Introduction to changepoint detection from first principles

Arnaud Liehrmann & Guillem Rigaill

Sorbonne Université & INRAE

February 25, 2025

Introduction

- Detecting and locating changes in distribution within time series data presents a fundamental statistical challenge.
- The first studies on changepoint detection emerged in the 1940s [Wald, 1945, Page, 1954]
- A significant increase in research activity in this area has occurred in recent decades (see [Venkatraman and Olshen, 2007, Killick et al., 2012, Fryzlewicz, 2014, Maidstone et al., 2017] among many others))

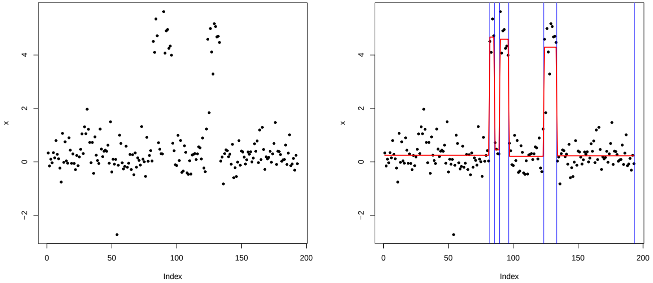
Plan

A Simple Example

An archetypical model

Conclusion

A Simple Example



- Data from [Lai et al., 2005]
- Comparative Genomic Hybridization (CGH) reveals chromosomal aberrations in DNA.
- Abrupt changes in signal intensity reveal these aberrations.

Data Analysis Questions

- Are there changes in the intensity that reveal chromosomal aberrations?
- If so,
    - How many are there?
        - A marker of genomic instability?
    - Where are they located?
        - To detect fusion and splits of certain genes
    - Which chromosomal regions are amplified or deleted?
        - Potential oncogenes and tumor suppressor genes

Statistical Goals

Detection Has a change occurred?

How many If there are changes, how many are there?

No Spurious Avoid the detection of false changes.

Estimation If there is a change, what is the intensity before and after?

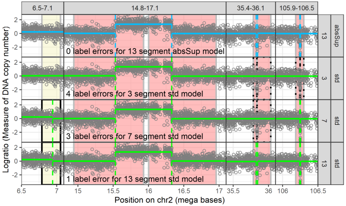
Localization If there are changes, where are they, and how confident are we about their locations?

Many Types of Changes

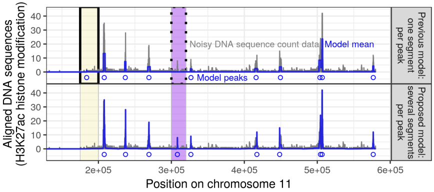
- Gaussian data: Change in mean.
- Genomics: Poisson or Negative Binomial models
- Slope (with a continuity constraint)
- Peaks
- Changes in variance
- Changes in multiple parameters.
- Multivariate settings: Regression or graphical models.
- Network structure changes.
- Variance/Covariance
- Non-parametric changes in distribution
- ...

Many Types of Changes

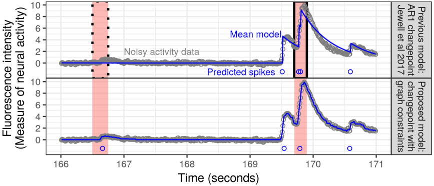
CGH



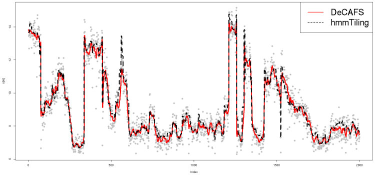
Chip-Seq



Neuroscience



RNA-seq



Plan

A Simple Example

An archetypical model

Conclusion

A Principled Approach

- Hope to provide a solid understanding of the core principles of changepoint detection (exercises...)
- Focus on the univariate change-in-mean model
- Key challenges and difficulties related to changepoint detection are already present in this univariate change-in-mean model
- Considering as a baseline a penalized maximum likelihood approach (with a constant per changepoint penalty)

The univariate change in mean model is not that simple!

- Despite its apparent simplicity, it remains an active area of research (see, for example,[Killick et al., 2012, Aue and Kirch, 2024, Fryzlewicz, 2014, Kovács et al., 2023, Verzelen et al., 2023, Yu et al., 2023])

- Why: Let us consider the "vanilla" approach for just one change.
  - Our data $y_1, y_2, \ldots, y_n$
  - Assuming the $y_t = \mu_t + \varepsilon_t$
  - Assuming ε_t are Gaussian $\mathcal{N}(0, 1)$
  - $\mu_t = 0$ for $t \leq \tau^*$ and 1 otherwise

At most one change “vanilla” approach

- Compare the mean square error with a change at τ :

$$\sum_{t=1}^{\tau}(y_t - \bar{y}_{1:\tau})^2 + \sum_{t=\tau+1}^{n}(y_t - \bar{y}_{\tau+1:n})^2,$$

- with the mean square error without a change

$$\sum_{t=1}^{n}(y_t - \bar{y}_{1:n})^2$$

- A large difference between these two indicates a change

Why is it not simple?

- We need to consider $n-1$ changepoints/models
- To decide/infer whether there is a change or not
  - Compute/Compare all $n-1$ mean squared errors
  - Seek to control their variation
- These squared errors are dependant
- How to exploit this statistically and computationally?

Course Focus

- Not a tutorial on a package.
- Rather aim to provide an understanding of what to look for and test in changepoint detection packages
- Focus on method optimizing globally or locally a penalized likelihood

Plan

A Simple Example

An archetypical model

Conclusion

Outline

1. Basics of At Most One Change offline
2. Basics of Multiple changepoints offline
3. Overview of more advanced topics
    - Model selection
    - Pruned dynamic programming algorithm
    - Local Search and Isolation techniques
    - Maximum Likelihood for online data

Conclusion

- Changepoint detection is a fundamental statistical challenge.
- Various types of changes can be detected in different settings
- Key challenges include detection, localization, and avoiding spurious changes.
- This course aims to provide a solid understanding of changepoint detection principles

Aue, A. and Kirch, C. (2024).
The state of cumulative sum sequential changepoint testing 70 years after page.
Biometrika, 111(2):367–391.

Fryzlewicz, P. (2014).
Wild binary segmentation for multiple change-point detection.

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

Kovács, S., Bühlmann, P., Li, H., and Munk, A. (2023).
Seeded binary segmentation: a general methodology for fast and optimal changepoint detection.
Biometrika, 110(1):249–256.

Lai, W. R., Johnson, M. D., Kucherlapati, R., and Park, P. J. (2005).

Comparative analysis of algorithms for identifying amplifications and deletions in array cgh data.
Bioinformatics, 21(19):3763–3770.

Maidstone, R., Hocking, T., Rigaill, G., and Fearnhead, P. (2017).
On optimal multiple changepoint algorithms for large data.
Statistics and computing, 27:519–533.

Page, E. S. (1954).
Continuous inspection schemes.
Biometrika, 41(1/2):100–115.

Venkatraman, E. and Olshen, A. B. (2007).
A faster circular binary segmentation algorithm for the analysis of array cgh data.
Bioinformatics, 23(6):657–663.

Verzelen, N., Fromont, M., Lerasle, M., and Reynaud-Bouret, P. (2023).
Optimal change-point detection and localization.
The Annals of Statistics, 51(4):1586–1610.

Wald, A. (1945).
Sequential tests of statistical hypotheses.
In *Ann. Math. Statist. (June, 1945)*, pages 117–186.

Yu, Y., Madrid Padilla, O. H., Wang, D., and Rinaldo, A. (2023).
A note on online change point detection.
Sequential Analysis, 42(4):438–471.