Optimal Detection of Changepoints With a Linear Computational Cost

Article *in* Journal of the American Statistical Association · December 2012

DOI: 10.1080/01621459.2012.737745 · Source: arXiv

3 authors, including:

Rebecca Killick
Lancaster University
71 PUBLICATIONS **3,741** CITATIONS

SEE PROFILE

Paul Fearnhead
Lancaster University
207 PUBLICATIONS **12,716** CITATIONS

SEE PROFILE

Optimal detection of changepoints with a linear computational cost

Killick, R.*, Fearnhead, P.* and Eckley, I.A.*

Abstract

We consider the problem of detecting multiple changepoints in large data sets. Our focus is on applications where the number of changepoints will increase as we collect more data: for example in genetics as we sequence larger regions of the genome, or in finance as we observe time-series over longer periods. We consider the common approach of detecting changepoints through minimising a cost function over possible numbers and locations of changepoints. We introduce a new method for finding the minimum of such cost functions and hence the optimal number and location of changepoints, that has a computational cost which, under mild conditions, is linear in the number of observations. This compares favourably with existing methods for the same problem whose computational cost can be quadratic, or even cubic. In simulation studies we show that our new method can be orders of magnitude faster than these alternative methods. We also compare with Binary Segmentation, a fast but approximate approach to detecting changepoints and show that the exactness of our approach can lead to substantial improvements in the accuracy of the inferred segmentation of the data.

1 Introduction

As increasingly longer data sets are being collected, a common inferential problem is to detect changes in the distributional properties of such data. Consider for example recent work in genomics, looking at detecting changes in gene copy numbers, or in the compositional structure of the genome (Braun et al., 2000; Olshen et al., 2004; Picard et al., 2005); and in finance where, for example, interest lies in detecting changes in the volatility of time series (Aggarwal et al., 1999; Andreou and Ghysels, 2002; Fernandez, 2004). There is therefore a growing need to be able to detect changes in the distributional properties over time and search for such changes efficiently. It is this latter, search problem which we consider in this paper. In particular we focus on applications where we expect the number of changepoints to increase as we collect more data. This is a natural assumption in many cases, for example as we sequence longer regions of the genome, or as we record financial time-series over longer time-periods. By comparison it does not necessarily apply to situations where we are obtaining data over a fixed time-period at a higher frequency.

At the time of writing Binary Segmentation proposed by Scott and Knott (1974) is arguably the most widely used changepoint search method but it is approximate in nature and has $\mathcal{O}(n \log n)$ computational cost, where n is the number of data points. While exact search algorithms exist for the most common forms of changepoint models, these have a much greater computational cost. Several exact search methods exist and are based on dynamic programming. For example the Segment Neighbourhood method proposed by Auger and Lawrence (1989) is $\mathcal{O}(Qn^2)$, where Q is the number of maximum number of changepoints you wish to search for. Note that in scenarios where the number of changepoints increases linearly with n, this can correspond to a computational cost that is cubic in the length of the data. An alternative dynamic programming algorithm is the Optimal Partitioning approach of Jackson et al. (2005). As we describe in Section 2.2.2 this can be applied to a slightly smaller class of problems and is an exact approach whose computational cost is $\mathcal{O}(n^2)$.

We present a new approach for searching for changepoints, which is exact and under mild conditions has a computational cost that is linear in the number of data points: the **P**runed **E**xact **L**inear **T**ime (PELT)

*Department of Mathematics & Statistics, Lancaster University, Lancaster, UK.

method. This approach is based on the algorithm of Jackson et al. (2005), but involves a pruning step within the dynamic programming. This pruning reduces the computational cost of the method, but does not affect the exactness of the resulting method. In simulations we compare PELT with both Binary Segmentation and with Optimal Partitioning. We show that PELT attains linear computational cost within these simulations and can be orders of magnitude faster than Optimal Partitioning, particularly for long data sets. We also show that PELT leads to a substantially more accurate segmentation of data than Binary Segmentation.

The paper is organised as follows. We begin in Section 2 by reviewing some basic changepoint notation and summarizing existing work in the area of search methods. The PELT method is introduced in Section 3 and the computational cost of this approach is considered in Section 3.1. In particular we show that under mild conditions this cost is linear in the number of observations. The efficiency and accuracy of the PELT method are demonstrated in Section 4. In particular we demonstrate the methods performance on large data sets coming from financial and genomic applications. The paper concludes with a discussion.

2 Background

Changepoint analysis can, loosely speaking, be considered to be the identification of points within a data set where the statistical properties change. More formally, let us assume we have an ordered sequence of data, $y_{1:n} = (y_1, \ldots, y_n)$. Our model will have a number of changepoints, m, together with their positions, $\tau_{1:m} = (\tau_1, \ldots, \tau_m)$. Each changepoint position is an integer between 1 and $n-1$ inclusive. We define $\tau_0 = 0$ and $\tau_{m+1} = n$, and assume that the changepoints are ordered such that $\tau_i < \tau_j$ if, and only if, $i < j$. Consequently the m changepoints will split the data into $m+1$ segments, with the ith segment containing $y_{(\tau_{i-1}+1):\tau_i}$.

One commonly used approach to identifying multiple changepoints is to minimise the following expression:

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

Here $\mathcal{C}$ is a cost function for a segment and $\beta f(m)$ is a penalty to guard against over fitting. The negative log likelihood is a commonly used cost function in the changepoint literature (see for example Horvath (1993); Chen and Gupta (2000)), although other cost functions such as quadratic loss and cumulative sums are also used (e.g. Rigaill (2010); Inclan and Tiao (1994)). Turning to choice of penalty, in practice by far the most common choice is one which is linear in the number of changepoints, i.e. $\beta f(m) = \beta m$. Examples of such penalties include Akaike's Information Criterion (AIC, Akaike (1974)) ($\beta = 2$) and Schwarz Information Criterion (SIC, Schwarz (1978)) ($\beta = \log n$). We briefly note that the SIC is also known as the Bayesian Information Criterion (BIC). The PELT method which we introduce in Section 3 also makes this assumption. Although this assumption is commonplace within the changepoint literature Guyon and Yao (1999), Picard et al. (2005) and Birge and Massart (2007) offer examples and discussion of alternative penalty choices. For a more comprehensive background to changepoint analysis, see Chen and Gupta (2000), Braun et al. (2000) or Eckley et al. (2010).

Several methods exist to calculate the optimal number and locations of these changepoints, $\{\tau_i\}$. These methods include (i) exact approaches such as those proposed by Auger and Lawrence (1989) and Jackson et al. (2005) and (ii) approximate approaches such as those proposed by Scott and Knott (1974) and Inclan and Tiao (1994). As we shall discuss in Section 4 the accuracy of the exact approaches comes at the expense of computational efficiency. Conversely the computational benefits of the approximate methods result in less accurate identification of changepoints. There is therefore a benefit in developing a search method which is both efficient and optimal.

The remainder of this section describes the two most commonly used methods for multiple changepoint detection; Binary Segmentation (Scott and Knott, 1974) and Segment Neighbourhoods (Auger and Lawrence, 1989). A third method proposed by Jackson et al. (2005) is also described as it forms the basis for the PELT method which we propose.

2.1 Binary Segmentation

Binary Segmentation is arguably the most established search method used within the changepoint literature. Early applications include Scott and Knott (1974) and Sen and Srivastava (1975). In essence the method extends any single changepoint method to multiple changepoints by iteratively repeating the method on different subsets of the sequence. It begins by initially applying the single changepoint method to the entire data set, i.e. we test if a τ exists that satisfies

$$\mathcal{C}(y_{1:\tau}) + \mathcal{C}(y_{(\tau+1):n}) + \beta < \mathcal{C}(y_{1:n}). \tag{2}$$

If (2) is false then no changepoint is detected and the method stops. Otherwise the data is split into two segments consisting of the sequence before and after the identified changepoint, τ_a say, and apply the detection method to each new segment, in effect testing,

$$\begin{aligned} & \mathcal{C}(y_{1:\tau}) + \mathcal{C}(y_{(\tau+1):\tau_a}) + \beta < \mathcal{C}(y_{1:\tau_a}) \\ \text{and} \qquad & \mathcal{C}(y_{\tau_a:\tau}) + \mathcal{C}(y_{(\tau+1):n}) + \beta < \mathcal{C}(y_{\tau_a:n}). \end{aligned}$$

If either or both tests are true, we split these into further segments at the newly identified changepoint(s), applying the detection method to each new segment. This procedure is repeated until no further changepoints are detected. For pseudo-code demonstrating the implementation of the Binary Segmentation method, see Eckley et al. (2010).

The advantage of the Binary Segmentation method is that it is computationally efficient, resulting in an $\mathcal{O}(n \log n)$ calculation. However this computational efficiency comes at the cost of exactness. In particular, this procedure is an approximate minimisation of equation (1) with $f(m) = m$. Note that if the cost function, $\mathcal{C}$ is the log likelihood then equation (2) is equivalent to a likelihood ratio test. For data sets where most of the changepoints are easy to detect this approximation will be negligible, but as the changepoints become more subtle the accuracy of Binary Segmentation will suffer. In practice this means that Binary Segmentation becomes less accurate with either smaller changes or changes that are close in proximity.

2.2 Exact methods

2.2.1 Segment Neighbourhood

Auger and Lawrence (1989) propose an alternative, exact search method for changepoint detection, namely the Segment Neighbourhood method. This approach searches the entire segmentation space using dynamic programming. It begins by setting an upper limit on the size of the segmentation space (i.e. the maximum number of changepoints) that is required – this is denoted Q. The method then continues by computing the cost function for all possible segments. From this all possible segmentations with between none and Q changepoints are considered.

In addition to being an exact search method, the Segment Neighbourhood approach has the ability to incorporate an arbitrary penalty function of the form, $\beta f(m)$. However, a consequence of the exhaustive search is that the method has significant computational cost, $\mathcal{O}(Qn^2)$. If as the observed data increases, the number of changepoints increases linearly, then $Q = \mathcal{O}(n)$ and hence the method will have a computational cost of $\mathcal{O}(n^3)$.

2.2.2 The optimal partitioning method

Yao (1984) and Jackson et al. (2005) propose a search method that aims to minimise

$$\sum_{i=1}^{m+1} \left[\mathcal{C}(y_{(\tau_{i-1}+1):\tau_i}) + \beta\right]. \tag{3}$$

This is equivalent to (1) where $f(m) = m$.

Following Jackson et al. (2005) the optimal partitioning method begins by first conditioning on the last point of change and calculating the optimal segmentation of the data up to that changepoint. Following this, the last changepoint is then moved through from the start to the end of the data and the optimal overall segmentation chosen as our final set of changepoints. More formally, let $F(n)$ denote the minimisation from (3). Thus

$$F(n) = \min_{\boldsymbol{\tau}} \left\{ \sum_{i=1}^{m+1} \left[\mathcal{C}(y_{(\tau_{i-1}+1):\tau_i}) + \beta \right] \right\}. \tag{4}$$

Setting $\tau_m = \tau^*$ to denote the last changepoint and condition on its location to obtain the following result.

$$F(n) = \min_{\tau^*} \left\{ \min_{\boldsymbol{\tau}|\tau^*} \sum_{i=1}^{m} \left[\mathcal{C}(y_{(\tau_{i-1}+1):\tau_i}) + \beta \right] + \mathcal{C}(y_{(\tau^*+1):n}) \right\}. \tag{5}$$

This could equally be repeated for the second to last, third to last, ... changepoints. The recursive nature of this conditioning becomes clearer as one notes that the inner minimisation is reminiscent of equation (4). In fact the inner minimisation is equal to $F(\tau^*)$ and as such (5) can be re-written as

$$F(n) = \min_{\tau^*} \left\{ F(\tau^*) + \mathcal{C}(y_{(\tau^*+1):n}) \right\}. \tag{6}$$

This result enables the calculation of the global optimal segmentation using optimal segmentations on subsets of the data. In particular it gives a recursive form to the method as the optimal segmentation for data $y_{1:\tau^*}$ is identified and then used to inform the optimal segmentation for data $y_{1:(\tau^*+1)}$.

At each step in the method we store the optimal segmentation up to τ^*. When we reach $F(n)$ the optimal segmentation for the entire data has been identified and the number and location of changepoints have been recorded. The recursions involved in the optimal partitioning method are described more formally in Algorithm 1.

Optimal Partitioning

Input: A set of data of the form, $(y_1, y_2, \ldots, y_n)$ where $y_i \in \mathbb{R}$.
A measure of fit $\mathcal{C}(\cdot)$ dependent on the data.
A penalty β which does not depend on the number or location of changepoints.

Initialise: Let n = length of data and set $F(0) = -\beta$, $cp(0) = 0$.

Iterate for $\tau^* = 1, \ldots, n$

1. Calculate $F(\tau^*) = \min_{0 \le \tau < \tau^*} \left[F(\tau) + \mathcal{C}(y_{(\tau+1):\tau^*}) + \beta \right]$.
2. Let $\tau' = \arg \left\{ \min_{0 \le \tau < \tau^*} \left[F(\tau) + \mathcal{C}(y_{(\tau+1):\tau^*}) + \beta \right] \right\}$.
3. Set $cp(\tau^*) = (cp(\tau'), \tau')$.

Output the change points recorded in $cp(n)$.

Algorithm 1: Optimal Partitioning.

2.3 Differences and Improvements

The implementation of the exact Segment Neighbourhood method and the approximate Binary Segmentation method differ in a subtle way. The key to this difference is that when a changepoint is added to the exact methods, the location of previous changepoints can change to accommodate the addition whereas in the Binary Segmentation method the locations of previously identified changepoints are fixed. The consequence of this is that the Binary Segmentation method is approximate rather than exact but this also accounts for the decrease in computational time as only the location of the new changepoint needs to be considered. We also note that the recursions presented in Algorithm 1 are more efficient than the method

presented by Auger and Lawrence (1989) as the location and number of changepoints are decided in one pass of the data. As Jackson et al. (2005) describe, this results in an $\mathcal{O}(n^2)$ method compared to $\mathcal{O}(Qn^2)$ if the number of changepoints Q were known.

Whilst the optimal partitioning method improves on the computational efficiency of the Segment Neighbourhoods method, it is still far from being competitive computationally with the Binary Segmentation method. Section 3 introduces a modification of the optimal partitioning method denoted PELT which results in an approach which is theoretically more efficient than Binary Segmentation whilst retaining exact results. This exact and efficient computation is achieved via a combination of dynamic programming and pruning methods. The key to the proposed approach is that the optimal number and location of changepoints are calculated efficiently in one pass of the data.

3 A Pruned Exact Linear Time Method

The optimal partitioning method described in Section 2.2.2 is a recursive method designed to identify the number and locations of multiple changepoints within a sequence of data. This improves on the Segment Neighbourhood method (Auger and Lawrence, 1989) but is still computationally more expensive than the approximate $\mathcal{O}(n \log n)$ Binary Segmentation method (Scott and Knott, 1974).

One way in which the efficiency could be improved is to 'prune' the optimal partitioning method. The essence of pruning in this context is to remove those values of τ which can never be minima from the minimisation performed at each iteration in step 1 of Algorithm 1. Consider, by way of example, a time s during the recursions. At this time point

$$F(s) = \min_{0 \leq \tau < s} \left[F(\tau) + \mathcal{C}(y_{(\tau+1):s}) + \beta \right].$$

Now let t be a time such that $0 \leq t < s$ and

$$F(t) + \mathcal{C}(y_{(t+1):s}) + \beta > F(s). \tag{7}$$

In other words, t is *not* the location of the last changepoint prior to s.

Now let us consider a future time $T > s$. Can we somehow use our knowledge of the difference of $F(t) + \mathcal{C}(y_{(t+1):s}) - F(s)$ to identify whether t is the location of the last changepoint prior to T? The answer is yes! Moreover *if* we can place a condition on t that assures it cannot be the future location of a last changepoint then we can remove t from the minimisation at each step of the proposed method. As we are removing changepoint sequences that cannot feature in the final set of changepoints, this pruning method remains exact.

More formally, the above condition and result can be stated as follows:

Theorem 3.1 *We assume that when introducing a changepoint into a sequence of observations the cost, $\mathcal{C}$, of the sequence reduces. More formally, we assume there exists a constant K such that for all $t < s < T$,*

$$\mathcal{C}(y_{(t+1):s}) + \mathcal{C}(y_{(s+1):T}) + K < \mathcal{C}(y_{(t+1):T}). \tag{8}$$

Then if

$$F(t) + \mathcal{C}(y_{(t+1):s}) + K > F(s) \tag{9}$$

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

Proof. Assume that (9) is true. Now

$$\begin{aligned} & F(t) + \mathcal{C}(y_{(t+1):s}) + \beta + K > F(s) + \beta \\ \implies \quad & F(t) + \mathcal{C}(y_{(t+1):s}) + \beta + K + \mathcal{C}(y_{(s+1):T}) > F(s) + \beta + \mathcal{C}(y_{(s+1):T}) \\ \implies \quad & F(t) + \mathcal{C}(y_{(t+1):T}) + \beta > F(s) + \beta + \mathcal{C}(y_{(s+1):T}) \qquad \text{(by (8).)} \end{aligned}$$

Hence, it follows that t cannot be a future minimum and can be removed from the set of τ for each future step. ■

The intuition behind this result is that if (9) holds then for any $T > s$ the best segmentation with the most recent changepoint prior to T being at s will be better than any which has this most recent changepoint at t. Note that almost all cost functions used in practice satisfy assumption (8). For example, if we take the cost function to be minus the log-likelihood then the constant K would equal zero and if we take it to be minus a penalised log-likelihood then K would equal the maximum penalisation factor.

The condition imposed in Theorem 3.1 for a candidate changepoint, t, to be discarded from future consideration is important as it removes computations that are not relevant for obtaining the final set of changepoints. This condition can be easily implemented into the so-called optimal partitioning method and the pseudo-code is given in Algorithm 2. This shows that at each step in the method the candidate changepoints, t, satisfying the condition are noted and removed from the next iteration. We show in the next section that under certain conditions the computational cost of this method will be linear in the number of observations, as a result we have called this the **P**runed **E**xact **L**inear **T**ime (PELT) method.

PELT Method

Input: A set of data of the form, $(y_1, y_2, \ldots, y_n)$ where $y_i \in \mathbb{R}$.
A measure of fit $\mathcal{C}(.)$ dependent on the data.
A penalty β which does not depend on the number or location of changepoints.

Initialise: Let $n =$ length of data and set $F(0) = -\beta$, $cp(0) = 0$, $pts = 0$.

Iterate for $\tau^* = 1, \ldots, n$

1. Calculate $F(\tau^*) = \min_{\tau \in (0, pts, \tau^* - 1)} \left[F(\tau) + \mathcal{C}(y_{(\tau+1):\tau^*}) + \beta\right]$.
2. Let $\tau^1 = \arg\left\{\min_{0 \leq \tau < \tau^*} \left[F(\tau) + \mathcal{C}(y_{(\tau+1):\tau^*}) + \beta\right]\right\}$.
3. Set $cp(\tau^*) = [cp(\tau^1), \tau^1]$.
4. Set $pts = \arg_\tau \{F(\tau) + \mathcal{C}(y_{\tau+1:\tau^*}) + \beta + K > F(\tau^*) + \beta\}$.

Output the change points recorded in $cp(n)$.

Algorithm 2: PELT Method.

3.1 Linear Computational Cost of PELT

We now investigate theoretically the computational cost of the PELT method. We focus on the most important class of changepoint models and penalties and provide sufficient conditions for the method to have a computational cost that is linear in the number of data points. Similar results will apply more generally and we discuss this briefly at the end of the proof of our main theorem. The case we focus on is the set of models where the data is IID within a segment, segment parameters are independent across segments and the cost function for a segment is minus the maximum log-likelihood value for the data in that segment.

Our result is for the expected computational cost of the method and how this depends on the number of data points we analyse. In order to define this we need an underlying stochastic model for the data generating process. We define such a process over positive-integer time points and then consider analysing the first n data points generated by this process. Our model will assume that the parameters associated with a given segment are IID with density function $\pi(\theta)$ and that given the parameter, θ, for a segment, the data points within the segment are IID with density function $f(y|\theta)$. Our cost function will be based on minus the maximum log-likelihood:

$$\mathcal{C}(y_{(t+1):s}) = -\max_\theta \sum_{i=t+1}^{s} \log f(y_i|\theta).$$

Note that for this loss-function, $K = 0$ in (8). Hence pruning in PELT will just depend on the choice of penalty β.

We also need a stochastic model for the location of the changepoints, via a model for the length of each segment. If the changepoint positions are $\tau_1, \tau_2, \ldots$, then define the segment lengths to be $S_1 = \tau_1$ and for $i = 2, 3, \ldots$, $S_i = \tau_i - \tau_{i-1}$. We assume the S_i are IID copies of a random variable S. Furthermore $S_1, S_2, \ldots$, are independent of the parameters associated with the segments.

Theorem 3.2 *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:

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

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)

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

(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 $L < \infty$.

Proof. See Appendix. □

Conditions (A1) and (A2) of Theorem 3.2 are weak technical conditions. For example, general asymptotic results for maximum likelihood estimation would give $B_n = \mathcal{O}_p(1)$, and (A1) is a slightly stronger condition which is controlling the probability of B_n taking values that are $\mathcal{O}(n^{1/2})$ or greater.

The other two conditions are more important. Condition (A3) is needed to control the probability of large segments. One important consequence of (A3) is that the expected number of changepoints will increase linearly with n. Finally condition (A4) is a natural one as it is required for the expected penalised likelihood value obtained with the true changepoint and parameter values is greater than the expected penalised likelihood value if we fit a single segment to the data with segment parameter θ^*.

4 Simulation and Data Examples

We now turn to two examples to demonstrate how the methods presented in Section 3 can be used in practice. Alongside the examples we present the results of a simulation study which highlights the attractive properties of the PELT method. The examples demonstrate the method's applicability in detecting (i) changes in variance within the Dow Jones Index returns and (ii) changes in mean and variance related to the analysis of the human chromosome. In each case we aim to show that the PELT method can be applied to a variety of data types with little effort and consistently produces quick and exact results.

Before proceeding further, we note that the Segment Neighbourhoods method produces identical results to both the optimal partitioning and PELT methods as they are all exact in nature. The only difference between the three approaches is their computational efficiency.

4.1 Changes in Variance within Normally Distributed Data

In general detecting changes in variance is more challenging than detecting changes in mean and is important in financial and environmental applications (see for example Killick et al. (2010)). In the following subsections we specifically consider multiple changes in variance within data sets that are assumed to follow a Normal distribution with a constant mean. We begin by showing the power of the PELT method in detecting multiple changes via a simulation study. This is followed by an application to the Dow Jones Index returns.

4.1.1 Simulation Study

This section demonstrates both the computational and statistical benefits of the PELT method. The computational benefit is the decrease from $\mathcal{O}(Qn^2)$ (Segment Neighbourhood) or $\mathcal{O}(n \log n)$ (Binary Segmentation) to linear computational time, $\mathcal{O}(n)$ for the PELT method introduced in Section 3. In addition the PELT method has a lower overall cost as it is an exact minimisation.

In order to demonstrate these theoretical properties in practice we shall construct sets of simulated data on which we shall run various multiple changepoint methods. It is reasonable to set the cost function, $\mathcal{C}$ as the negative log-likelihood. The cost of a segment is then

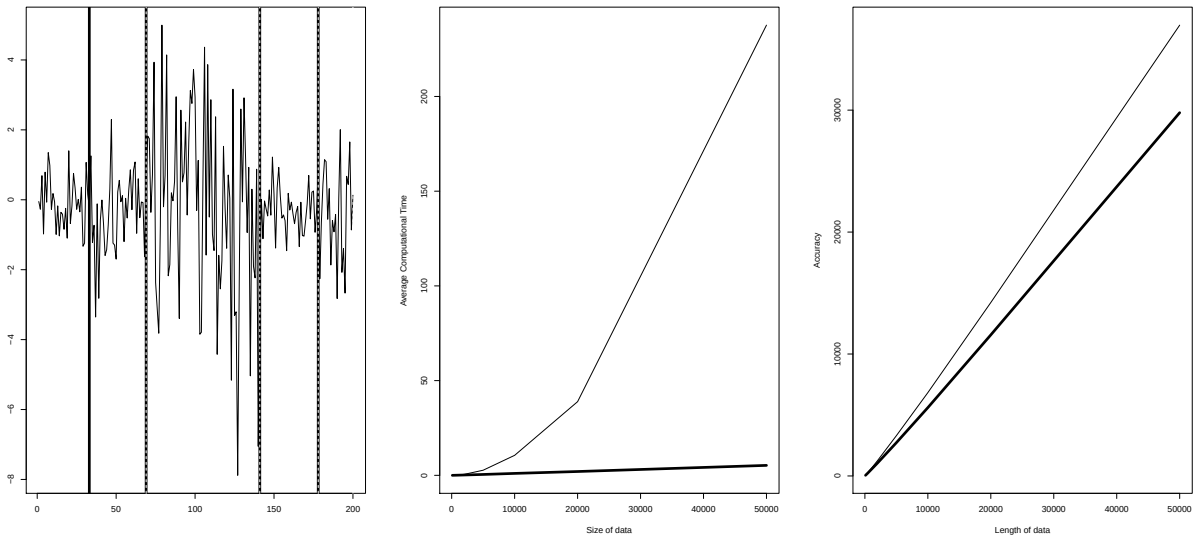
$$\mathcal{C}(y_{(\tau_{i-1}+1):\tau_i}) = (\tau_i - \tau_{i-1}) \left(\log(2\pi) + \log \left(\sum_{j=\tau_{i-1}+1}^{\tau_i} (y_j - \mu)^2 \right) + 1 \right). \tag{10}$$

Our simulated data will consist of 9 scenarios with varying lengths, n =(100, 200, 500, 1000, 2000, 5000, 10000, 20000, 50000) and each scenario will contain $\frac{n}{50}$ changepoints. These changepoints are distributed uniformly across $(2, n-2)$ with the only constraint being that there must be at least 30 observations between changepoints. Within each of these 9 scenarios we have 1,000 repetitions where the mean is fixed at 0 and the variance parameters for each segment are assumed to have a Log-Normal distribution with mean 0 and standard deviation $\frac{\log(10)}{2}$. These parameters are chosen so that 95% of the simulated variances are within the range $\left[\frac{1}{10}, 10\right]$. An example realisation is shown in Figure 1(a).

Using the simulation study structure above, Figure 1(b) shows the average computational time over the 1,000 repetitions. This figure clearly shows that the PELT method has linear computational time and is a vast improvement over the Optimal Partitioning method (which in itself is a vast improvement over the Segment Neighbourhoods method).

The PELT method is exact and so produces lower overall likelihood values than the Binary Segmentation method which is approximate. The degree of decrease in likelihood values depends on many factors but due to the nature of the methods, an equal or decreased value of the likelihood is guaranteed.

An alternative measure of fit is the average cumulative difference between the estimated and true variance



(a) (b) (c)

Figure 1: (a) Plot showing a realisation of multiple changes in variance. The true changepoint locations are shown by thick black lines and the dotted lines show changes identified by both the Binary Segmentation and the PELT method. (b) Average Computational Time for a change in variance (thin line: Optimal Partitioning, thick line: PELT). (c) Average cumulative difference between parameter estimates and true values for PELT (thick line) and Binary Segmentation (thin line) methods.

at each data point. The average cumulative difference for a single dataset is defined as

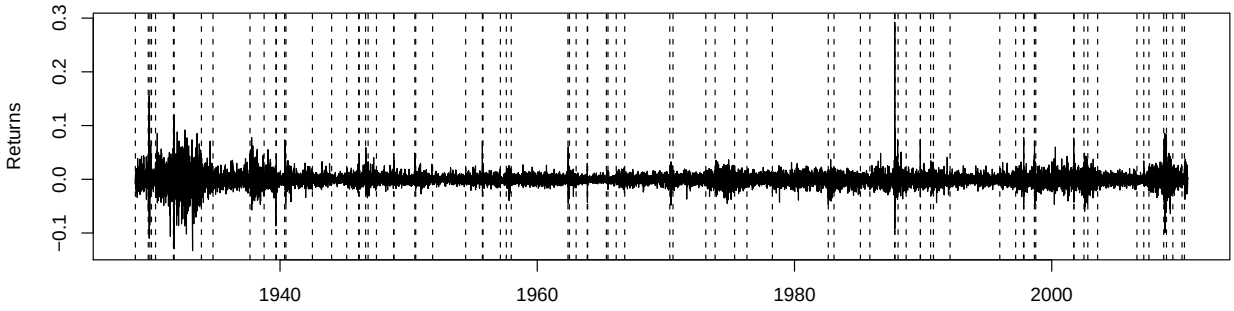
$$\text{ACD} = \frac{\sum_{i=1}^{n} |\hat{\theta}_i - \theta_i|}{n}, \tag{11}$$

where $\hat{\theta}_i$ is the estimated parameter and θ_i is the true parameter at data point y_i. Figure 1(c) shows the average cumulative difference between the estimated and true variance at each data point over the 1,000 data sets. There are separate lines for the PELT and Binary segmentation estimates which shows that the average difference is small when n is small but shows a very large difference between the method estimates as n becomes large. These values use the same penalty of $2 \log n$ for both methods.

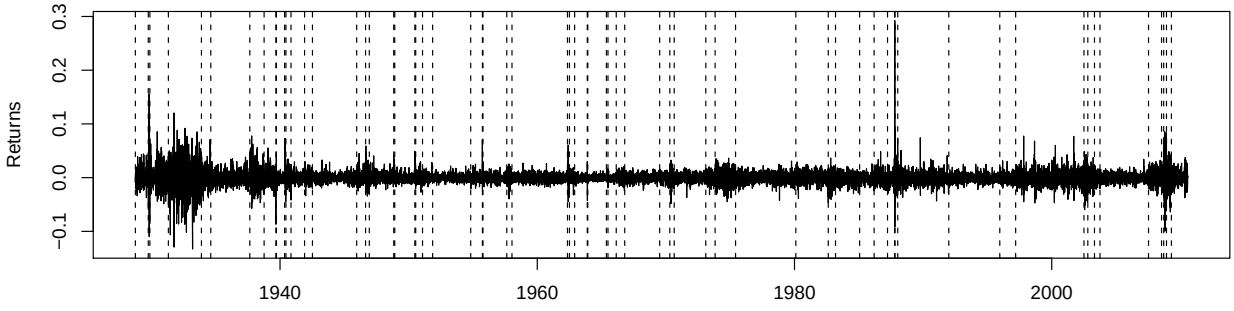
This subsection has shown that the theoretical properties of the optimal partitioning and PELT methods can be shown in practice to identify changes in variance within data that follow a Normal distribution. The following subsection illustrates the above analysis techniques for the Dow Jones Index daily returns.

4.1.2 Application to Dow Jones Index

The Dow Jones index has been studied in part by many authors including Hsu (1977) and Berkes et al. (2006). Following these authors, Figure 2 shows the daily closing returns from 1st October 1928 to 30th July 2010 defined as $R_t = c_{t+1}/c_t - 1$, where c_t is the closing price on day t. Previous authors have modelled the Dow Jones daily returns as a change in variance which seems reasonable from Figure 2. We also take this approach to analysing the data and set the cost function as the negative log likelihood. As in previous analyses we shall assume that the returns follow a Normal distribution with constant mean and piecewise stationary (unknown) variance. The changes in variance identified by the PELT and Binary Segmentation methods using the SIC penalty are shown as vertical lines in Figure 2. These results are consistent with previous studies on portions of the data; the 3rd week in March 1973 found by Hsu (1977) and October 1987 found by Berkes et al. (2006). The PELT method identifies 82 changepoints and the Binary Segmentation method identifies 65. The log-likelihood + penalty for the PELT method is -135849.2 and for Binary Segmentation is -135549.1. For this data, the decrease in speed from the Optimal Partitioning method to the PELT method is by a factor of 14.



(a)



(b)

Figure 2: The Dow Jones index daily returns data with changepoints marked using (a) PELT and (b) Binary Segmentation methods with the SIC penalty.

4.2 Changes in Mean and Variance within Normally Distributed Data

Traditionally the changepoint problem is concerned with a change in one underlying parameter. This may not always be appropriate for a given dataset and so methods for detecting changes in multiple parameters at a single changepoint were constructed. The simplest change in multiple parameters is a change in both the mean and variance of data that are assumed to follow a Normal distribution. This has been considered by many authors including Horvath (1993) and Picard et al. (2005). Below we aim to show how the optimal partitioning and PELT methods can be applied to multiple parameter changepoint problems just as easily as the single parameter case. This subsection is structured as in Section 4.1 with a simulation study to show the properties of the optimal partitioning and PELT methods under known conditions and then concluding with an analysis of a real dataset.

4.2.1 Simulation Study

The simulation study here will be constructed in a similar way to that of Section 4.1.1. It is assumed that the data follow a Normal distribution with mean and variance depending on the segment, as previously we shall take the cost function as twice the negative of the log likelihood. For a specific segment the cost is

$$\mathcal{C}(y_{(\tau_{i-1}+1):\tau_i}) = (\tau_i - \tau_{i-1})\left(\log(2\pi) + \log\left(\sum_{j=\tau_{i-1}+1}^{\tau_i}\left(y_j - \frac{\sum_{k=\tau_{i-1}+1}^{\tau_i} y_k}{(\tau_i - \tau_{i-1})}\right)^2\right) + 1\right). \qquad (12)$$

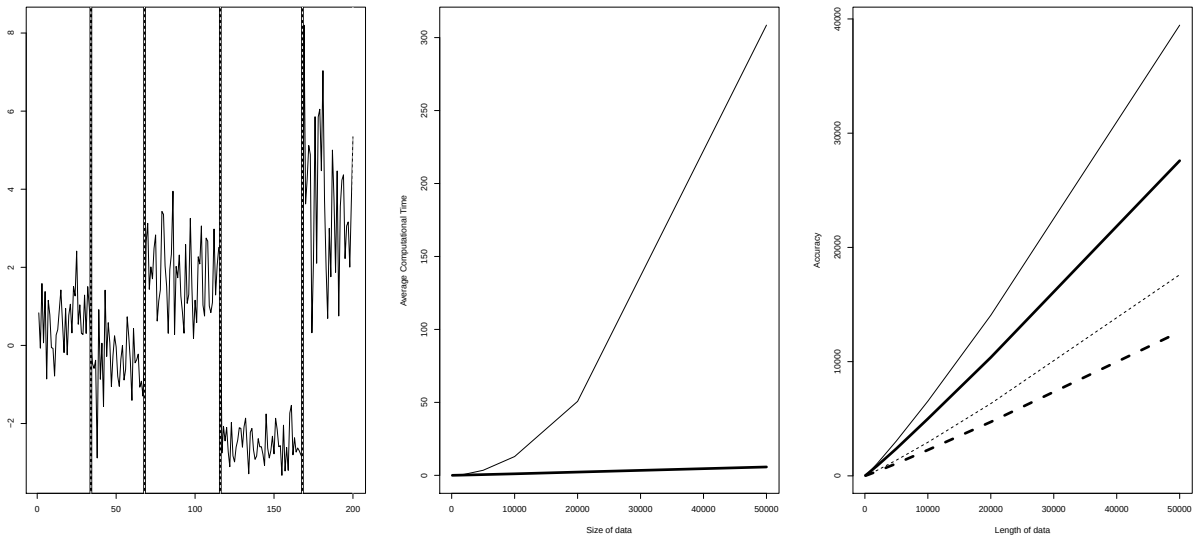
The simulated data will consist of 9 scenarios with varying lengths, $n = c(100, 200, 500, 1000, 2000, 5000, 10000, 20000, 50000)$ and each scenario will contain $\frac{n}{50}$ changepoints. These changepoints are distributed uniformly across $(2, n-2)$ with the only constraint being that there must be at least 30 observations between changepoints. Within each of these 9 scenarios we have 1,000 repetitions where the mean for each segment is a realisation from a Normal distribution with mean 0 and standard deviation 2.5 and the variance parameters for each segment are a realisation from a Log-Normal distribution with mean 0 and standard deviation $\frac{\log(10)}{2}$. These parameters are chosen so that 95% of the simulated means are within the range [-5,5] and 95% of the simulated variances are within the range $\left[\frac{1}{10}, 10\right]$. A realisation is shown in Figure 3(a), where all the changepoints are easily detected.

Figure 3(b) shows the average computational time for a change in mean and variance over various lengths of data. As for the change in variance simulation study, the PELT method runs in linear time. Figure 3(c) shows the average cumulative difference between the estimated and true parameters at each data point, as defined in (11), for the 1,000 data sets. There are separate lines for the PELT and Binary segmentation estimates and for the mean and variance parameters. Just as was shown with the change in variance simulations, the average difference is small when n is small but the difference increases as n becomes large. The penalty used to obtain these results is $4\log n$.

4.2.2 Application to Human Chromosome 1

The Human Genome project has been highly successful in sequencing the DNA of humans. The interest is now in studying how similar the DNA sequence of difference organisms are and if there is reasoning behind the structures that the project has already found. This section considers the proportion of C+G content within part of Human Chromosome 1, see Oliver et al. (2004) for background to the data. Our approach models each segment as following a Normal distribution with its own mean and variance as in the simulation study above. A similar model has been fitted using Binary Segmentation (Oliver et al., 2004), though they work on the raw sequence data and introduce an extra step to filter out fine-scale variation in C+G content prior to deciding whether to introduce a changepoint. We filter out this fine-scale variation by analysing the C+G content in 3kb windows (as in Fearnhead and Vasileiou, 2009).

We summarize the data by partitioning the 23 Mb region into 3.0 kb long windows. For each window we retain the proportion of G and C content within that window. Figure 4 shows the proportion of C+G



(a) (b) (c)

Figure 3: (a) Plot showing a realisation of multiple changes in mean and variance. The true changepoint locations are shown by thick black lines and the dotted lines show changes identified by both the Binary Segmentation and the PELT method. (b) Average Computational Time for a change in mean and variance (thin line: Optimal Partitioning, thick line: PELT). (c) Average cumulative difference between parameter estimates (full line: variance, dotted line: mean) and true values for PELT (thick line) and Binary Segmentation (thin line) methods.

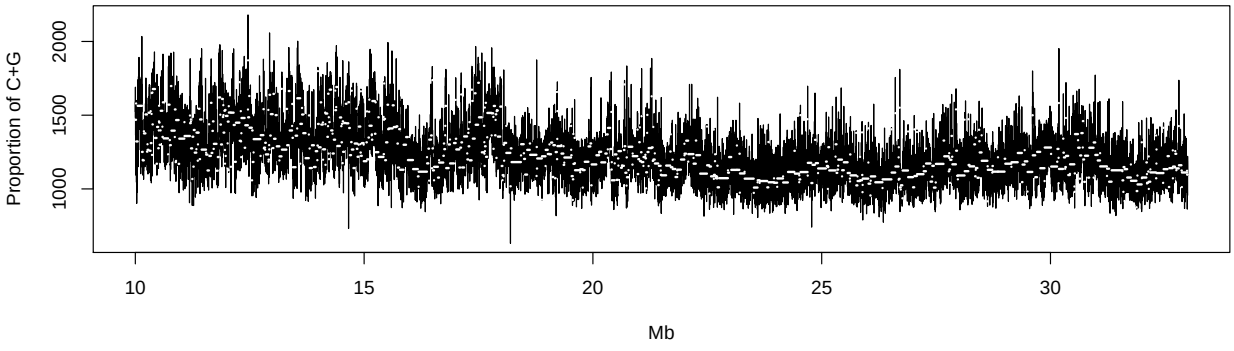
content for part of Human Chromosome 1 along with the changepoints identified by the PELT and Binary Segmentation methods. The changepoints are identified using a penalty of 14. There is a large difference in the number of changepoints identified by the two methods with the PELT method identifying 805 changes and the binary segmentation method identifying 378. The log-likelihood + penalty for the PELT method is 284497 and for Binary Segmentation is 289253. For this data, the decrease in speed from the Optimal Partitioning method to the PELT method is by a factor of 47.

5 Discussion

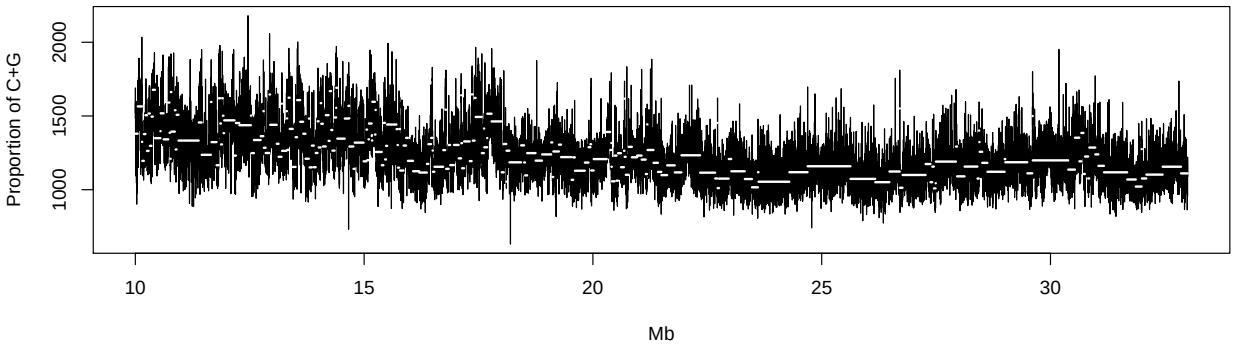
In this paper we have presented the PELT method; an alternative exact multiple changepoint method that is both computationally efficient and versatile in its application. It has been shown that under certain conditions the computational efficiency is $\mathcal{O}(n)$. The simulation study and real data examples demonstrate that the assumptions and conditions are not restrictive and a wide class of cost functions can be implemented. In a world of increasing lengths of data, a reduction in computational time from the $\mathcal{O}(Qn^2)$ of the Segment Neighbourhood method to the potentially $\mathcal{O}(n)$ of the PELT method is important.

The approximate Binary Segmentation method of $\mathcal{O}(n \log n)$ is typically used instead of the computationally expensive Segment Neighbourhood method. However, we have shown that this no longer need be the case since, for specific cost functions, the PELT method produces exact results in potentially $\mathcal{O}(n)$ time. Not only is the order of computational time of the PELT method smaller but the overall value of the cost function can be reduced using this approach.

A recent paper, Rigaill (2010), presents an alternative approach to detecting changepoints, which is exact and is shown empirically to have a linear computational cost. While also based on pruning a dynamic programming algorithm, their approach involves working with a cost function defined on the joint space of the most recent changepoint and the parameter value for the most recent segment. The idea is that the optimal most recent changepoint depends on the current segment parameter and often there are just a small number of intervals of the parameter space for which the optimal value of the most recent changepoint



(a)



(b)

Figure 4: The proportion of C+G for 23 Mb of Human Chromosome 1 with changepoints marked using the (a) PELT method and (b) Binary Segmentation, using a penalty of 14.

will vary. Thus the algorithm aims to store and update these intervals, together with the optimal value of the most recent changepoint and the functional form of the cost as the parameter value varies across this interval. We believe that our method is both easier to implement, being a simple modification of existing dynamic programming algorithms, and will have good computational performance more generally (such as scaling to models with multiple parameters per segment).

Code implementing PELT is available within the `R` library `changepoint`, and is available on request from the first author.

6 Acknowledgement

R. Killick acknowledges financial support from Shell Research Limited and the Engineering and Physical Sciences Research Council (EPSRC).

Appendix: Proof of Theorem 3.2

The proof of Theorem 3.2 has two parts. Firstly we show that the expected computational cost is bounded by nL_n, where L_n is the expected number of changepoint-times stored (i.e. not pruned) when analysing the nth observation. Secondly we show that under assumptions (A1)–(A4) $\lim_{n\to\infty} L_n < \infty$.

Introduce notation $G(y_{(s+1):t})$ to be the minimum value of the cost function (3) for data $y_{(s+1):t}$. So, as previously defined, $F(t) = G(y_{1:t})$. We will consider pruning on the more stringent condition that changepoint $t-j$ is removed at iteration t if

$$\mathcal{C}(y_{(t-j+1):t}) > G(y_{(t-j+1):t}). \tag{13}$$

That fact that this is a more stringent condition, comes from (9), noting that $F(t) \leq F(t-j)+G(y_{(t-j+1):t})$ as the latter is the smallest overall cost for segmentations that include a changepoint at $t-j$, and remembering that for our choice of $\mathcal{C}(\cdot)$, $K = 0$. Furthermore the computational cost of PELT will be bounded above by the method which prunes using this condition.

Assume we are pruning with condition (13). For a positive integer $j \leq t$, let $I_{t,j}$ be an indicator of whether a changepoint at time $t-j$ is stored after processing the observation at time t. The overall computational cost of processing the observation at time $(t+1)$ is $1 + \sum_{j=1}^{t} I_{t,j}$ Now as the data-generating process is time-invariant, and our condition (13) just depends on data $y_{(t-j+1):t}$, we have $\mathbb{E}\left(I_{t,j}\right) = E_j$, independent of t. So the expected computational cost is bounded by nL_n where

$$L_n = 1 + \sum_{j=1}^{n-1} E_j,$$

the expected computational cost of processing the last observation.

Now we define L as the limit of L_n as $n \to \infty$. We need to show this is finite. If it is, the computational cost of a method using (13) to prune will have a computational cost that is linear in n, and hence so will PELT. We will do this by showing that E_j decays to 0 sufficiently quickly as $j \to \infty$.

Now (by choosing $t = j$) E_j is the probability that $I_{j,j} = 1$, which is that a changepoint at time 0 is not pruned after processing the jth observation. For $I_{j,j} = 1$ we need

$$\mathcal{C}(1, j) \leq F(j),$$

where $\mathcal{C}(1, j)$ is the cost associated with assuming a single segment for observation $y_{1:j}$ and $F(j)$ is the minimum cost possible for segmenting $y_{1:j}$. Now we will define m_j to be the number of actual changepoints before time j, and $\tau_1, \ldots, \tau_{m_j}$ their positions. Furthermore we define $\tau_0 = 0$ and, with a slight abuse of notation, $\tau_{m_j+1} = j$. Then

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

So

$$E_j \leq \Pr\left(\mathcal{C}(1,j) \leq \sum_{i=1}^{m_j+1}\left[\mathcal{C}(y_{(\tau_{i-1}+1):\tau_i}) + \beta\right]\right).$$

Now define θ_i to be the value of the parameter associated with the true segment of observation i; and $\tilde{\theta}_i$ the value of the maximum likelihood estimate of the parameter associated with the true segment of observation i:

$$\tilde{\theta}_i = \arg\max \sum_{k=\tau_{l-1}+1}^{\tau_l} \log f(y_k|\theta),$$

where l is defined so that $\tau_{l-1} < i \leq \tau_l$.

Now $\mathcal{C}(1,j) = -\sum_{i=1}^{j} \log f(y_i|\hat{\theta}_j)$, where $\hat{\theta}_j$ is defined in Theorem 3.2 to be the maximum likelihood estimate θ for data $y_{1:j}$ under an assumption of a single segment. By the definition above, we also have

$$\sum_{i=1}^{m_j+1} \mathcal{C}(y_{(\tau_{i-1}+1):\tau_i}) = -\sum_{i=1}^{j} \log f(y_i|\tilde{\theta}_i)$$

So we can re-write

$$\overbrace{\mathcal{C}(1,j) - \sum_{i=1}^{m_j+1}\left[\mathcal{C}(y_{(\tau_{i-1}+1):\tau_i}) + \beta\right]}^{A_j} = \overbrace{\sum_{i=1}^{j}[\log f(y_i|\theta^*) - \log f(y_i|\hat{\theta}_j)]}^{B_j}$$

$$\overbrace{\sum_{i=1}^{j}[\log f(y_i|\theta_i) - \log f(y_i|\theta^*)]}^{D_j} - (m_j+1)\beta + \overbrace{\sum_{i=1}^{j}[\log f(y_i|\tilde{\theta}_i) - \log f(y_i|\theta_i)]}^{R_j}. \quad (14)$$

First note that $R_j \geq 0$. So $E_j = \Pr(A_j \leq 0) \leq \Pr(B_j + D_j \leq 0)$. Now we will bound this probability using Markov's inequality.

By (A1), and using that the expected number of changepoints is related to the expected segment length, $\mathbb{E}(M_j) = j/\mathbb{E}(S) + o(j)$ (elementary renewal theorem), we have

$$\mathbb{E}(B_j + D_j) = \mathbb{E}(B_j) + \mathbb{E}(D_j) = \mathbb{E}\left(\sum_{i=1}^{j}[\log f(Y_i|\theta_i) - \log f(Y_i|\theta^*)]\right) - \beta\frac{j}{\mathbb{E}(S)} + o(j)$$

Thus, using (A4), we have that there exists $c > 0$ such that for sufficiently large j

$$\mathbb{E}(B_j + D_j) > cj.$$

Now let $B_j^* = B_j - \mathbb{E}(B_j)$ and $D_j^* = D_j - \mathbb{E}(D_j)$. We now consider $\mathbb{E}\left((B_j^* + D_j^*)^4\right)$, and show that this is $\mathcal{O}(j^2)$. Now the Minkowski Inequality gives that

$$\mathbb{E}\left((B_j^* + D_j^*)^4\right) \leq \left[\mathbb{E}\left((B_j^*)^4\right)^{1/4} + \mathbb{E}\left((D_j^*)^4\right)^{1/4}\right]^4.$$

Now by (A1) $\mathbb{E}\left((B_j^*)^4\right) = \mathcal{O}(j^2)$, so we need only to show that $\mathbb{E}\left((D_j^*)^4\right)$ is $\mathcal{O}(j^2)$, in order for $\mathbb{E}\left((B_j^* + D_j^*)^4\right)$ to be $\mathcal{O}(j^2)$.

Define

$$Z_i = \log f(Y_i|\theta_i) - \log f(Y_i|\theta^*) - \mathbb{E}(\log f(Y_i|\theta_i) - \log f(Y_i|\theta^*)),$$

so $D_j^* = \sum_{i=1}^{j} Z_i$. Note that Z_i has the same distribution for all i, and we will let Z denote a further random variable with this distribution. We have

$$\begin{aligned}\mathbb{E}\left(\left(D_j^*\right)^4\right) &= \mathbb{E}\left(\left[\sum_{i=1}^{j} Z_i\right]^4\right) \\ &= \sum_{i_1=1}^{j}\sum_{i_2=1}^{j}\sum_{i_3=1}^{j}\sum_{i_4=1}^{j}\mathbb{E}\left(Z_{i_1}Z_{i_2}Z_{i_3}Z_{i_4}\right).\end{aligned}$$

If we condition on the position of the changepoints we have that, by independence across segments:

$$\mathbb{E}\left(Z_{i_1}Z_{i_2}Z_{i_3}Z_{i_4}\right) \leq \begin{cases} \mathbb{E}\left(Z^4\right) & \text{if each segment contains an even number of } i_1,\ldots,i_4 \\ 0 & \text{otherwise.}\end{cases}$$

Thus we get a bound on the fourth moment of D^* in terms of the expected value of the segment lengths of our changepoint process. Denote $S_i^{(j)} = \min\{S_i, j\}$, and note that for each segment to contain an even number of $i_1,\ldots,i_4$ we need one segment to contain all four values, or two segments to contain two each. If we know $S_1,\ldots,S_j$ this involves at most

$$3\sum_{i=1}^{m_j+1}\sum_{k=1,k\neq i}^{m_j+1}\left(S_i^{(j)}\right)^2\left(S_k^{(j)}\right)^2 + \sum_{i=1}^{m_j+1}\left(S_i^{(j)}\right)^4$$

possible combinations of $i_1,\ldots,i_4$. Thus taking expectations with respect to $S_1,\ldots,S_n$ we get:

$$\begin{aligned}\mathbb{E}\left(\left(D_j^*\right)^4\right) &\leq \mathbb{E}\left(Z^4\right)\mathbb{E}\left(3\sum_{i=1}^{m_j+1}\sum_{k=1,k\neq i}^{m_j+1}\left(S_i^{(j)}\right)^2\left(S_k^{(j)}\right)^2 + \sum_{i=1}^{m_j+1}\left(S_i^{(j)}\right)^4\right) \\ &\leq \mathbb{E}\left(Z^4\right)\mathbb{E}\left(3\sum_{i=1}^{j}\sum_{k=1,k\neq i}^{j}\left(S_i^{(j)}\right)^2\left(S_k^{(j)}\right)^2 + \sum_{i=1}^{j}\left(S_i^{(j)}\right)^4\right) \\ &\leq \mathbb{E}\left(Z^4\right)\left\{3\mathbb{E}\left(\sum_{i=1}^{j}\left(S_i^{(j)}\right)^2\right)\mathbb{E}\left(\sum_{k=1}^{j}\left(S_k^{(j)}\right)^2\right) + \mathbb{E}\left(\sum_{i=1}^{j}\left(S_i^{(j)}\right)^4\right)\right\} \\ &\leq \mathbb{E}\left(Z^4\right)\left[3j^2\mathbb{E}\left(S^2\right) + j^2\mathbb{E}\left(S^3\right)\right].\end{aligned}$$

The last inequality uses that $\mathbb{E}\left(\left(S_i^{(j)}\right)^4\right) \leq j\mathbb{E}\left(\left(S_i^{(j)}\right)^3\right) \leq j\mathbb{E}\left(S_i^3\right)$.

This shows that there exists a $K < \infty$ such that $\mathbb{E}\left(\left(B_j^* + D_j^*\right)^4\right) < Kn^2$. Now using Markov's inequality we have, for j large enough that $\mathbb{E}\left(B_j + D_j\right) > cj$

$$E_j \leq \Pr(B_j + D_j \leq 0) \leq \Pr(|B_j^* + D_j^*| \geq \mathbb{E}\left(B_j + D_j\right)) \leq \frac{\mathbb{E}\left(\left(B_j^* + D_j^*\right)^4\right)}{\left[\mathbb{E}\left(B_j + D_j\right)\right]^4} \leq \frac{Kj^2}{c^4 j^4}.$$

Thus we have $E_j = \mathcal{O}(j^{-2})$, and hence $L = \lim_{n\to\infty}\sum_{j=1}^{n} E_j$ is finite, as required. □

Discussion

The basic idea of the proof is to show that the probability of pruning $t - j$ as the value of the most recent changepoint before t goes to zero sufficiently quickly. This in turn required considering the cost function we are trying to minimise, and considering the distribution of the difference of this cost function assuming no changepoint between $t - j$ and t, and the cost associated with the true changepoint positions between

$t - j$ and t. For more general cost functions and changepoint models, if we can show that the expected value of this difference decreases linearly with j, but its fourth moment increases only quadratically with j, then the same proof will show that PELT has a linear computational cost.

By a similar argument, if we relax conditions (A2) and (A3) to involve just finite second moments, we can use Chebyshev's inequality to prove that $L_n = \mathcal{O}(\log n)$, and hence the computational cost of PELT is at worse $\mathcal{O}(n \log n)$. Also, note that the term R_j that we ignored has an expectation that is $\mathcal{O}(j)$. If we do not ignore R_j in the above argument, we will need a slightly weaker version of (A4) to prove that PELT has linear computational cost.

References

Aggarwal, R., C. Inclan, and R. Leal (1999). Volatility in emerging stock markets. *The Journal of Financial and Quantitative Analysis 34*(1), 33–55.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*(6), 716–723.

Andreou, E. and E. Ghysels (2002). Detecting multiple breaks in financial market volatility dynamics. *Journal of Applied Econometrics 17*(5).

Auger, I. E. and C. E. Lawrence (1989). Algorithms for the optimal identification of segment neighborhoods. *Bulletin of Mathematical Biology 51*(1), 39–54.

Berkes, I., L. Horvath, P. Kokoszka, and Q. Shao (2006). On discriminating between long-range dependence and changes in mean. *The Annals of Statistics 34*, 1140–1165.

Birge, L. and P. Massart (2007). Minimal Penalties for Gaussian Model Selection. *Probability Theory and Related Fields 138*, 33–73.

Braun, J. V., R. K. Braun, and H. G. Muller (2000). Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation. *Biometrika 87*(2), 301–314.

Chen, J. and A. K. Gupta (2000). *Parametric statistical change point analysis*. Birkhauser.

Eckley, I. A., P. Fearnhead, and R. Killick (2010). Analysis of changepoint models. In D. Barber, T. Cemgil, and S. Chiappa (Eds.), *Bayesian Time Series Models*. Cambridge University Press.

Fearnhead, P. and D. Vasileiou (2009). Bayesian analysis of isochores. *Journal of the American Statistical Association 104*, 132–141.

Fernandez, V. (2004). Detection of breakpoints in volatility. *Estudios de Administracion 11*(1), 1–38.

Guyon, X. and J.-f. Yao (1999). On the underfitting and overfitting sets of models chosen by order selection criteria. *Journal of Multivariate Analysis 70*(2), 221 – 249.

Horvath, L. (1993). The maximum likelihood method of testing changes in the parameters of normal observations. *Annals of Statistics 21*(2), 671–680.

Hsu, D. A. (1977). Tests for variance shift at an unknown time point. *Applied Statistics 26*(3), 279–284.

Inclan, C. and G. C. Tiao (1994). Use of cumulative sums of squares for retrospective detection of changes of variance. *Journal of the American Statistical Association 89*(427), 913–923.

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

Killick, R., I. A. Eckley, P. Jonathan, and K. Ewans (2010). Detection of changes in the characteristics of oceanographic time-series using statistical change point analysis. *Ocean Engineering 37*(13), 1120–1126.

Oliver, J. L., P. Carpena, M. Hackenberg, and P. Bernaola-Galvan (2004). Isofinder: computational prediction of isochores in genome sequences. *Nucleic Acid Research 32*, W287–W292.

Olshen, A. B., E. S. Venkatraman, R. Lucito, and M. Wigler (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. . *Biostatistics 5*, 557–72.

Picard, F., S. Robin, M. Lavielle, C. Vaisse, and J. J. Daudin (2005). A statistical approach for array cgh data analysis. *Bioinformatics 6*.

Rigaill, G. (2010, April). Pruned dynamic programming for optimal multiple change-point detection. *ArXiv e-prints*.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*(2), 461–464.

Scott, A. J. and M. Knott (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics 30*(3), 507–512.

Sen, A. and M. S. Srivastava (1975). On tests for detecting change in mean. *The Annals of Statistics 3*(1), 98–108.

Yao, Y. (1984). Estimation of a noisy discrete-time step function: bayes and empirical bayes approaches. *The Annals of Statistics 12*, 1434–1447.