Supplementary Material for Detecting changes in slope with an L_0 penalty

A Proof of Theorem 2.1

Throughout this section m will denote the true number of changepoints. When we consider possible segmentations with a general number of changepoints, we will tend to let d denote the number of changepoints. For data $\mathbf{Y}_{1:n}$ denote the penalised cost of segmenting the data with d changepoints $\hat{\tau}_{1:d}$ by

$$Q(\mathbf{Y}_{1:n};\hat{\tau}_{1:d}) = \min_{\boldsymbol{\phi}} \left[\sum_{i=0}^{d} \left\{ \mathcal{C}(\mathbf{Y}_{\hat{\tau}_i+1:\hat{\tau}_{i+1}}, \phi_{\hat{\tau}_i}, \phi_{\hat{\tau}_{i+1}}) + h(\hat{\tau}_{i+1}-\hat{\tau}_i) \right\} + \beta_n(d+1) \right]. \quad (10)$$

Further, denote the unpenalised cost by

$$Q_0(\mathbf{Y}_{1:n};\hat{\tau}_{1:d}) = \min_{\boldsymbol{\phi}} \left\{ \sum_{i=0}^{d} \mathcal{C}(\mathbf{Y}_{\hat{\tau}_i+1:\hat{\tau}_{i+1}}, \phi_{\hat{\tau}_i}, \phi_{\hat{\tau}_{i+1}}) \right\}. \quad (11)$$

We will allow the second argument of both of these functions to be an unordered vector of changepoints, in which case the penalised, or unpenalised, cost is calculated in the obvious way: we remove any duplicate changepoints, order the changepoints and use either (10) or (11) for the ordered changepoints. We also allow the vector of changepoints to include times outside the time-interval for the data – in which case those changepoints are ignored. We write $Q_0(\mathbf{Y}_{1:n})$ for the unpenalised cost if we fit a model with no changepoints.

We base our proof on related proofs for consistency of the number and location of changepoints for change in mean (e.g. Yao, 1988). The extra complication comes from the cost associated with a given segment depending on the location of the other changepoints. To overcome this issue we will use the property of our model that if we add two changepoints at consecutive time-points then the costs associated with segmenting the data before and the data after the pair of changepoints can be calculated independently of each others. So given a set of d_1 changepoints prior to t, $\hat{\boldsymbol{\tau}}_{1:d_1}$ and a set of d_2 changepoints after $t+1$, $\hat{\boldsymbol{\tau}}_{(d_1+1):(d_1+d_2)}$, then

$$Q_0(\boldsymbol{Y}_{1:n}; \hat{\boldsymbol{\tau}}_{1:(d_1+d_2)}, t, t+1) = Q_0(\boldsymbol{Y}_{1:t}, \hat{\boldsymbol{\tau}}_{1:d_1}) + Q_0(\boldsymbol{Y}_{(t+1):n}, \hat{\boldsymbol{\tau}}_{(d_1+1):(d_1+d_2)}). \tag{12}$$

This can be shown by a simple reparameterisation between the change in slope model fitted for the left-hand side of the equation and the two change in slope models fitted on the right-hand side. As adding changepoints can only lead to a reduction in the unpenalised cost, this gives the following way of bounding the residual sum of squares associated with a given segmentation, which we repeatedly use. For any $s = 1, \ldots, n$,

$$Q_0(\boldsymbol{Y}_{1:n}; \hat{\boldsymbol{\tau}}_{1:d}) - \sum_{t=1}^{n} Z_t^2 \geq \left\{ Q_0(\boldsymbol{Y}_{1:s}; \hat{\boldsymbol{\tau}}_{1:d}) - \sum_{t=1}^{s} Z_t^2 \right\} + \left\{ Q_0(\boldsymbol{Y}_{(s+1):n}; \hat{\boldsymbol{\tau}}_{1:d}) - \sum_{t=s+1}^{n} Z_t^2 \right\}, \tag{13}$$

where, as defined above, we interpret $Q_0(\boldsymbol{Y}_{1:s}, \hat{\boldsymbol{\tau}}_{1:d})$, say, as the unpenalised cost for segmenting $\boldsymbol{Y}_{1:s}$ using just the subset of the changepoints $\hat{\boldsymbol{\tau}}_{1:d}$ that lie between time 1 and time $s-1$.

We define three events which depend on $Y_{1:n}$. The first of these, which we call E_n^1, is the event that, for suitable constants $\alpha > 0$ and $\alpha' > 0$,

$$\max_{i=0,\ldots,m} \left[\max_{d, \hat{\boldsymbol{\tau}}_{1:d}} \left\{ Q_0(\boldsymbol{Y}_{(\tau_i+1):\tau_{i+1}}; \hat{\boldsymbol{\tau}}_{1:d}) - \sum_{t=\tau_i+1}^{\tau_{i+1}} Z_t^2 + d\alpha \log n + \alpha' \sqrt{\log n} \right\} \right] > 0.$$

This event states that if you consider any segment, then the unpenalised cost for fitting just the data in that segment with changepoints $\hat{\boldsymbol{\tau}}_{1:d}$ is less than $d\alpha \log n + \alpha'(\log n)^{1/2}$ lower than the sum of the square of the true residuals for that segment. This holds for all segments and all choices of changepoints.

The second event, E_n^2, is that for $l_n = \lfloor \delta_n/2 \rfloor$

$$\min_{i=1,\ldots,m} \left\{ Q_0(\boldsymbol{Y}_{(\tau_i-l_n+1):(\tau_i+l_n)}) - Q_0(\boldsymbol{Y}_{(\tau_i-l_n+1):(\tau_i+l_n)}; \tau_i) \right\} > \frac{1}{50} l_n^3 \Delta_n^2$$

This event states that if you consider the l_n data points either side of any changepoint, then the reduction in the unpenalised cost of fitting a model with the true change, as compared to fitting a model with no change, to this data is greater than a term proportional to $l_n^3\Delta_n^2$. This holds for all m changepoints.

The final event, E_n^3 is similar to that for E_n^2 but with a different number of data points associated with each changepoint. For $i = 1, \dots, m$ let $l_n^i = \lfloor C_2(\log n)^{1/3}(\Delta_n^i)^{-2/3}\rfloor$ with C_2 as defined in the statement of Theorem 2.1. The event E_n^3, is that

$$\min_{i=1,\dots,m}\left[\left\{Q_0(\boldsymbol{Y}_{(\tau_i-l_n^i+1):(\tau_i+l_n^i)}) - Q_0(\boldsymbol{Y}_{(\tau_i-l_n^i+1):(\tau_i+l_n^i)};\tau_i)\right\} - \frac{1}{50}(l_n^i)^3(\Delta_n^i)^2\right] > 0.$$

Lemmas B.1 and B.3, which are stated and proved in Section B, show that each of these three events occurs with probability tending to 1. Thus in the following we will assume they hold, and show that if they do then, for sufficiently large n, the event in the statement of Theorem 2.1 must also hold. We will do this in three stages.

First we show that, for sufficiently large n, $\hat{m}_n \geq m$ if E_n^2 occurs. To do this we consider an arbitrary segmentation of the data $\hat{\boldsymbol{\tau}}_{1:d}$ with $d < m$ changepoints, and show that the penalised cost for this segmentation must be higher than the cost of another segmentation.

For such a segmentation, there must exist at least one true changepoint such that no estimated changepoint lies within half the minimum segment length, $l_n = \lfloor\delta_n/2\rfloor$, of it. Denoting such a changepoint by τ_i,

$$\begin{aligned}
&Q_0(\boldsymbol{Y}_{1:n};\hat{\boldsymbol{\tau}}_{1:d}) \geq Q_0(\boldsymbol{Y}_{1:(\tau_i-l_n)};\hat{\boldsymbol{\tau}}_{1:d}) + Q_0(\boldsymbol{Y}_{(\tau_i-l_n+1):(\tau_i+l_n)}) + Q_0(\boldsymbol{Y}_{(\tau_i+l_n+1):n};\hat{\boldsymbol{\tau}}_{1:d})\\
&\quad > \; Q_0(\boldsymbol{Y}_{1:(\tau_i-l_n)};\hat{\boldsymbol{\tau}}_{1:d}) + Q_0(\boldsymbol{Y}_{(\tau_i-l_n+1):(\tau_i+l_n)};\tau_i) + Q_0(\boldsymbol{Y}_{(\tau_i+l_n+1):n};\hat{\boldsymbol{\tau}}_{1:d}) + l_n^3\Delta_n^2/50\\
&\quad = \; Q_0(\boldsymbol{Y}_{1:n};\hat{\boldsymbol{\tau}}_{1:d},\tau_i-l_n,\tau_i-l_n+1,\tau_i,\tau_i+l_n,\tau_i+l_n+1) + l_n^3\Delta_n^2/50
\end{aligned}$$

The first inequality comes from (13). We have then used (12) and the bound on the change of unpenalised cost from adding a true changepoint that comes from event E_n^2. The penalised cost

$$Q(\boldsymbol{Y}_{1:n};\hat{\boldsymbol{\tau}}_{1:d}) - Q(\boldsymbol{Y}_{1:n};\hat{\boldsymbol{\tau}}_{1:d},\tau_i-l_n,\tau_i-l_n+1,\tau_i,\tau_i+l_n,\tau_i+l_n+1)$$

is thus bounded below by $\Delta_n^2 l_n^3/50 - 5|\gamma|\log n - 5\beta_n$. By the assumptions on Δ_n and δ_n, $\log n = o(\Delta_n^2 l_n^3)$. If $\beta_n = o(\Delta_n^2 l_n^3)$ this will be positive for sufficiently large n. This argument applies for any segmentation with fewer than m changepoints, and hence for sufficiently large n, if E_n^2 occurs then no segmentation with fewer than m changepoints can minimise the penalised cost.

Next we show that $\hat{m}_n \leq m$ if the event E_n^1 occurs. To do this we consider an arbitrary segmentation of the data $\hat{\boldsymbol{\tau}}_{1:d}$ with $d > m$ changepoints, and show that the penalised cost for this segmentation must be higher than the cost of the true segmentation. First note that

$$\sum_{t=1}^{n} Z_t^2 \geq Q_0(\boldsymbol{Y}_{1:n}; \boldsymbol{\tau}_{1:m}).$$

Hence

$$Q(\boldsymbol{Y}_{1:n}; \hat{\boldsymbol{\tau}}_{1:d}) - Q(\boldsymbol{Y}; \boldsymbol{\tau}_{1:m}) \geq Q_0(\boldsymbol{Y}_{1:n}; \hat{\boldsymbol{\tau}}_{1:d}) - \sum_{t=1}^{n} Z_t^2 - d|\gamma|\log n + (d-m)\beta_n,$$

where we have used a simple bound on the difference in the contribution of the $h(\cdot)$ terms to the two penalised costs. We can bound the first part of the right-hand side by repeated application of (13):

$$\begin{aligned} Q_0(\boldsymbol{Y}_{1:n}; \hat{\boldsymbol{\tau}}_{1:d}) - \sum_{t=1}^{n} Z_t^2 &\geq \sum_{i=0}^{m} \left\{ Q_0(\boldsymbol{Y}_{(\tau_i+1):\tau_{i+1}}; \hat{\boldsymbol{\tau}}_{1:d}) - \sum_{t=\tau_i+1}^{\tau_{i+1}} Z_t^2 \right\} \\ &> -\alpha d \log n - \alpha'(m+1)\sqrt{\log n}. \end{aligned}$$

The last inequality comes from using event E_n^1 to bound the contribution from each term in the sum. If $\beta_n > C_1 \log n$ then

$$Q(\boldsymbol{Y}_{1:n}; \hat{\boldsymbol{\tau}}_{1:d}) - Q(\boldsymbol{Y}; \boldsymbol{\tau}_{1:m}) > \{C_1(d-m) - d|\gamma| - \alpha d\} \log n - \alpha'(m+1)\sqrt{\log n}.$$

For $C_1 > m(|\gamma| + \alpha)$ this is positive for all $d > m$ for sufficiently large n. Hence there exists a constant C_1 such that if $\beta_n > C_1 \log n$ a segmentation with $d > m$ will never minimise the penalised cost.

Taken together, the results shown so far show that $\hat{m}_n = m$ with probability tending to 1. The final part of the proof is to show that there exists a constant, C_2, such that with probability tending to 1

$$\max_{i=1,\dots,m} \left\{ |\hat{\tau}_i - \tau_i| \, (\Delta_n^i)^{2/3} \right\} \le C_2 (\log n)^{1/3}. \tag{14}$$

We show that this is guaranteed, for sufficiently large n, if all events occur. Similar to before, our proof will be to consider an arbitrary segmentation for which (14) does not hold, and show that it cannot minimise the penalised cost. We will consider only n large enough that l_n^i is greater than δ_n for all i. This must be occur for large enough n as l_n increases at rate that is bounded above by a constant times $(\log n/\Delta_n)^{1/3}$, while by the assumptions of the Theorem δ_n increases at a strictly faster rate.

As $\hat{m}_n = m$ with probability tending to 1, we need only consider segmentations with m changes. Let $\hat{\boldsymbol{\tau}}_{1:m}$ be such a segmentation for which (14) does not hold, and let τ_i be a changepoint for which

$$|\hat{\tau}_i - \tau_i| \, (\Delta_n^i)^{2/3} > C_2 (\log n)^{1/3}.$$

Define an event, E_n^4, to be the event that both

$$\max_{\hat{\boldsymbol{\tau}}_{1:d}} \left\{ Q_0(\boldsymbol{Y}_{(\tau_{i-1}+1):(\tau_i - l_n^i)}; \hat{\boldsymbol{\tau}}_{1:d}) - \sum_{t=\tau_{i-1}+1}^{\tau_i - l_n^i} Z_t^2 \right\} + d\alpha \log n + \alpha' \sqrt{\log n} > 0,$$

and

$$\max_{\hat{\boldsymbol{\tau}}_{1:d}} \left\{ Q_0(\boldsymbol{Y}_{(\tau_i + l_n^i + 1):(\tau_{i+1})}; \hat{\boldsymbol{\tau}}_{1:d}) - \sum_{t=\tau_i + l_n^i + 1}^{\tau_{i+1}} Z_t^2 \right\} + d\alpha \log n + \alpha' \sqrt{\log n} > 0,$$

occur for all i. This will occur with probability tending to 1 by Lemma B.1.

We have

$$Q(\boldsymbol{Y}_{1:n}; \hat{\boldsymbol{\tau}}_{1:m}) - Q(\boldsymbol{Y}_{1:n}; \boldsymbol{\tau}_{1:m}) \ge Q_0(\boldsymbol{Y}_{1:n}; \hat{\boldsymbol{\tau}}_{1:m}) - \sum_{t=1}^{n} Z_t^2 - m|\gamma| \log n.$$

Now using (13)

$$Q_0(\boldsymbol{Y}_{1:n};\hat{\boldsymbol{\tau}}_{1:m}) - \sum_{t=1}^{n} Z_t^2 \geq \sum_{j=0}^{i-2}\left\{Q_0(\boldsymbol{Y}_{(\tau_j+1):\tau_{j+1}};\hat{\boldsymbol{\tau}}_{1:m}) - \sum_{t=\tau_j+1}^{\tau_{j+1}} Z_t^2\right\}$$
$$\sum_{j=i+1}^{m}\left\{Q_0(\boldsymbol{Y}_{(\tau_j+1):\tau_{j+1}};\hat{\boldsymbol{\tau}}_{1:m}) - \sum_{t=\tau_j+1}^{\tau_{j+1}} Z_t^2\right\} + \left\{Q_0(\boldsymbol{Y}_{(\tau_{i-1}+1):(\tau_i-l_n^i)};\hat{\boldsymbol{\tau}}_{1:m}) - \sum_{t=\tau_{i-1}+1}^{\tau_i-l_n^i} Z_t^2\right\} +$$
$$\left\{Q_0(\boldsymbol{Y}_{(\tau_i+l_n^i+1):\tau_{i+1}};\hat{\boldsymbol{\tau}}_{1:m}) - \sum_{t=\tau_i+l_n^i+1}^{\tau_{i+1}} Z_t^2\right\} + \left\{Q_0(\boldsymbol{Y}_{(\tau_i+l_n^i+1):(\tau_i+l_n^i)}) - \sum_{t=\tau_i-l_n^i+1}^{\tau_i+l_n^i} Z_t^2\right\}, \quad (15)$$

where we interpret a sum from $j = 0$ to -1, or from $j = m + 1$ to m as having the value 0. If E_n^1 and E_n^4 occur then we can lower bound the sum of all terms except the final one by $-m\alpha \log n - (m + 1)\alpha' \sqrt{\log n}$

The final term on the right-hand side of (15) can be written as

$$\{Q_0(\boldsymbol{Y}_{(\tau_i-l_n^i+1):(\tau_i+l_n^i)}) - Q_0(\boldsymbol{Y}_{(\tau_i-l_n^i+1):(\tau_i+l_n^i)};\tau_i)\} + \left\{Q_0(\boldsymbol{Y}_{(\tau_i-l_n^i+1):(\tau_i+l_n^i)};\tau_i) - \sum_{t=\tau_i-l_n^i+1}^{\tau_i+l_n^i} Z_t^2.\right\}$$

Using events E_n^3 and E_n^1, the two bracketed terms on the right-hand side can be bounded below by $\frac{1}{50}(l_n^i)^3(\Delta_n^i)^2$ and $-\alpha \log n - \alpha' \sqrt{\log n}$ respectively.

Thus

$$Q(\boldsymbol{Y}_{1:n};\hat{\boldsymbol{\tau}}_{1:m}) - Q(\boldsymbol{Y}_{1:n};\boldsymbol{\tau}_m) > \frac{1}{50}(l_n^i)^3(\Delta_n^i)^2 - (m+1)\alpha \log n - (m+2)\alpha' \sqrt{\log n} - |\gamma| m \log n. \quad (16)$$

By the definition of l_n^i,

$$(l_n^i)^3(\Delta_n^i)^2 = (C_2)^3 \log n + o(\log n),$$

and thus we can choose C_2 such that (16) is positive for large enough n. □

B Lemmas for Proof of Theorem 2.1

Throughout this section $Z_1, Z_2, \ldots$ will denote an infinite set of independent, identically distributed standard Gaussian random variables.

The following lemmas show that each of E_n^1, E_n^2 and E_n^3 occur with probability tending to 1.

Lemma B.1 *Consider data from a segment of length l,*

$$Y_t = \phi_0 + \frac{\phi_1 - \phi_0}{l} t + Z_t, \text{ for } t = 1, \ldots, l.$$

where, without loss of generality, we have assumed this is the first segment. Fix $\epsilon > 0$ and choose any constant $\alpha > 2(1+\epsilon)$. For any set of $d \geq 1$ changepoints $\tau_{1:d}$ with $0 < \tau_1 < \cdots < \tau_d < l$, there exists a constant C independent of l, d and the changepoint locations such that

$$\Pr\left(\sum_{t=1}^{l} Z_t^2 - Q_0(Y_{1:l}; \tau_{1:d}) > d\alpha \log l\right) \leq C l^{-d(1+\epsilon)}; \tag{17}$$

and for any $\alpha' > 0$,

$$\Pr\left(\sum_{t=1}^{l} Z_t^2 - Q_0(Y_{1:l}) > \alpha' \sqrt{\log l}\right) \to 0 \tag{18}$$

as $l \to \infty$.

Furthermore as $l \to \infty$,

$$\Pr\left\{\max_{d, \tau_{1:d}} \left(\sum_{t=1}^{l_n} Z_t^2 - Q_)(Y_{1:l}; \tau_{1:d}) - d\alpha \log l - \alpha' \sqrt{\log l}\right) > 0\right\} \to 0 \tag{19}$$

Proof. For the first set of results $\tau_{1:d}$ is a fixed set of d changepoints. Standard results for the normal linear model give,

$$\sum_{t=1}^{l} Z_t^2 - Q_0(Y_{1:l}; \tau_{1:d}) \sim \chi^2_{d+2},$$

as we are fitting a model with $d+2$ parameters. We can bound the upper tail of this random variable using (see e.g. Lemma 8.1 of Birgé, 2001)

$$\Pr\left\{\chi^2_{d+2} > (d+2) + 2\sqrt{(d+2)x} + 2x\right\} \leq \exp(-x). \tag{20}$$

For any $\alpha > 2(1+\epsilon)$, for large enough l and any integer $d > 0$

$$d\alpha \log l > (d+2) + 2\sqrt{(d+2)d(1+\epsilon)\log l} + 2d(1+\epsilon)\log l,$$

and hence there exists an L_0 such that for $l > L_0$, using (20) with $x = d(1+\epsilon)\log l$,

$$\Pr(\chi^2_{d+2} > d\alpha \log l) \leq \exp\{-d(1+\epsilon)\log l\} = l^{-d(1+\epsilon)}.$$

As we can choose an L_0 independent of d, this is sufficient to prove (17).

To show (18) we use (20) with $d = 0$. For any $\alpha' > 0$

$$\alpha' \sqrt{\log l} > 2 + 2\sqrt{2x} + 2x,$$

where $x = (\alpha'/3)(\log l)^{1/2}$, for large enough l. Hence for large enough l

$$\Pr\left(\sum_{t=1}^{l} Z_t^2 - Q_0(Y_{1:l}) > \alpha' \sqrt{\log l}\right) \leq \exp\left(a\frac{\alpha'}{3}\sqrt{\log l}\right),$$

and the right-hand side tends to 0 as $l \to \infty$.

To show (19) holds it is sufficient to sum the probabilities in (17) over all segmentations of $Y_{1:l_n}$ and show this sum tends to 0. To do this note that we can bound the number of segmentations with d changepoints by l^d. Thus

$$\Pr\left\{\max_{d,\tau_{1:d}}\left(\sum_{t=1}^{l} Z_t^2 - Q_0(Y_1 : l; \tau_{1:d}) - d\alpha \log l - \alpha'\sqrt{\log l}\right) > 0\right\} \leq \sum_{d=1}^{l-1} l^d C l^{-d(1+\epsilon)} < C\sum_{d=1}^{\infty} l^{-d\epsilon}.$$

This is just $Cl^{-\epsilon}/(1 - l^{-\epsilon})$ which, as $\epsilon > 0$, tends to 0 as $l \to \infty$ as required. □

Corollary B.2 *Event $E_n^{(1)}$ occurs with probability tending to 1 as $n \to \infty$.*

Proof. This follows immediately from using (19) for each of the $m + 1$ segments. □

Lemma B.3 *For a given l and any ϕ_0, ϕ_1 and ϕ_2 with*

$$\Delta = \left|\frac{\phi_1 - \phi_0}{l} - \frac{\phi_2 - \phi_1}{l}\right|$$

let

$$Y_t = \phi_0 + \frac{\phi_1 - \phi_0}{l} t + Z_t, \text{ for } t = 1, \dots, l, \text{ and}$$
$$Y_t = \phi_1 + \frac{\phi_2 - \phi_1}{l}(t - l) + Z_t, \text{ for } t = l + 1, \dots, 2l.$$

Then for $l > 2$

$$\Pr\left(Q_0(\boldsymbol{Y}_{1:2l}) - Q_0(\boldsymbol{Y}_{1:2l}; l) < \frac{1}{50}\Delta^2 l^3\right) \le \exp\left\{-\frac{1}{800}\Delta^2 l^3\right\}.$$

Proof. Standard results for the normal linear model (e.g Theorem 15.8 of Muller and Stewart, 2006) give that, for $l > 2$, $Q_0(\boldsymbol{Y}_{1:2l}) - Q_0(\boldsymbol{Y}_{1:2l}; l)$ has a non-central chi-squared distribution with 1 degree of freedom, and non-centrality parameter

$$\nu = \Delta^2 \frac{l(l+1)(l-1)}{24}\left\{\frac{4l^2+2}{4l^2-1}\right\}.$$

For $l > 2$, $\nu > \Delta^2 l^3/25$. We can bound the lower tail of such a random variable, $\chi^2_1(\nu)$, using (see e.g. Lemma 8.1 of Birgé, 2001)

$$\Pr\left(\chi^2_1(\nu) < 1 + \nu - 2\sqrt{(1+2\nu)x}\right) \le \exp\{-x\}.$$

Taking $x = (1+2\nu)/64$, and noting that for such an x, $(\nu+1) - 2\sqrt{(1+2\nu)x} > \nu/2$, we get

$$\Pr\left(Q_0(\boldsymbol{Y}_{1:2l}) - Q_0(\boldsymbol{Y}_{1:2l}; l) < \frac{1}{50}\Delta^2 l^3\right) \le \Pr\left(Q_0(\boldsymbol{Y}_{1:2l}) - Q_0(\boldsymbol{Y}_{1:2l}; l) < \nu/2\right) \le \exp\{-\nu/32\}.$$

The result follows by noting that $\nu > l^3\Delta^2/25$ for $l > 2$. □

Corollary B.4 *Events $E_n^{(2)}$ and $E_n^{(3)}$ occur with probability tending to 1 as $n \to \infty$.*

Proof. We can apply Lemma B.3 to each region around a changepoint as $l_n > 2$ for sufficiently large n. For event E_n^2, as $\Delta_n^2 l_n^3 \to \infty$ the probability of

$$Q_0(\boldsymbol{Y}_{\tau_i - l_n + 1:\tau_i + l_n}) - Q_0(\boldsymbol{Y}_{\tau_i - l_n + 1:\tau_i + l_n}; \tau_i) > \frac{1}{50} l_n^3 \Delta_n^2$$

for a given changepoint, τ_i, tends to 1. As there are a fixed number of changepoints, we get that this must hold for all changepoints with probability tending to 1, as required. A similar argument holds for event E_n^3. □

C Updates for Quadratic Functions

In Section 3 (equation 5) we define a function, $f^t_{\boldsymbol{\tau}}(\phi)$, as the minimum cost of segmenting $\mathbf{y}_{1:t}$ with changepoints at $\boldsymbol{\tau} = \tau_1, \ldots, \tau_k$ and fitted value $\phi_t = \phi$ at time t. We then derived a recursion for these functions as follows

$$f^t_{\boldsymbol{\tau}}(\phi) = \min_{\phi'} \left\{ f^{\tau_k}_{\tau_1,\ldots,\tau_{k-1}}(\phi') + \mathcal{C}(y_{\tau_k+1:t}, \phi', \phi) + \beta + h(\tau_{i+1} - \tau_i) \right\}. \tag{21}$$

The functions $f^t_{\boldsymbol{\tau}}(\phi)$ are quadratics in ϕ, and we denote $f^t_{\boldsymbol{\tau}}(\phi)$ as follows

$$f^t_{\boldsymbol{\tau}}(\phi) = a^t_{\boldsymbol{\tau}} + b^t_{\boldsymbol{\tau}}\phi + c^t_{\boldsymbol{\tau}}\phi^2, \tag{22}$$

for some constants $a^t_{\boldsymbol{\tau}}$, $b^t_{\boldsymbol{\tau}}$ and $c^t_{\boldsymbol{\tau}}$. We then wish to calculate these coeffcients by updating the coefficients that make up $f^{\tau_k}_{\tau_1,\ldots,\tau_{k-1}}(\phi')$ using (21). To do this we need to write the cost for the segment from $\tau_k + 1$ to t in quadratic form. Defining the length of the segment as $s = t - \tau_k$ this cost can be written as

$$\begin{aligned}\mathcal{C}(\mathbf{y}_{\tau_k+1:t}, \phi', \phi) =& \frac{(s+1)(2s+1)}{6s\sigma^2}\phi^2 + \left(\frac{(s+1)}{\sigma^2} - \frac{(s+1)(2s+1)}{3s\sigma^2} \right)\phi'\phi \\ & - \left(\frac{2}{s\sigma^2} \sum y_j(j-\tau_k) \right)\phi + \left(\frac{1}{\sigma^2} \sum y_i^2 \right) \\ & + 2\left(\frac{1}{s\sigma^2} \sum y_j(j-\tau_k) - \frac{1}{\sigma^2} \sum y_i \right)\phi' + \frac{(s-1)(2s-1)}{6s\sigma^2}\phi'^2.\end{aligned} \tag{23}$$

Writing (23) as $A\phi^2 + B\phi'\phi + C\phi + D + E\phi' + F\phi'^2$ for constants A, B, C, D and E, substituting (23) into (21) and minimising out ϕ' we can get the formula for the updating the coefficients of the quadratic $f^t_{\boldsymbol{\tau}}(\phi)$:

$$
\begin{aligned}
a_{\boldsymbol{\tau}}^{t} &= A - \frac{B^2}{4\left(a_{(\tau_1,\ldots,\tau_{k-1})}^{\tau_k} + F\right)},\\
b_{\boldsymbol{\tau}}^{t} &= C - \frac{\left(b_{(\tau_1,\ldots,\tau_{k-1})}^{\tau_k} + E\right) B}{2\left(a_{(\tau_1,\ldots,\tau_{k-1})}^{\tau_k} + F\right)},\\
c_{\boldsymbol{\tau}}^{t} &= c_{(\tau_1,\ldots,\tau_{k-1})}^{\tau_k} + D - \frac{\left(b_{(\tau_1,\ldots,\tau_{k-1})}^{\tau_k} + E\right)^2}{4\left(a_{(\tau_1,\ldots,\tau_{k-1})}^{\tau_k} + F\right)} + \beta + h(t-\tau_k).
\end{aligned} \tag{24}
$$

D Proofs from Section 3

D.1 Proof of Theorem 3.1

The proof of Theorem 3.1 works by contrapositive. We show that if $(\boldsymbol{\tau}, s) \in \overset{*}{\mathcal{T}}_t$ then a necessary condition of this is that $\boldsymbol{\tau} \in \overset{*}{\mathcal{T}}_s$, taking the contrapositive of this gives Theorem 3.1.

proof Assume $(\boldsymbol{\tau}, s) \in \overset{*}{\mathcal{T}}_t$, then there exists ϕ such that

$$
f^t(\phi) = f^t_{(\boldsymbol{\tau},s)}(\phi),
$$

Now for any ϕ^*,

$$
\begin{aligned}
f^s(\phi^*) + \mathcal{C}(\mathbf{y}_{s+1:t}, \phi^*, \phi) + \beta &\geq \min_{\phi',r}\left[f^r(\phi') + \mathcal{C}(\mathbf{y}_{r+1:t}, \phi', \phi) + \beta\right],\\
&= f^t(\phi),\\
&= f^t_{(\boldsymbol{\tau},s)}(\phi),\\
&= \min_{\phi''}\left\{f^s_{\boldsymbol{\tau}}(\phi'') + \mathcal{C}(\mathbf{y}_{s+1:t}, \phi'', \phi) + \beta\right\}, \qquad (25)\\
&= f^s_{\boldsymbol{\tau}}(\phi^A) + \mathcal{C}(\mathbf{y}_{s+1:t}, \phi^A, \phi) + \beta,
\end{aligned}
$$

where ϕ^A is the value of ϕ'' which minimises (25). As ϕ^* can be chosen as any value, we can choose it as ϕ^A. By cancelling terms we get $f^s(\phi^A) \geq f^s_{\boldsymbol{\tau}}(\phi^A)$ and hence

$f^s(\phi^A) = f^s_{\boldsymbol{\tau}}(\phi^A)$ and therefore $\boldsymbol{\tau} \in \overset{*}{\mathcal{T}}_s$. We have shown that if $(\boldsymbol{\tau}, s) \in \overset{*}{\mathcal{T}}_t$ then $\boldsymbol{\tau} \in \overset{*}{\mathcal{T}}_s$, by taking the contrapositive the theorem holds. $\square$

D.2 Proof of Theorem 3.2

The proof for Theorem 3.2 follow a similar argument to the corresponding proof in Killick et al. (2012). However we have to add a segment consisting of the single point y_{t+1} to deal with the dependence between the segments.

Proof Let $\boldsymbol{\tau}^*$ denote the optimal segmentation of $\mathbf{y}_{1:t}$. We will repeatedly use the fact that

$$\mathcal{C}(y_{t+1}, \phi', \phi) = \frac{1}{\sigma^2}(y_{t+1} - \phi)^2,$$

and this does not depend on ϕ'.

First consider $T = t+1$. As adding a changepoint without penalty will always reduce the cost, it is straightforward to show

$$\begin{aligned}
f^T_{\boldsymbol{\tau}}(\phi) &\geq \min_{\phi'} \left[f^t_{\boldsymbol{\tau}}(\phi') + \mathcal{C}(y_{t+1}, \phi', \phi) \right], \\
&= \min_{\phi'}[f^t_{\boldsymbol{\tau}}(\phi')] + \min_{\phi'}[\mathcal{C}(y_{t+1}, \phi', \phi)], \\
&> \min_{\phi'} \left[f^t(\phi') \right] + K + \min_{\phi'}[\mathcal{C}(y_{t+1}, \phi', \phi)], \\
&\geq \min_{\phi'} \left[f^t(\phi') + \mathcal{C}(y_{t+1}, \phi', \phi) + \beta + h(1) \right].
\end{aligned}$$

Thus segmenting $\mathbf{y}_{1:T}$ with changepoints $\boldsymbol{\tau}$ always has a greater cost than segmenting $\mathbf{y}_{1:T}$ with changepoints $(\boldsymbol{\tau}^*, t)$.

Now we consider $T > t+1$. We start by noting that by adding changes, at any point, without the penalty term and minimising over the corresponding ϕ values will also decrease the cost. Therefore

$$f^T_{\boldsymbol{\tau}}(\phi) \geq \min_{\phi', \phi''} \left[f^t_{\boldsymbol{\tau}}(\phi') + \mathcal{C}(y_{t+1}, \phi', \phi'') + \mathcal{C}(\mathbf{y}_{t+2:T}, \phi'', \phi) \right]. \tag{26}$$

So from (26) and using (8),

$$
\begin{aligned}
f_{\boldsymbol{\tau}}^{T}(\phi) &\geq \min_{\phi',\phi''} \left[f_{\boldsymbol{\tau}}^{t}(\phi') + \mathcal{C}(y_{t+1}, \phi', \phi'') + \mathcal{C}(\mathbf{y}_{t+2:T}, \phi'', \phi) \right], \\
&\geq \min_{\phi'} [f_{\boldsymbol{\tau}}^{t}(\phi')] + \min_{\phi',\phi''} [\mathcal{C}(y_{t+1}, \phi', \phi'') + \mathcal{C}(\mathbf{y}_{t+2:T}, \phi'', \phi)], \\
&> \min_{\phi'} \left[f^{t}(\phi') \right] + K + \min_{\phi',\phi''} [\mathcal{C}(y_{t+1}, \phi', \phi'') + \mathcal{C}(\mathbf{y}_{t+2:T}, \phi'', \phi)], \\
&\geq \min_{\phi',\phi''} \left[f^{t}(\phi') + \mathcal{C}(y_{t+1}, \phi', \phi'') + \beta + h(1) + \mathcal{C}(\mathbf{y}_{t+2:T}, \phi'', \phi) + \beta + h(T-t+1) \right].
\end{aligned}
$$

Therefore the cost of segmenting $\mathbf{y}_{1:T}$ with changepoints $\boldsymbol{\tau}$ is always greater than the cost of segmenting $\mathbf{y}_{1:T}$ with changepoints $(\boldsymbol{\tau}^*, t, t+1)$ (where $\boldsymbol{\tau}^*$ is the optimal segmentation of $\mathbf{y}_{1:t}$) and this holds for all $T > t+1$ and hence $\boldsymbol{\tau}$ can be pruned. $\square$

E Pseudo-Code for CPOP

The CPOP algorithm uses Algorithm 2 to calculate the intervals on which each function is optimal. This then enables the functions that are not optimal for any value of ϕ to be removed. The idea of this algorithm is as follows.

We initialise the algorithm by setting the current parameter value as $\phi_{curr} = -\infty$ and comparing the cost functions in our current set of candidates (which we initialise as $\mathcal{T}_{temp} = \hat{\mathcal{T}}_t$) to get the optimal segmentation for this value, $\boldsymbol{\tau}_{curr}$. This can be optimisation can be done my noting that the quadratic with smallest cost will have the smallest coefficient of the quadratic term. If more than one quadratic has the smallest coefficient, we then choose the quadratic with the largest coefficient of the linear term; and if necessary, then choose the quadratic with the smallest constant term.

For each $\boldsymbol{\tau} \in \mathcal{T}_{curr}$ we calculate where $f_{\boldsymbol{\tau}}^t$ next intercepts with $f_{\boldsymbol{\tau}_{curr}}^t$ (smallest value of ϕ for which $f_{\boldsymbol{\tau}}^t(\phi) = f_{\boldsymbol{\tau}_{curr}}^t(\phi)$ and $\phi > \phi_{curr}$) and store this as $x_{\boldsymbol{\tau}}$. If for a $\boldsymbol{\tau} \in \mathcal{T}_{temp}$ we have $x_{\boldsymbol{\tau}} = \emptyset$ (i.e. $f_{\boldsymbol{\tau}}^t$ doesn't intercept with $f_{\boldsymbol{\tau}_{curr}}^t$ for any $\phi > \phi_{curr}$) then we

Algorithm 1: Algorithm for Continuous Piecewise-linear Optimal Partitioning (CPOP)

Input : Set of data of the form $\mathbf{y}_{1:n} = (y_1, \ldots, y_n)$.

A positive penalty constant, β, and a non-negative, non-decreasing penalty function $h(\cdot)$.

Let $n =$ length of data;

set $\hat{\mathcal{T}}_1 = \{0\}$;

and set $K = 2\beta + h(1) + h(n)$;

for $t = 1, \ldots, n$ **do**

  for $\boldsymbol{\tau} \in \hat{\mathcal{T}}_t$ **do**

    if $\tau = \{0\}$ **then**

      $f^t_{\boldsymbol{\tau}}(\phi) = \min_{\phi'} \mathcal{C}(\mathbf{y}_{1:t}, \phi', \phi) + h(t)$;

    else

      $f^t_{\boldsymbol{\tau}}(\phi) = \min_{\phi'} \left\{ f^{\tau_k}_{\tau_1,\ldots,\tau_{k-1}}(\phi') + \mathcal{C}(\mathbf{y}_{\tau_k+1:t}, \phi', \phi) + h(t-\tau_k) + \beta \right\}$;

  for $\boldsymbol{\tau} \in \hat{\mathcal{T}}_t$ **do**

    $Int^t_{\boldsymbol{\tau}} = \left\{ \phi : f^t_{\boldsymbol{\tau}}(\phi) = \min_{\boldsymbol{\tau}' \in \hat{\mathcal{T}}_t} f^t_{\boldsymbol{\tau}'}(\phi) \right\}$;

    $\overset{*}{\mathcal{T}}_t = \{\boldsymbol{\tau} : Int^t_{\boldsymbol{\tau}} \neq \emptyset\}$;

    $\hat{\mathcal{T}}_{t+1} = \hat{\mathcal{T}}_t \cup \left\{ (\boldsymbol{\tau}, t) : \boldsymbol{\tau} \in \overset{*}{\mathcal{T}}_t \right\}$;

  $\hat{\mathcal{T}}_{t+1} = \left\{ \boldsymbol{\tau} \in \hat{\mathcal{T}}_{t+1} : \min_{\phi} f^t_{\boldsymbol{\tau}}(\phi) \leq \min_{\phi', \boldsymbol{\tau}'} \left[f^t_{\boldsymbol{\tau}'}(\phi') \right] + K \right\}$;

$f_{opt} = \min_{\boldsymbol{\tau}, \phi} f^n_{\boldsymbol{\tau}}(\phi)$;

$\boldsymbol{\tau}_{opt} = \arg\min_{\boldsymbol{\tau}} \left[\min_{\phi} f^n_{\boldsymbol{\tau}}(\phi) \right]$;

Output: The optimal cost, f_{opt}, and the corresponding changepoint vector, $\boldsymbol{\tau}_{opt}$.

remove $\boldsymbol{\tau}$ from $\mathcal{T}_{temp}$. We take the minimum of $x_{\boldsymbol{\tau}}$ (the first of the intercepts) and set it as our new ϕ_{curr} and the corresponding changepoint vector that produces it as $\boldsymbol{\tau}_{curr}$. We repeat this procedure until the set $\mathcal{T}_{temp}$ consists of only a single value $\boldsymbol{\tau}_{curr}$

which is the optimal segmentation for all future $\phi > \phi_{curr}$.

As written, our algorithm assumes there is a unique quadratic that is optimal for each interval – which we believe will happen with probability 1. If this is not the case, we can interpret the algorithm as choosing one of the optimal quadratics, and outputing an optimal, as opposed to the unique optimal, segmentation. Obviously the algorithm could be re-written to store and output multiple optimal segmentations if they exist.

Algorithm 2: Algorithm for calculation of $Int^t_{\boldsymbol{\tau}}$ at time t

---

Input : Set of changepoint candidate vectors $\hat{\mathcal{T}}_t$ for current timestep, t,
Optimal segmentation functions $f^t_{\boldsymbol{\tau}}(\phi)$ for current time step t and $\boldsymbol{\tau} \in \hat{\mathcal{T}}_t$.

$\mathcal{T}_{temp} = \hat{\mathcal{T}}_t$;

$Int^t_{\boldsymbol{\tau}} = \emptyset$ for $\boldsymbol{\tau} \in \hat{\mathcal{T}}_t$;

$\phi_{curr} = -\infty$;

$\boldsymbol{\tau}_{curr} = \arg\min_{\boldsymbol{\tau} \in \mathcal{T}_{temp}} \left[f^t_{\boldsymbol{\tau}}(\phi_{curr}) \right]$;

while $\mathcal{T}_{temp} \backslash \{\boldsymbol{\tau}_{curr}\} \neq \emptyset$ **do**

  for $\boldsymbol{\tau} \in \mathcal{T}_{temp} \backslash \{\boldsymbol{\tau}_{curr}\}$ **do**

    $x_{\boldsymbol{\tau}} = \min\{\phi : f^t_{\boldsymbol{\tau}}(\phi) - f^t_{\boldsymbol{\tau}_{curr}}(\phi) = 0 \;\; \& \;\; \phi > \phi_{curr}\}$;

    if $x_{\boldsymbol{\tau}} = \emptyset$ **then**

      $\mathcal{T}_{temp} = \mathcal{T}_{temp} \backslash \{\boldsymbol{\tau}\}$

  $\boldsymbol{\tau}_{new} = \arg\min_{\boldsymbol{\tau}} (x_{\boldsymbol{\tau}})$;

  $\phi_{new} = \min_{\boldsymbol{\tau}} (x_{\boldsymbol{\tau}})$;

  $Int^t_{\boldsymbol{\tau}_{curr}} = [\phi_{curr}, \phi_{new}] \cup Int^t_{\boldsymbol{\tau}_{curr}}$;

  $\boldsymbol{\tau}_{curr} = \boldsymbol{\tau}_{new}$;

  $\phi_{curr} = \phi_{new}$;

Output: The intervals $Int^t_{\boldsymbol{\tau}}$ for $\boldsymbol{\tau} \in \hat{\mathcal{T}}_t$

References

Birgé, L. (2001). An alternative point of view on Lepski's method. *Lecture Notes-Monograph Series*, pages 113–133.

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

Muller, K. E. and Stewart, P. W. (2006). *Linear model theory: univariate, multivariate, and mixed models*. John Wiley & Sons.

Yao, Y.-C. (1988). Estimating the number of change-points via Schwarz' criterion. *Statistics & Probability Letters*, 6(3):181–189.