Optimal detection of changepoints with a linear computational cost

Paul Fearnhead

Department of Mathematics & Statistics, Lancaster University
Joint work with Rebecca Killick and Idris Eckley

April 2012

Summary of Talk

- Motivation
- Review existing methods
- PELT (Pruned Exact Linear Time) method
- Simulation Study
- Oceanographic Example

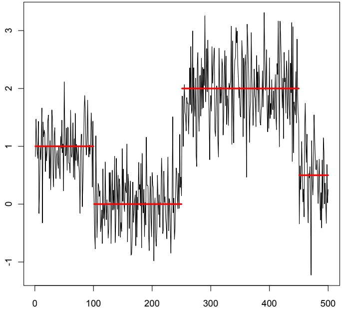
Motivation

Example: Change in mean

Assume we have time-series data where

$$Y_t|\theta_t \sim \mathrm{N}(\theta_t, 1),$$

but where the means, θ_t, are piecewise constant through time.



Example: Inferring Changepoints

We want to infer the number and position of the points at which the mean changes. There are a number of approaches:

Likelihood Ratio Test

To detect a single changepoint we can use the likelihood ratio test statistic:

$$LR = 2 \max_{\tau} \{\ell(y_{1:\tau}) + \ell(y_{\tau+1:n}) - \ell(y_{1:n})\}.$$

We infer a changepoint if $LR > 2\beta$ for some (suitably chosen) β. If we infer a changepoint its position is estimated as

$$\tau = \arg\max \{\ell(y_{1:\tau}) + \ell(y_{\tau+1:n}) - \ell(y_{1:n})\}.$$

This can test can be repeatedly applied to new segments to find multiple changepoints.

Inferring Changepoints: Likelihood Ratio Tests

Define m to be the number of changepoints, with positions $\boldsymbol{\tau} = (\tau_0, \tau_1, \ldots, \tau_{m+1})$ where $\tau_0 = 0$ and $\tau_{m+1} = n$.

Then one application of the Likelihood ratio test can be viewed as aiming

$$\min_{m \in \{0,1\}, \boldsymbol{\tau}} \left\{ \sum_{i=1}^{m+1} \left[-\ell(y_{\tau_{i-1}:\tau_i}) \right] + \beta m \right\}$$

Repeated application is thus aiming to minimise

$$\min_{\boldsymbol{\tau}} \left\{ \sum_{i=1}^{m+1} \left[-\ell(y_{\tau_{i-1}:\tau_i}) \right] + \beta m \right\}$$

Inferring Changepoints: Penalised Likelihood

The above can be viewed as a special case of penalised likelihood. Here the aim is to maximise the *likelihood* over the number and position of the changepoints, but *subject to* a penalty, that depends on the number of changepoints. The penalty is to avoid over-fitting.

This is equivalent to minimising

$$\min_{\boldsymbol{\tau}} \left\{ \sum_{i=1}^{m+1} \left[-\ell(y_{\tau_{i-1}:\tau_i}) \right] + \beta f(m) \right\}$$

for a suitable penalty function $f(m)$ and penalty constant β.

Inferring Changepoints: Bayesian MAP

A Bayesian approach would involve introducing a prior for θ within each segment, and a prior for the number and position of the changepoints.

If the priors for θ are *independent across segments*, and the prior for the changepoints is based on an *independent* geometric distribution for segment lengths. Then the *Bayesian MAP* estimate would satisfy

$$\min_{\boldsymbol{\tau}} \left\{ \sum_{i=1}^{m+1} \left[-\mathrm{ML}(y_{\tau_{i-1}:\tau_i}) \right] + \beta m \right\},$$

where $\mathrm{ML}(\cdot)$ is the segment marginal likelihood; and β depends on the parameter of the geometric distribution.

An approach (from computer science) is to estimate the changepoints via minimising the *description length* of the model for the data.

Davis et al. (2006) derive the MDL criteria for changepoint models. For a change in mean this is

$$\min_{\boldsymbol{\tau}} \left\{ \sum_{i=1}^{m+1} \left[-\ell(y_{\tau_{i-1}:\tau_i}) + \frac{1}{2}(\tau_i - \tau_{i-1} + 1) \right] + m \log n + \log(m+1) \right\},$$

How do we identify changepoints?

All these methods can be cast in terms of minimising a function of $\boldsymbol{\tau}$ of the form:

$$\sum_{i=1}^{m+1} \left[\mathcal{C}(y_{(\tau_{i-1}+1):\tau_i}) \right] + \beta f(m).$$

This function depends on the data just through a sum of a *cost* for each segment.

There is then a penalty term that depends on the number of segments.

The Challenge

- What are the values of $\tau_1, \ldots, \tau_m$?
- What is m?

The Challenge

- What are the values of $\tau_1, \ldots, \tau_m$?
- What is m?

- For n data points there are 2^{n-1} possible solutions

- If m is known there are still $\binom{n-1}{m-1}$ solutions
- If $n = 1000$ and $m = 10$, 2.634096×10^{21} solutions

The Challenge

- What are the values of $\tau_1, \ldots, \tau_m$?
- What is m?

- For n data points there are 2^{n-1} possible solutions

- If m is known there are still $\binom{n-1}{m-1}$ solutions
- If $n = 1000$ and $m = 10$, 2.634096×10^{21} solutions

- How do we search the solution space efficiently?

Existing Search Methods

Existing Methods

Existing methods are either *fast* but *approximate*.

Such as Binary Segmenation (Scott and Knott (1974)). Binary segmentation is $\mathcal{O}(n \log n)$ in CPU time.

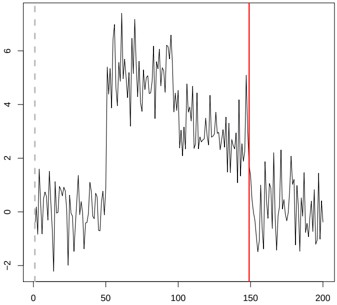
Or they are *slower* but *exact*.

These method used dynamic programming. For example, Segment Neighbourhood (Auger and Lawrence (1989)) is $\mathcal{O}(n^3)$.

For linear penalties $f(m) = m$, Optimal Partitioning (Jackson et al. (2005)) is $\mathcal{O}(n^2)$

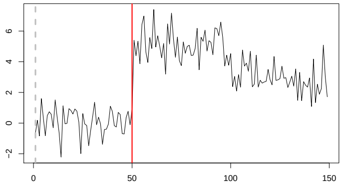
Binary Segmentation

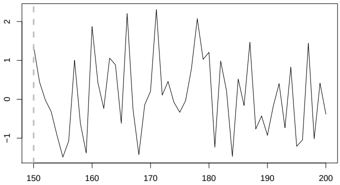
- Start by finding the optimal location for one changepoint



Binary Segmentation

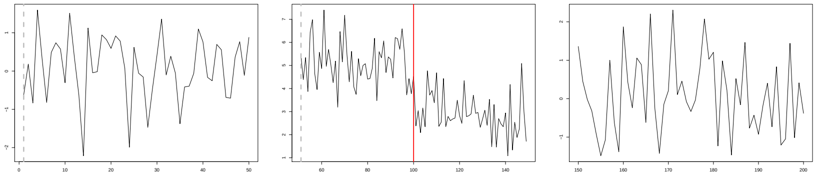
- Then before and after the changepoint are treated as separate datasets





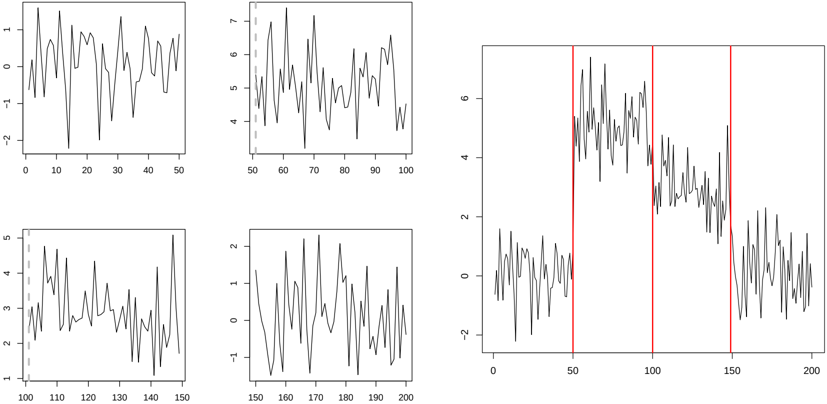
Binary Segmentation

- This continues until no more changepoints are found



Binary Segmentation

- This continues until no more changepoints are found



Optimal Partitioning

Applies to $f(m) = m$.

Consider $y_{1:2}$, either

1. There is no changepoint, or
2. There is a changepoint at y_1

Both scenarios are calculated and the optimal kept

Optimal Partitioning

Now consider $y_{1:3}$,

1. No changepoint
2. A changepoint at y_1
3. A changepoint at y_2
4. A changepoint at y_1 and y_2

But the decision between the latter two has already been decided (at the previous iteration)!

Optimal Partitioning

The decision for $y_{1:3}$ becomes

1. No changepoint
2. Most recent changepoint at y_1,
   i.e. a single change at y_1
3. Most recent changepoint at y_2, and the optimal partition of $y_{1:2}$.

Optimal Partitioning

In a similar fashion, the decision for $y_{1:4}$ becomes

1. No changepoint
2. Most recent changepoint at y_1,
   i.e. a single change at y_1
3. Most recent changepoint at y_2, and the optimal partition of $y_{1:2}$.
4. Most recent changepoint at y_3, and the optimal partition of $y_{1:3}$.

Optimal Partitioning

If we define

$$\mathcal{P}_t = \{\boldsymbol{\tau} : 0 < \tau_1 < \cdots < \tau_m < t\}$$

$$F(t) = \min_{\boldsymbol{\tau} \in \mathcal{P}_t} \left\{ \sum_{i=1}^{m+1} \left[\mathcal{C}(y_{(\tau_{i-1}+1):\tau_i}) + \beta \right] \right\}$$

i.e. $f(m) = m$ in original minimisation.

So

$$F(t) = \min_{\tau^*} \left\{ \min_{\boldsymbol{\tau} \in \mathcal{P}_{\tau^*}} \left[\sum_{i=1}^{m} \mathcal{C}(y_{(\tau_{i-1}+1):\tau_i}) + \beta \right] + \mathcal{C}(y_{(\tau^*+1):t}) + \beta \right\}$$

Optimal Partitioning

Thus we minimise,

$$F(t) = \min_{\tau^*} \left\{ F(\tau^*) + \mathcal{C}(y_{(\tau^*+1):t}) + \beta \right\}$$

Recursively solving the minimisation for $t = 1, \ldots, n$ gives an algorithm that is $\mathcal{O}(n^2)$.

The PELT Method

(Pruned Exact Linear Time)

PELT: The idea

By eye there is often an obvious changepoint at (or by) a time-point s.

This means that for any $T > s$ the most recent changepoint cannot be at time $t < s$.

Thus we could prune the search step: and avoid searching over $t < s$.

PELT: Assumption

ASSUMPTION: adding a changepoint reduces the overall cost

This means that for $t < s < T$:

$$C(y_{t+1:T}) \geq C(y_{t+1:s}) + C(y_{s+1:T})$$

This holds in for costs based on the negative log-likelihood; and often can be made to hold for costs based on minus the log-marginal-likelihood.

PELT: Pruning

Let $0 < t < s < T$

Theorem

If

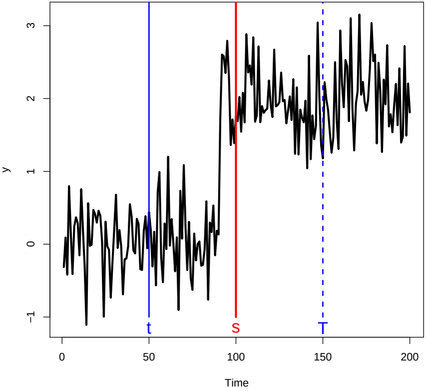
$$F(t) + \mathcal{C}(y_{(t+1):s}) > F(s)$$

then at any future time $T > s$, t can never be the optimal last changepoint prior to T.

PELT: Intuition

The condition in the theorem just means that for any $T > s$ the best partition which involves a changepoint at s is will be better than one which has $[t, T]$ as a single segment.

Thus t can never be the (optimal) most recent changepoint prior to T for ant $T > s$.



PELT: Computational Complexity

If many t are pruned and excluded from the minimisation then computational time will be drastically reduced.

We can prove that, under certain regularity conditions, the expected computational complexity will be $\mathcal{O}(n)$.

The most important condition is that *the number of changepoints increases linearly with* n.

This is natural in many applications: e.g. as you collect time-series data over larger time-periods; or genomic data or larger regions of the genome.

PELT Algorithm Recap

At some time t in the algorithm

- Calculate $F(t) = \min_{\tau \in (pts, t-1)} \left[F(\tau) + \mathcal{C}(y_{(\tau+1):t}) + \beta \right]$.
- Let τ^* be the optimum last changepoint prior to t.
- Calculate the potential changepoints to be included in the next iteration:

  Set $pts = \arg_\tau \{F(\tau) + \mathcal{C}(y_{\tau+1:t}) > F(\tau^*)\}$.

PELT: Non-linear penalties

Can we do anything if the penalty function $f(m)$ is non-linear?

You can show that for a concave penalty (such as that in MDL $f(m) = m \log n + \log(m+1)$) there exists a β such that the optimal segmentation under penalty $f(m)$ is the same as under penalty βm.

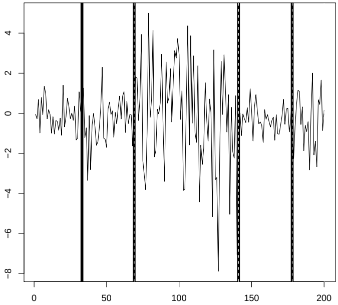
Thus we can apply iteratively (for different β) to find the optimal segmentation in these cases.

Simulation Study

Change in Variance: Simulation Structure

- 9 scenarios with lengths
  100, 200, 500, 1000, 2000, 5000, 10000, 20000, 50000
- Uniform distributed changepoints, subject to >30 observations per segment
- Each scenario has 1,000 repetitions

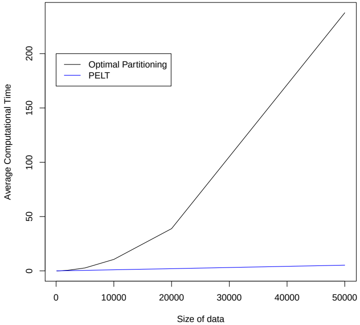
- Cost function: Negative log-likelihood
- Mean set to 0
- Variances simulated from a Log-Normal distribution

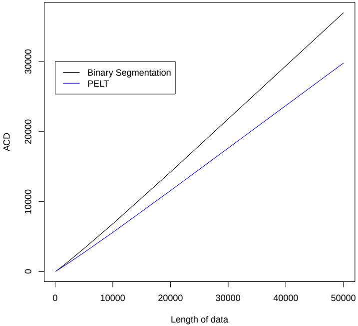


Change in Mean and Variance

θ is a parameter that changes, i.e. the variance

$$\text{ACD} = \frac{\sum_{i=1}^{n} |\hat{\theta}_i - \theta_i|}{n}$$





MDL fit for AR(p) models

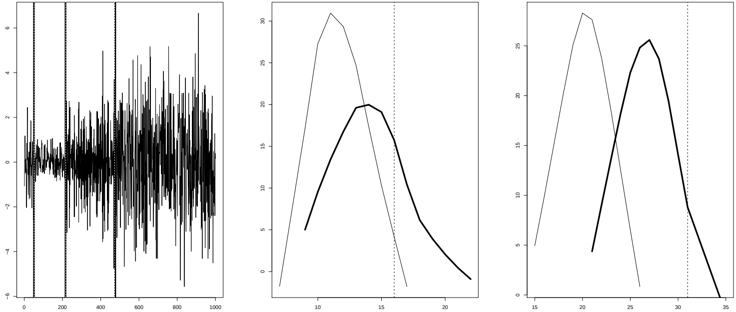
We compared the accuracy of PELT with a genetic algorithm approach (Davis et al.) for optimising under an MDL criteria.

The underlying model within each segment was AR(p), with p unknown. Average Improvement in MDL for varying lengths of data.

n	1,000	2,000	5,000	10,000	20,000
Improvement	9	14	60	250	900

MDL fit for AR(p) models

Realisation of a piecewise stationary autoregressive process. Smoothed number of segments identified by PELT (thick line) and Auto-PARM (thin line) algorithms for (b) $n = 5,000$ and (c) $n = 10,000$.



Ocean Engineering Application

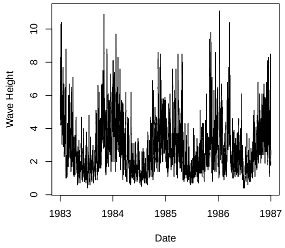
Ocean Engineering

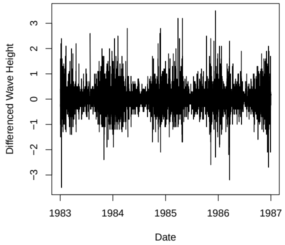
- The wave heights at a particular location change over time

- Understanding wave heights is key to
  - security of sea structures
  - development of new techniques for harnessing sea energy

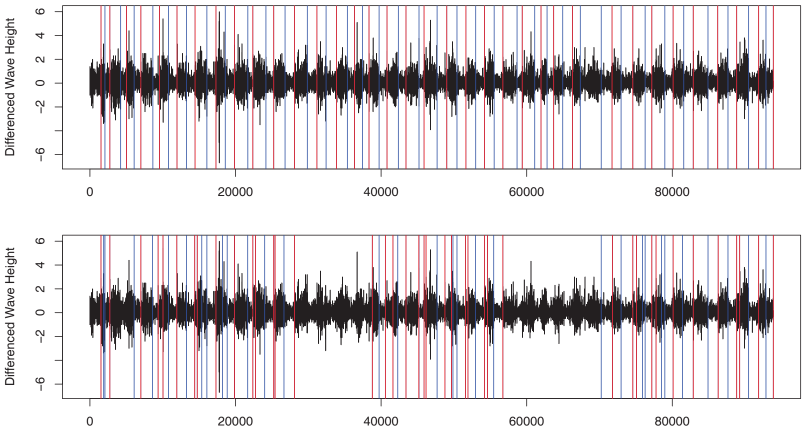
- Data is 3-hourly wave heights from a location in the North North Sea from 1973–2009.

- Assume first difference of wave heights is Normal (μ, σ_i^2).





Wave Heights



Red - increase; Blue - decrease in variance in following segment

Summary

- Being able to find changepoints quickly is important
- Existing methods are either inefficient or approximate
- PELT is $\mathcal{O}(n)$ under certain conditions and is exact
- Code is available within the R package changepoint on CRAN

References I

I.E. Auger and C.E. Lawrence.
Algorithms for the Optimal Identification of Segment Neighborhoods.
Bulletin of Mathematical Biology, 51(1):39–54, 1989.

A.J. Scott and M. Knott.
A Cluster Analysis Method for Grouping Means in the Analysis of Variance.
Biometrics, 30(3):507–512, 1974.

B. Jackson, J.D. Sargle, D. Barnes, S. Arabhi, A. Alt, P. Gioumousis, E. Gwin, P. Sangtrakulcharoen, L. Tan and T.T. Tsai.
An Algorithm for optimal partitioning of data on an interval.
IEEE, Signal Processing Letters, 12(2):105–108, 2005.

R. Killick, P. Fearnhead and I.A. Eckley.
Optimal detection of changepoints with a linear computational cost.
Submitted, http://arxiv.org/abs/1101.1438.

Assumptions of PELT

- Independence between segments
- IID within a segment
- Additivity of the cost function over segments
- Penalty that is linear in the number of changepoints

Theorem

Theorem

Define θ^* to be the value that maximises the expected log-likelihood

$$\theta^* = \arg\max \int \int f(y|\theta) f(y|\theta_0) dy \pi(\theta_0) d\theta_0.$$

Let θ_i be the true parameter associated with the segment containing y_i and $\hat{\theta}_n$ be the maximum likelihood estimate for θ given data $y_{1:n}$ and an assumption of a single segment:

$$\hat{\theta}_n = \arg\max_{\theta} \sum_{i=1}^{n} \log f(y_i|\theta).$$

Theorem cont.

Theorem

Then if

(A1) *denoting* $B_n = \sum_{i=1}^{n} \log \left[f(y_i|\hat{\theta}_n) - \log f(y_i|\theta^*) \right]$, *we have* $\mathbb{E}(B_n) = o(n)$ *and* $\mathbb{E}\left([B_n - \mathbb{E}(B_n)]^4\right) = \mathcal{O}(n^2)$;

(A2) $\mathbb{E}\left([\log f(Y_i|\theta_i) - \log f(Y_i|\theta^*)]^4 \right) < \infty$;

(A3) $\mathbb{E}\left(S^3\right) < \infty$; *and*

(A4) $\mathbb{E}\left(\log f(Y_i|\theta_i) - \log f(Y_i|\theta^*)\right) > \frac{\beta}{\mathbb{E}(S)}$;

the expected CPU cost of PELT for analysing n data points is bounded above by Ln for some constant $L < \infty$.