Detection of changes in variance using binary segmentation and optimal partitioning

Christian Rohrbeck

Abstract

This work explores the performance of binary segmentation and optimal partitioning in the context of detecting changes in variance for time-series. Both, binary segmentation and optimal partitioning, are based on cost functions that penalise a high amount of changepoints in order to avoid overfitting. Analysis is performed on simulated time-series; first on Normal data with constant but unknown mean and changing variance and second on Exponential data with changing parameter. Results suggest a good performance of both approaches.

1 Introduction

In a wide range of sciences, it is a great issue to investigate the consequences of the climate change. For example, hydrologists study the relation between the climate change and significant wave heights. Similarly, meteorologists survey the coherency between the increasing average temperature and the number and intensity of storms. In both sciences, analysis is often based on data collected other the past decades. The appearance and intensity of a storm at sea and the variability of the sea level in form of waves have a high positive correlation. Consequently, detecting changes in the variability of the sea level allows conclusions on the number and intensity of storms over the considered period at sea.

In medicine, particularly in hospitals, parameters of intensive care patients are measured and analysed continuously by a medical monitor. If one parameters drops below a specific value or the frequency of the heart changes, the medical staff are alerted. In order to offer the best chance of survival, it is necessary to detect changes in mean and variability of the parameters immediately. Therefore, the speed of detection is of high interest in this context, in contrast to the first example. Further areas for which the detection of change points is of interest are bioinformatics (Lio and Vanucci, 2000) and econometrics (Zhou et al., 2010). Given a time-series $\{y_t : t \in 1, \dots n\}$, a changepoint occurs at time τ if the distributions of $\{y_1, \dots, y_\tau\}$ and $\{y_{\tau+1}, \dots, y_n\}$ differ with respect to at least one criterion such as mean, variance or regression structure. For example:

1. Change in mean: y_t has mean
$$\mu_t = \begin{cases} \mu_1 & ,t \leq \tau \\ \mu_n & ,t > \tau \end{cases},$$
where $\mu_1 \neq \mu_n$.
2. Change in variance: y_t has variance
$$\sigma_t = \begin{cases} \sigma_1 & ,t \leq \tau \\ \sigma_n & ,t > \tau \end{cases},$$
where $\sigma_1 \neq \sigma_n$.

In this work, we are interested in detecting changes in variance of time-series. The task is to decide for time-series whether changepoints exist and if so, to detect their location. In Section 2, we explore the methods of binary segmentation (Scott and Knott, 1974) and optimal partitioning (Jackson et al., 2005) for this purpose. In Section 3, both methods are compared and discussed by evaluating the performance of the approaches on simulated Normal and Exponential data with multiple changepoints. For the Normal data, we assume that the mean is constant but unknown and only the variance changes.

2 Detecting changes in variance

The detection of changes in variance has been well studied in the past years and several methods exist, including cumulative sums of square (Inclan and Tiao, 1994), penalised likelihood (Yao, 1988) and Bayesian posterior odds (Fearnhead, 2006). Formally speaking, for an ordered sequence of data $(y_1, \ldots, y_n)$, we aim to determine an unknown number, m, of changepoints $\tau_1, \ldots, \tau_m$, where each changepoint is an integer value between 2 and $n-1$. Let us define that the sequence of changepoints is ordered, such that $\tau_i < \tau_j$ if, and only if, $i < j$. Further, we denote $\tau_0 = 1$ and $\tau_{m+1} = n$. Consequently, the data is split into $m+1$ segments with the ith segment containing $y_{(\tau_{i-1}+1):\tau_i}$.

One approach, mentioned in several publications, is to identify multiple changepoints by minimising

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

with respect to m and $\tau_1, \ldots, \tau_m$. Here, C is a cost function for a segment and $\beta f(m)$ a penalty term in order to avoid overfitting. The choice of C as twice the negative log-likelihood is commonly used in the changepoint literature (see for example Chen and Gupta (2000)). However, Inclan and Tiao (1994), for example, propose a different cost function. In the following, we consider the penalty term as to be linear in m, i.e. $\beta f(m) = \beta m$. The naive approach of testing all possible changepoint locations is hardly practicable for large n as the number of possible partitions is 2^{n-1}. Therefore, computationally effective algorithms are required. In the following binary segmentation by Scott and Knott (1974) and optimal partitioning by Jackson et al. (2005) are explained as approaches to this problem. The notation in Section 2.2 is based on Killick et al. (2012).

2.1 Binary segmentation

Binary segmentation is the standard method in changepoint literature (Killick et al., 2012). Basically, it iteratively applies a single changepoint method to different subsets of the sequence $y_1, \ldots, y_n$ in order to detect multiple changepoints. It starts by applying the single changepoint method to the entire sequence, i.e. it tests if a split of the sequence exists such that the cost function over the two sub sequences plus the penalty term is smaller than the cost function on the entire sequence. Formally, in the context of (1), the single changepoint method tests whether there exist an integer $\tau \in \{1, \ldots, n-1\}$ that satisfies

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

If such an τ does not exist, no changepoint is detected and the algorithm stops. Otherwise, the corresponding value of τ is identified as a changepoint and the sequence is split up into two subsequences $y_{1:\tau}$ and $y_{(\tau+1):n}$, i.e. the sequences before and after the changepoint. Then the single changepoint method is applied to each of these subsequences. The procedure continues until no further changepoint is detected.

Binary segmentation can be seen as an approach to minimise (1) by iteratively deciding whether a changepoint should be added or not. Binary segmentation is computationally efficient, with $\mathcal{O}(n \log n)$ calculations. However, binary segmentation does not automatically lead to the global minimum of equation (1) and is thus only approximative.

2.2 Optimal partitioning

Optimal partitioning by Jackson et al. (2005) is in contrast to binary segmentation an exact method to solve the minimisation problem described by (1) with linear penalty term, i.e. $f(m) = m$. The method uses dynamic programming which can be used because the principle of optimality holds which states that any subpartition of an optimal partition is optimal; for details and proof see Jackson et al. (2005). Dynamic programming is a recursive method which allows to solve several kinds of combinatorial optimisation problems, here, the minimisation problem (1).

Starting with the first observation, the algorithm iteratively determines the optimal partition of the first $t+1$ data points by using the optimal partition of the first t. At each iteration step $t+1$, the algorithm considers all possibilities $j \in \{0, \ldots, t\}$ for the last changepoint of the optimal partition. Consequently, due to the principle of optimality, the cost of the partition is given by the cost of the optimal partition prior to j and the cost of the sequence from the last changepoint to the end of the data, i.e. $C(y_{j:(t+1)})$. The cost of the optimal partition prior to j was calculated in previous iteration steps. At the end of each iteration step, the partition which minimises the cost is stored. The algorithm runs until $t+1=n$. More formally, we denote the set of possible changepoints for the first t data points by $\mathcal{P}_t = \{\boldsymbol{\tau} : 0 < \tau_1 < \cdots < \tau_m < t\}$. Further, we define by $F(t)$ the cost of the optimal partition of the data until time point t and set $F(0) = -\beta$. Based on equation (1), F(t) is determined as

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

Using the principle of optimality, we can write $F(t)$ in (3) as

$$F(t) = \min_{\tau^*} \left[F(\tau^*) + C(y_{(\tau^*+1):t}) + \beta \right] \quad (4)$$

which implies that in iteration step t only $t-1$ calculations are necessary. Consequently, the number of calculations for a sequence of n observations is $\mathcal{O}(n^2)$ and hence it is not as computationally efficient as binary segmentation. However, the approach of optimal partitioning leads to the calculation of the global minimum of the minimsation problem (1); see Theorem 2 of Jackson et al. (2005) for details. Therefore, optimal partitioning is more accurate than binary segmentation. Steps for implementing the optimal partitioning approach are given in Algorithm 1.

Algorithm 1 Optimal Partitioning by Killick et al. (2012) based on Jackson et al. (2005)

Require: Set of data $(y_1, \ldots, y_n)$ where $y_i \in \mathbb{R}$
Require: Cost function $C(\cdot)$
Require: Penalty constant β
1: Set n= length of data.
2: Set $F(0) = -\beta$ and $t = 1$.
3: Set $cp(0) = NULL$ and $\mathbf{F} = F(0)$.
4: **while** $t < n$ **do**
5: Get $F(t) = \min_{\tau^*} \left[F(\tau^*) + C(y_{(\tau^*+1):t}) + \beta \right]$.
6: Get $\tau' = \operatorname{argmin}_{\tau^*} \left[F(\tau^*) + C(y_{(\tau^*+1):t}) + \beta \right]$.
7: Set $cp(t) = (cp(\tau'), \tau')$.
8: Set $\mathbf{F} = (\mathbf{F}, F(t))$.
9: t:= t+1
10: **end while**
11: **return** Changepoints recorded in $cp(n)$.
12: **return** Optimal Costs recorded in $\mathbf{F}$.

Line 6 in Algorithm 1 enables us to get the the global minimum of (1) by backtracking, as all necessary information is recorded in $cp(n)$. The optimal costs for each time point are then recorded in $\mathbf{F}$.

3 Simulation study

3.1 Application to Normal data

In the following simulation study, the approaches of binary segmentation and optimal partitioning are applied to a sequence $\{y_i\}_{i=1}^n$ of independent and normally distributed random variables with unknown parameters $\{\mu, \sigma_i\}_{i=1}^n$. Such samples occur in a lot of applications like oceanology (Killick et al., 2010) and finance (Chen and Gupta, 1997). We consider the penalised likelihood approach by Yao (1988) (Killick et al., 2010). Therefore, the cost function for a sequence $\{y_i\}_{i=s}^t$, $1 \le s \le t \le n$, is chosen as twice the negative log likelihood. Formally,

$$\begin{aligned} C(y_{s:t}) &= -2\ell(\mu, \sigma_s, \ldots, \sigma_t | y_s, \ldots, y_t) \\ &= \sum_{i=s}^{t} \left[\frac{(y_i-\mu)^2}{\sigma_i^2} + \log\left(2\pi\sigma_i^2\right) \right]. \end{aligned} \quad (5)$$

As the exact parameters μ and $\{\sigma_i\}_{i=s}^t$ are unknown, they are replaced by their maximum likelihood estimates. Because the mean is constant over the entire sequence, the estimate on the mean, $\hat{\mu}$, simplifies to the average over all observations. For the penalty term β, we consider the Schwarz Information Criterion (SIC), proposed by Yao (1988) and set $\beta = \log n$. Further, we select $f(m)$ as to be linear in m, i.e. $f(m) = m$. After defining all parameters and functions for (1), binary segmentation and optimal partitioning are performed.

As mentioned in Section 2.2, binary segmentation iteratively performs a single changepoint method on a subsequence $\{y_i\}_{i=s}^t$, $1 \le s \le t \le n$. For the considered case of normally distributed data, the single changepoint method can be viewed as an approach to test the hypothesis H_0 of no changepoint (Gupta and Tang, 1987)

$$H_0: \ \sigma_s^2 = \sigma_{s+1}^2 = \cdots = \sigma_t^2$$

versus the alternative of an unknown changepoint $\tau \in \{s, \ldots, t-1\}$

$$H_1: \ \sigma_s^2 = \cdots = \sigma_\tau^2 \neq \sigma_{\tau+1}^2 \cdots = \sigma_t^2.$$

Following the test formulation, the cost function (5) for a sequence $\{y_i\}_{i=s}^t$ determines to

$$C(y_{s:t}) = \sum_{i=s}^{t} \left[\frac{(y_i-\mu)^2}{\hat{\sigma}_{s:t}^2} + \log\left(2\pi\hat{\sigma}_{s:t}^2\right) \right], \quad (6)$$

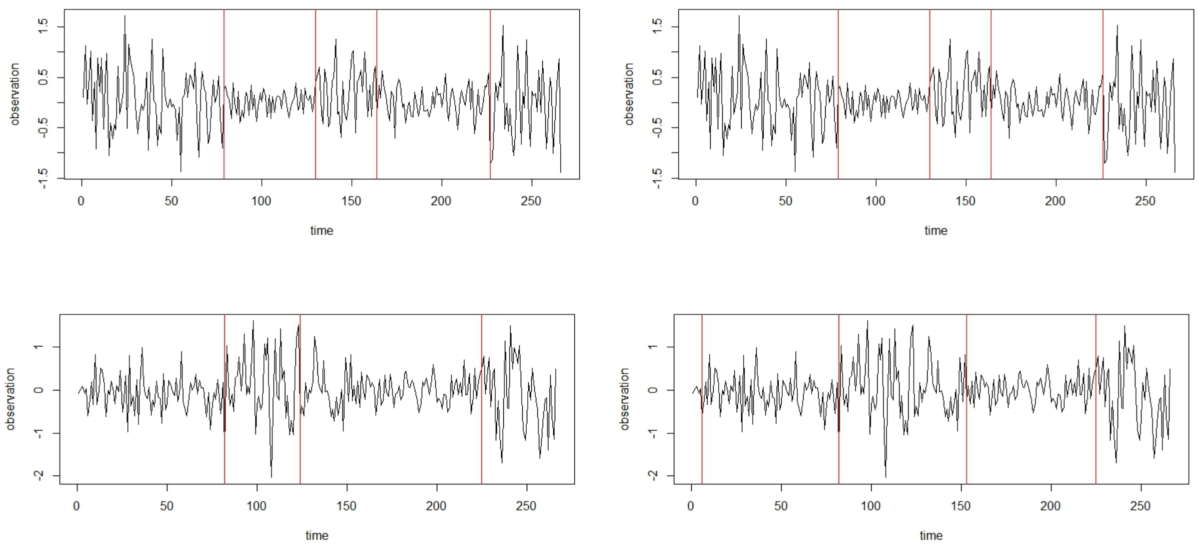

Figure 1: Two simulated time series with Normal data (top and bottom) and detected changepoints by binary segmentation(left) and optimal partitioning (right).

where $\hat{\sigma}_{s:t}$ is the maximum likelihood estimate of the variance based on the data $\{y_i\}_{i=s}^{t}$. The decision whether to reject H_0 or not is decided by equation (2). This approach is also considered by Killick et al. (2010) and Chen and Gupta (1997). Thus, we first determine the integer value τ^* which minimises the left-hand side in (2). More formally,

$$\tau^* = \underset{\tau}{\operatorname{argmin}} \left[C(y_{s:\tau}) + C(y_{(\tau+1):t}) + \log(n) \right]. \quad (7)$$

Second, we check whether

$$C(y_{s:\tau^*}) + C(y_{(\tau^*+1):t}) + \log(n) < C(y_{s:t}). \quad (8)$$

If so, the algorithm repeats the procedure considered above for the subsequences $\{y_i\}_{i=s}^{\tau^*}$ and $\{y_i\}_{i=\tau*+1}^{t}$. If (8) is not fulfilled, we conclude that the variance is constant between the time points s and t and can be estimated by maximum likelihood. For the method of optimal partitioning, we set the variance constant in each subsequence and estimate it by maximum likelihood, as for the binary segmentation, and perform Algorithm 1.

For the simulation study, we generate two time series with $n = 266$ data points and perform the two approaches considered above. The variance changes for each time-series four times at the same time points. In particular, we select $\tau_1 = 81$, $\tau_2 = 130$, $\tau_3 = 162$ and $\tau_4 = 226$ as change points and the variances for the first time-series are $\boldsymbol{\sigma}^{(1)} = (1.3, 0.3, 0.8, 0.4, 1.1)$ and for the second $\boldsymbol{\sigma}^{(2)} = (0.4, 0.7, 0.5, 0.3, 0.8)$. The estimated changepoints for the two time-series and the two methods are illustrated in Figure 1. For the first time-series, the results are identical for the two methods and get the real changepoints. For the second time-series, the results are different and each method does not detect one changepoint. Further, optimal partitioning detects one changepoint, at $t = 6$, at which the true variance does not change.

3.2 Application to Exponential data

In the second simulation study, we consider a sequence $\{y_i\}_{i=1}^{n}$ of n independent and exponentially distributed random variables with unknown parameters $\{\lambda_i\}_{i=1}^{n}$. Such cases occur in queuing modelling, where we have to decide whether the arrival rates are constant over a time period or not. In contrast to Normal data, the mean is determined by the variance and thus a change of the variance results in a change of the mean. Nevertheless, as both functionals only depend on one variable λ_i, binary segmentation and optimal partition are applicable. As for Normal data, we set the cost function $C(y_{s:t})$ equal to twice the negative log-likelihood of the data $\{y_i\}_{i=s}^{t}$, $1 \leq s \leq t \leq n$, $\beta = \log n$ and

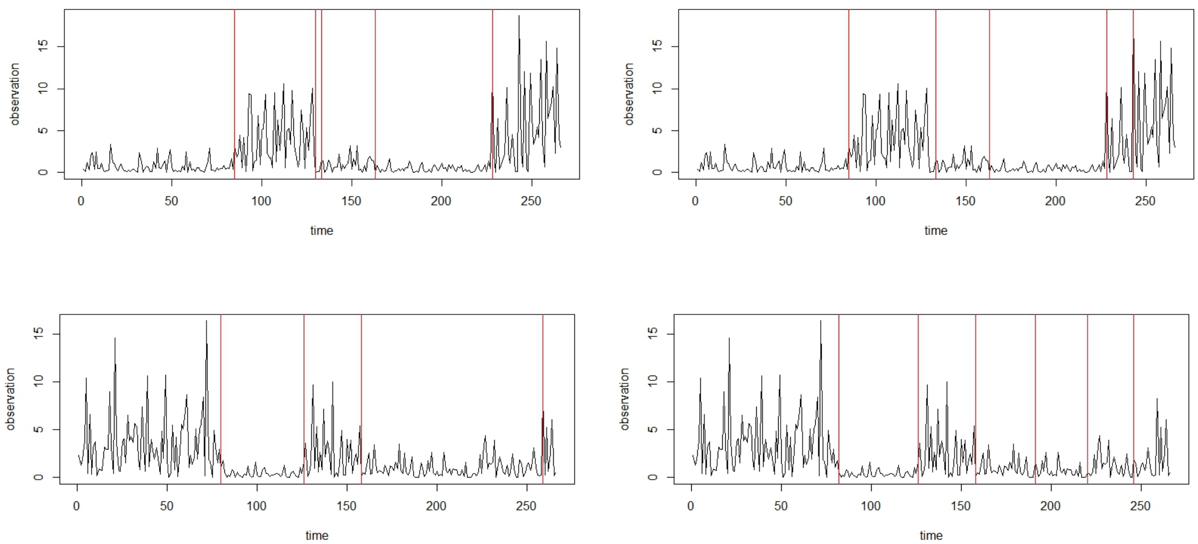

Figure 2: Two simulated time series with Exponential data (top and bottom) and detected changepoints by binary segmentation(left) and optimal partitioning (right).

$f(m) = m$. Consequently, the cost function for the sequence $\{y_i\}_{i=s}^t$ determines to

$$C(y_{s:t}) = 2\sum_{i=s}^{t}\left[\lambda_i y_i - \log\lambda_i\right] \tag{9}$$

and as the true parameter values are unknown, they are set to their maximum likelihood estimates.

The approach of formulating the single changepoint method as a test, as done for Normal data, is also applicable for Exponential data. Hence, we test for a sequence $\{y_i\}_{i=s}^t$ the hypothesis

$$H_0 : \lambda_s = \cdots = \lambda_t = \lambda$$

versus the alternative of an unknown changepoint τ

$$H_1 : \lambda_s = \cdots = \lambda_\tau \neq \lambda_{\tau+1} = \cdots = \lambda_t.$$

Consequently, this results in a cost function of

$$C(y_{s:t}) = 2\sum_{i=s}^{t}\left[\hat{\lambda}_{s:t} y_i - \log\hat{\lambda}_{s:t}\right], \tag{10}$$

where $\hat{\lambda}_{s:t}$ is the maximum likelihood estimate of the parameter λ_s based on the data $\{y_i\}_{i=s}^t$. The algorithms of binary segmentation and optimal partitioning are than applied as for the previous case of Normal data under consideration of equation (10).

We generate two time-series with multiple changepoints in order to evaluate the performance of binary segmentation and optimal partitioning. The set of changepoints is the same in both time-series, $\tau_1 = 81$, $\tau_2 = 130$, $\tau_3 = 162$ and $\tau_4 = 226$. For the different values of the parameter, we select $\boldsymbol{\lambda}^{(1)} = (1.4, 0.3, 0.1, 1.9, 0.1)$ for the first and $\boldsymbol{\lambda}^{(2)} = (0.3, 1.9, 0.4, 1.2, 0.7)$ for the second time-series. The results presented in Figure 2 illustrate the detected changepoints for the two time-series by binary segmentation and optimal partitioning.

The results for the first time-series are quite good as all changepoints are detected by both methods. However, each method detects one changepoint where the parameter stays constant. For the second time-series binary segmentation detects the exact number of changepoints, but it detects a wrong changepoint at the end and not τ_4. Further, optimal partitioning detects all true changepoints and two changepoints at which the true parameter value stays constant. Nevertheless, in all cases at least three changepoints are correctly detected.

4 Discussion

This work aimed to present methods to detect changepoints in variance for time-series. Both methods considered, binary segmentation and optimal partitioning, are based on a minimisation problem that includes a penalty term to avoid overfitting. Optimal partitioning solves the related minimisation problem exactly. In contrast, binary segmentation only approximates the exact solution. Another example for an exact method is segment neighbourhoods by Auger and Lawrence (1989); however it has a lower computational efficiency than optimal partitioning (Killick et al., 2012). In contrast, binary segmentation is more computationally efficient than optimal partitioning. Both approaches fit the changepoints in simulated time-series quite well but often missed at least one changepoint.

In the context of applications to detect changes in parameters of intensive care patients, both approaches seem applicable as the number of computations is at maximum quadratic in the number of observations. If we keep all observations, optimal partitioning only needs to perform as many computations as we have observations to detect whether the second to last observation is a changepoint. As the computational efficiency for binary segmentation is $\mathcal{O}(n \log(n))$, optimal partition is even faster than binary segmentation for continuously incoming new observations because for binary segmentation all calculations have to be done again. For the case of changes over an already observed time period, like for the GOMOS hindcast time-series considered by Killick et al. (2010), binary segmentation is more computationally efficient.

Recent literature explores methods to improve the computational efficiency of the exact solution of the minimisation problem for a linear penalty term. Killick et al. (2012) introduce the Pruned Exact Linear Time (PELT) method which allows to compute the global minimum of the optimisation problem (1) with a linear computational cost by pruning. In this approach, the computationally efficiency of optimal partitioning is improved by removing values which can never be minima at each iteration step (see Section 3 of Killick et al. (2012) for details).

References

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

Chen, J. and Gupta, A. (1997). Testing and locating variance changepoints with application to stock prices. *Journal of the American Statistical Association*, 92(438):739–747.

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

Fearnhead, P. (2006). Exact and efficient bayesian inference for multiple changepoint problems. *Statistical Computing*, 16:203–213.

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

Inclan, C. and Tiao, G. (1994). Use of cumulative sum of squares for retrospective detection of changes of variance. *Journal of the American Statistical Association*, 89:912–923.

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

Killick, R., Eckley, I., Ewans, K., and Jonathan, P. (2010). Detection of changes in variance of oceanographic time-series using changepoint analysis. *Ocean Engineering*, 37:1120–1126.

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

Lio, P. and Vanucci, M. (2000). Wavelet changepoint prediction of transmembrane proteins. *Bioinformatics*, 16:376–382.

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

Yao, Y. (1988). Estimating the number of changepoints via schwarz' criterion. *Statistics and Probability Letters*, 6:181–189.

Zhou, Y., Wan, A., Xie, S., and Wang, X. (2010). Wavelet analysis of change-points in a nonparametric regression with heteroscedastic variance. *Journal of Econometrics*, 259:183–201.