Research methods for data-driven analysis of non-equilibrium systems

Davide Pigoli - King's College London

28 February 2018

Introduction

Overall plan:

- 2 introductory lectures (today and tomorrow)
- 3 group tutorials (TBC – approx. middle of June)
- Keats page
- Project report – by beginning of September
- Oral presentation

Introduction

Project topic:

Changepoints detection and segmentation for electroencephalograms (EEG) data

Davide Pigoli, davide.pigoli@kcl.ac.uk

Summary

1. A short intro to EEG data

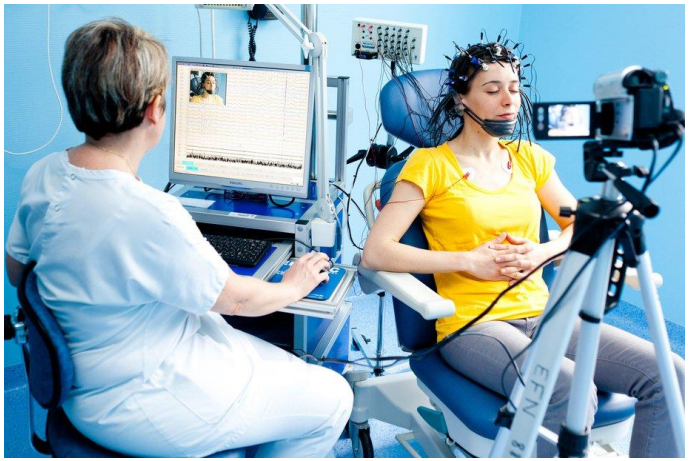
2. The changepoint(s) detection problem

3. Three approaches:
   - Segment statistics
   - Bayes
   - Hidden Markov Models

4. Sources for data and software

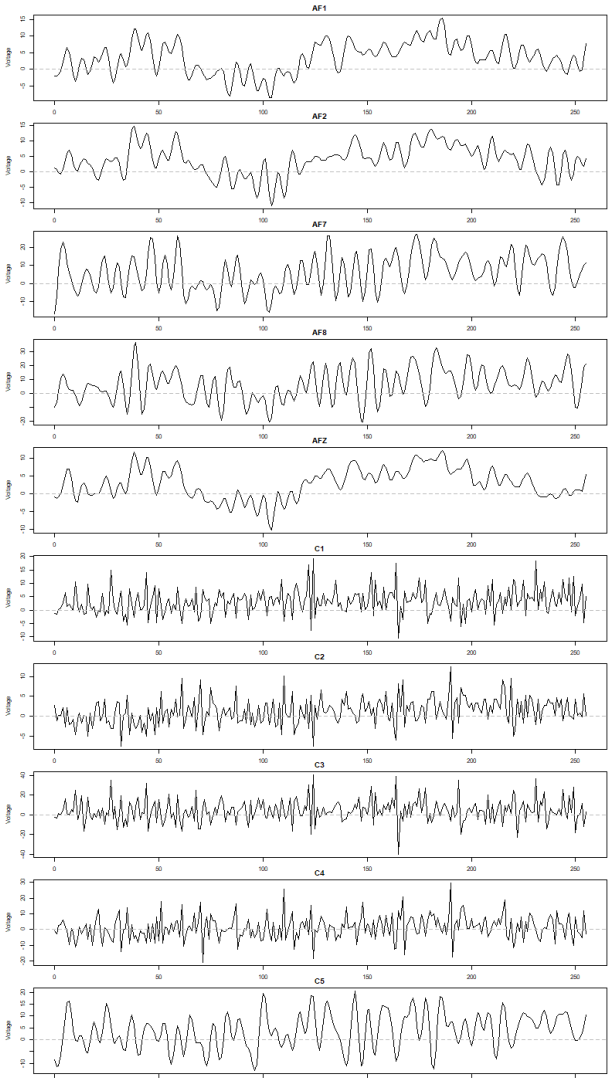
5. Project report requirements

EEG Data

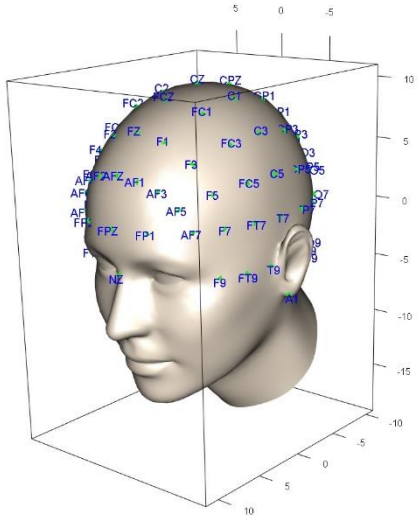


[Image taken from NHS website]

Subsample of EEG channels



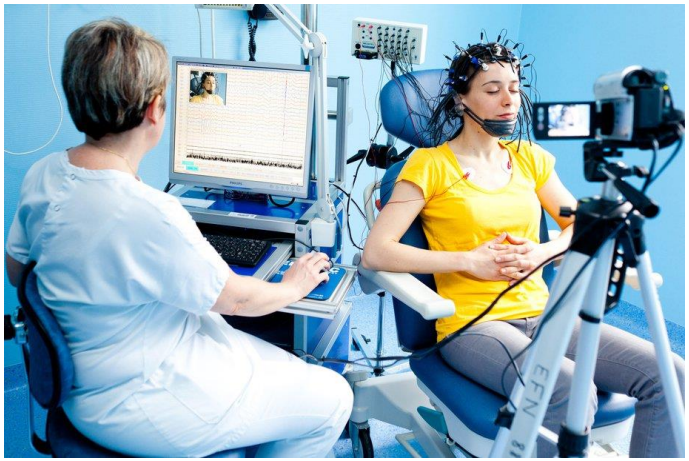
Electrodes positions



Data from
R package eegkit

Davide Pigoli, davide.pigoli@kcl.ac.uk

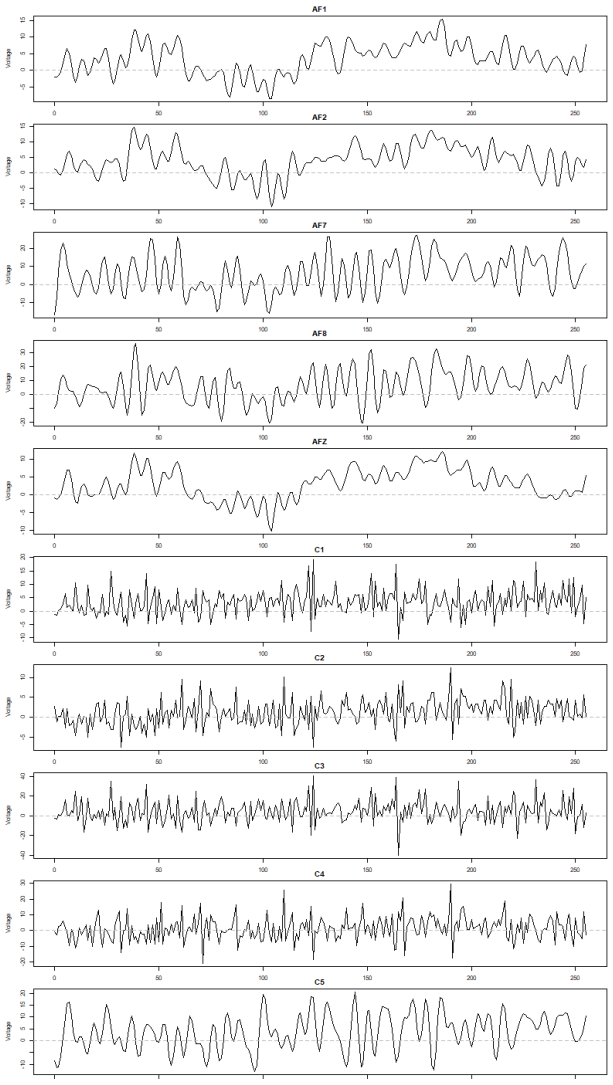
EEG Data



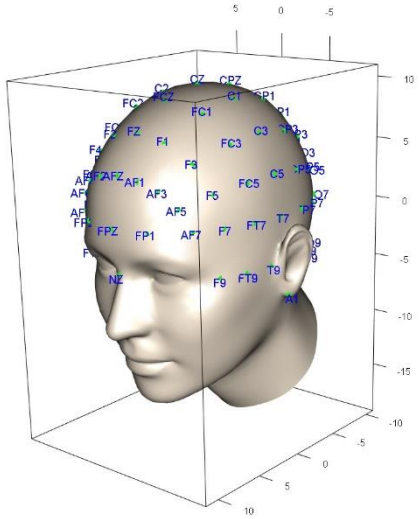
[Image taken from NHS website]

- The channels are correlated
- There are changes in time both in the mean and in the covariance.

Subsample of EEG channels



Electrodes positions

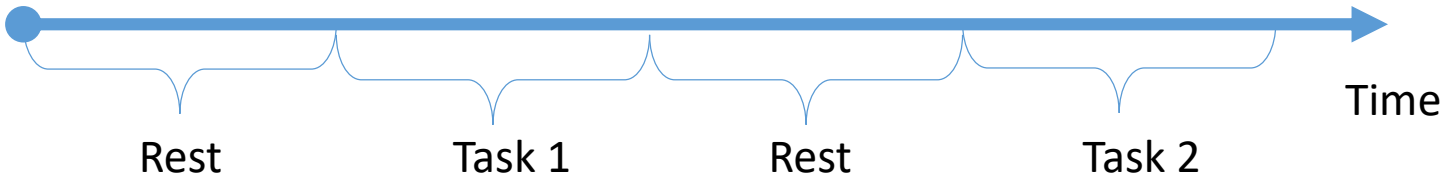


Data from
R package eegkit

Davide Pigoli, davide.pigoli@kcl.ac.uk

EEG Data

- Applications in neuroscience, to investigate how the brain reacts to external stimuli (music, tasks to perform..)



- Medical applications:
  - Monitor alertness and coma.
  - Investigate epilepsy and locate seizure events.
  - Check drugs effect.
  - Investigate sleep disorders.
  - …

Summary

1. A short intro to EEG data

2. The changepoint(s) detection problem

3. Three approaches:
   - Segment statistics
   - Bayes
   - Hidden Markov Models

4. Sources for data and software

5. Project report requirements

Time series

Time series are sets of observations $\{Y_t, t \in T\}$ collected/observed over time.

A complete description of the time series is given by the joint distribution of the observed variables but that is usually not available, since we observe only one replicate from their distribution.

Thus, modelling assumptions are used to simplify the problem, usually on:

- The mean $\mu_t = E[Y_t]$
- The autocovariance function:

$$\gamma_Y(s,t) = cov(Y_s, Y_t) = E[(Y_s - E[Y_s])(Y_t - E[Y_t])]$$

- (for multivariate data) the (cross) covariance matrix.

Stationary time series

(Strictly or strong)

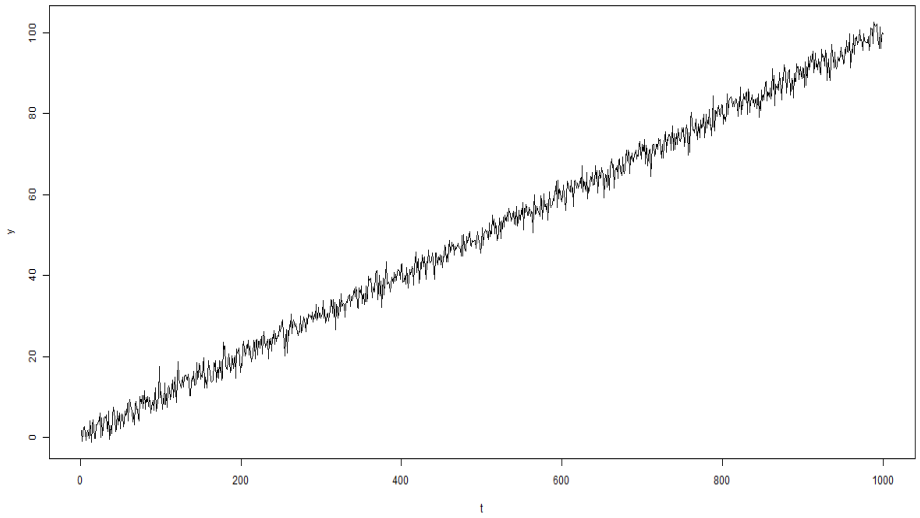
A time series is called **strictly stationary** if the joint distribution of every collection $(Y_{t'_1}, \ldots, Y_{t'_k})$ is equal to the joint distribution of the time shifted set $(Y_{t'_1+h}, \ldots, Y_{t'_k+h})$, for any $h \in \mathbb{Z}$ such that $t'_1 + h \geq t_1$ and $t'_k + h \leq t_n$.

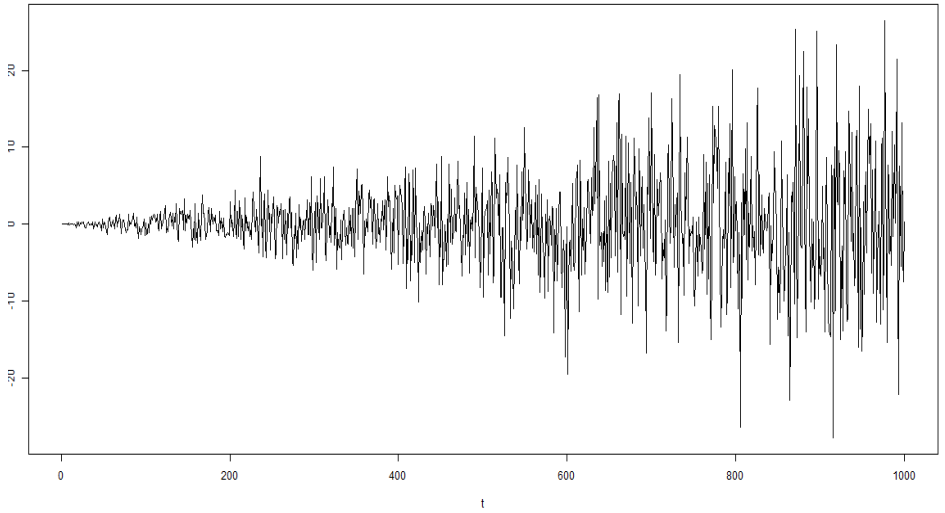
(Weak)

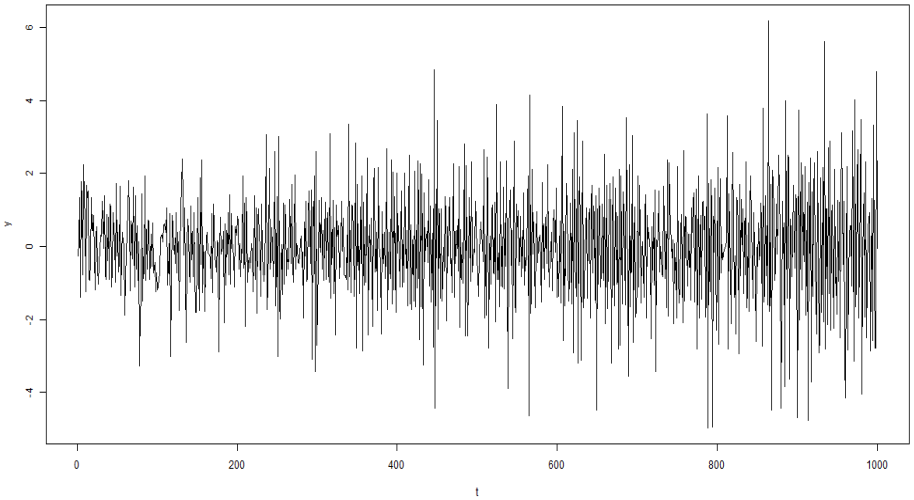
A time series is called **weakly stationary** (or simply **stationary**) if Y_t has finite variance for all t and

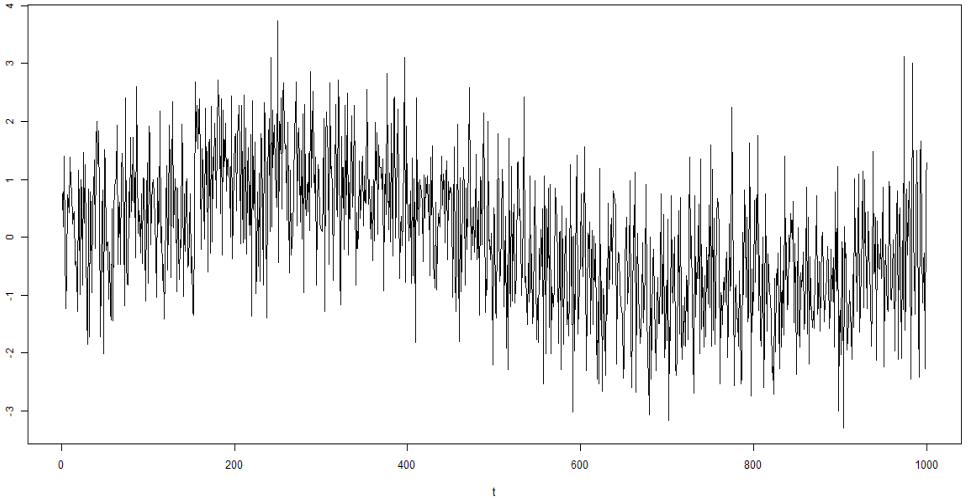
1. the mean $E[Y_t]$ is constant for all t and
2. the autocovariance function $\gamma_Y(s,t)$ depends on s and t only through their difference $h = |s - t|$.

Non stationary time series





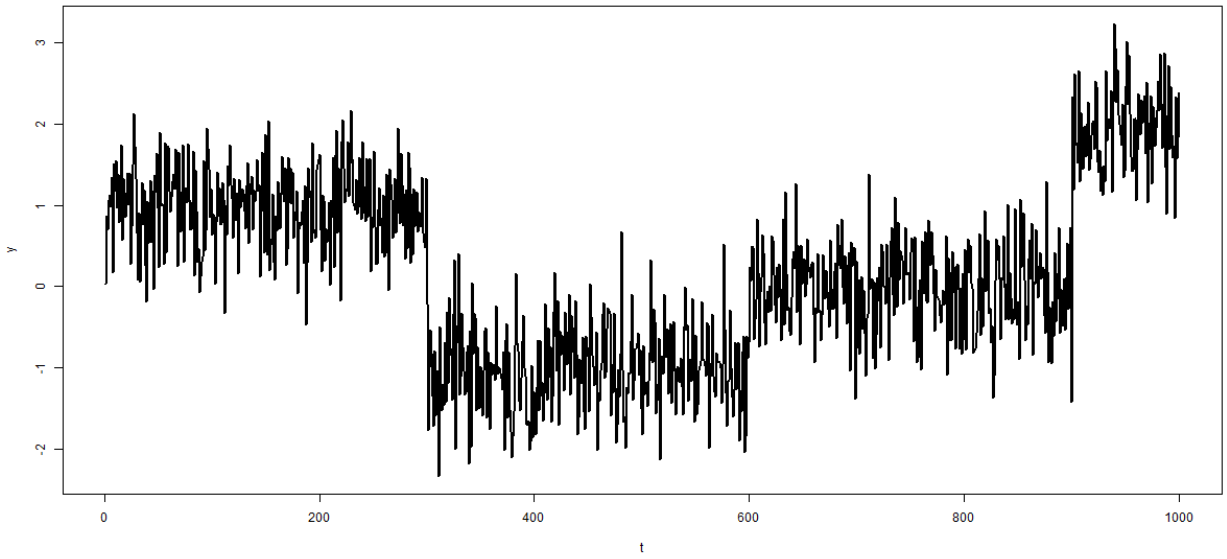




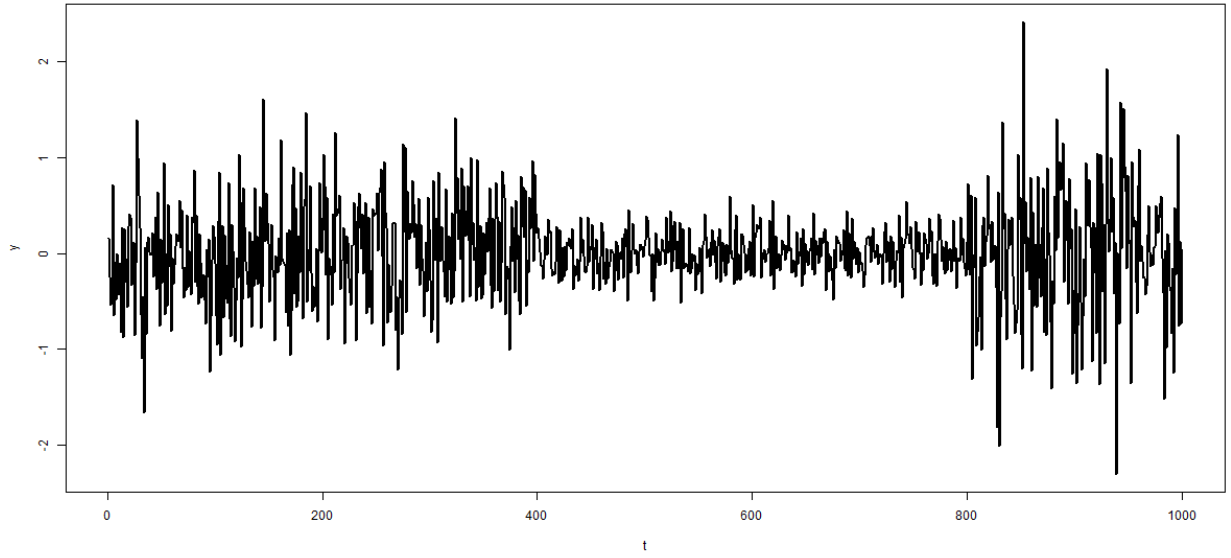
Davide Pigoli, davide.pigoli@kcl.ac.uk

KING'S College LONDON

Changepoints detection (segmentation)



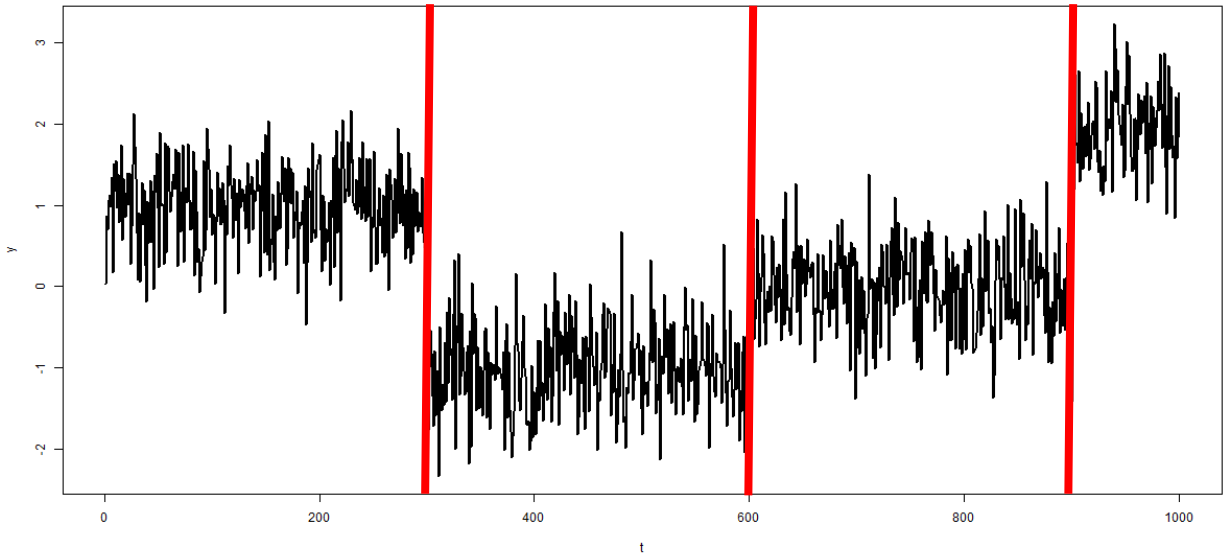
Change in the mean



Change in the variance

Davide Pigoli, davide.pigoli@kcl.ac.uk

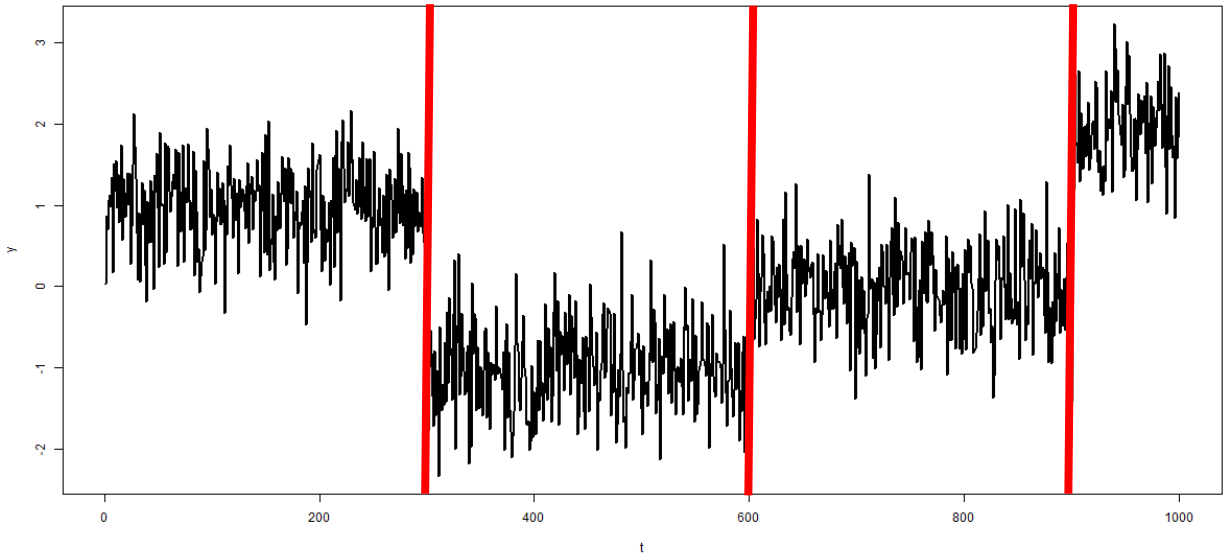
Changepoints detection (segmentation)



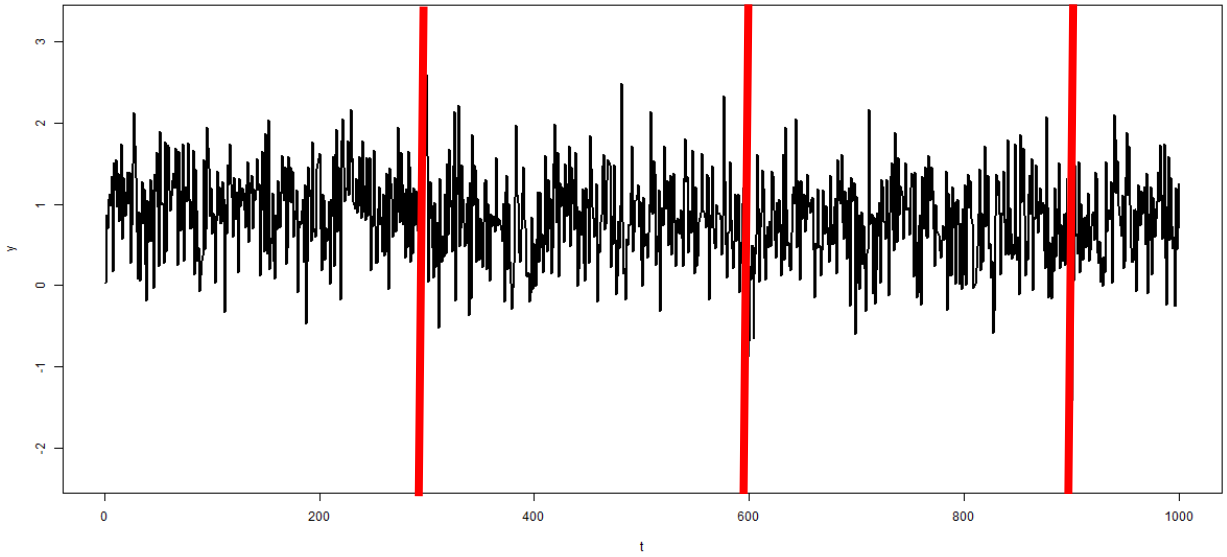
Change in the mean

Davide Pigoli, davide.pigoli@kcl.ac.uk

Changepoints detection (segmentation)



Change in the mean



Adjusted time series

Davide Pigoli, davide.pigoli@kcl.ac.uk

Summary

1. A short intro to EEG data

2. The changepoint(s) detection problem

3. Three approaches:
   - Segment statistics
   - Bayes
   - Hidden Markov Models

4. Sources for data and software

5. Project report requirements

Single changepoint in the mean

Model:

$$\mathbf{Y}_t = \boldsymbol{\mu} + \boldsymbol{\epsilon}_t, \text{ for } t = 1, \ldots, r$$

$$\mathbf{Y}_t = \boldsymbol{\mu}' + \boldsymbol{\epsilon}_t, \text{ for } t = r+1, \ldots, N$$

- $\boldsymbol{\epsilon}_t$ are zero mean errors
- Sometimes assumed independent, sometimes weakly stationary.
- In the following, we are considering the errors as independent (usually not true)
- The mean can in general be modelled as a (known) function of time in each segment

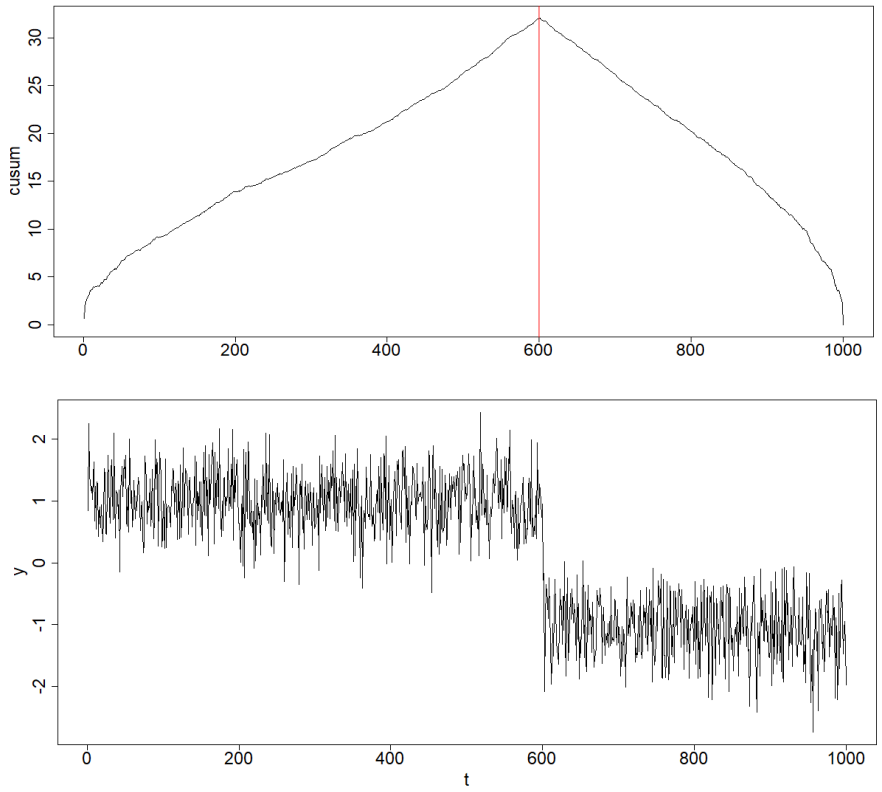
CUSUM statistics

[Fryzlewicz, P. (2014)]

$$\mathrm{CUSUM}_t = \sqrt{\frac{N-t}{Nt}} \sum_{i=1}^{t} Y_i - \sqrt{\frac{t}{N(N-t)}} \sum_{i=t+1}^{N} Y_i$$

CUSUM statistics

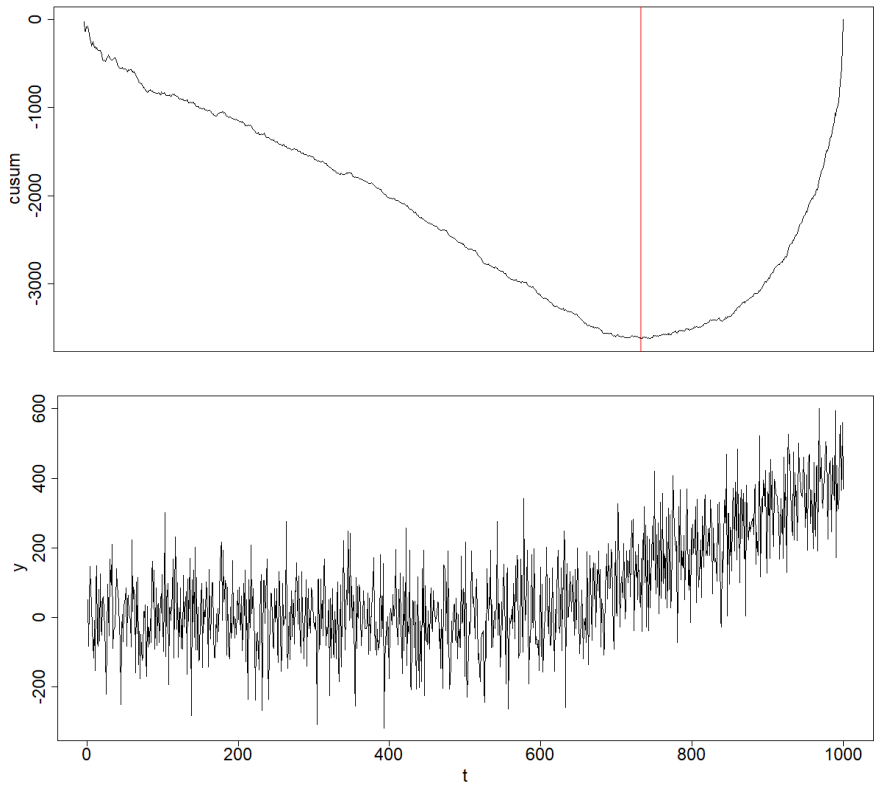
$$\text{CUSUM}_t = \sqrt{\frac{N-t}{Nt}} \sum_{i=1}^{t} Y_i - \sqrt{\frac{t}{N(N-t)}} \sum_{i=t+1}^{N} Y_i$$



Davide Pigoli, davide.pigoli@kcl.ac.uk

CUSUM statistics

$$\text{CUSUM}_t = \sqrt{\frac{N-t}{Nt}} \sum_{i=1}^{t} Y_i - \sqrt{\frac{t}{N(N-t)}} \sum_{i=t+1}^{N} Y_i$$



Davide Pigoli, davide.pigoli@kcl.ac.uk

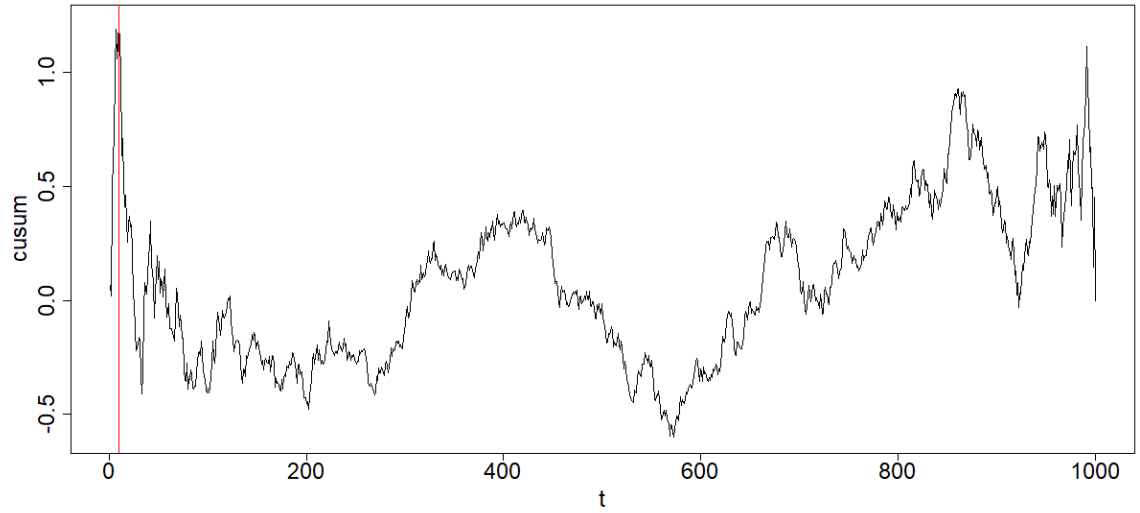
Changepoint estimation

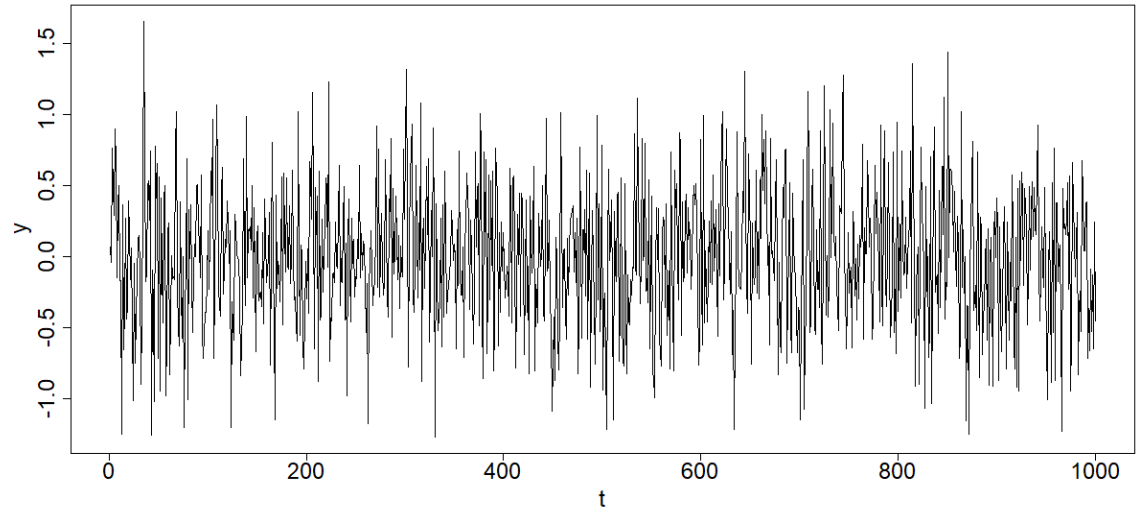
$$\text{CUSUM}_t = \sqrt{\frac{N-t}{Nt}} \sum_{i=1}^{t} Y_i - \sqrt{\frac{t}{N(N-t)}} \sum_{i=t+1}^{N} Y_i$$

$$\widehat{k} = \arg \max_{t=1,\ldots,N} |\text{CUSUM}_t|$$

Davide Pigoli, davide.pigoli@kcl.ac.uk

Is that truly a changepoint?





Davide Pigoli, davide.pigoli@kcl.ac.uk

Is that truly a changepoint?

- Threshold based on Gaussian (or asymptotic) approximation:

$$Y_t \sim N(\mu, \sigma^2) \Rightarrow \text{CUSUM}_t \sim N(0, \sigma^2)$$

$$Z_i \sim N(0,1), i = 1, \ldots, N, P[\max_i |Z_i| < \sqrt{2\log(N)}] \to 0 \text{ for } N \to +\infty$$

There is a changepoint if $|CUSUM_{\widehat{k}}| > \widehat{\sigma}\sqrt{2\log(N)}$

- Non-parametric methods (permutations [Antoch and Hušková, 2001], bootstrap etc.)

Likelihood ratio statistic approach

If we are ready to assume a parametric model for the observation (for example Gaussian), and we look for the changepoint k such as:

$$Y_t \sim f_0, \text{ for } t = 1, \ldots, k$$

$$Y_t \sim f_1 \text{ for } t = k+1, \ldots, N$$

The likelihood function for this model is:

$$L = \prod_{i=1}^{k} f_0(Y_i) \prod_{i=k+1}^{N} f_1(Y_i)$$

Davide Pigoli, davide.pigoli@kcl.ac.uk

Likelihood ratio statistic approach

Then we can test the hypothesis that there is a changepoint in k using the likelihood-ratio test statistics

$$LR = \frac{\prod_{i=1}^{k} f_0(Y_i) \prod_{i=k+1}^{N} f_1(Y_i)}{\prod_{i=1}^{N} f_0(Y_i)}$$

which is usually easier to deal with in the logarithmic scale:

$$\log LR = \sum_{i=1}^{k} \log \frac{f_0(Y_i)}{f_0(Y_i)} + \sum_{i=k+1}^{N} \log \frac{f_1(Y_i)}{f_0(Y_i)} = \sum_{i=k+1}^{N} \log \frac{f_1(Y_i)}{f_0(Y_i)}$$

Likelihood ratio statistic approach

Under the null hypothesis of no changepoint and with mild regularity conditions on the parametric distribution, we have that asymptotically for N that goes to infinity

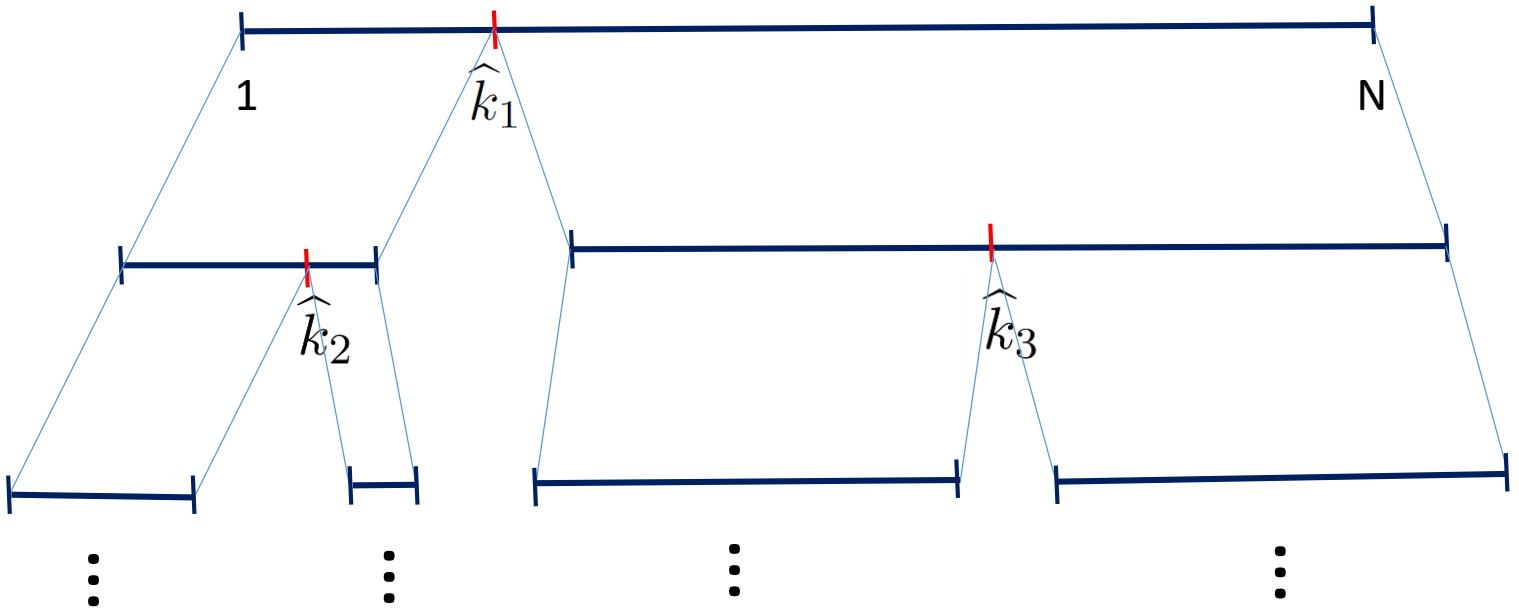
$$-2 \log LR \sim \chi^2_p$$

where p is the difference in numbers of parameters between the model with and without changepoint. This can be used to set a threshold to determine if there is a changepoint in k

The location of the changepoint can be estimated by maximizing the log LR statistics. The above approximation is not correct anymore, ad hoc results need to be used or the distribution of the max of log LR can be derived via simulations.

Multiple changepoints search

Binary segmentation:



Stop when no changepoints are found (i.e. candidates do not overcome the threshold)

Multiple changepoints search

Alternatives:

- Wild binary segmentation: Looking for the changepoint in random intervals and pick the one(s) with higher max |CUSUM|. [Fryzlewicz, P. (2014)]

- Direct optimization of the cost function for multiple changepoints (usually time consuming).

More general cost functions

A different approach consist in minimising a cost function applied to the segments to find the changepoint(s).

$$(k_1, \ldots, k_m) = \arg\min \sum_{i=1}^{m+1} \mathcal{C}([Y_{k_i}; Y_{k_{i+1}}])$$

with $k_0 = 1$, $k_{m+1} = N$

Examples of cost functions include:

- negative log-likelihood [Chen and Gupta (2000)]

- divergence measures [R package ecp; Matteson and James (2014).]

More general cost functions

This set up also allows us to search directly for multiple changepoints by minimising a penalised cost function:

$$\min \sum_{i=1}^{m+1} \mathcal{C}([Y_{k_i}; Y_{k_{i+1}}]) + \beta(p, N)m$$

Efficient algorithms have been developed to solve this problem (PELT – Killick et al, 2012).

Multivariate case

CUSUM statistics for a change in the mean [Horváth et al. (1999)]:

$$\Delta_t = \sqrt{\frac{N-t}{Nt}} \sum_{i=1}^{t} \mathbf{Y}_i - \sqrt{\frac{t}{N(N-t)}} \sum_{i=t+1}^{N} \mathbf{Y}_i$$

$$\widehat{k} = \arg \max_{t=1,\ldots,N} \Delta_t \widehat{\Sigma}^{-1} \Delta_t^T$$

The threshold for selecting changepoints needs to be adjusted.

Change in covariance

A similar idea can be applied to look for changepoints in the covariance matrix, by defining a modified CUSUM statistics [Galeano and Peña (2007)]:

$$A_k = \sum_{i=1}^{k} (\mathbf{Y}_i - \overline{\mathbf{Y}}_k) \widehat{\Sigma}^{-1} (\mathbf{Y}_i - \overline{\mathbf{Y}}_k)^T$$

$$C_t = \frac{t}{N} \left(\frac{A_t}{t} - \frac{A_N}{N} \right)$$

$$\widehat{k} = \arg \max_{t=1,\ldots,N} |C_t|$$

Davide Pigoli, davide.pigoli@kcl.ac.uk

Summary

1. A short intro to EEG data

2. The changepoint(s) detection problem

3. Three approaches:
   - Segment statistics
   - Bayes
   - Hidden Markov Models

4. Sources for data and software

5. Project report requirements

A Bayesian approach (single changepoint example)

A prior distribution needs to be specified over the set of possible changepoints, i.e.

$$p_t = \text{ probability of time } t \text{ being a changepoint, } \sum_{t=1}^{N} p_t = 1$$

For a single changepoint model , the likelihood of the observed sequence given a changepoint at time t is:

$$L(Y|k=t) = \prod_{i=1}^{t} f_0(Y_i) \prod_{i=t+1}^{N} f_1(Y_i)$$

A Bayesian approach (single changepoint example)

Bayes theorem gives us the posterior probability of the changepoint being in t:

$$P(t \text{ is a changepoint } |Y) \propto L(Y|k = t)p_t$$

The estimate for the changepoint location can be found by maximizing the posterior probability.

Things become more complicated when the parameters of the two distribution need to be estimated, see Smith (1975) and Chaturvedi and Shrivastava (2016) for more details.

Summary

1. A short intro to EEG data

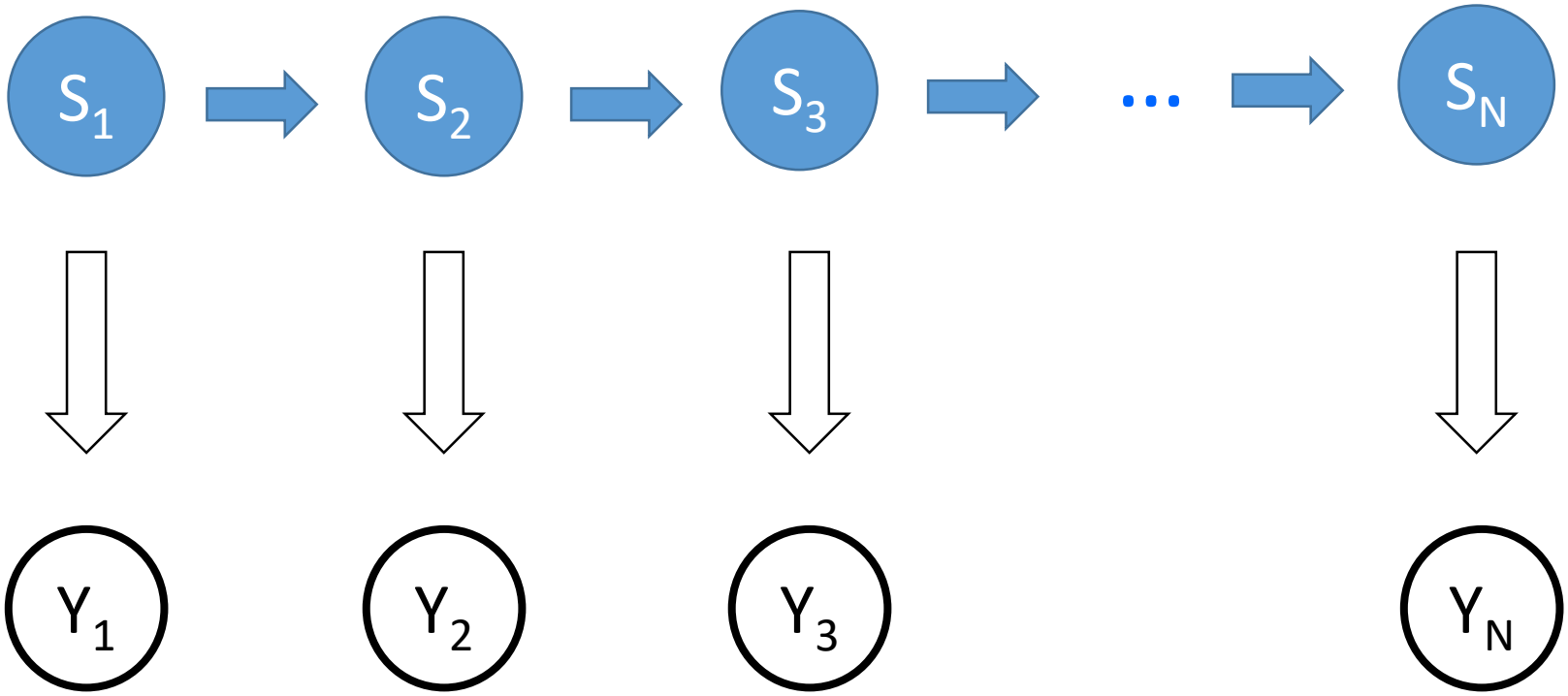
2. The changepoint(s) detection problem

3. Three approaches:
   - Segment statistics
   - Bayes
   - Hidden Markov Models

4. Sources for data and software

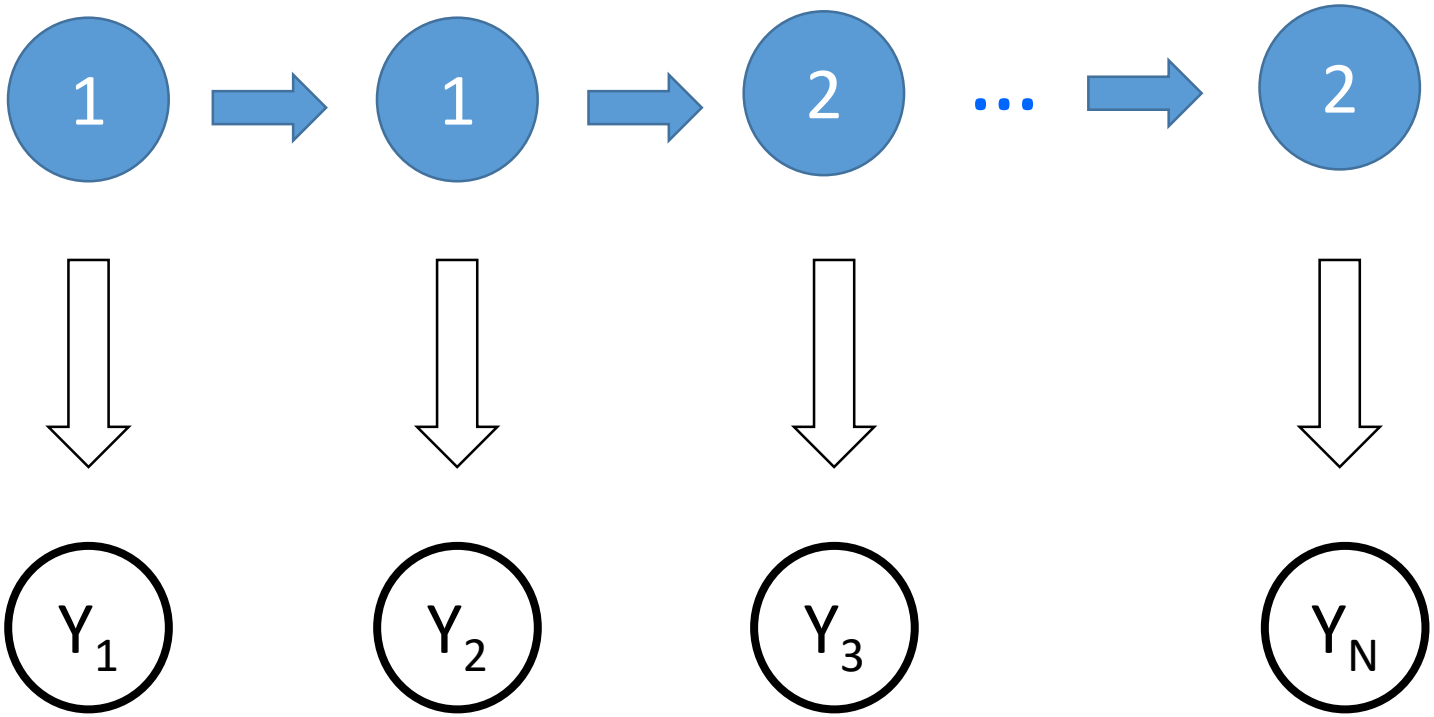
5. Project report requirements

Hidden-Markov models



The hidden states $S_i \in \{1, \ldots, K\}$ represent the segments between changepoints in the signal.

Hidden-Markov models



Davide Pigoli, davide.pigoli@kcl.ac.uk

Hidden-Markov models

Markov chain for the hidden states:

$$P[S_1 = i] = \pi_i$$

$$P[S_{t+1} = i | S_t = j] = \lambda_{ij}$$

The law of the observed variables depend on the hidden state:

$$Y_t \sim f_{S_t}$$

Example: change in the mean for a Gaussian process

$$Y_t \sim N(\mu_{S_t}, \sigma^2)$$

Hidden-Markov models

[Rabiner (1989)]

Likelihood problem: Given the parameters of the HMM, the likelihood of a specific observed sequence is computed using the **forward algorithm.**

Learning problem: The unknown parameters λ_{ij} , $\mu_1, \ldots$ are estimated using a expectation-maximization (EM) algorithm (**Baum-Welsh** or **forward-backward algorithm**).

(Decoding problem: Given the parameters of the model, the sequence of hidden states that maximize the probability of the observed sequence is obtained using the **Viterbi algorithm.)**

Forward algorithm

In principle, we could compute the probability/likelihood of the observed sequence by averaging over all the possible sequences:

$$P(Y) = \sum_{S} P(Y|S) = \sum_{S} \prod_{i=1}^{N} f_{s_i}(Y_i) P[S_i = s_i]$$

$$P[S_i = s_i] = \prod_{j=2}^{i} P[S_j = s_j | S_{j-1} = s_{j-1}] P[S_1 = s_1]$$

$$S = (s_1, \ldots, s_N)$$

In practice, it is computationally unfeasible to explore all possible sequences of states and this is where the forward algorithm is helpful.

Forward algorithm

Let $\alpha_t(k)$ be the probability of the observed sequence up to time t and of the state t to be k, i.e.,

$$\alpha_t(k) = P(Y_1, \ldots, Y_y, S_t = k)$$

Thanks to the Markov property,

$$\alpha_t(i) = \sum_{j=1}^{K} \alpha_{t-1}(j) \lambda_{ij} f_i(Y_t)$$

We can then initialise the sequence as:

$$\alpha_1(i) = \pi_i f_i(Y_1)$$

And get the probability/likelihood for the full observed sequence by averaging over the possible states in the final time:

$$P(Y) = \sum_{j=1}^{K} \alpha_N(j)$$

Baum-Welsh Algorithm

The learning problem can be solved using an expectation-maximization algorithm to maximise the likelihood function. However, this guarantees only to get a local minimum, therefore starting values of the parameters may influence the result.

Let us start by introducing the backward probability

$$\beta_t(i) = P[Y_{t+1}, \ldots, Y_N | S_t = i, \lambda, \pi]$$

and its Markov update:

$$\beta_t(i) = \sum_{j=1}^{K} \beta_{t+1}(j) \lambda_{ij} f_t(Y_{j+1})$$

We can *arbitrarily* initialise the sequence as

$$\beta_N(i) = 1 \text{ for } i = 1, \ldots, K$$

Baum-Welsh Algorithm

Let us then define the probability of being in states i and j at two contiguous time steps, given the parameters of the model,

$$\varsigma_t(i,j) = P[S_t = i, S_{t+1} = j | \lambda, \pi]$$

which can be obtained from the forward and backward probabilities as

$$\varsigma_t(i,j) = \frac{\alpha_t(i)\lambda_{ij} f_j(Y_{t+1})\beta_{t+1}(j)}{P(Y|\lambda,\pi)}$$

and the probability of being in state i at time t is

$$\gamma_t(i) = P[S_t = i | Y, \lambda, \pi] = \sum_{j=1}^{K} \varsigma_t(i,j)$$

Baum-Welsh Algorithm

Averaging in time,

$$\sum_{t=1}^{N-1} \gamma_t(i) = \text{“average number of times with state } i \text{”}$$

$$\sum_{t=1}^{N-1} \varsigma_t(i,j) = \text{“average number of transictions from state } i \text{ to state } j \text{”}$$

we can estimate the parameters of the model as

$$\widehat{\pi}_i = \gamma_1(i) \qquad \widehat{\lambda}_{ij} = \frac{\sum_{t=1}^{N-1} \varsigma_t(i,j)}{\sum_{t=1}^{N-1} \gamma_t(i)}$$

Baum-Welsh Algorithm

Finally, the algorithm:

1) Initialize λ, π

2) **E-Step:** Use forward and backward algorithm to get $\alpha_t(i)$, $\beta_t(i)$. Then,

$$\varsigma_t(i,j) = \frac{\alpha_t(i)\lambda_{ij} f_j(Y_{t+1})\beta_{t+1}(j)}{P(Y|\lambda,\pi)} \qquad \gamma_t(i) = \sum_{j=1}^{K} \varsigma_t(i,j)$$

3) **M-Step**:

$$\widehat{\pi}_i = \gamma_1(i) \qquad \widehat{\lambda}_{ij} = \frac{\sum_{t=1}^{N-1} \varsigma_t(i,j)}{\sum_{t=1}^{N-1} \gamma_t(i)}$$

(Optional) – Estimate parameters in the distributions of the observations by weighting them with corresponding states probabilities.

4) Iterate steps 2)-3) until convergence.

Baum-Welsh Algorithm

Finally, the algorithm:

1) Initialize λ, π

2) **E-Step:** Use forward and backward algorithm to get $\alpha_t(i)$, $\beta_t(i)$. Then,

$$\varsigma_t(i,j) = \frac{\alpha_t(i)\lambda_{ij} f_j(Y_{t+1})\beta_{t+1}(j)}{P(Y|\lambda,\pi)} \qquad \gamma_t(i) = \sum_{j=1}^{K} \varsigma_t(i,j)$$

3) **M-Step**:

$$\widehat{\pi}_i = \gamma_1(i) \qquad \widehat{\lambda}_{ij} = \frac{\sum_{t=1}^{N-1} \varsigma_t(i,j)}{\sum_{t=1}^{N-1} \gamma_t(i)}$$

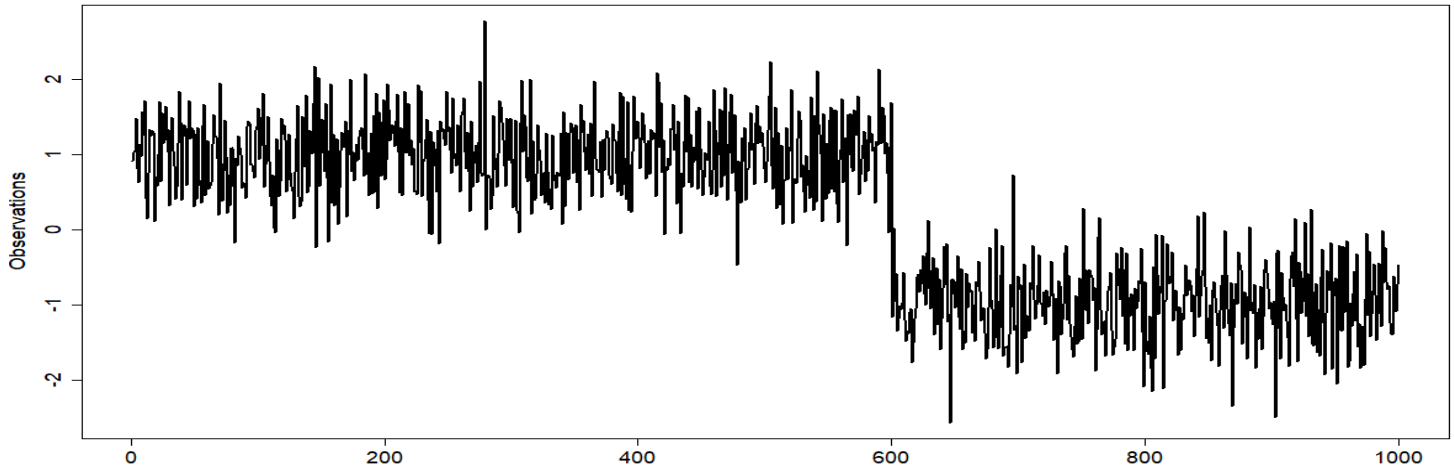
(Optional) – Estimate parameters in the distributions of the observations by weighting them with corresponding states probabilities.

4) Iterate steps 2)-3) until convergence.

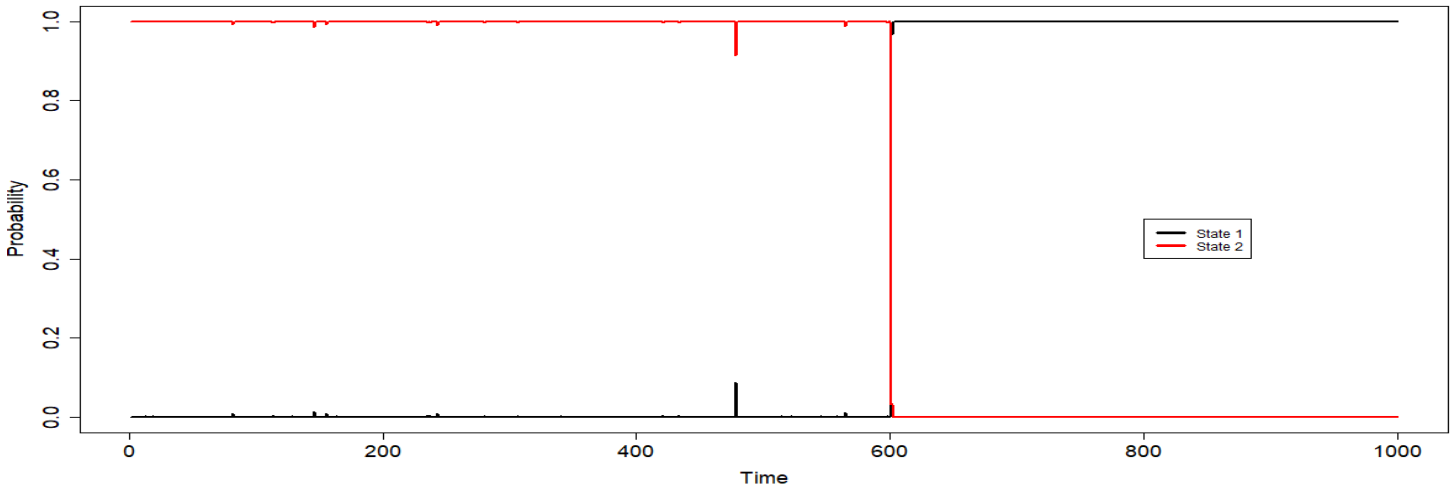
Baum-Welsh Algorithm

Example with Gaussian observations (model fitted with R package depmixS4):

$$Y_t \sim N(\mu_{S_t}, \sigma^2)$$

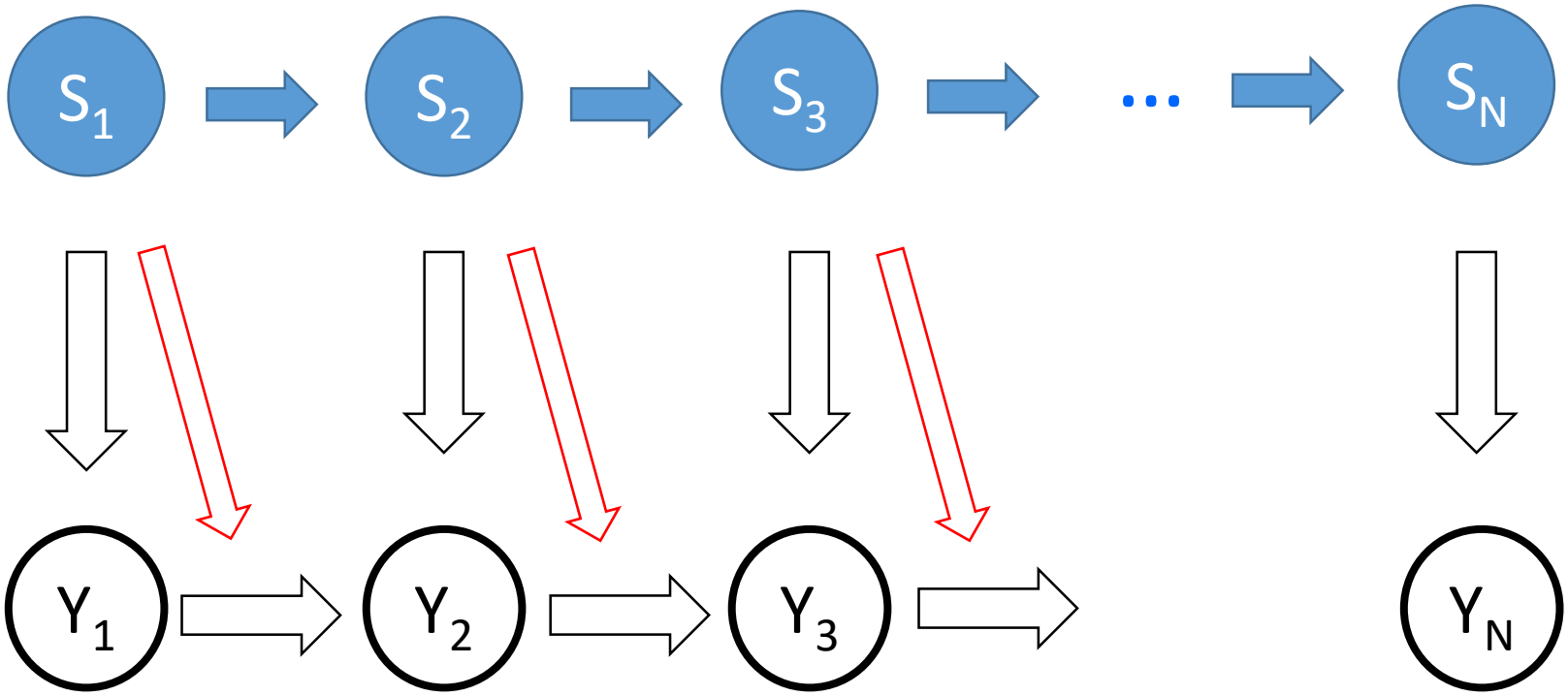


$$P[S_t = i | Y, \lambda, \pi]$$



Davide Pigoli, davide.pigoli@kcl.ac.uk

Alternative HMM designs…



Summary

1. A short intro to EEG data

2. The changepoint(s) detection problem

3. Three approaches:
   - Segment statistics
   - Bayes
   - Hidden Markov Models

4. Sources for data and software

5. Project report requirements

References and resources

Online resources:

Methods and software: www.changepoint.info

R packages: eegkit, eegUtils, wbs, changepoint, ecp, depmixS4, HMM.

Similar toolbox in Matlab…

Data:

- Task experiments:
http://archive.ics.uci.edu/ml/datasets/EEG+Eye+State

https://sccn.ucsd.edu/~arno/fam2data/publicly_available_EEG_data.html

- Medical EEG:
https://www.physionet.org/pn6/chbmit/

https://archive.ics.uci.edu/ml/datasets/Epileptic+Seizure+Recognition
… and many more!

References and resources

Changepoints detection:

Antoch, J., & Hušková, M. (2001). Permutation tests in change point analysis. *Statistics & probability letters*, *53*(1), 37-46.

Chaturvedi, A., & Shrivastava, A. (2016). Bayesian analysis of a linear model involving structural changes in either regression parameters or disturbances precision. *Communications in Statistics-Theory and Methods*, *45*(2), 307-320.

Chen, J., & Gupta, A. K. (2000), *Parametric Statistical Change Point Analysis*, New York.

Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, *42*(6), 2243-2281.

Galeano, P., & Peña, D. (2007). Covariance changes detection in multivariate time series. *Journal of Statistical Planning and Inference*, *137*(1), 194-211.

Horváth, L., Kokoszka, P., & Steinebach, J. (1999). Testing for changes in multivariate dependent observations with an application to temperature changes. *Journal of Multivariate Analysis*, *68*(1), 96-119.

References and resources

Killick, R., Fearnhead, P., & Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, *107*(500), 1590-1598.

Matteson, D. S., & James, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, *109*, 334-345.

Smith, A. F. M. (1975). A Bayesian approach to inference about a change-point in a sequence of random variables. *Biometrika*, *62*(2), 407-416.

Hidden Markov Models:

Bilmes, J. A. (1998). A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute*, *4*(510), 126.

Luong, T. M., Perduca, V., & Nuel, G. (2012). Hidden Markov Model Applications in Change-Point Analysis. *arXiv preprint arXiv:1212.1778*.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257-286.

References and resources

Scott, S. L. (2002). Bayesian methods for hidden Markov models: Recursive computing in the 21st century. *Journal of the American Statistical Association*, *97*(457), 337-351.

EEG:

Blythe, D. A., Meinecke, F. C., Von Bünau, P., & Müller, K. R. (2013). Explorative data analysis for changes in neural activity. *Journal of neural engineering*, *10*(2), 026018.

Kirch, C., Muhsal, B., & Ombao, H. (2015). Detection of changes in multivariate time series with application to EEG data. *Journal of the American Statistical Association*, *110*(511), 1197-1216.

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of neuroscience methods*, *134*(1), 9-21.

Summary

1. A short intro to EEG data

2. The changepoint(s) detection problem

3. Three approaches:
   - Segment statistics
   - Bayes
   - Hidden Markov Models

4. Sources for data and software

5. Project report requirements

Project structure

Essential components:

- Choose a dataset of EEG, describe it and explore it graphically and via summary statistics
- Look for changepoints in the mean and/or covariance (number and locations)
- Diagnostics of the segmented sequences.

Possible developments:

- Consider multiple datasets that showcase different challenges in the analysis.
- Compare results from multiple methods.
- Consider extensions of the discussed methods (or additional methods!) that take into account more challenging characteristics of your dataset, such as autocorrelation of the errors, trends, non Gaussian models, etc…