 21.78422 0.02246056 0.006571347 0.04226636
##              4-5
## [1,] 0.001779711
## [2,] 0.002606399
## [3,] 0.018273127
```

pairs are the unique pairs which are formed by using $k = 3$ neighbors.

Remark: If there are more than k neighbors with the same distance to a certain location, then more than k neighbors are considered. In this example, location 2,..,5 all have the distance 1 to location 1 (see distance matrix), therefore all these 4 locations are paired with location 1. This means that all 4 difference series are used by pairwiseSNHT. Moreover, station 1's 3 closest neighbours may be stations 2, 3, and 4 but the pair station1-station5 may appear. In this case, station 5's 3 closest neighbours must include station 1. In that case, this pair would only be used for detecting changepoints in station 5 and not in station 1.

out1 is the matrix with the calculated SNHT statistic for each time point and each pair. Since period was chosen to be 200, the first and last 200 values in each time series of *out1* will be NA as there is not assigned any SNHT statistic. This is due to the definition of the SNHT statistic (see: Haimberger (2007)).

Let us plot the obtained SNHT statistics for each pair.

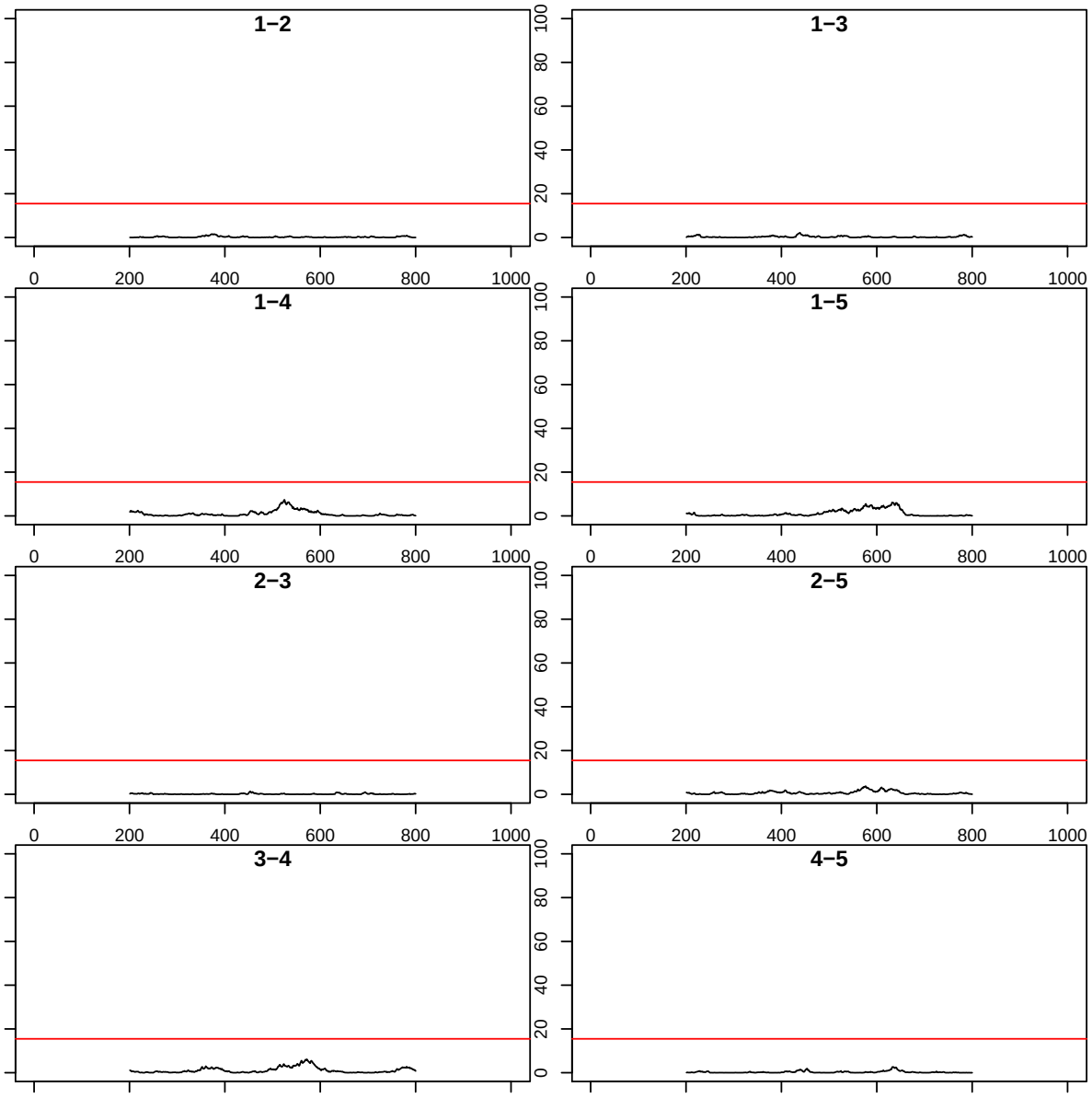


One can now clearly see that it is most likely that a change point occurred at time 400. This change point as well as its magnitude is also obtained through setting returnStat=FALSE in pairwiseSNHT. Furthermore, the corrected data can be accessed through *out2$breaks* respectively *out2$data*.

```
out2 <- pairwiseSNHT(baseData, dist, k=3, period=200,
crit=qchisq(1-0.05/600,df=1), returnStat=F)
out2$breaks

##   time location     shift
## 1  402        1 0.5857773

str(out2$data)

## 'data.frame': 5000 obs. of  3 variables:
##  $ data    : num  0.507 1.945 1.816 0.249 3.232 ...
##  $ location: Factor w/ 5 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ time    : int  1 2 3 4 5 6 7 8 9 10 ...
```

One can now check that the corrected data, which is stored in *out2$data* provides better results, i.e., the SNHT statistic does no longer exceed the critical value.

```
newPair2 <- out2$data
outNew1 <- pairwiseSNHT(newPair2,dist,k=3,period=200,
                        crit=qchisq(1-0.05/600,df=1),returnStat=T)
```



References

Haimberger L (2007). "Homogenization of radiosonde temperature time series using innovation statistics." *Journal of Climate*, **20**(7), 1377–1403.

Menne, Williams (2009). "Homogenization of Temperature Series via Pairwise Comparisons." *Journal of Climate*, **22**(7), 1700–1717.

Affiliation:

Carina Schneider
Master's Student
University of Zurich, Switzerland
E-mail: carinamariaschneider@gmail.com