Efficient Detection Of Multiple Changepoints Within An Oceanographic Time Series

Killick, Rebecca
Department of Mathematics & Statistics
Lancaster University
Lancaster, LA1 4YF, UK
E-mail: r.killick@lancs.ac.uk

Eckley, Idris A.
Department of Mathematics & Statistics
Lancaster University
Lancaster, LA1 4YF, UK
E-mail: i.eckley@lancs.ac.uk

Jonathan, Philip
Shell Technology Centre Thornton
P.O. Box 1
Chester, UK
E-mail: philip.jonathan@shell.com

Introduction

Oceanographic time series are commonly used by engineers to characterise the ocean environment. Efficient and accurate analysis of such data is central to reliable design and operation of marine structures. Detecting the presence of changepoints in oceanographic time-series is of particular importance as statistical and engineering modelling of the ocean environment, structural loading and response typically assume that the variability is stationary in time. In this paper we attempt to verify this assumption for a location within the North Sea. We show that the variability is not constant through time; instead it varies between stormy and calm seasons. We use this opportunity to compare and contract the recently proposed PELT search algorithm with the long established Binary Segmentation algorithm. We show that the PELT method clearly detects the start and end of storm seasons whereas the Binary Segmentation algorithm does not achieve this.

The paper is organised as follows, we begin with an introduction to changepoint detection and then review of multiple changepoint search methods. The search methods are applied to wave heights from the North Sea and the results compared. The paper concludes with a discussion of future work.

Changepoint Detection

When considering oceanographic time series the assumption that statistical properties remain the same throughout the series may be unrealistic. One possible way of overcoming this is to identify a set of changepoints, between which the statistical properties of the series remain constant. A range of different test statistics can be used to identify specific types of changes, such as changes in mean or variance. We briefly recap the likelihood-based approach below. For a more comprehensive review please refer to Chen and Gupta (2000) and Eckley et al. (2011).

By way of introduction, 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, $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_m)$. 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 $y_{(\tau_{i-1}+1):\tau_i}$.

Single Changepoint Detection Before considering the general problem of identifying the $\boldsymbol{\tau}$ changepoint positions, we first consider the detection of a single changepoint. Detecting a single changepoint is similar to performing a hypothesis test. The null hypothesis, H_0, corresponds to no changepoint

($m = 0$) whilst the alternative hypothesis, H_1, consists of a single changepoint ($m = 1$).

We now introduce the general likelihood-ratio based approach to test this hypothesis. The potential for using a likelihood based approach to detect changepoints was first proposed by Hinkley (1970) who derives the asymptotic distribution of the likelihood ratio test statistic for a change in the mean within normally distributed observations. The likelihood based approach has been extended to changes in variance within normally distributed observations by Gupta and Tang (1987).

We can construct a test statistic which will decide whether a change has occurred. The likelihood ratio method requires calculating the maximum log-likelihood value under both null and alternative hypotheses. For the null hypothesis the maximum log-likelihood value is $\log p(y_{1:n}|\hat{\theta})$, where $p(\cdot)$ is a probability density function and $\hat{\theta}$ is the maximum likelihood estimate of the parameters.

Under H_1, consider a model with a changepoint at τ_1, with $\tau_1 \in \{1, 2, \ldots, n-1\}$. Then the maximum log likelihood for a given τ_1 is

$$ML(\tau_1) = \log p(y_{1:\tau_1}|\hat{\theta}_1) + \log p(y_{(\tau_1+1):n}|\hat{\theta}_2). \quad (1)$$

The maximum log-likelihood value under the alternative is just $\max_{\tau_1} ML(\tau_1)$. The test statistic is

$$\lambda = 2\left[\max_{\tau_1} ML(\tau_1) - \log p(y_{1:n}|\hat{\theta})\right].$$

The test involves choosing a threshold, c, such that we reject the null hypothesis if $\lambda > c$. If we reject the null hypothesis, corresponding to detecting a changepoint, then we estimate its position as $\hat{\tau}_1$ the value of τ_1 that maximises $ML(\tau_1)$.

It is clear that the likelihood test statistic can be extended to multiple changes simply by summing the likelihood for each of the m segments. The problem becomes one of identifying the maximum of $ML(\boldsymbol{\tau})$ over all possible combinations of $\boldsymbol{\tau}$. The next section explores existing search methods that address this problem.

Multiple Changepoint Detection

Whilst an approach for detecting a single changepoint is very useful, in practice the assumption of only one change within a series may be unrealistic. With the amount of data collected increasing and time series becoming longer, it is likely that multiple changes exist within a single series. As such, methods that can efficiently extend single changepoint methodology to search the, often very large, solution space for multiple changepoints are vital. The search methods presented in this paper aim to minimise,

$$\sum_{i=1}^{m+1} \left[\mathcal{C}(y_{(\tau_{i-1}+1):\tau_i})\right] + \beta f(m). \quad (2)$$

Here $\mathcal{C}$ is a cost function for a segment and $\beta f(m)$ is a penalty to guard against over fitting. Using the notation from the previous section, $\mathcal{C}$ may be the negative log-likelihood and $\beta f(m)$ may be cm. In the following paragraphs we explore two methods that attempt to minimise (2); Binary Segmentation and Pruned Exact Linear Time (PELT) proposed by Scott and Knott (1974) and Killick et al. (2011) respectively.

Binary Segmentation In essence the method extends a single changepoint method to multiple changepoints by repeating the method on varying subsets of the series iteratively. The method begins by applying a single changepoint method to the entire data. If a changepoint is found, the data are split at the changepoint to create two new sub-sequences. The single changepoint method is applied to each sub-sequence and if new changepoints are found, another split is applied. The method ends

when no new sub-sequences are created and the final set of changepoints is the location of all the split points. The pseudo-code demonstrating the implementation of the Binary Segmentation method is given in Algorithm 1.

The Binary Segmentation search method is computationally efficient, resulting in an $\mathcal{O}(n \log n)$ calculation. However computational efficiency comes at the cost of exactness. The location of a changepoint is conditional on the locations of previous changepoints. The consequence of this is that the method does not search the entire solution space and as such is an approximation. When the series contains few changepoints the consequence of this approximation is small. However, when there are many changepoints or small distances between changepoints the effect of the approximation can be pronounced.

Input: A time series of the form, $(y_1, y_2, \ldots, y_n)$.
A test statistic $\Lambda(\cdot)$ dependent on the time series.
An estimator of changepoint position $\hat{\tau}(\cdot)$.
A rejection threshold C.

Initialise: Let $\mathcal{C} = \emptyset$, and $\mathcal{S} = \{[1, n]\}$

Iterate: While $\mathcal{S} \neq \emptyset$

1. Choose an element of $\mathcal{S}$; denote this element as $[s, t]$.
2. If $\Lambda(y_{s:t}) < C$, remove $[s, t]$ from $\mathcal{S}$.
3. If $\Lambda(y_{s:t}) \geq C$ then:
   (a) remove $[s, t]$ from $\mathcal{S}$;
   (b) calculate $r = \hat{\tau}(y_{s:t}) + s - 1$, and add r to $\mathcal{C}$;
   (c) if $r \neq s$ add $[s, r]$ to $\mathcal{S}$;
   (d) if $r \neq t - 1$ add $[r + 1, t]$ to $\mathcal{S}$.

Output: The set of change points recorded $\mathcal{C}$.

Algorithm 1: Binary Segmentation search method for multiple changepoint analysis.

Pruned Exact Linear Time (PELT) This search method was introduced by Killick et al. (2011) and balances the competing computational cost and accuracy properties. The PELT algorithm is $\mathcal{O}(n)$ under certain assumptions (see Killick et al. (2011, §3.1)) and, in contrast to Binary Segmentation, the search is exact.

The PELT method considers the data sequentially and searches the solution space exhaustively. Computational efficiency is achieved by removing solution paths that are known not to lead to optimality. The assumptions and theorems which allow removal of solution paths are explained further in Killick et al. (2011, §3). A key assumption is that of a penalty, c, linear in the number of changepoints m. As such the optimal segmentation is $F(n)$ where,

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

Conditioning on the last point of change, τ_m and calculating the optimal segmentation of the data up to that changepoint gives,

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

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 (3). In fact the inner minimisation is equal to $F(\tau_m)$ and as such (3) can be re-written as

$$F(n) = \min_{\tau_m} \left\{ F(\tau_m) + \mathcal{C}(y_{(\tau_m+1):n}) \right\}. \quad (5)$$

We start by calculating $F(1)$ and then recursively calculate $F(2)$, ..., $F(n)$. At each step we store the optimal segmentation up to τ_{m+1}. 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.

At each step the minimisation over τ_m covers all previous values e.g. when calculating $F(3)$ the minimisation covers $\tau_m = 0, 1, 2$. The computational efficiency of the PELT method is achieved by removing candidate values of τ_m from the minimisation at each step using Theorem 1.

Theorem 1 (Killick et al., 2011)

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 $t < s < T$, $\mathcal{C}(y_{(t+1):s}) + \mathcal{C}(y_{(s+1):T}) + K < \mathcal{C}(y_{(t+1):T})$. Then if

$$F(t) + \mathcal{C}(y_{(t+1):s}) + K > F(s) \quad (6)$$

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

If many candidate values of τ_m can be removed then the efficiency is greatly improved. In fact, Killick et al. (2011) show that using the log-likelihood cost function for $\mathcal{C}$, under mild assumptions the computational efficiency is $\mathcal{O}(n)$. Pseudo-code for the PELT method is given in Algorithm 2.

Input: A time series 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 time series 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 $pts = \arg_\tau \{ F(\tau) + \mathcal{C}(y_{\tau+1:\tau^*}) + \beta + K > F(\tau^*) + \beta \}$.

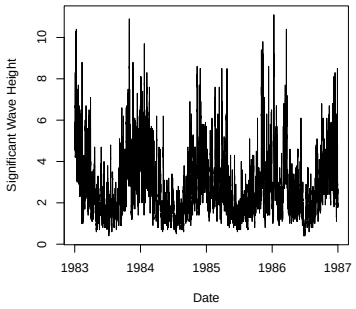
Output: The change points recorded in $cp(n)$.

Algorithm 2: PELT search method for multiple changepoint analysis.

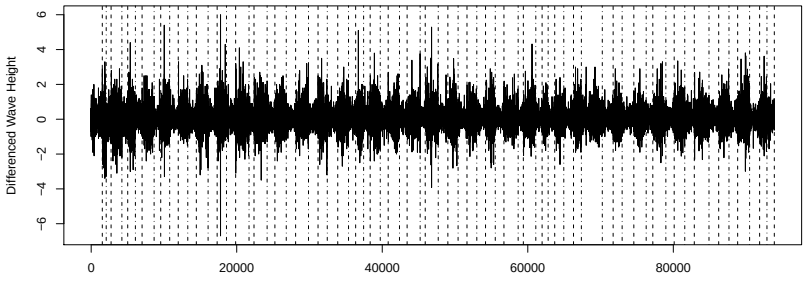
Application to Oceanographic Data

With our test statistic and search methods defined, we now investigate potential changes in variance for time series of measured wave height, H_S. Data are recorded at several locations across the North Sea. We focus on one specific location which we shall refer to as North North Sea (NNS). The NNS data are recorded at a time interval of 3 hours starting on 22^{nd} February 1973 and ending on 8^{th} June 2009. Figure 1(a) shows the measured wave heights for a 4 year period from 1983-1987. The cyclic nature is clear but there also appears to be a change in variability across the series. Peaks represent winter storm periods where wave heights are expected to increase. In particular we note that the variability seems larger during the peaks (winters) than in the troughs (summers).

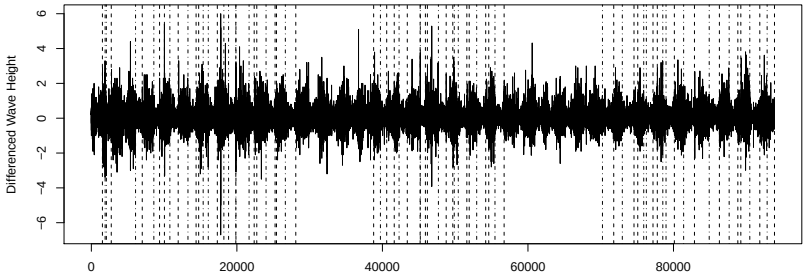
To identify if a change in variability has occurred, we first remove cycles from the data as this affects estimation. We take first differences of the data so as not to make further assumptions. Figure 1(b) shows the (first order) difference data where it is clear that changes in variability are present.



(a) Part of the original data



(b) Differenced Data with PELT changepoints



(c) Differenced Data with Binary Segmentation changepoints

Figure 1: North North Sea (NNS) Analysis (−− increase, − · − decrease in variability)

The simulation study by Eckley et al. (2011) demonstrates the ability of the Normal likelihood method to identify changes in variability. In addition, the simulation study by Killick et al. (2011) shows that due to the approximate nature of the Binary Segmentation method the locations of the estimated changepoints can vary greatly from the exact PELT method.

We use the *changepoint* package in R (Killick and Eckley, 2010) to perform the analysis. For a change in variance assuming a Normal distribution the alternative likelihood from equation (1) becomes,

$$ML(\boldsymbol{\tau}) = \sum_{i=1}^{m+1} \left[(\tau_i - \tau_{i-1}) \left(\log(2\pi) + \log \left(\sum_{j=\tau_{i-1}+1}^{\tau_i} (y_j - \hat{\mu})^2 \right) + 1 \right) \right].$$

Year	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984
Start	15-11	17-09	16-11	31-08	28-08	11-09	10-09	29-09	05-09	08-08	18-09	18-09
End	18-03	30-05	18-04	03-05	29-03	27-03	21-04	29-03	09-05	21-04	31-03	28-04
Year	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996
Start	10-10	27-09	13-09	29-09	15-09	13-09	09-09	08-10	04-10	14-09	22-09	23-09
End	28-03	03-05	19-04	13-04	27-04	22-05	25-04	21-04	09-04	13-04	06-06	08-05
Year	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Start	01-12	08-10	22-10	–	23-10	16-10	11-09	12-09	24-08	17-10	09-09	26-09
End	14-04	30-04	15-03	19-04	29-03	18-05	22-04	13-03	03-05	12-04	07-04	01-02

Table 1: Start and end (dd-mm) of storm season for NNS by year

The changepoint location estimates from applying the PELT method are shown in Figure 1(b). The vertical lines identify the estimated changepoint locations and on closer inspection it is clear that there is a pattern of increased volatility followed by a decrease volatility. The dates of these changepoints appear to align with the start and end of the storm period in each year. Table 1 details the start and end dates identified for NNS, similar results were found at other North Sea locations.

We compare the above results with those obtained from the long established Binary Segmentation method, see Figure 1(c). Due to the nature of the algorithm, for the same number of changepoints detected as the PELT method, the Binary Segmentation method finds no changes between 1995 and 2006 favouring multiple changes within storm periods in earlier years. Increasing the number of changepoints identified rectifies this anomaly but produces multiple changes per storm season. Whilst these extra changepoints may be of interest, for the purpose of identifying the start and end of storm seasons, the Binary Segmentation method produces very poor results. This is due to the approximate nature of the method and Inclan and Tiao (1994) advocate a post-processing step to try to avoid this.

Conclusion

Changepoint analysis is a useful tool in analysing oceanographic time series. Using the exact, computationally efficient PELT method we have identified the start and end of the storm season automatically. The PELT method is more efficient and more accurate than Binary Segmentation for our penalty choice. PELT produces quicker and more consistent results than identification 'by eye' or assuming that the variability is constant. In addition, we find no evidence to suggest that onset or duration of winter storm season varies systematically with time.

This analysis has focussed on changes in variability within oceanographic data, this is not the only type of change that may occur. There are many oceanographic time series that may exhibit changes in mean, regression or rates of storms. For example, in extreme value analysis, temporal de-clustering of dependent data is essential for reliable inference, yet can be problematic in practice. Similarly, whilst the work here has focussed on likelihood based cost functions, we could have equally used alternative approaches such as quadratic loss. The changepoint search approach presented here may also prove useful for partitioning time series into a sequence of storm events for subsequent modelling. This is left as an avenue for future research.

Acknowledgements

The authors are grateful to Graham Feld for useful discussions concerning the North Sea time series data. R. Killick is funded by Shell Research Limited and EPSRC. P. Jonathan acknowledges financial support from Shell International Exploration Production.

References

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

Eckley, I. A., Fearnhead, P., and Killick, R. (2011). Analysis of changepoint models. In Barber, D., Cemgil, T., and Chiappa, S., editors, *Bayesian Time Series Models*. Cambridge University Press.

Gupta, A. K. and Tang, J. (1987). On testing homogeneity of variances for Gaussian models. *Journal of Statistical Computation and Simulation*, 27:155–173.

Hinkley, D. V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika*, 57:1–17.

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(427):913–923.

Killick, R. and Eckley, I. A. (2010). *changepoint*. `http://CRAN.R-project.org/pacakge=changepoint`.

Killick, R., Fearnhead, P., and Eckley, I. A. (2011). Optimal detection of changepoints with a linear computational cost. *In Submission*.

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