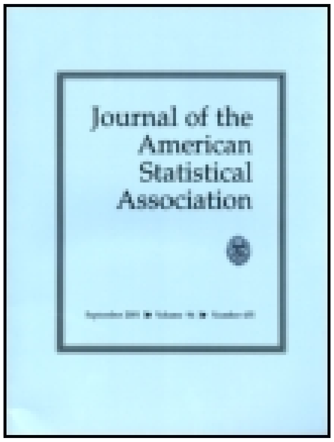

Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

Optimal Detection of Changepoints With a Linear Computational Cost

R. Killick [a] , P. Fearnhead [a] & I. A. Eckley [a]
[a] Department of Mathematics and Statistics , Lancaster University , Lancaster , UK

Accepted author version posted online: 17 Oct 2012.Published online: 21 Dec 2012.

To cite this article: R. Killick , P. Fearnhead & I. A. Eckley (2012) Optimal Detection of Changepoints With a Linear Computational Cost, Journal of the American Statistical Association, 107:500, 1590-1598, DOI: 10.1080/01621459.2012.737745

To link to this article: http://dx.doi.org/10.1080/01621459.2012.737745

Optimal Detection of Changepoints With a Linear Computational Cost

R. KILLICK, P. FEARNHEAD, and I. A. ECKLEY

In this article, we consider the problem of detecting multiple changepoints in large datasets. Our focus is on applications where the number of changepoints will increase as we collect more data: for example, in genetics as we analyze 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 minimizing a cost function over possible numbers and locations of changepoints. This includes several established procedures for detecting changing points, such as penalized likelihood and minimum description length. 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 favorably 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 exact methods. We also compare with the binary segmentation algorithm for identifying changepoints, showing that the exactness of our approach can lead to substantial improvements in the accuracy of the inferred segmentation of the data. This article has supplementary materials available online.

KEY WORDS: Dynamic programming; PELT; Segmentation; Structural change.

1. INTRODUCTION

As increasingly longer datasets are being collected, more and more applications require the detection of 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, Braun, and Muller 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, Inclan, and Leal 1999; Andreou and Ghysels 2002; Fernandez 2004). Typically, such series will contain several changepoints. There is therefore a growing need to be able to search for such changes efficiently. It is this search problem that we consider in this article. 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 analyze 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 (BS) proposed by Scott and Knott (1974) is arguably the most widely used changepoint search method. It is approximate in nature with an $\mathcal{O}(n \log n)$ 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 are based on dynamic programming. For example, the segment neighborhood (SN) method proposed by Auger and Lawrence (1989) is $\mathcal{O}(Qn^2)$, where Q is the maximum number of changepoints one wishes 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 provided by the optimal partitioning (OP) approach of Jackson et al. (2005). As we describe in Section 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 to search for changepoints, which is exact and, under mild conditions, has a computational cost that is linear in the number of data points: the pruned exact linear time (PELT) method. This approach is based on the algorithm of Jackson et al. (2005), but involves a pruning step within the dynamic program. This pruning reduces the computational cost of the method, but does not affect the exactness of the resulting segmentation. It can be applied to find changepoints under a range of statistical criteria such as penalized likelihood, quasi-likelihood (Braun, Braun, and Muller 2000), and cumulative sum of squares (Inclan and Tiao 1994; Picard et al. 2011). In simulations, we compare PELT with both BS and OP. We show that PELT can be calculated orders of magnitude faster than OP, particularly for long datasets. While asymptotically PELT can be quicker, we find that, in practice, BS is quicker on the examples we consider, and we believe this would be the case in almost all applications. However, we show that PELT leads to a substantially more accurate segmentation than BS.

The article is organized 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. The efficiency and accuracy of the PELT method are demonstrated in Section 4. In particular, we demonstrate the methods' performance on large

R. Killick is Senior Research Associate, Department of Mathematics and Statistics, Lancaster University, Lancaster, UK (E-mail: *r.killick@lancs.ac.uk*). P. Fearnhead is Professor, Department of Mathematics and Statistics, Lancaster University, Lancaster, UK (E-mail: *p.fearnhead@lancs.ac.uk*). I. A. Eckley is Senior Lecturer, Department of Mathematics and Statistics, Lancaster University, Lancaster, UK (E-mail: *i.eckley@lancs.ac.uk*). The authors are grateful to Richard Davis and Alice Cleynen for providing the Auto-PARM and PDPA software, respectively. Part of this research was conducted while R. Killick was a jointly funded Engineering and Physical Sciences Research Council (EPSRC)/Shell Research Ltd. graduate student at Lancaster University. Both I. A. Eckley and R. Killick also gratefully acknowledge the financial support of the Research Councils UK Energy Programme grant number EP/I016368/1.

Journal of the American Statistical Association
December 2012, Vol. 107, No. 500, Theory and Methods
DOI: 10.1080/01621459.2012.737745

datasets coming from oceanographic (Section 4.2) and financial (online supplementary materials) applications. Results show the speed gains over other exact search methods and the increased accuracy relative to approximate search methods such as BS. The article concludes with a discussion in Section 5.

2. BACKGROUND

Changepoint analysis can, loosely speaking, be considered to be the identification of points within a dataset 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 identify multiple changepoints is to minimize

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

Here $\mathcal{C}$ is a cost function for a segment and $\beta f(m)$ is a penalty to guard against overfitting. Twice the negative log-likelihood is a commonly used cost function in the changepoint literature (see, e.g., Horvath 1993; Chen and Gupta 2000), although other cost functions such as quadratic loss and cumulative sums are also used (e.g., Inclan and Tiao 1994; Rigaill 2010), or those based on both the segment log-likelihood and the length of the segment (Zhang and Siegmund 2007). Turning to the choice of penalty, in practice, by far the most common choice is one which is linear in the number of changepoints, that is, $\beta f(m) = \beta m$. Examples of such penalties include Akaike's information criterion (AIC; Akaike 1974) ($\beta = 2p$) and Schwarz information criterion (SIC, also known as BIC; Schwarz 1978) ($\beta = p \log n$, where p is the number of additional parameters introduced by adding a changepoint). The PELT method that we introduce in Section 3 is designed for such linear cost functions. Although linear cost functions are 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. In Section 3.2, we show how PELT can be applied to some of these alternative choices.

The remainder of this section describes two commonly used methods for multiple changepoint detection: BS (Scott and Knott 1974) and SN (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 that we propose here. For notational simplicity, we describe all the algorithms (including PELT) assuming that the minimum segment length is a single observation, that is, $\tau_{i-1} - \tau_i \geq 1$. A larger minimum segment length is easily implemented when appropriate; see, for example, Section 4.

2.1 Binary Segmentation (BS)

BS 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 dataset, that is, 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 are 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. 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 pseudocode of the BS method, see, for example, Eckley, Fearnhead, and Killick (2011).

BS can be viewed as attempting to minimize Equation (1) with $f(m) = m$: each step of the algorithm attempts to introduce an extra changepoint if and only if it reduces (1). The advantage of the BS method is that it is computationally efficient, resulting in an $\mathcal{O}(n \log n)$ calculation. However, this comes at a cost as it is not guaranteed to find the global minimum of (1).

2.2 Exact Methods

2.2.1 Segment Neighborhood (SN). Auger and Lawrence (1989) proposed an alternative, exact search method for changepoint detection, namely the SN method. This approach searches the entire segmentation space using dynamic programming (Bellman and Dreyfus 1962). 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 as Q. The method then continues by computing the cost function for all possible segments. From this, all possible segmentations with between 0 and Q changepoints are considered.

In addition to being an exact search method, the SN approach has the ability to incorporate an arbitrary penalty 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 the method will have a computational cost of $\mathcal{O}(n^3)$.

2.2.2 The OP Method. Yao (1984) and Jackson et al. (2005) proposed a search method that aims to minimize

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

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

Following Jackson et al. (2005), the OP method begins by first conditioning on the last point of change. It then relates the optimal value of the cost function to the cost for the optimal partition of the data prior to the last changepoint plus the cost for the segment from the last changepoint to the end of the data. More formally, let $F(s)$ denote the minimization from (3) for data $y_{1:s}$ and $\mathcal{T}_s = \{\boldsymbol{\tau} : 0 = \tau_0 < \tau_1 < \cdots < \tau_m < \tau_{m+1} = s\}$ be the set of possible vectors of changepoints for such data.

Finally, set $F(0) = -\beta$. It therefore follows that

$$
\begin{aligned}
F(s) &= \min_{\tau\in\mathcal{T}_s} \left\{ \sum_{i=1}^{m+1} [\mathcal{C}(y_{(\tau_{i-1}+1):\tau_i}) + \beta] \right\}, \\
&= \min_{t} \left\{ \min_{\tau\in\mathcal{T}_t} \sum_{i=1}^{m} [\mathcal{C}(y_{(\tau_{i-1}+1):\tau_i}) + \beta] + \mathcal{C}(y_{(t+1):n}) + \beta \right\}, \\
&= \min_{t} \{F(t) + \mathcal{C}(y_{(t+1):n}) + \beta\}.
\end{aligned}
$$

This provides a recursion that gives the minimal cost for data $y_{1:s}$ in terms of the minimal cost for data $y_{1:t}$ for $t < s$. This recursion can be solved in turn for $s = 1, 2, \ldots, n$. The cost of solving the recursion for time s is linear in s, so the overall computational cost of finding $F(n)$ is quadratic in n. Steps for implementing the OP method are given 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 constant $\beta$ which does not depend on the number or
                 location of changepoints.
 Initialise:  Let $n$ = length of data and set $F(0) = -\beta$, $cp(0) = NULL$.
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.

While OP improves on the computational efficiency of the SN method, it is still far from being competitive computationally with the BS method. Section 3 introduces a modification of the OP method denoted as PELT, which results in an approach whose computational cost can be linear in n while retaining an exact minimization of (3). This exact and efficient computation is achieved via a combination of OP and pruning.

3. A PELT METHOD

We now consider how pruning can be used to increase the computational efficiency of the OP method while still ensuring that the method finds a global minimum of the cost function (3). The essence of pruning in this context is to remove those values of τ that can never be minima from the minimization performed at each iteration in (1) of Algorithm 1.

The following theorem gives a simple condition under which we can do such pruning.

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 $t < s < T$,

$$\mathcal{C}(y_{(t+1):s}) + \mathcal{C}(y_{(s+1):T}) + K \le \mathcal{C}(y_{(t+1):T}). \tag{4}$$

Then, if

$$F(t) + \mathcal{C}(y_{(t+1):s}) + K \ge F(s) \tag{5}$$

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

Proof. See Section 5 of the online supplementary materials. □

The intuition behind this result is that if (5) 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 that has this most recent changepoint at t. Note that almost all cost functions used in practice satisfy assumption (4). For example, if we take the cost function to be minus the log-likelihood, then the constant $K = 0$ and if we take it to be minus a penalized log-likelihood, then K would equal the penalization 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 OP method and the pseudocode is given in Algorithm 2. This shows that at each step in the method, the candidate changepoints 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 call this the 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 constant $\beta$ which does not depend on the number or
                 location of changepoints.
              A constant $K$ that satisfies Equation 4.
 Initialise:  Let $n$ = length of data and set $F(0) = -\beta$, $cp(0)$ = NULL,
                 $R_1 = \{0\}$.
Iterate for $\tau^* = 1, \ldots, n$
  1. Calculate $F(\tau^*) = \min_{\tau\in R_{\tau^*}} \left[F(\tau) + \mathcal{C}(y_{(\tau+1):\tau^*}) + \beta\right]$.
  2. Let $\tau^1 = \arg\left\{\min_{\tau\in R_{\tau^*}} \left[F(\tau) + \mathcal{C}(y_{(\tau+1):\tau^*}) + \beta\right]\right\}$.
  3. Set $cp(\tau^*) = [cp(\tau^1), \tau^1]$.
  4. Set $R_{\tau^*+1} = \{\tau^* \cap \{\tau \in R_{\tau^*} : F(\tau) + \mathcal{C}(y_{\tau+1:\tau^*}) + K < F(\tau^*)\}\}$.
Output the change points recorded in $cp(n)$.
```

Algorithm 2: PELT method.

3.1 Linear Computational Cost of PELT

We now investigate the theoretical 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. The case we focus on is the set of models where the 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.

More formally, our result relates to the expected computational cost of the method and how this depends on the number of data points we analyze. To this end, we define an underlying stochastic model for the data-generating process. Specifically, we define such a process over positive-integer time points and then consider analyzing the first n data points generated by this process. Our result assumes that the parameters associated with a given segment are iid with density function $\pi(\theta)$. For notational simplicity, we assume that given the parameter, θ, for a segment, the data points within the segment are iid with density function $f(y|\theta)$, although extensions to dependence within a segment is trivial. Finally, as previously stated, 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 (4). Hence, pruning in PELT will just depend on the choice of penalty constant β.

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

Theorem 3.2. Define θ^* to be the value that maximizes the expected log-likelihood

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

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

$$B_n = \sum_{i=1}^{n} [\log f(y_i|\hat{\theta}_n) - \log f(y_i|\theta^*)],$$

we have $\mathbb{E}(B_n) = o(n)$ and $\mathbb{E}([B_n - \mathbb{E}(B_n)]^4) = \mathcal{O}(n^2)$,

(A2)

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

(A3)

$$\mathbb{E}(S^4) < \infty, \text{ and}$$

(A4)

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

where S is the expected segment length, then the expected CPU cost of PELT for analyzing n data points is bounded above by Ln for some constant $L < \infty$.

Proof. See Section 6 of the online supplementary materials. □

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 that 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 penalized likelihood value obtained with the true changepoint and parameter values to be greater than the expected penalized likelihood value if we fit a single segment to the data with segment parameter θ^*.

In all cases, the worst-case complexity of the algorithm is where no pruning occurs and the computational cost is $\mathcal{O}(n^2)$.

3.2 PELT for Concave Penalties

There is a growing body of research (see Guyon and Yao 1999; Picard et al. 2005; Birge and Massart 2007) that considers nonlinear penalty forms. In this section, we address how PELT can be applied to penalty functions that are concave,

$$\beta f(m) + \sum_{i=1}^{m+1} \mathcal{C}(y_{(\tau_{i-1}+1):\tau_i}), \tag{6}$$

where $f(m)$ is concave and differentiable.

For an appropriately chosen γ, the following result shows that the optimum segmentation based on such a penalty corresponds to minimizing

$$m\gamma + \sum_{i=1}^{m+1} \mathcal{C}(y_{(\tau_{i-1}+1):\tau_i}). \tag{7}$$

Theorem 3.3. Assume that f is concave and differentiable, with derivative denoted as f'. Further, let $\hat{m}$ be the value of m for which the criteria (6) is minimized. Then the optimal segmentation under this set of penalties is the segmentation that minimizes

$$m f'(\hat{m}) + \sum_{i=1}^{m+1} \mathcal{C}(y_{(\tau_{i-1}+1):\tau_i}). \tag{8}$$

Proof. See Section 7 of the online supplementary materials. □

This suggests that we can minimize penalty functions based on $f(m)$ using PELT—the correct penalty constant just needs to be applied. A simple approach is to run PELT with an arbitrary penalty constant, say $\gamma = f'(1)$. Let m_0 denote the resulting number of changepoints estimated. We then run PELT with penalty constant $\gamma = f'(m_0)$, and get a new estimate of the number of changepoints m_1. If $m_0 = m_1$ we stop. Otherwise, we update the penalty constant and repeat until convergence. This simple procedure is not guaranteed to find the optimal number of changepoints. Indeed, more elaborate search schemes may be better. However, as tests of this simple approach in Section 4.3 show, it can be quite effective.

4. SIMULATION AND DATA EXAMPLES

We now compare PELT with both OP and BS on a range of simulated and real examples. Our aim is to see empirically (1)

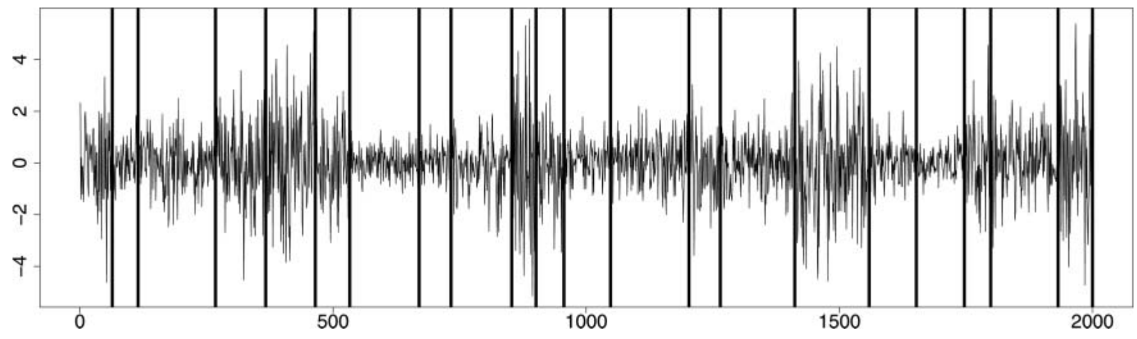

Figure 1. A realization of multiple changes in variance where the true changepoint locations are shown by vertical lines.

how the computational cost of PELT is affected by the amount of data, (2) to evaluate the computational savings that PELT gives over OP, and (3) to evaluate the increased accuracy of exact methods over BS. Unless otherwise stated, we used the SIC penalty. In this case, the penalty constant increases with the amount of data, and as such the application of PELT lies outside the conditions of Theorem 3.2. We also consider the impact of the number of changepoints not increasing linearly with the amount of data, a further violation of the conditions of Theorem 3.2.

4.1 Changes in Variance Within Normally Distributed Data

In the following sections, we consider multiple changes in variance within datasets that are assumed to follow a normal distribution with a constant (unknown) mean. We begin by showing the power of the PELT method in detecting multiple changes via a simulation study, and then use PELT to analyze oceanographic data and Dow Jones Index returns (Section 2 in the online supplementary materials).

4.1.1 Simulation Study. To evaluate PELT, 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 twice the negative log-likelihood. Note that for a change in variance (with unknown mean), the minimum segment length is two observations. The cost of a segment is then

$$\mathcal{C}(y_{(\tau_{i-1}+1):\tau_i}) = (\tau_i - \tau_{i-1})\left(\log(2\pi) + \log\left(\frac{\sum_{j=\tau_{i-1}+1}^{\tau_i}(y_j-\mu)^2}{\tau_i-\tau_{i-1}}\right)+1\right). \quad (9)$$

Our simulated data consist of scenarios with varying lengths, $n = (100, 200, 500, 1000, 2000, 5000, 10{,}000, 20{,}000, 50{,}000)$. For each value of n, we consider a linearly increasing number of changepoints, $m = n/100$. In each case, the 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 scenarios, we have 1000 repetitions where the mean is fixed at 0 and the variance parameters for each segment are assumed to have a lognormal 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 $[\frac{1}{10}, 10]$. An example realization is shown in Figure 1. Additional simulations considering a wider range of options for the number of changepoints (square root: $m = \lfloor\sqrt{n}/4\rfloor$ and fixed: $m = 2$) and parameter values are given in Section 1 of the online supplementary materials.

Results are shown in Figure 2 where we denote the BS method, which identifies the *same* number of changepoints as PELT, as subBS. Conversely, the number of changepoints BS would optimally select is called optimal BS. First, Figure 2(a) shows that when the number of changepoints increases linearly

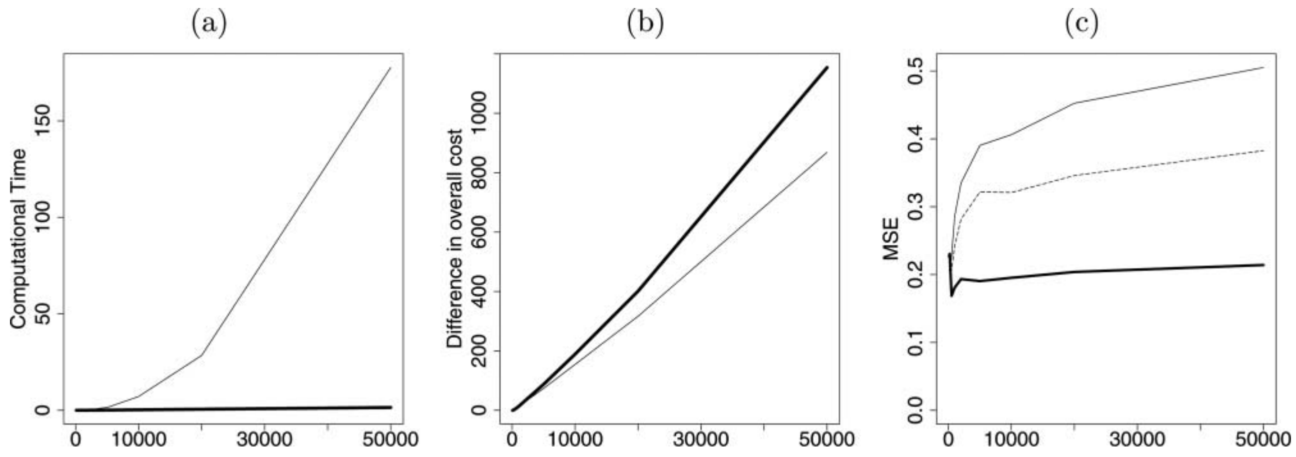

Figure 2. (a) Average computational time (in seconds) for a change in variance (thin: OP, thick: PELT). (b) Average difference in cost between PELT and BS for subBS (thin) and optimal BS (thick). (c) MSE for PELT (thick), optimal BS (thin), and subBS (dotted).

with n, PELT does indeed have a CPU cost that is linear in n. By comparison, figures in the online supplementary materials show that if the number of changepoints increases at a slower rate, for example, square root or even fixed number of changepoints, the CPU cost of PELT is no longer linear. However, even in the latter two cases, substantial computational savings are attained relative to OP. Comparison of times with BS are also given in the online supplementary materials. These show that PELT and BS have similar computational costs for the case of linearly increasing number of changepoints, but BS can be orders of magnitude quicker for other situations.

The advantage of PELT over BS is that PELT is guaranteed to find the optimal segmentation under the chosen cost function, and as such is likely to be preferred, provided sufficient computational time is available to run it. Figure 2(b) shows the improved fit to the data that PELT attains over BS in terms of the smaller values of the cost function that are found. If one considers using the log-likelihood to choose between competing models, the value for $n = 50{,}000$ is over 1000, which is very large. An alternative comparison is to look at how well each method estimates the parameters in the model. We measure this using mean square error (MSE):

$$\text{MSE} = \frac{\sum_{i=1}^{n} (\hat{\theta}_i - \theta_i)^2}{n}. \qquad (10)$$

Figure 2(c) shows the increase in accuracy in terms of MSE of estimates of the parameter. The figures in the online supplementary materials show that for the fixed number of changepoints scenario, the difference is negligible, but for the linearly increasing number of changepoints scenario, the difference is relatively large.

A final way to compare the accuracy of PELT with that of BS is to look at how accurately each method detects the actual times at which changepoints occurred. For the purposes of this study, a changepoint is considered to be correctly identified if we infer its location within a distance of 10 time points of the true position. If two changepoints are identified in this window, then one is counted as correct and the other as false. The number of false changepoints is then the total number of changepoints identified minus the number correctly identified. The results are depicted in Figure 3 for a selection of data lengths, n, for the case $m = n/100$. As n increases, the difference between the PELT and BS algorithm becomes clearer with PELT correctly identifying more changepoints than BS. Qualitatively, similar results are obtained if we change how close an inferred changepoint has to be to a true changepoint to be classified as correct. Figures for square root increasing and fixed numbers of changepoints are given in the online supplementary materials. As the number of changepoints decreases, a higher proportion of true changepoints is detected with fewer false changepoints.

The online supplementary materials also contain an exploration of the same properties for changes in both mean and variance. The results are broadly similar to those described above. We now demonstrate increased accuracy of the PELT algorithm compared with BS on an oceanographic dataset; a financial application is given in the online supplementary materials.

4.2 Application to Canadian Wave Heights

There is interest in characterizing the ocean environment, particularly in areas where there are marine structures, for example, offshore wind farms or oil installations. Short-term operations, such as inspection and maintenance of these marine structures, are typically performed in periods where the sea is less volatile to minimize risk.

Here we consider publicly available data for a location in the North Atlantic where data have been collected on wave heights at hourly intervals from January 2005 until September 2012; see Figure 4(a). Our interest is in segmenting the series into periods of lower and higher volatility. The data we use are obtained from Fisheries and Oceans Canada, East Scotian Slop buoy ID C44137 and have been reproduced in the "changepoint" R package (Killick and Eckley 2010).

The cyclic nature of larger wave heights in the winter and small wave heights in the summer is clear. However, the transition point from periods of higher volatility (winter storms) to lower volatility (summer calm) is unclear, particularly in some years. To identify these features, we work with the first difference data. Consequently, a natural approach is to use the change in variance cost function of Section 4.1. Of course, this is but one of several ways in which the data could be segmented.

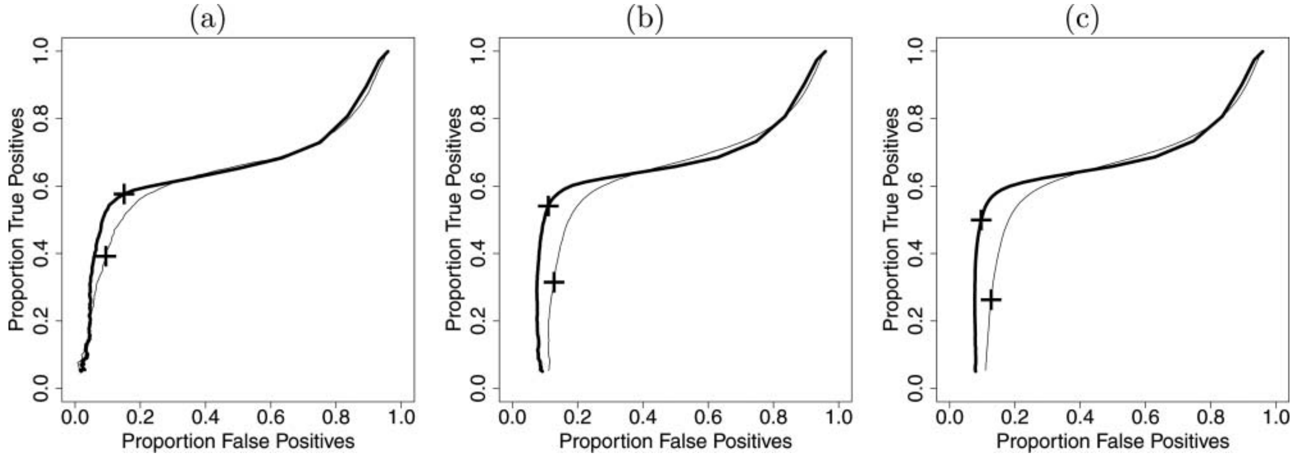

Figure 3. Proportion of correctly identified changepoints against the proportion of falsely detected changepoints. Change in variance with $m = n/100$ where (a) $n = 500$, (b) $n = 5000$, and (c) $n = 50{,}000$ (PELT: thick line, BS: thin line, +: SIC penalty).

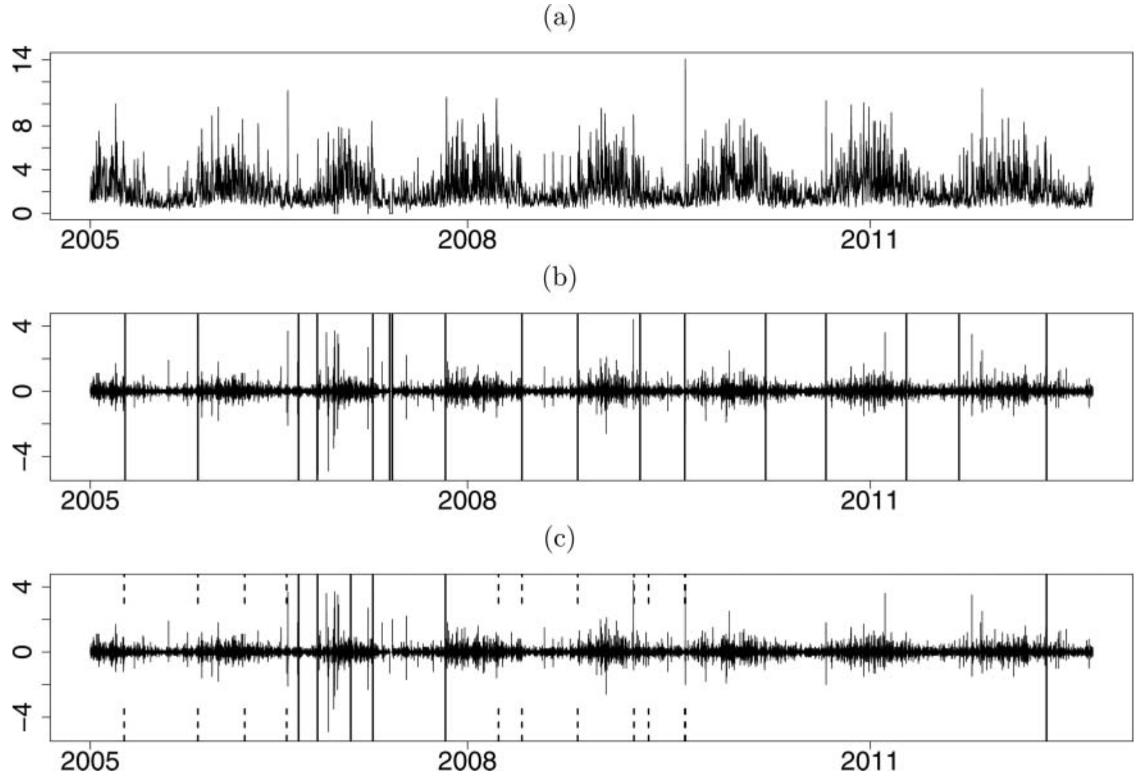

Figure 4. North Atlantic wave heights: (a) original data, (b) differenced data with PELT changepoints, and (c) differenced data with optimal BS changepoints and additional subBS changepoints (dotted lines).

For the data we consider [Figure 4(a)], there is quite a difference in the number of changepoints identified by PELT (17) and optimal BS (6). However, the location of the detected changepoints is quite similar. The difference in likelihood between the inferred segmentations is 3851. PELT chooses a segmentation that, by-eye, segments the series well into the different volatility regions [Figure 4(b)]. Conversely, the segmentation produced by BS does not [Figure 4(c)]; most notably, it fails to detect any transitions between 2008 and 2012. If we increase the number of changepoints BS finds to equal that of PELT, the additional changepoints still fail to capture the regions appropriately.

4.3 Changes in Autocovariance Within Autoregressive Data

Changes in autoregressive (AR) models have been considered by many authors, including Davis, Lee, and Rodriguez-Yam (2006), Huskova, Praskova, and Steinebach (2007), and Gombay (2008). This section describes a simulation study that compares the properties of PELT and the genetic algorithm used by Davis, Lee, and Rodriguez-Yam (2006) to implement the minimum description length (MDL) test statistic.

4.3.1 MDL for AR Models. The simulation study here will be constructed in a similar way to that of Section 4.1. It is assumed that the data follow an AR model with order and parameter values depending on the segment. We shall take the cost function to be the MDL, and consider allowing AR models of order $1, \ldots, p_{\max}$, for some chosen $p_{\max}$ within each segment. The associated cost for a segment is

$$\mathcal{C}(y_{(\tau_{i-1}+1):\tau_i}) = \min_{p\in\{1,\ldots,p_{\max}\}} \left\{ \log p + \frac{p+2}{2}\log(\tau_i - \tau_{i-1}) + \frac{\tau_i-\tau_{i-1}}{2}\log\left(2\pi\hat{\sigma}(p,\tau_{i-1}+1,\tau_i)^2\right)\right\}, \quad (11)$$

where $\hat{\sigma}(p, \tau_{i-1}+1, \tau_i)^2$ is the Yule–Walker estimate of the innovation variance for data $y_{(\tau_{i-1}+1):\tau_i}$ and order p. When implementing PELT, we set $K = -[2\log(p_{\max}) + (p_{\max}/2)\log(n)]$, to ensure that (4) is satisfied.

4.3.2 Simulation Study. The simulated data consist of five scenarios with varying lengths, $n = c(1000, 2000, 5000, 10{,}000, 20{,}000)$ and each scenario contains $0.003n$ changepoints. These changepoints are distributed uniformly across $(2, n-2)$ with the constraint that there must be at least 50 observations between changepoints. Within each of these five scenarios, we have 200 repetitions where the segment order is selected randomly from $\{0, 1, 2, 3\}$ and the AR parameters for each segment are a realization from a standard normal distribution subject to stationarity conditions. We compare the output from PELT with an approximate method proposed by Davis, Lee, and Rodriguez-Yam (2006) for minimizing the MDL criteria, which uses a genetic algorithm. This was implemented in the program Auto-PARM (Automatic Piecewise AutoRegressive Modeling), made available by the authors. We used the recommended settings except that for both methods we assumed $p_{\max} = 7$.

Table 1 shows the average difference in MDL over each scenario for each fitted model. It is clear that on average PELT achieves a lower MDL than the Auto-PARM algorithm, and

Table 1. Average MDL and number of PELT iterations over 200 repetitions

n	1000	2000	5000	10,000	20,000
No. iterations	2.470	2.710	2.885	2.970	3.000
Auto-PARM–PELT	8.856	13.918	59.825	252.796	900.869

that this difference increases as the length of the data increases. Overall, for 91% of datasets, PELT gave a lower value of MDL than Auto-PARM. In addition, the average number of iterations required for PELT to converge is small in all cases.

Previously, it was noted that the PELT algorithm for the MDL penalty is no longer an exact search algorithm. For $n = 1000$, we evaluated the accuracy of PELT by calculating the optimal segmentation in each case using SN. The average difference in MDL between the SN and the PELT algorithm is 1.01 (to 2dp). However, SN took an order of magnitude longer to run than PELT—its computational cost increasing with the cube of the data size making it impracticable for large n. A better approach to improve on the results of our analysis would be to improve the search strategy for the value of penalty function to run PELT with.

5. DISCUSSION

In this article, 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, most importantly that the number of changepoints is increasing linearly with n, the computational efficiency of PELT 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. The empirical results show a resulting computational cost for PELT that can be orders of magnitude smaller than alternative exact search methods. Furthermore, the results show substantial increases in accuracy by using PELT compared with BS. While PELT is not, in practice, computationally quicker than BS, we would argue that the statistical benefits of an exact segmentation outweigh the relatively small computational costs. There are other fast algorithms for segmenting data that improve upon BS (Gey and Lebarbier 2008; Harchaoui and Levy-Leduc 2010), although these do not have the guarantee of exactness that PELT does.

Rigaill (2010) developed a competing exact method called pruned dynamic programming algorithm (PDPA). This method also aims to improve the computational efficiency of an exact method, this time SN, through pruning, but the way pruning is implemented is very different from PELT. The methods are complementary. First, they can be applied to different problems, with PDPA able to cope with a nonlinear penalty function for the number of changepoints, but restricted to models with a single parameter within each segment. Second, the applications under which they are computationally efficient are different, with PDPA best suited to applications with few changepoints. While unable to compare PELT with PDPA on the change in variance or the change in mean and variance models considered in the results section, we have done a comparison between them on a change in mean. Results are presented in Table 1 of the online supplementary materials. Our comparison was for both a linearly increasing number of changepoints and a fixed number of changepoints scenario. For the former, PELT was substantially quicker, by a factor of between 300 and 40,000 as the number of data points varied between 500 and 500,000. When we fixed the number of changepoints to 2, PDPA was a factor of 2 quicker for data with 500,000 changepoints, though often much slower for smaller datasets.

Code implementing PELT is contained within the R library "changepoint," which is available on CRAN (Killick and Eckley 2010).

SUPPLEMENTARY MATERIALS

The supplementary material contains additional simulations, examples and proofs from the main article. Section 1 contains additional simulations for square root increasing and fixed change points as well as different variance ranges for the change in variance example. Changes in variance within the Dow Jones Index are examined in Section 2. Section 3 repeats the simulation study from the paper and Section 1 of the supplementary material for changes in both mean and variance. Section 4 describes a comparison of PELT and PDPA under changes in mean. Finally Sections 5, 6, 7 contains the proofs for Theorems 3.1, 3.2 and 3.3 respectively.

[Received July 2011. Revised June 2012.]

REFERENCES

Aggarwal, R., Inclan, C., and Leal, R. (1999), "Volatility in Emerging Stock Markets," *The Journal of Financial and Quantitative Analysis*, 34, 33–55. [1590]

Akaike, H. (1974), "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, 19, 716–723. [1591]

Andreou, E., and Ghysels, E. (2002), "Detecting Multiple Breaks in Financial Market Volatility Dynamics," *Journal of Applied Econometrics*, 17, 579–600. [1590]

Auger, I. E., and Lawrence, C. E. (1989), "Algorithms for the Optimal Identification of Segment Neighborhoods," *Bulletin of Mathematical Biology*, 51, 39–54. [1590,1591]

Bellman, E., and Dreyfus, S. E. (1962), *Applied Dynamic Programming*, Princeton, NJ: Princeton University Press. [1591]

Birge, L., and Massart, P. (2007), "Minimal Penalties for Gaussian Model Selection," *Probability Theory and Related Fields*, 138, 33–73. [1591,1593]

Braun, J. V., Braun, R. K., and Muller, H. G. (2000), "Multiple Changepoint Fitting via Quasilikelihood, With Application to DNA Sequence Segmentation," *Biometrika*, 87, 301–314. [1590]

Chen, J., and Gupta, A. K. (2000), *Parametric Statistical Change Point Analysis*, New York: Birkhäuser. [1591]

Davis, R. A., Lee, T. C. M., and Rodriguez-Yam, G. A. (2006), "Structural Break Estimation for Nonstationary Time Series Models," *Journal of the American Statistical Association*, 101, 223–239. [1596]

Eckley, I. A., Fearnhead, P., and Killick, R. (2011), "Analysis of Changepoint Models," in *Bayesian Time Series Models*, eds. D. Barber, T. Cemgil, and S. Chiappa, Cambridge: Cambridge University Press, pp. 205–224. [1591]

Fernandez, V. (2004), "Detection of Breakpoints in Volatility," *Estudios de Administracion*, 11, 1–38. [1590]

Gey, S., and Lebarbier, E. (2008), "Using CART to Detect Multiple Change-Points in the Mean for Large Samples," Research Report No.12, Statistics for Systems Biology (SSB) Preprint. [1597]

Gombay, E. (2008), "Change Detection in Autoregressive Time Series," *Journal of Multivariate Analysis*, 99, 451–464. [1596]

Guyon, X., and Yao, J.-F. (1999), "On the Underfitting and Overfitting Sets of Models Chosen by Order Selection Criteria," *Journal of Multivariate Analysis*, 70 , 221–249. [1591,1593]

Harchaoui, Z., and Levy-Leduc, C. (2010), "Multiple Change-Point Estimation With a Total-Variation Penalty," *Journal of the American Statistical Association*, 105, 1480–1493. [1597]

Horvath, L. (1993), "The Maximum Likelihood Method of Testing Changes in the Parameters of Normal Observations," *The Annals of Statistics*, 21, 671–680. [1591]

Huskova, M., Praskova, Z., and Steinebach, J. (2007), "On the Detection of Changes in Autoregressive Time Series," *Journal of Statistical Planning and Inference*, 137, 1243–1259. [1596]

Inclan, C., and Tiao, G. C. (1994), "Use of Cumulative Sums of Squares for Retrospective Detection of Changes of Variance," *Journal of the American Statistical Association*, 89, 913–923. [1590,1591]

Jackson, B., Sargle, J. D., Barnes, D., Arabhi, S., Alt, A., Gioumousis, P., Gwin, E., Sangtrakulcharoen, P., Tan, L., and Tsai, T. T. (2005), "An Algorithm for Optimal Partitioning of Data on an Interval," *IEEE Signal Processing Letters*, 12, 105–108. [1590,1591]

Killick, R., and Eckley, I. A. (2010), *Changepoint: Analysis of Changepoint Models*, Lancaster: Lancaster University. [1595,1597]

Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004), "Circular Binary Segmentation for the Analysis of Array-Based DNA Copy Number Data," *Biostatistics*, 5, 557–572. [1590]

Picard, F., Lebarbier, E., Hoebeke, M., Rigaill, G., Thiam, B., and Robin, S. (2011), "Joint Segmentation, Calling and Normalization of Multiple cgh Profiles," *Biostatistics*, 12, 413–428. [1590]

Picard, F., Robin, S., Lavielle, M., Vaisse, C., and Daudin, J. J. (2005), "A Statistical Approach for Array cgh Data Analysis," *Bioinformatics*, 6: DOI:10.1186/1471-2105-6-27. [1590,1591,1593]

Rigaill, G. (2010), "Pruned Dynamic Programming for Optimal Multiple Change-Point Detection," Technical Report, arXiv:1004.0887. [1591,1597]

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464. [1591]

Scott, A. J., and Knott, M. (1974), "A Cluster Analysis Method for Grouping Means in the Analysis of Variance," *Biometrics*, 30, 507–512. [1590,1591]

Sen, A., and Srivastava, M. S. (1975), "On Tests for Detecting Change in Mean," *The Annals of Statistics*, 3, 98–108. [1591]

Yao, Y. (1984), "Estimation of a Noisy Discrete-Time Step Function: Bayes and Empirical Bayes Approaches," *The Annals of Statistics*, 12, 1434–1447. [1591]

Zhang, N. R., and Siegmund, D. O. (2007), "A Modified Bayes Information Criterion With Applications to the Analysis of Comparative Genomic Hybridization Data," *Biometrics*, 63, 22–32. [1591]