

# Introduction to change point detection: lab worksheet

Euan McGonigle

2024-09-12

## Question 1: Creating a test statistic to estimate change points

- (a) Without using pre-existing libraries, write a function to calculate the CUSUM statistic vector  $\{\mathcal{T}_{0,k,n}\}_{k=1}^{n-1}$  for a change point in the mean.
- (b) Without using pre-existing libraries, write a function to calculate the MOSUM statistic vector  $\{\mathcal{T}_G(k)\}_{k=G}^{n-G}$  for a change point in the mean, for a given bandwidth  $G$ .
- (c) To set a threshold for declaring changes, we need to know the noise level  $\sigma$ . In reality, this is unknown; how could we estimate it?

*Hint:* the sample standard deviation would be positively biased by the changes. Medians are more robust than means: how can we transform the data to remove most of the effect of the change points, before using a median-like analogue of the standard deviation? If in doubt, Google is your friend.

- (d) Implement your chosen method of estimating  $\sigma$ , and add the functionality to your CUSUM and MOSUM functions so that the test statistic calculations in parts (a) and (b) are scaled by your estimate  $\hat{\sigma}$ .
- (e) (Harder) To find multiple change points with the MOSUM approach, we can calculate all  $\hat{\theta}$  that satisfy:

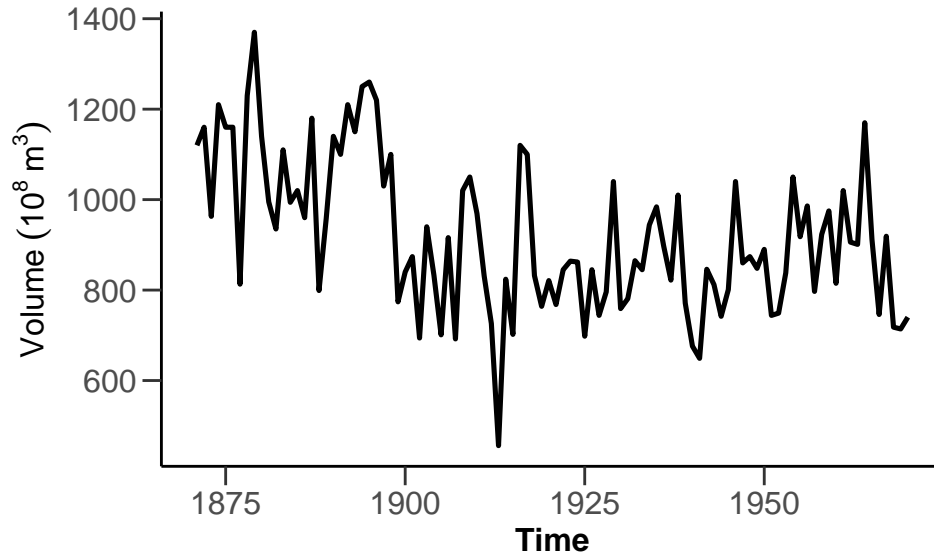
$$\mathcal{T}_\ell(G, \hat{\theta}) > D \quad \text{and} \quad \hat{\theta} = \operatorname{argmax}_{k: |k - \hat{\theta}| \leq \eta G} \mathcal{T}_G(k).$$

for some window fraction  $\eta \in (0, 1)$  and a threshold  $D$ . That is,  $\hat{\theta}$  is declared a change point if it is a local maximiser of  $\mathcal{T}_G(k)$  over a sufficiently large interval of size  $\eta G$ , at which the threshold  $D$  is exceeded.

Extend your MOSUM function to return as output the change point estimators satisfying this criterion.

## Question 2: Nile annual river flow

In this question you will use the code you've written so far to analyse data collected on the river Nile. Data in the file `nile_volume.txt` records measurements of the annual volume (in units  $10^8 m^3$ ) of discharge from the Nile River at Aswan for the years 1871 to 1970. The measurements are of meteorological importance as evidence of a possible abrupt change in the rainfall levels around the turn of the 20th century.



- Load the data into R from the `nile_volume.txt` file, and plot it. By eye, where does it look like there could be a change in mean?
- Calculate the CUSUM statistic for the Nile volume data. Using the threshold  $D = \sqrt{2 \log(n)}$ , perform a test to decide if there is a change point. If there is one, where is it?
- Calculate the MOSUM statistic for the Nile volume data using a bandwidth  $G = 25$ . Using the threshold  $D = 3.4$ , compute the change point estimators as in q1 (e). Does this agree with your answer from part (b)?
- Now use the MOSUM method with bandwidth  $G = 40$ . How does your answer differ from part (c)?
- The estimated mean signal  $\hat{f}_t$  can be calculated using sample means of the segments defined by the estimated change points. For the CUSUM method, add the estimated  $\hat{f}_t$  to your plot from part (a).

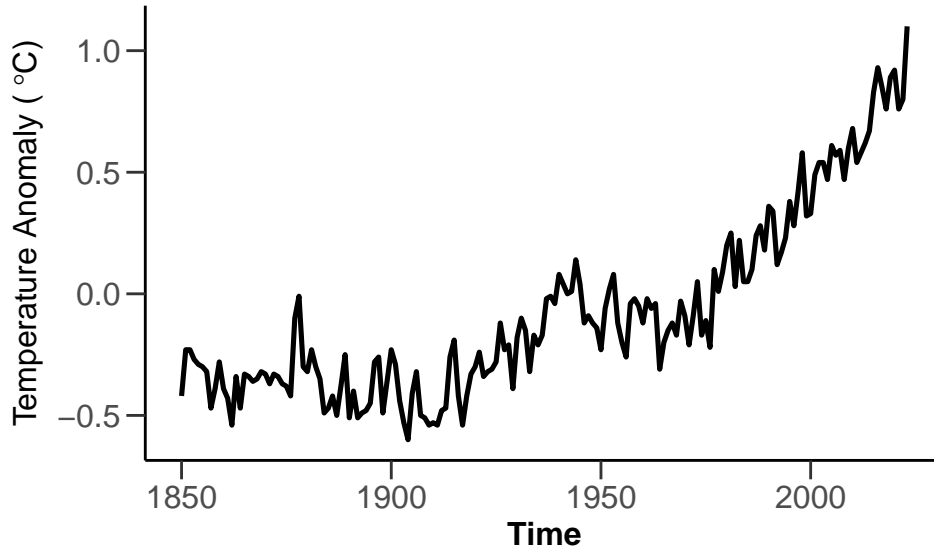
### Question 3: Using the `mosum` and `changepoint` packages

- Load the `changepoint` and `mosum` R packages into your R workspace. Using the help files, get acquainted with the `cpt.mean` and `mosum` functions.
- Using the default settings, use the `cpt.mean` function on the Nile data set. Do the results agree with your answer from question 1(a)?
- By altering the input parameters, apply the PELT algorithm using the `cpt.mean` function on the Nile data set. Is the answer the same as part (b)?
- By investigating the output returned by the `cpt.mean` function, what was the value of the penalty function used?
- Using the bandwidth  $G = 25$ , use the `mosum` function on the Nile data set. Do the results agree with your answer from question 1(b)?
- By investigating the output returned by the `mosum` function, what are the p-value(s) of the detected change points?

*Python* users: you can use the `ruptures` function `Pelt` and the `mosum` function `mosum` functions for this question. See <https://centre-borelli.github.io/ruptures-docs/> and <https://pypi.org/project/mosum/> for help.

#### Question 4: Global yearly mean sea temperature anomalies

Global mean surface temperature series help enable the monitoring of global warming. The warming can be quantified by, for example, a change from a base period used as reference. The Hadley Centre/Climatic Research Unit (HadCRUT) surface temperature data set provides the annual global temperature anomalies from 1850-2023, where anomalies are calculated relative to the 1961–1990 period. The data are shown below, which can be seen to exhibit a gradual upward trend.



- (a) Load the data into R from the `temp_anomalies.txt` file, and plot it. Using e.g. the `lm` function, fit two linear trends to the data: one from 1850 to 1963, and one from 1964 to 2023, and add these lines to the plot, as well as a vertical line through year 1963.
- (b) Using your own functions, and functions from the `changept` and `mosum` packages, fit 4 models to the data set;
  - 1. a constant mean model with no change points,
  - 2. a mean change point model,
  - 3. a linear trend model with no change points.
  - 4. a linear trend change point model.

Looking at the data, are there any other models that might be appropriate?

*Hint:* if you are struggling, the `EnvCpt` R package has everything you need.

*Python users:* you can use the `Pelt` function and change the `model` parameter.

- (c) Pick the “best” change point model from your candidates computed in part (b), and justify your choice. Add the estimated mean/trend function onto a plot of the data.

*Hint:* the Akaike information criterion (AIC) is given by  $AIC = 2k - 2\log(\hat{L})$ , where  $k$  is the number of parameters of the model, and  $\hat{L}$  is the maximised value of the likelihood function for the model.

#### Question 5: Multivariate data

- (a) Suppose we want to find mean change points in a time series  $\{X_t\}_{t=1}^n$ , where  $X_t = (X_{1t}, \dots, X_{pt})^\top$  is a  $p$ -dimensional vector. One approach is to calculate a test statistic for each variable, and then combine

the results across variables to give a single test statistic. For example, for the  $i$ -th variable of  $\{X_t\}_{t=1}^n$ , denoted  $\{X_{it}\}_{t=1}^n$ , the MOSUM test statistic is given by

$$\mathcal{T}_G(k, i) = \frac{1}{\sqrt{2G}} \left( \sum_{t=k+1}^{k+G} X_{it} - \sum_{t=k-G+1}^k X_{it} \right).$$

Then, the MOSUM statistic  $\mathcal{T}_G(k)$  for a change point in  $\{X_t\}_{t=1}^n$  can be computed by aggregating the  $\{\mathcal{T}_G(k, i)\}_{i=1}^p$  using some aggregating function  $f$ , so that

$$\mathcal{T}_G(k) = f(\mathcal{T}_G(k, 1), \dots, \mathcal{T}_G(k, p)).$$

What possibilities could we use for the aggregating function  $f$ ?

- (b) Using your aggregating function  $f$ , write a function to compute the MOSUM or CUSUM statistic for a change in the mean vector of a multivariate time series.
- (c) Simulate a data set of length  $n = 500$  and dimension  $p = 40$  as follows. Generate  $\{X_t\}_{t=1}^{250}$  from the standard  $p$ -dimensional normal distribution, and generate  $\{X_t\}_{t=251}^{500}$  from the  $p$ -dimensional normal distribution with identity covariance matrix, and mean vector given by  $\mu = 1.4 \times (1/\sqrt{p}, \dots, 1/\sqrt{p})^\top$ . Use the method from part (b), with  $G = 80$ , to compute the test statistic for a change in mean. Is the change point easy to see?
- (d) Re-do part (c), but generate  $\{X_t\}_{t=251}^{500}$  from the  $p = 200$ -dimensional normal distribution with identity covariance matrix, and mean vector given  $\mu$  with first 2 entries equal to 0.8, and last  $p - 2$  entries equal to 0. Is the change point easy to see?