SUPPLEMENTARY MATERIAL: Detecting Abrupt Changes in the Presence of Local Fluctuations and Autocorrelated Noise

A Proof of Proposition 1

The initial condition for $Q_1(\mu)$ follows immediately from its definition.

Then, for $t \in \{2, ..., n\}$, we need to condition the problem separately on whether or not we have a changepoint. If we consider no change in the mean of the signal, then we can we can re-arrange the cost at time t based on the cost at time $t-1$ in the following way:

$$Q_t(\mu|\delta_t = 0) = \min_u \left\{ Q_{t-1}(u) + \lambda(\mu - u)^2 + \gamma\Big((y_t - \mu) - \phi(y_{t-1} - u)\Big)^2 \right\} .$$

Similarly, when we have a change:

$$\begin{aligned} Q_t(\mu|\delta_t \neq 0) &= \min_{u,\delta} \left\{ Q_{t-1}(u) + \lambda(\mu - u - \delta)^2 + \gamma\Big((y_t - \mu) - \phi(y_{t-1} - u)\Big)^2 + \beta \right\} \\ &= \min_u \left\{ Q_{t-1}(u) + \gamma\Big((y_t - \mu) - \phi(y_{t-1} - u)\Big)^2 + \beta \right\} \end{aligned}$$

where the second equality comes from minimising over δ.

Lastly, to obtain the whole cost at time t we take the minimum of these two functions:

$$\begin{aligned} Q_t(\mu) &= \min\left\{Q_t(\mu|\delta_t = 0), Q_t(\mu|\delta_t \neq 0)\right\} \\ &= \min_u \left\{ Q_{t-1}(u) + \min\{\lambda(\mu - u)^2, \beta\} + \gamma\Big((y_t - \mu) - \phi(y_{t-1} - u)\Big)^2 \right\} . \end{aligned}$$

□

B Proof of Proposition 2

From the result obtained in Appendix A, simple, albeit tedious, algebraic manipulation enables us to re-write the recursions for $Q_t(\mu|\delta_t \neq 0)$ and $Q_t(\mu|\delta_t = 0)$ in terms of the infimal convolution operator. Let $z_t = y_t - \phi y_{t-1}$.

For $Q_t(\mu|\delta_t \neq 0)$, we can rearrange

$$\begin{aligned}
&\gamma\Big((y_t - \mu) - \phi(y_{t-1} - u)\Big)^2 = \gamma(z_t - \mu + \phi u)^2 \\
&= \gamma(z_t - \mu)^2 + \gamma\phi^2 u^2 + 2\gamma\phi u z_t - 2\gamma\phi u\mu \\
&= \gamma(z_t - \mu)^2 + \gamma\phi^2 u^2 + 2\gamma\phi u z_t + \gamma\phi(u - \mu)^2 - \gamma\phi u^2 - \gamma\phi\mu^2 \\
&= \gamma\phi(u - \mu)^2 - \gamma\phi(1 - \phi)\left(u - \frac{z_t}{1 - \phi}\right)^2 + \gamma\phi\frac{z_t^2}{1 - \phi} + \gamma(z_t - \mu)^2 - \gamma\phi\mu^2
\end{aligned}$$

Hence, we have

$$\begin{aligned}
Q_t(\mu|\delta_t \neq 0) &= \min_{u\in\mathbb{R}}\left[Q_{t-1}(u) - \gamma\phi(1 - \phi)\left(u - \frac{z_t}{1 - \phi}\right)^2 + \gamma\phi(u - \mu)^2\right] \\
&\quad + \frac{\gamma}{1 - \phi}(z_t - (1 - \phi)\mu)^2 + \beta \\
&= \mathsf{INF}_{\mathbb{Q}_{t-1},\gamma\phi}(\mu) + \frac{\gamma}{1 - \phi}\Big(z_t - (1 - \phi)\mu\Big)^2 + \beta = Q_t^{\neq}(\mu),
\end{aligned}$$

where

$$\mathbb{Q}_{t-1}(u) = Q_{t-1}(u) - \gamma\phi(1 - \phi)\left(u - \frac{z_t}{1 - \phi}\right)^2.$$

Similar, for $Q_t(\mu|\delta_t = 0)$, we can rearrange

$$\begin{aligned}
&\lambda(\mu - u)^2 + \gamma\Big((y_t - \mu) - \phi(y_{t-1} - u)\Big)^2 \\
&= (\gamma\phi + \lambda)(u - \mu)^2 - \gamma\phi(1 - \phi)\left(u - \frac{z_t}{1 - \phi}\right)^2 + \gamma\phi\frac{z_t^2}{1 - \phi} + \gamma(z_t - \mu)^2 - \gamma\phi\mu^2.
\end{aligned}$$

Hence

$$Q_t(\mu|\delta_t = 0) \;=\; \mathsf{INF}_{\mathbb{Q}_{t-1},\gamma\phi+\lambda}(\mu) + \frac{\gamma}{1-\phi}\Big(z_t - (1-\phi)\mu\Big)^2 = Q_t^{=}(\mu),$$

where $\mathbb{Q}_{t-1}$ is defined above.

□

If $\phi < 0$ then $\gamma\phi < 0$ and the infimal convolution $\mathsf{INF}_{\mathbb{Q}_{t-1},\gamma\phi}(\mu)$ is not defined. In this case we make a transformation of variable $\tilde{u} = -u$ and $\tilde{\phi} = -\phi$ so that

$$\begin{aligned}
&\gamma\Big((y_t - \mu) - \phi(y_{t-1} - u)\Big)^2 = \gamma(z_t - \mu + \phi u)^2 = \gamma(z_t - \mu + \tilde{\phi}\tilde{u})^2 \\
&= \;\gamma\phi(\tilde{u} - \mu)^2 - \gamma\tilde{\phi}(1-\tilde{\phi})\left(\tilde{u} - \frac{z_t}{1-\tilde{\phi}}\right)^2 + \gamma\tilde{\phi}\frac{z_t^2}{1-\tilde{\phi}} + \gamma(z_t - \mu)^2 - \gamma\tilde{\phi}\mu^2,
\end{aligned}$$

by the same manipulation as given at the start of this section.

Thus using that $Q_{t-1}(u) = Q_{t-1}(\tilde{u})$ we obtain

$$\begin{aligned}
Q_t(\mu|\delta_t \neq 0) \;&=\; \min_{\tilde{u}\in\mathbb{R}}\left[Q_{t-1}(-\tilde{u}) - \gamma\tilde{\phi}(1-\tilde{\phi})\left(\tilde{u} - \frac{z_t}{1-\tilde{\phi}}\right)^2 + \gamma\tilde{\phi}(\tilde{u} - \mu)^2\right] \\
&\quad + \frac{\gamma}{1-\tilde{\phi}}(z_t - (1-\tilde{\phi})\mu)^2 + \beta \\
&=\; \mathsf{INF}_{\tilde{\mathbb{Q}}_{t-1},\gamma\tilde{\phi}}(\mu) + \frac{\gamma}{1-\tilde{\phi}}\Big(z_t - (1-\tilde{\phi})\mu\Big)^2 + \beta = Q_t^{\neq}(\mu),
\end{aligned}$$

where

$$\tilde{\mathbb{Q}}_{t-1}(u) = Q_{t-1}(-u) - \gamma\tilde{\phi}(1-\tilde{\phi})\left(\tilde{u} - \frac{z_t}{1-\tilde{\phi}}\right)^2.$$

Similarly, using the same transformation $\tilde{u} = -u$ and $\tilde{\phi} = -\phi$ we also derive

$$Q_t(\mu|\delta_t = 0) \;=\; \mathsf{INF}_{\tilde{\mathbb{Q}}_{t-1},\gamma\tilde{\phi}+\lambda}(\mu) + \frac{\gamma}{1-\tilde{\phi}}\Big(z_t - (1-\tilde{\phi})\mu\Big)^2 = Q_t^{=}(\mu),$$

where $\tilde{\mathbb{Q}}_{t-1}$ is defined above.

C Proof of Theorem 1

An important property of the convolution is its stability for quadratics: the infimal transformation of a quadratic is a quadratic. Indeed, one can easily prove that the quadratic $q : \mu \mapsto a\mu^2 + b\mu + c$ with $(a, b, c) \in \mathbb{R}^+ \times \mathbb{R}^2$ is transformed into

$$\mathsf{INF}_{q,\omega} : \mu \mapsto \frac{a\omega}{a+\omega}\mu^2 + \frac{b\omega}{a+\omega}\mu + c - \frac{b^2}{4(a+\omega)} .$$

We can also show that q and $\mathsf{INF}_{q,\omega}$ have the same minimum and argminimum. Moreover, $\mathsf{INF}_{q,\omega} \leq q$, resulting in a flattening of the quadratics.

The proof is based on the following lemmas.

Lemma 1 *For any lower-bounded function $Q : \mathbb{R} \to \mathbb{R}$, we define the proxy operator*

$$\hat{u}_\omega : \left\{ \begin{array}{l} \mathbb{R} \to \mathbb{R} \\ \theta \ \mapsto \ \min\Big\{ \underset{u\in\mathbb{R}}{\operatorname{argmin}} \Big(Q(u) + \omega(u-\theta)^2 \Big) \Big\} . \end{array} \right.$$

The function $\hat{u}_\omega$ is non-decreasing on $\mathbb{R}$.

Notice that we use a minimum in the definition of $\hat{u}_\omega$ only to get a single-valued function (we could have done another choice). Indeed, taking $Q = \min(q_1, q_2)$ with $q_1(\theta) = (\theta+1)^2$ and $q_2(\theta) = (\theta-1)^2$, we have $\hat{u}_1(0) = \underset{u\in\mathbb{R}}{\operatorname{argmin}}(Q(u) + u^2) = \{-\frac{1}{2}, \frac{1}{2}\}$ and we need to make a choice (here the smallest value) to get a well-defined function.

Proof: We consider $\theta_1, \theta_2 \in \mathbb{R}$ such that $\theta_1 < \theta_2$ and define $\hat{u}_1 = \hat{u}_\omega(\theta_1)$, $\hat{u}_2 = \hat{u}_\omega(\theta_2)$. Using the definition of $\hat{u}_1$ and $\hat{u}_2$ we can write

$$Q(\hat{u}_1) + \omega(\hat{u}_1 - \theta_1)^2 \leq Q(\hat{u}_2) + \omega(\hat{u}_2 - \theta_1)^2 ,$$

$$Q(\hat{u}_2) + \omega(\hat{u}_2 - \theta_2)^2 \leq Q(\hat{u}_1) + \omega(\hat{u}_1 - \theta_2)^2 .$$

Summing the two inequalities, the Q terms cancel out and we get

$$(\hat{u}_2 - \hat{u}_1)(\theta_2 - \theta_1) \geq 0\,,$$

which shows that $\hat{u}_1 \leq \hat{u}_2$ and the result is proven. □

In our stochastic models the function Q is described by a list of functions $Q = (q_1, ..., q_s)$ with $Q|_{D_i} = q_i$ where $D_i = [d_i, d_{i+1}[\subset \mathbb{R}$ is an interval and $\{D_i\}_{i=1,...,s}$ a partition of the real line. To compute the convolution, we define the functions

$$\overline{q}_i(u) = \begin{cases} q_i(u)\,, & \text{if} \quad u \in D_i\,, \\ +\infty\,, & \text{if} \quad u \notin D_i\,. \end{cases}$$

The infimal convolution of this kind of functions can be analytically described.

Lemma 2 *The infimal convolution of a function $\overline{q}$ given by*

$$\overline{q}(u) = \begin{cases} q(u)\,, & \textit{if} \quad u \in [m_1, m_2]\,, \\ +\infty\,, & \textit{if} \quad u \notin [m_1, m_2]\,, \end{cases}$$

with any function q continuously differentiable (C^1) on $[m_1, m_2]$ is given by

$$\mathsf{INF}_{\overline{q},\omega}(\theta) = \begin{cases} \min\limits_{u \in [m_1, m_2]} \Big(q(u) + \omega(u - \theta)^2 \Big)\,, & \textit{if} \quad \theta \in [m_1^*, m_2^*]\,, \\ q(m_1) + \omega(m_1 - \theta)^2\,, & \textit{if} \quad \theta < m_1^*\,, \\ q(m_2) + \omega(m_2 - \theta)^2\,, & \textit{if} \quad \theta > m_2^*\,, \end{cases} \tag{9}$$

with $[m_1^*, m_2^*] = [\frac{1}{2\omega} q'(m_1) + m_1, \frac{1}{2\omega} q'(m_2) + m_2]$.

Proof: Using Lemma 1 we know that the proxy operator $\hat{u}_\omega$ with $Q = \overline{q}$ is a non-decreasing function in θ. Thus, there exist $m_1^*, m_2^* \in \mathbb{R}$ such that for all $\theta \in [m_1^*, m_2^*]$, the argminimum of $\overline{q}_\omega : u \mapsto \overline{q}(u) + \omega(u - \theta)^2$ belongs to the interval $[m_1, m_2]$ and $\overline{q} = q$ on this interval. As q is C^1, the stationary points of $\overline{q}_\omega$ are solutions of the equation $\frac{1}{2\omega} q'(u) + u = \theta$. At

point m_1 (resp. m_2) we have the argminimum m_1^* with $m_1^* = \frac{1}{2\omega}q'(m_1) + m_1$ (resp. $m_2^* = \frac{1}{2\omega}q'(m_2) + m_2$). If we have $\theta < m_1^*$, then the argminimum of $\overline{q}_\omega$ is less than m_1 and then attained at $u = m_1$ (as $\overline{q}(u) = +\infty$ if $u < m_1$) and we get $\mathsf{INF}_{\overline{q},\omega}(\theta) = q(m_1) + \omega(m_1 - \theta)^2$. With the same reasoning in case $\theta > m_2^*$ the lemma is proven. □

Using these two lemmas, we can prove the following proposition.

Proposition 7 *The infimal convolution of the functional cost $Q = (q_1, ..., q_s)$ is given by $\mathsf{INF}_{Q,\omega} = (\mathsf{INF}_{q_1,\omega}, ..., \mathsf{INF}_{q_s,\omega})$.*

Proof: With previously introduced notations we have $Q(\theta) = \min_{i=1,...,s}\{\overline{q}_i(\theta)\}$. Then

$$\mathsf{INF}_{Q,\omega}(\theta) = \min_{u\in\mathbb{R}}\left(\min_{i=1,...,s}\{\overline{q}_i(\theta)\} + \omega(u-\theta)^2\right) = \min_{u\in\mathbb{R}}\left(\min_{i=1,...,s}\{\overline{q}_i(\theta) + \omega(u-\theta)^2\}\right)$$

$$= \min_{i=1,...,s}\left\{\min_{u\in\mathbb{R}}\left(\overline{q}_i(\theta) + \omega(u-\theta)^2\right)\right\},$$

which gives us $\mathsf{INF}_{Q,\omega}(\theta) = \min_{i=1,...,s}\{\mathsf{INF}_{\overline{q}_i,\omega}(\theta)\}$ for all $\theta \in \mathbb{R}$. $\mathsf{INF}_{Q,\omega}$ can be described by a list $(\mathsf{INF}_{\overline{q}_{\nu(1)},\omega}, \mathsf{INF}_{\overline{q}_{\nu(2)},\omega}, ..., \mathsf{INF}_{\overline{q}_{\nu(r)},\omega})$ with $\nu(i) \in \{1, ..., s\}$. The function $i \mapsto \nu(i)$ is increasing due to Lemma 1 (and $\nu(r) = s$). □

In order to prove Theorem 1 we only need to show that we can remove the overline sign in $(\mathsf{INF}_{\overline{q}_{\nu(1)},\omega}, \mathsf{INF}_{\overline{q}_{\nu(2)},\omega}, ..., \mathsf{INF}_{\overline{q}_{\nu(r)},\omega})$ without consequences. We assume that Q is continuously differentiable (C^1) except at the points d_i for $i = 2, ..., s$. The left and right derivatives at point θ are respectively designated by $Q'_-(\theta)$ and $Q'_+(\theta)$. With these assumptions we can prove the following result.

Lemma 3 *If at points $\theta = d_i$ we have $Q'_-(d_i) > Q'_+(d_i)$ then d_i is never an argminimum for the convolution.*

Proof: We study the stationary points of $Q_\omega : u \mapsto Q(u) + \omega(u-\theta)^2$. The necessary condition for optimality $Q_\omega(u) \leq Q_\omega(u+\epsilon)$ for all ϵ leads to the inequalities

$$\frac{1}{2\omega}Q'_-(u) + u \leq \theta \leq \frac{1}{2\omega}Q'_+(u) + u\,.$$

In case $Q'_-(u) > Q'_+(u)$ there exists no such θ satisfying the two inequalities so that this u can not be used in any minimization of Q_ω and $\hat{u}_\omega$ never takes this value. □

With this result the d_i never appear as an argminimum for the convolution and using Lemma 2, we get $(\mathsf{INF}_{\overline{q}_{\nu(1)},\omega}, \mathsf{INF}_{\overline{q}_{\nu(2)},\omega}, ..., \mathsf{INF}_{\overline{q}_{\nu(r)},\omega}) = (\mathsf{INF}_{q_{\nu(1)},\omega}, \mathsf{INF}_{q_{\nu(2)},\omega}, ..., \mathsf{INF}_{q_{\nu(r)},\omega})$ in Proposition 7.

By looking at updates in Propositions 1 and 2, it remains to prove that at any time step, no slope discontinuity at $\theta = d$ in $Q_t = Q$ satisfies the inequality $Q'_-(d) < Q'_+(d)$. We prove this result by recursion: at the initialisation step, there is no such breakpoint in the cost function and all the min operators involved can not produce them. We eventually have to prove that the infimal transformation in Lemma 2 can not introduce these discontinuities.

Around m_1^* in (9) we have:

$$\frac{d}{d\theta}\mathsf{INF}_{Q,\omega}(\theta) = \begin{cases} \frac{d\hat{u}_\omega(\theta)}{d\theta}q'(\hat{u}_\omega(\theta)) + 2\omega(\frac{d\hat{u}_\omega(\theta)}{d\theta} - 1)(\hat{u}_\omega(\theta) - \theta)\,, & \text{if} \quad \theta \geq m_1^*\,, \\ -2\omega(m_1 - \theta)\,, & \text{if} \quad \theta < m_1^*\,, \end{cases}$$

with the function $\theta \mapsto \hat{u}_\omega(\theta)$ being the argminimum of the infimal convolution (see Lemma 1). By direct computation with $\hat{u}_\omega(m_1^*) = m_1$ and $m_1^* = \frac{1}{2\omega}q'(m_1) + m_1$ we get $\mathsf{INF}'_{Q,\omega-}(m_1^*) = q'(m_1) = \mathsf{INF}'_{Q,\omega+}(m_1^*)$. This result achieves the proof of Theorem 1.

D Algorithm for $\mathsf{INF}_{\mathsf{Q}_t,\omega}$

Algorithm 2 shows how we can now calculate $\mathsf{INF}_{\mathsf{Q}_t,\omega}$ in a linear-in-piece $O(s)$ time complexity. In this algorithm we have input $q_*^i = \mathsf{INF}q_t^i$, where q_t^i is the i^{th} piece-wise quadratic from Q_t with $i \in \{1, ..., s\}$. Algorithm 2 computes the intervals, DOM_*^i such that $\{\text{DOM}_*^{u_i}, i = 1, ..., s^*\}$ is the partition of the real line for $\mathsf{INF}_{\mathsf{Q}_t,\omega}$, with Q_* storing the associated quadratics for each interval in this partition. In Algorithm 2 we use the list-operator $Last(l)$ to designate the last element of the list l; $index\, Last(l)$, $delete\, Last(l)$ to get the associated index of the last element or to delete this element.

Algorithm 2: $\mathsf{INF}_{\mathsf{Q}_{t,\omega}}$ pruning

Input: List of ordered quadratics $(q_*^1, q_*^2, \ldots, q_*^{s-1}, q_*^s)$

begin Initialization: Q_* means "Remaining quadratics" and LB "Left Bound"
  $Q_* \leftarrow (q_*^1)$; $LB \leftarrow (-\infty)$
end
for $i = 2$ *to* s **do**
  $j \leftarrow index\, Last(Q_*)$
  $\mu_i : q_*^i(\mu_i) - q_*^j(\mu_i) = 0$ with $q_*^i(\mu) < q_*^j(\mu)$ for $\mu > \mu_i$ close to μ_i
  while $\mu_i < Last(LB)$ **do**
    $delete\, Last(Q_*)$; $delete\, Last(LB)$
    $j \leftarrow index\, Last(Q_*)$
    $\mu_i : q_*^i(\mu_i) - q_*^j(\mu_i) = 0$ with $q_*^i(\mu) < q_*^j(\mu)$ for $\mu > \mu_i$ close to μ_i
  end
  $Q_* \leftarrow (Q_*, q_*^i)$; $LB \leftarrow (LB, \mu_i)$
end
$s^* = \#LB$ (the number of element in LB)
for $i = 1$ *to* $s^* - 1$ **do**
  $\text{DOM}_*^i =]LB(i), LB(i+1)]$
end
$\text{DOM}_*^{s^*} =]LB(s^*), +\infty[$
Return Q_* and $(\text{DOM}_*^1, ..., \text{DOM}_*^{s^*})$

E Proofs for Section 5

By definition of the random-walk model for $\tilde{\eta}_{1:n}$ in Equation (2) and the auto-regressive model for $\epsilon_{1:n}$ in Equation (3) we have that the covariance matrices have entries

$$[\Sigma_{\mathsf{AR}}]_{ij} = \frac{\sigma_\nu^2}{1-\phi^2}\phi^{|i-j|}, \qquad [\Sigma_{\mathsf{RW}}]_{ij} = \sigma_\eta^2 \min\{i,j\}.$$

It is straightforward to find that their inverses have entries

$$[\Sigma_{\mathsf{AR}}^{-1}]_{ij} = \begin{cases} 1/\sigma_\nu^2 & \text{if } i=j=1 \text{ or } n, \\ (1+\phi^2)/\sigma_\nu^2 & \text{if } i=j\neq 1 \text{ or } n, \\ -\phi/\sigma_\nu^2 & \text{if } |i-j|=1, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$[\Sigma_{\mathsf{RW}}^{-1}]_{ij} = \begin{cases} 1/\sigma_\eta^2 & \text{if } i=j=n, \\ 2/\sigma_\eta^2 & \text{if } i=j\neq n, \\ -1/\sigma_\eta^2 & \text{if } |i-j|=1, \\ 0 & \text{otherwise,} \end{cases}$$

The unpenalised cost conditional on the set of changepoints is

$$\begin{aligned} \mathcal{C}(\tau_{1:m}) &= \min\left\{(1-\phi^2)\gamma(y_1-\mu_1)^2 + \sum_{t=2}^{n}\left[\lambda(\mu_t-\mu_{t-1}-\delta_t)^2 + \gamma\Big((y_t-\mu_t)-\phi(y_{t-1}-\mu_{t-1})\Big)^2\right]\right\} \\ &= \min\left\{(1-\phi^2)\gamma(y_1-\mu_1)^2 + \sum_{t=2}^{n}\left[\lambda(\tilde{\eta}_t-\tilde{\eta}_{t-1})^2 + \gamma\Big((y_t-\mu_t)-\phi(y_{t-1}-\mu_{t-1})\Big)^2\right]\right\} \end{aligned}$$

where the minimisation is over $\mu_{1:n}$, and $\delta_{2:n}$ consistent with the set of changepoints; and we have made a change of variables such that $\tilde{\eta}_i - \tilde{\eta}_{i-1} = \mu_i - \mu_{i-1} - \delta_i$ for $i = 2, \ldots, n$ in the second equality.

This change of variables is not unique, and we get the same value for any choice of $\tilde{\eta}_1$. Thus we trivially have that

$$\begin{aligned}
&\mathcal{C}(\tau_{1:m}) \\
&= \min\left\{(1-\phi^2)\gamma(y_1-\mu_1)^2 + \sum_{t=2}^{n}\left[\lambda(\tilde{\eta}_t-\tilde{\eta}_{t-1})^2 + \gamma\Big((y_t-\mu_t)-\phi(y_{t-1}-\mu_{t-1})\Big)^2\right] + \lambda\tilde{\eta}_1^2\right\},
\end{aligned}$$

where the minimisation is now also over $\tilde{\eta}_1$, and the minimum is attained with $\tilde{\eta}_1 = 0$.

By our definition of the matrix $X_{\tau_{1:m}}$ we have that if $\Delta = (\mu_1 - \tilde{\eta}_1, \delta_{\tau_{1:m}})$ we can write $\mu_{1:n} = X_{\tau_{0:m}}\Delta + \tilde{\eta}_{1:n}$. Thus by re-writing the sums, e.g.

$$\sum_{t=2}^{n}\lambda\{\tilde{\eta}_t - \tilde{\eta}_{t-1}\}^2 + \lambda\tilde{\eta}_1{}^2 = \tilde{\eta}_{1:n}^T\Sigma_{\mathsf{RW}}^{-1}\tilde{\eta}_{1:n},$$

as $\lambda = 1/\sigma_\eta^2$, gives that

$$\mathcal{C}(\tau_{1:m}) = \min_{\Delta,\tilde{\eta}_{1:n}}\left[(y_{1:n} - X_{\tau_{0:m}}\Delta - \tilde{\eta}_{1:n})^T\Sigma_{\mathsf{AR}}^{-1}(y_{1:n} - X_{\tau_{0:m}}\Delta - \tilde{\eta}_{1:n}) + \tilde{\eta}_{1:n}^T\Sigma_{\mathsf{RW}}^{-1}\tilde{\eta}_{1:n}\right]. \quad (10)$$

Proof of Proposition 4. To simplify notation we will write $\tilde{\eta}$ for $\tilde{\eta}_{1:n}$, y for $y_{1:n}$ and X for $X_{\tau_{0:m}}$. Re-writing right-hand side of (10) gives

$$\begin{aligned}
&\min_{\Delta,\tilde{\eta}}\left[(y - X\Delta - \tilde{\eta})^T\Sigma_{\mathsf{AR}}^{-1}(y - X\Delta - \tilde{\eta}) + \tilde{\eta}^T\Sigma_{\mathsf{RW}}^{-1}\tilde{\eta}\right] \\
&= \min_{\Delta,\tilde{\eta}}\Big[\{\tilde{\eta} - (\Sigma_{\mathsf{AR}}^{-1} + \Sigma_{\mathsf{RW}}^{-1})^{-1}\Sigma_{\mathsf{AR}}^{-1}(y - X\Delta)\}^T(\Sigma_{\mathsf{AR}}^{-1} + \Sigma_{\mathsf{RW}}^{-1})\{\tilde{\eta} - (\Sigma_{\mathsf{AR}}^{-1} + \Sigma_{\mathsf{RW}}^{-1})^{-1}\Sigma_{\mathsf{AR}}^{-1}(y - X\Delta)\} \\
&\quad + (y - X\Delta)^T\left\{\Sigma_{\mathsf{AR}}^{-1} - \Sigma_{\mathsf{AR}}^{-1}(\Sigma_{\mathsf{AR}}^{-1} + \Sigma_{\mathsf{RW}}^{-1})^{-1}\Sigma_{\mathsf{AR}}^{-1}\right\}(y - X\Delta)\Big] \\
&= \min_{\Delta}\left[(y - X\Delta)^T\left\{\Sigma_{\mathsf{AR}}^{-1} - \Sigma_{\mathsf{AR}}^{-1}(\Sigma_{\mathsf{AR}}^{-1} + \Sigma_{\mathsf{RW}}^{-1})^{-1}\Sigma_{\mathsf{AR}}^{-1}\right\}(y - X\Delta)\right].
\end{aligned}$$

Finally using the Woodbury matrix identity, for symmetric invertible matrices A and B, $(A+B)^{-1} = A^{-1} - A^{-1}(A^{-1} + B^{-1})^{-1}A^{-1}$. Thus we have

$$\left\{\Sigma_{\mathsf{AR}}^{-1} - \Sigma_{\mathsf{AR}}(\Sigma_{\mathsf{AR}}^{-1} + \Sigma_{\mathsf{RW}}^{-1})^{-1}\Sigma_{\mathsf{AR}}^{-1}\right\} = (\Sigma_{\mathsf{AR}} + \Sigma_{\mathsf{RW}})^{-1}.$$

The result follows immediately. □

Proof of Corollary 1.

As before write y for $y_{1:n}$ and X for $X_{\tau_{0:d}}$; further let $X_0 = X_{\tau_0}$. The value of Δ that minimises the right-hand side of (8) is

$$\hat{\Delta} = \{X^T(\Sigma_{\mathsf{AR}} + \Sigma_{\mathsf{RW}})^{-1}X\}^{-1}X^T(\Sigma_{\mathsf{AR}} + \Sigma_{\mathsf{RW}})^{-1}y.$$

To further simplify notation let $A = (\Sigma_{\mathsf{AR}} + \Sigma_{\mathsf{RW}})^{-1}$ and let Φ be such that $A = \Phi\Phi^T$ with Φ invertible; and let Ψ be a matrix such that $\Sigma = \Psi\Psi^T$. Then the reduction in cost over fitting no change is

$$\begin{aligned} C_0 - C(\tau_{0:d}) &= y^T\Big(AX(X^TAX)^{-1}X^TA - AX_0(X_0^TAX_0)^{-1}X_0^TA\Big)y \\ &= y^T\Phi^T\Phi^{-T}\Big(AX(X^TAX)^{-1}X^TA - AX_0(X_0^TAX_0)^{-1}X_0^TA\Big)\Phi^{-1}\Phi y = y^T\Phi^TB\Phi y, \end{aligned}$$

for the matrix $B = \Phi^{-T}\Big(AX(X^TAX)^{-1}X^TA - AX_0(X_0^TAX_0A)^{-1}X_0^T\Big)\Phi^{-1}$. By standard properties of linear models, as our model includes an intercept term this quadratic form is invariant to adding a constant to all entries of y. Thus as our model assumes no change we can, without loss of generality assume the mean of y is the zero vector.

Now it is straightforward to show that $B^2 = B$ and that B has rank d. Furthermore as under our assumptions y is Gaussian with variance Σ, Φy has variance $\Phi\Sigma\Phi^T$. From standard results for quadratic forms of Gaussian random variables, see for example Theorem 9.5 of Muller & Stewart (2006), the distribution of our quadratic form, $y^T\Phi^TB\Phi y$ is

$$\sum_{i=1}^{d}\alpha_iZ_i^2,$$

where α_i are the non-zero eigenvalues of $\Phi^T\Psi^TB\Psi\Phi$, and each Z_i^2 are independent χ_1^2 distributed random variables.

The result follows by first noting that as B is a projection its eigenvalues are 1 or 0. Thus $\alpha_i \leq \alpha^+$, where α^+ is the largest eigenvalue of $\Phi^T\Psi^T\Psi\Phi$, which by standard results is also the largest eigenvalue of $\Phi\Phi^T\Psi\Psi^T = (\Sigma_{\mathsf{AR}} + \Sigma_{\mathsf{RW}})^{-1}\Sigma$. Thus

$$\sum_{i=1}^{d} \alpha_i Z_i^2 \leq \sum_{i=1}^{d} \alpha^+ Z_i^2 = \alpha^+ \sum_{i=1}^{d} Z_i^2,$$

and the right-hand side has the same distribution as α^+ times a χ^2_d random variable. If $\Sigma = \Sigma_{\mathsf{AR}} + \Sigma_{\mathsf{RW}}$ then we further have that $\alpha_i = 1$ and hence the distribution is χ^2_d.

To prove the consistency of $\hat{m}$ we need to show that the probability of

$$\mathcal{C}_0 - \mathcal{C}(\tau_{1:d}) < d\beta$$

jointly for all d and $\tau_{1:d}$ tends to 1. A standard argument (see the proof of Proposition 3.1 in Zheng et al. 2019), is to use a union bound:

$$\begin{aligned}
\Pr(\hat{m} = 0) &\geq 1 - \sum_{d=1}^{n} \frac{n!}{d!(n-d)!} \Pr\left(\chi^2_d > \frac{d\beta}{\alpha^+}\right) \\
&\geq 1 - \sum_{d=1}^{n} \frac{n!}{d!(n-d)!} \Pr\left(\chi^2_d > dC\log(n)\right) \\
&\geq 1 - \sum_{d=1}^{n} n^d \exp\left\{-d\left(\frac{C\log(n) - \sqrt{2C\log(n) - 1}}{2}\right)\right\} \\
&\geq 1 - \sum_{d=1}^{n} \exp\left\{-d\left(\frac{(C-2)\log(n) - \sqrt{2C\log(n) - 1}}{2}\right)\right\}
\end{aligned}$$

with the second inequality using a tail bound for a χ^2_d random variable (Lemma 1 in Laurent & Massart 2000). The final expression will tend to 1 as $n \to \infty$ as $C > 2$. □

Proof of Proposition 5.

We use the notations $A = (\Sigma_{\mathsf{AR}} + \Sigma_{\mathsf{RW}})^{-1}$, $u_1 = u_{\tau_1}$, and write $c_0 = u_0^T A u_0$, $c_{0,1} = u_0^T A u_1$ and $c_1 = u_1^T A u_1$.

The optimal cost is equal to $y^TAy-(X^TAy)^T(X^TAX)^{-1}X^TAy$. If X is simply a column of ones, $X=u_0$ then $(X^TAX)^{-1}=\frac{1}{c_0}$, and $X^TAy=u_0^TAy$.

If X is the concatenation of u_0 and u_1, $X=(u_0\ u_1)$ we can compute

$$X^TAX=\begin{bmatrix} c_0 & c_{0,1} \\ c_{0,1} & c_1 \end{bmatrix} \quad \text{and} \quad (X^TAX)^{-1}=\frac{1}{c_0c_1-c_{0,1}^2}\begin{bmatrix} c_1 & -c_{0,1} \\ -c_{0,1} & c_0 \end{bmatrix}.$$

We also have

$$\begin{bmatrix} U_0 \\ U_1 \end{bmatrix}=\begin{bmatrix} u_0^TAy \\ u_1^TAy \end{bmatrix}=X^TAy\,.$$

Finally

$$C(\tau_1)=y^TAy-\frac{1}{c_0c_1-c_{0,1}^2}\Big(U_0c_1U_0-2U_0c_{0,1}U_1+U_1c_0U_1\Big).$$

Hence we can write the reduction in cost for fitting a change as

$$\begin{aligned} C_0-C(\tau_1) &= \frac{1}{c_0c_1-c_{0,1}^2}\Big(c_1U_0^2-2c_{0,1}U_0U_1+c_0U_1^2\Big)-\frac{1}{c_0}U_0^2 \\ &= \frac{1}{c_0^2c_1-c_0c_{0,1}^2}\Big(c_{0,1}^2U_0^2-2c_{0,1}c_0U_0U_1+c_0^2U_1^2\Big). \end{aligned}$$

Simple algebraic rearrangement gives the result in (i).

For part (ii) note that $\sum_{i=1}^n v_i=u_0^Tv$, using the definition of v gives

$$u_0^Tv=\frac{1}{\sqrt{c_1-c_{0,1}^2/c_0}}\left\{c_{0,1}-\frac{c_{0,1}}{c_0}c_0\right\}=0.$$

Similarly

$$v^T(\Sigma_{\mathsf{AR}}+\Sigma_{\mathsf{RW}})v=\frac{1}{c_1-c_{0,1}^2/c_0}\left\{c_1-2\frac{c_{0,1}}{c_0}c_{0,1}+\left(\frac{c_{0,1}}{c_0}\right)^2c_0\right\}=1.$$

Part (iii) is a standard result on the optimality of the weighted least squares estimator. To show it we can directly solve the constrained optimisation problem of maximising $(u_1^Tw)^2$

subject to $u_0^T w = 0$ and $w^T(\Sigma_{\mathsf{AR}} + \Sigma_{\mathsf{RW}})w = 1$. Using Lagrange multipliers we have that for constants α and δ

$$2(u_1^T w)u_1 = \alpha u_0 + 2\delta(\Sigma_{\mathsf{AR}} + \Sigma_{\mathsf{RW}})w.$$

Defining $\delta' = (u_1^T w)/\delta$, and $\alpha' = -\alpha/(2\delta)$, we get

$$w = \delta'(\Sigma_{\mathsf{AR}} + \Sigma_{\mathsf{RW}})^{-1}u_1 + \alpha'(\Sigma_{\mathsf{AR}} + \Sigma_{\mathsf{RW}})^{-1}u_0.$$

This means that w is a linear combination of the vectors $(\Sigma_{\mathsf{AR}} + \Sigma_{\mathsf{RW}})^{-1}u_1$ and $(\Sigma_{\mathsf{AR}} + \Sigma_{\mathsf{RW}})^{-1}u_0$, with the constants uniquely defined by the constraints. However this is the form that v as defined in part (i) takes, hence part (iii) of the proposition holds. □

Proof of Theorem 2

We will first consider the case where $\phi = 0$. For each n introduce the following sets of segmentations of the data:

$$\mathcal{A}_{i,m}^n = \left\{ \tau_{1:m} : \min_{j=1,\ldots,m} |\tau_j - \tau_i^0| > (\log n)^2 \right\};\ i = 1, \ldots, m^0,\ m = 1, \ldots, m_{\max};$$

$$\mathcal{B}_m^n = \left\{ \tau_{1:m} : \max_{i=1,\ldots,m^0} \left(\min_{j=1,\ldots,m} |\tau_j - \tau_i^0| \right) \leq (\log n)^2 \right\};\ m = m^0 + 1, \ldots, m_{\max}.$$

Thus $\mathcal{A}_{i,m}^n$ is the set of segmentations with m changepoints which do not contain a change within a distance $(\log n)^2$ of the ith actual changepoint; and $\mathcal{B}_m^n$ is the set of segmentations with $m > m^0$ changepoints and that have one changepoint within a distance of $(\log n)^2$ of each true changepoint. If a segmentation is in none of these sets then it must have the correct number of chanepoints, and one changepoint within a distance $(\log n)^2$ of each true change. As there are fixed number of these sets, to prove our result we need to show that $\Pr(\hat{\tau}_{1:\hat{m}} \in \mathcal{A}_{i,m}^n) \to 0$ for each i and m; and $\Pr(\hat{\tau}_{1:\hat{m}} \in \mathcal{B}_m^n) \to 0$ for each m.

Let $\mathcal{C}(\tau_{1:m})$ denote the unpenalised cost for the segmentation $\tau_{1:m}$, with, for example, $\mathcal{C}(\tau_{1:m}, \tau_{1:m^0}^0)$ the unpenalised cost from the segmentation that has the changepoints in the

union of $\tau_{1:m}$ and $\tau^0_{1:m^0}$. We first show that for any $m = m^0+1, \ldots, m_{\max}$, $\Pr(\hat{\tau}_{1:\hat{m}} \in \mathcal{B}^n_m) \to 0$. To do this consider a $\tau_{1:m} \in \mathcal{B}^n_m$, we will compare the cost of this segmentation with that of the true segmentation. As adding changepoints can only reduce the unpenalised cost we have the difference in penalised costs is

$$\mathcal{C}(\tau_{1:m}) + m\beta - \mathcal{C}(\tau^0_{1:m^0}) - m^0\beta \geq (m - m^0)\beta - \left(\mathcal{C}(\tau^0_{1:m^0}) - \mathcal{C}(\tau_{1:m}, \tau^0_{1:m^0})\right).$$

Furthermore, by the same argument used in Corollary 1, $(\mathcal{C}(\tau^0_{1:m^0}) - \mathcal{C}(\tau_{1:m}, \tau^0_{1:m^0}))/\alpha$ is stochastically bounded by a χ^2_m distribution.

As there are fewer than $(2(\log n)^2)^{m^0} n^{m-m^0}$ segmentations in $\mathcal{B}^n_m$ we have

$$\begin{aligned}
&\Pr\left(\min_{\tau_{1:m} \in \mathcal{B}^n_m} \mathcal{C}(\tau_{1:m}) + m\beta < \mathcal{C}(\tau^0_{1:m^0}) + m^0\beta)\right) \\
&\leq \quad (2(\log n)^2)^{m^0} n^{m-m^0} \Pr(\chi^2_m > (m - m^0)\beta/\alpha) \\
&= \quad (2(\log n)^2)^{m^0} n^{m-m^0} \Pr(\chi^2_m > (m - m^0)C \log n).
\end{aligned}$$

By a similar argument to that used in the proof of Corollary 1, this probability tends to 0 as required.

Now we consider $\tau_{1:m} \in \mathcal{A}^n_{i,m}$. Again we will compare the cost of such a segmentation with that of the true segmentation. Let τ^0_{-i} denote the set of true changepoints excluding τ^0_i.

$$\begin{aligned}
&\mathcal{C}(\tau_{1:m}) + m\beta - \mathcal{C}(\tau^0_{1:m^0}) - m^0\beta \geq \mathcal{C}(\tau_{1:m}, \tau^0_{-i}) - \mathcal{C}(\tau^0_{1:m^0}) + (m - m^0)\beta \\
&= \quad \{\mathcal{C}(\tau_{1:m}, \tau^0_{-i}) - \mathcal{C}(\tau_{1:m}, \tau^0_{1:m^0}) - m^0\beta\} + \{\mathcal{C}(\tau_{1:m}, \tau^0_{1:m^0}) - \mathcal{C}(\tau^0_{1:m^0}) + m\beta\}
\end{aligned}$$

There are fewer than n^m segmentations in $\mathcal{A}^n_{i,m}$, and $(\mathcal{C}(\tau_{1:m}, \tau^0_{1:m^0}) - \mathcal{C}(\tau^0_{1:m^0}))/\alpha$ is stochastically bounded by a χ^2_m random variable. Thus by the same argument as above we have that

$$\Pr\left(\min_{\tau_{1:m} \in \mathcal{A}^n_{i,m}} \mathcal{C}(\tau_{1:m}, \tau^0_{1:m^0}) - \mathcal{C}(\tau^0_{1:m^0}) + m\beta < 0\right) \to 0.$$

To show $\Pr(\hat{\tau}_{1:m} \in \mathcal{A}^n_{i,m}) \to 0$ we only need to show

$$\Pr\left(\min_{\tau_{1:m} \in \mathcal{A}^n_{i,m}} \mathcal{C}(\tau_{1:m}, \tau^0_{-i}) - \mathcal{C}(\tau_{1:m}, \tau^0_{1:m^0}) - m^0\beta < 0\right) \to 0.$$

By the same argument as used in Proposition 5(i), $\mathcal{C}(\tau_{1:m}, \tau^0_{-i}) - \mathcal{C}(\tau_{1:m}, \tau^0_{1:m^0}) = (v^T y_{1:n})$ for some vector $v = v_{1:n}$. By standard properties of linear models, it is straightforward to show that v has the following properties: (i) $v^T \Sigma^*_n v = 1$, where $\Sigma^*_n = \Sigma_{\mathsf{RW}} + \Sigma_{\mathsf{AR}}$ is the variance of the noise in the fitted model; (ii) v is orthogonal to the column-space of the X matrix for the linear model (7) corresponding to the changepoints $\tau_{1:m}, \tau^0_{1:m^0}$; (iii) among vectors v that satisfy (i) and (ii) it is the one that maximises the signal for a change at τ_i, i.e. that maximises $(\sum_{t=1}^{\tau_i} v_i)^2$.

If we define $\nu = (\sum_{t=1}^{\tau_i} v_i)^2$, we can bound ν by choosing any vector $w = w_{1:n}$ that satisfies (ii) and then, after normalising using (i), property (iii) gives $\nu \geq (\sum_{t=1}^{\tau_i} w_i)^2/(w^T \Sigma^*_n w)$. Let $h = \lfloor (\log n)^2 \rfloor$. We choose such a w defined as $w_j = 1$ for $j = \tau_i - h + 1, \ldots, \tau_i$, $w_j = -1$ for $\tau_i + 1, \ldots, \tau_i + h$, and $w_j = 0$ otherwise. The column space of the X matrix in property (ii) contains vectors whose jth entries are either identically 0 or identically 1 for for $j = \tau_i - h + 1, \ldots, \tau_i + h$, and hence this vector satisfies property (ii).

Now using the fact that we run DeCAFS with $\phi = 0$ and so Σ_{AR} is the identity: $w^T \Sigma^*_n w = w^T \Sigma_{\mathsf{AR}} w + w^T \Sigma_{\mathsf{RW}} w \leq 2hc_\nu + h^3 c_\eta/n$, and $\nu \geq h^2/(2hc_\nu + h^3 c_\eta/n)$. Thus there exists $c_1 > 0$ such that for large enough n, $v^T y_{1:n}$ is normally distributed with $|\mathrm{E}(v^T y_{1:n})| \geq c_1 \log n$ and $\mathrm{Var}(v^T y_{1:n}) \leq \alpha$. So, for large enough n,

$$\begin{aligned} &\Pr\left(\min_{\tau_{1:m} \in \mathcal{A}^n_{i,m}} \mathcal{C}(\tau_{1:m}, \tau^0_i) - \mathcal{C}(\tau_{1:m}, \tau^0_{1:m^0}) - m^0\beta < 0\right) \\ &\leq \quad n^m \Pr\left(Z < \frac{1}{\sqrt{\alpha}}\{\sqrt{C\alpha \log nm^0} - c_1 \log n\}\right), \end{aligned}$$

where Z is a standard normal random variable. Using standard tail bounds we get that this probability tends to 0 as $n \to \infty$ as required.

The argument for the case where $\phi > 0$ is similar. The differences are just in the definition of the sets $\mathcal{A}^n_{i,m}$ and $\mathcal{B}^n_m$ which are now

$$\mathcal{A}^n_{i,m} = \left\{\tau_{1:m} : \min_{j=1,\ldots,m} |\tau_j - \tau^0_i| > 0\right\};\ \mathcal{B}^n_m = \left\{\tau_{1:m} : \max_{i=1,\ldots,m^0} \left(\min_{j=1,\ldots,m} |\tau_j - \tau^0_i|\right) = 0\right\};$$

and the final part of the argument that shows

$$\Pr\left(\min_{\tau_{1:m}\in\mathcal{A}^n_{i,m}} \mathcal{C}(\tau_{1:m}, \tau^0_i) - \mathcal{C}(\tau_{1:m}, \tau^0_{1:m^0}) - m^0\beta < 0\right) \to 0. \tag{11}$$

For this last part we use a different vector w to bound the distribution of $\mathcal{C}(\tau_{1:m}, \tau^0_i) - \mathcal{C}(\tau_{1:m}, \tau^0_{1:m^0}) = (v^T y)^2$. Our choice of w has $w_{\tau_i} = 1$, $w_{\tau_i+1} = -1$ and $w_j = 0$ otherwise. We then have $w^T\Sigma^*_n w = w^T\Sigma_{\mathsf{AR}} w + w^T\Sigma_{\mathsf{RW}} w = 2(1-\phi)c_\nu(1-\phi^2) + c_\eta/n$. Now as $\phi = \exp\{-c_\phi/n\} \geq 1 - c_\phi/n$ we have $w^T\Sigma^*_n w \leq c_1/n$ for some constant c_1. Thus $\nu \geq n/c_1$. As this is $O(n)$ it is straightforward to use the same tail bounds of a normal random variable to show (11)

Proof of Proposition 6

If we fix n, and let Σ^0 be the covariance matrix of the generated data then in case (i), $[\Sigma^0]_{ij} = \mathrm{Cov}(\zeta(i/n), \zeta(j,n)) = c^0_\eta \min i, j/n$ if $i \neq j$ and $[\Sigma_0]_{ii} = \mathrm{Var}(\zeta(i/n)) = c^0_\eta i + c^0_\nu$. Whilst in case (ii),

$$[\Sigma^0]_{ij} = \mathrm{Cov}(\zeta(i/n), \zeta(j,n)) = c^0_\eta \min i, j/n + c^0_\nu(\exp\{-c^0_\phi/n\})^{|i-j|}.$$

In both cases we can write $\Sigma^0 = \Sigma^0_{\mathsf{AR}} + \Sigma^0_{\mathsf{RW}}$ where Σ^0_{AR} is the covariance matrix of an AR(1) process with auto-correlation parameter, $\phi^0 = \exp\{-c^0_\phi/n\}$, and marginal variance c^0_ν and Σ^0_{RW} is the covariance matrix of a random walk process with variance parameter c^0_η/n.

We proceed by calculating a bound for the maximum eigenvalue of $\Sigma^{-1}\Sigma^0$, where $\Sigma = \Sigma_{\mathsf{AR}} + \Sigma_{\mathsf{RW}}$ and $\Sigma^0 = \Sigma^0_{\mathsf{AR}} + \Sigma^0_{\mathsf{RW}}$ are respectively the covariance assumed by DeCAFS and

the covariance of the data. We then further bound this as we vary n for the given parameter regimes for the two covariance matrices. We do this first for case (i) where $\phi = \phi^0 = 0$, then for the case where both autocorrelation parameters are non-zero.

Standard manipulations give that the maximum eigenvalues of $\Sigma^{-1}\Sigma^0$ is also the maximum eigenvalue of $\Sigma^{-1/2}\Sigma^0\Sigma^{-1/2}$, where $\Sigma^{-1/2}$ is a symmetric square root of Σ^{-1}. If v is an eigenvector of $\Sigma^{-1/2}\Sigma^0\Sigma^{-1/2}$ with eigenvalue ρ, then

$$v^T\Sigma^{-1/2}\Sigma^0\Sigma^{-1/2}v = \rho v^T v.$$

Writing $w = \Sigma^{-1/2}v$, we have

$$\frac{w^T\Sigma^0 w}{w^T\Sigma w} = \rho,$$

from which we have that we can bound the maximum eigenvalue by

$$\begin{aligned}\max_{w:|w|=1} \frac{w^T\Sigma^0 w}{w^T\Sigma w} &= \max_{w:|w|=1} \frac{w^T\Sigma^0_{\mathsf{AR}} w + w^T\Sigma^0_{\mathsf{RW}} w}{w^T\Sigma_{\mathsf{AR}} w + w^T\Sigma_{\mathsf{RW}} w} \\ &\leq \max\left\{\max_{w:|w|=1} \frac{w^T\Sigma^0_{\mathsf{AR}} w}{w^T\Sigma_{\mathsf{AR}} w}, \max_{w:|w|=1} \frac{w^T\Sigma^0_{\mathsf{RW}} w}{w^T\Sigma_{\mathsf{RW}} w}\right\}. \qquad (12)\end{aligned}$$

The first part of the Proposition follows by noting that $\Sigma^0_{\mathsf{RW}} = (c^0_\eta/c_\eta)\Sigma_{\mathsf{RW}}$, and, if $\phi = \phi^0 = 0$, $\Sigma^0_{\mathsf{AR}} = (c^0_\nu/c_\nu)\Sigma_{\mathsf{AR}}$. Hence,

$$\max_{w:|w|=1} \frac{w^T\Sigma^0_{\mathsf{AR}} w}{w^T\Sigma_{\mathsf{AR}} w} = \frac{c^0_\nu}{c_\nu}, \quad \max_{w:|w|=1} \frac{w^T\Sigma^0_{\mathsf{RW}} w}{w^T\Sigma_{\mathsf{RW}} w} = \frac{c^0_\eta}{c_\eta}.$$

For the case where $\phi^0 \neq 0$ and $\phi \neq 0$ we use a similar argument but, in addition, need to bound $\max_{w:|w|=1} w^T\Sigma^0_{\mathsf{AR}} w / w^T\Sigma_{\mathsf{AR}} w$. Now by similar arguments to above, we have that this is just the largest eigenvalue of $\Sigma^{-1/2}_{\mathsf{AR}}\Sigma^0_{\mathsf{AR}}\Sigma^{-1/2}_{\mathsf{AR}}$, which in turn is

$$\max_{w:|w|=1} \frac{w^T\Sigma^{-1}_{\mathsf{AR}} w}{w^T(\Sigma^0_{\mathsf{AR}})^{-1} w}.$$

To simplify notation and exposition, fix n and let $r = \phi^0$. Then

$$\Sigma^{-1}_{\mathsf{AR}} = \frac{1}{c_\nu(1 - \exp\{-2c_\phi/n\})} K_\phi, \text{ and} (\Sigma^0_{\mathsf{AR}})^{-1} = \frac{1}{c^0_\nu(1 - \exp\{-2c^0_\phi/n\})} K_r,$$

where K_ϕ is an $n \times n$ matrix with entries

$$[K_\phi]_{ij} = \begin{cases} 1 & \text{if } i = j = 1 \text{ or } n, \\ 1+\phi^2 & \text{if } i = j \neq 1 \text{ or } n, \\ -\phi & \text{if } |i-j| = 1, \\ 0 & \text{otherwise,} \end{cases}$$

and similarly for K_r. Clearly we have

$$\max_{w:|w|=1} \frac{w^T \Sigma_{\mathsf{AR}}^{-1} w}{w^T (\Sigma_{\mathsf{AR}}^0)^{-1} w} = \frac{c_\nu^0 (1 - \exp\{-c_\phi^0/n\})}{c_\nu (1 - \exp\{-c_\phi/n\})} \max_{w:|w|=1} \frac{w^T K_\phi w}{w^T K_r w}. \tag{13}$$

Let $v^{(i)}$, for $i = 1, \ldots, n$ be the eigenvectors of K_r. Standard results, (see, e.g., "Spectral decomposition of Kac-Murdock-Szego Matrices", a technical report by William F Trench available at `https://works.bepress.com/william_trench/133/`), are that the eigenvalues are of the form $1 - 2r\cos\theta_i + r^2$, for some angles $\theta_1, \ldots, \theta_n$. Furthermore the entries of $v^{(i)}$ satisfy

$$v_{j-1}^{(i)} - 2\cos\theta_i v_j^{(i)} + v_{j+1}^{(i)} = 0, \quad \text{for } j = 2, \ldots, n,$$

with $(2\cos\theta_i - r)v_1^{(i)} = v_2^{(i)}$ and $(2\cos\theta_i - r)v_n^{(i)} = v_{n-1}^{(i)}$.

Straightforward calculations then give

$$K_\phi v^{(i)} = (1 - 2\phi\cos\theta_i + \phi^2)v^{(i)} + \phi(r-\phi)(v_1^{(i)} e_1 + v_n^{(i)} e_n),$$

where e_1 and e_n are the n-vectors of 0s with a 1 in, respectively, the first and nth entries.

Now writing $w = \sum_{i=1}^n d_i v^{(i)}$, we have

$$\frac{w^T K_\phi w}{w^T K_r w} = \frac{\sum_{i=1}^n d_i^2 (1 - 2\phi\cos\theta_i + \phi^2) + \phi(r-\phi)(w_1^2 + w_n^2)}{\sum_{i=1}^n d_i^2 (1 - 2r\cos\theta_i + r^2)}.$$

For any w with $|w| = 1$ we trivially have that

$$\frac{\sum_{i=1}^n d_i^2 (1 - 2\phi\cos\theta_i + \phi^2)}{\sum_{i=1}^n d_i^2 (1 - 2r\cos\theta_i + r^2)} \leq \max_\theta \frac{(1 - 2\phi\cos\theta + \phi^2)}{(1 - 2r\cos\theta + r^2)} = \max\left\{ \frac{(1-\phi)^2}{(1-r)^2}, \frac{(1+\phi)^2}{(1+r)^2} \right\}.$$

Now if we write $\rho_i = (1 - 2r\cos\theta_i + r^2)$ for the ith eigenvalue of K_r, then

$$\max_{w:|w|=1} \frac{w_1^2}{\sum_{i=1}^n d_i^2 \rho_i} = \max_{d:|d|=1} \frac{\left(\sum_{i=1}^n d_i v_1^{(i)}\right)^2}{\sum_{i=1}^n d_i^2 \rho_i} = \left(\sum_{i=1}^n (v_1^{(i)})^2/\rho_i\right),$$

where we have first rewritten w and w_1 in terms of its expansion in the basis of the eigenvectors of K_r, and then used the fact that the maximum is achieved with $d_i \propto v_1^{(i)}/\rho_i$. Using the fact that each $v^{(i)}$ is an eigenvector of K_r^{-1} with eigenvalue $1/\rho_i$,

$$\left(\sum_{i=1}^n (v_1^{(i)})^2/\rho_i\right) = [K_r^{-1}]_{11} = \frac{1}{1-r^2}.$$

By a similar argument for the term involving w_n^2 we have

$$\max_{w:|w|=1} \frac{w^T K_\phi w}{w^T K_r w} \leq \max\left\{\frac{(1-\phi)^2}{(1-r)^2}, \frac{(1+\phi)^2}{(1+r)^2}\right\} + 2\max\left\{\phi\frac{(r-\phi)}{1-r^2}, 0\right\}.$$

Now using $\phi = \exp\{-c_\phi/n\}$ and $r = \exp\{-c_\phi^0/n\}$ we have this bound is $(c_\phi/c_\phi^0)^2 + (c_\phi - c_\phi^0)/c_\phi^0 + O(1/n)$ if $c_\phi > c_\phi^0$ and $1 + O(1/n)$ if $c_\phi \leq c_\phi^0$. The result follows trivially by combining this with (12) and (13).

F Additional Empirical Results

F.1 Parameter Estimation

We provide a simulation study to highlight the behavior of our estimators described in Section 4 for parameters σ_η, σ_ν and ϕ. We simulate 2000 time-series of length 5000 for each couple (ϕ, ω^2) on a grid for $\phi \in \{0, 0.05, 0.1, \ldots, 0.8, 0.85\}$ and $\omega^2 = \sigma_\eta^2/\sigma_\nu^2 \in [0, 2]$ with a log scale of 40 elements. We consider the *rand1* scenario with 1, 20 and 40 segments and use $K = 10$ for our estimation. In Figure 10 we used the same color scale bounds to highlight changes with different segment structures. The sign of the bias for σ_η and

σ_ν tend to depend on the number of changes and an underestimation of one variance is correlated to an overestimation of the other one. The standard deviation of the random walk variance increased with the ϕ parameter due to the unidentificability of case $\phi = 1$. The number of changes seems to have less impact on the estimation of the variance of the AR(1) when compared to the variance of the random walk. Notice also that the observed standard deviation for ϕ is often greater than 0.1 and an important deviation to the true parameter of order $0.1 - 0.2$ is not uncommon.

In order to improve our parameter estimation, we tested a two-stage estimator by first using our estimator, second running DeCAFS, and then using again our estimator on each obtained segment of length greater than 50. The weighted mean (by segment length) of the three parameters is presented in Figure 11 in a scenario with 40 segments. With this approach, we slightly reduced all the standard deviations in particular for the random walk variance with ϕ close to 1.

To see what might happen in case of a distorted parameter estimation, as mentioned in the simulation study of Section 6, please refer to Figure 12. We can see there, how even when misspecifying the model, in this case via fitting a pure AR(1) when there was some drift in the signal, we find a distorted signal μ estimation, however we are still able to reconstruct the changepoint locations relatively well.

F.2 Additional well-log data segmentation

In Figure 13 we report some additional segmentations of the log-well data described in Section 1.

F.3 Comparison of DeCAFS with Trend Filtering

We now compare DeCAFS with Trend Filtering (Kim et al. 2009). Our comparison is for the sinusoidal mean model from Figure 6, though to make it easier to see the differences in fit we consider just the first $n = 1000$ data points. Furthermore we compare the model with and without changepoints.

Figure 14 shows a comparison of the fit for a single data set for Trend Filtering with different degrees of smoothness to DeCAFS. The left-hand column shows the case where there are no abrupt changes in the mean, in this case Trend Filtering, particular of order 1 or 2 (fitting piecewise linear and piecewise quadratic functions respectively) can be seen to much better estimate the smooth mean than DeCAFS.

However, when we introduce abrupt changes, Trend Filtering of order 1 and 2 smooths over the abrupt change. This results in a poor estimate of the mean at time points close to the change. By comparison DeCAFS is able to detect these changes (in this example it fits the first two changes as abrupt changes, though does miss the final change and fits this with the random walk component of model for the mean). Trend Filtering of order 0, which fits a piecewise constant model, is better able to fit to the abrupt change, though has to also fit the smoothly varying component of the mean through smaller abrupt changes.

In terms of ability to estimate the underlying mean function, the qualitative picture we get from Figure 14 is borne out by a comparison of the mean square error of the fits over 100 replication, see Table 1. The smoother versions of Trend filtering perform best, by an order of magnitude when there are no changes, but are less accurate when we introduce the abrupt changes.

However, perhaps the starkest difference between Trend Filtering and DeCAFS comes if we wish to detect the abrupt changes. Whilst such changes can be detected by DeCAFS, this is not possible with the Trend Filtering. If the order of Trend Filtering is 1 or greater

	Trend Filter order 0	Trend Filter order 1	Trend Filter order 2	DeCAFS
No Changepoints	0.16	0.04	0.03	0.28
Changepoints	0.20	0.25	0.28	0.19

Table 1: Mean Square Error for estimating the mean for the sinusoidal example with or without changepoins.

it never fits an abrupt change; whereas if the order is 0 it overfits the number of changes as they are used for both the abrupt changes and to fit the smoothly varying mean.

F.4 Comparison of DeCAFS and AR1Seg on a Ornstein-Uhlenbeck process

We compare performances of both DeCAFS and AR1Seg from Chakar et al. (2017) on a discrete Ornstein-Uhlenbeck process with abrupt changes. Let $y_{1:n} = (y_1, \dots, y_n) \in \mathbb{R}^n$ a sequence of n realizations of the process:

$$y_t = \mu_t + \epsilon_t \quad t = 1, \dots, n$$

where for $t = 2, \dots, n$

$$\mu_t = f_t + \nu_t$$

and $\epsilon_t \underset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$, f_t is a piecewise constant signal we wish to infer the changes of whether $f_t \neq f_{t-1}$, and finally ν_t is a discrete Ornstein–Uhlenbeck process defined by:

$$\nu_t = \nu_{t-1} - \theta\nu_{t-1} + \sigma_\nu \eta_t; \quad \text{with } \eta_t \underset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\eta^2).$$

Differently from the RW process introduced in the main model in Equation 1 the OU process is a mean reverting process, which rather then diverging as a pure Random Process would do, it reverts to its original initial value. This is regulated by the parameter θ, where it can be seen that for $\theta = 0$ we observe a pure Random Walk process.

We performed a small simulation study comparable to the previous ones, which is summarised in Figure 15, where we report the average F1 scores of DeCAFS and AR1Seg over 100 replicates of each experiment. Separate figures for precision and recall can be found in Appendix G, Figure 22.

We denote how DeCAFS is relatively robust to this kind of model misspecification, producing good changepoints estimates overall, especially for larger values of θ. As a matter of fact, for $\theta \approx 1$ we have in fact a simple AR(1) noise with changes: in this scenario AR1Seg matches DeCAFS performances.

F.5 DeCAFS penalty for the gene expression data.

F.5.1 Robustness to $R(\delta)$

For the analysis of the Gene Expression data in *Bacilus subtilis* as described in section 7 we learned the optimal penalty (maximising $M(\delta)$ for a fixed $R(\delta)$) on the minus strand and show the results on the plus strand in Figure 9.

In Figure 16 we represent for the plus strand and minus strand $M(\delta)$ as a function of β for various value of $R(\delta)$.

F.5.2 Model checking for Gene Expression Data

In Figure 17 we show various model checking plots for the residuals, $Y_i - \hat{\mu}_i$, and the AR-residuals, $(Y_i - \hat{\mu}_i) - \hat{\phi}(Y_{i-1} - \hat{\mu}_{i-1}$. As shown in the qq-plots they are not Gaussian, but

the distribution is roughly symmetric; the middle plots suggests that distribution of the residuals does not vary substantially with the estimated signal ($\hat{mu}_i$) for the residuals, or with the fitted values ($\hat{\mu}_i - \hat{\phi}\hat{\mu}_{i-1}$) for the AR-residuals. Finally based on the acf plots it is clear that there is some level of auto-correlation in the data and that our AR-model capture part of it as the second coefficient is much smaller for the AR-residuals.

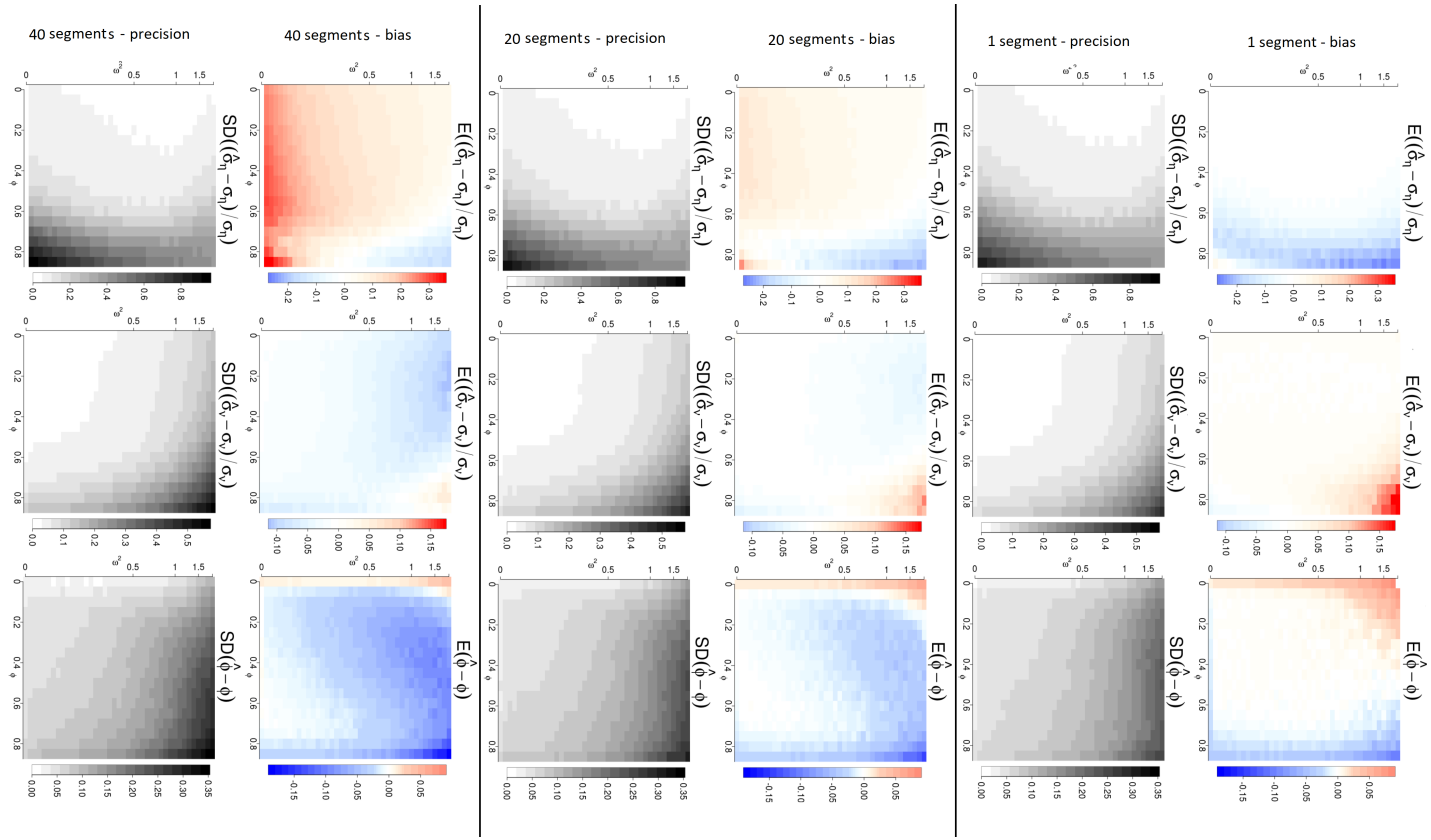


Figure 10: For each cell 2000 time-series of length 5000 have been generated under our model (1) – (3). We plot an accuracy measure: a percent error for the variances, the bias for the ϕ parameter and the standard deviation in gray scale. We chose $K = 10$ for our estimation procedure and compare scenarios *rand1* with 1, 20 and 40 segments

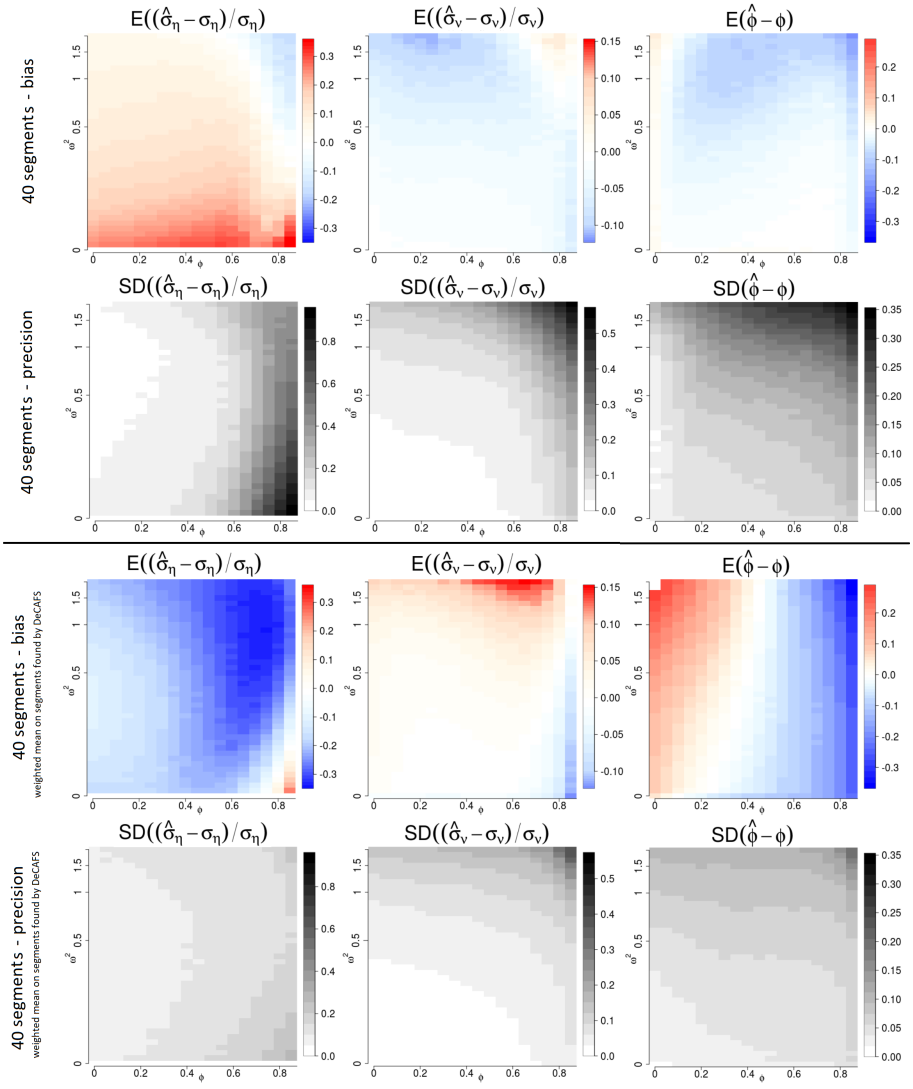


Figure 11: For each cell 2000 time-series of length 5000 have been generated under our model (1) – (3) with 40 changes (two first rows, same than the two last rows in Figure 10) and with the 2-stage estimator (two last rows). As in previous Figure, we plot accuracy and precision.

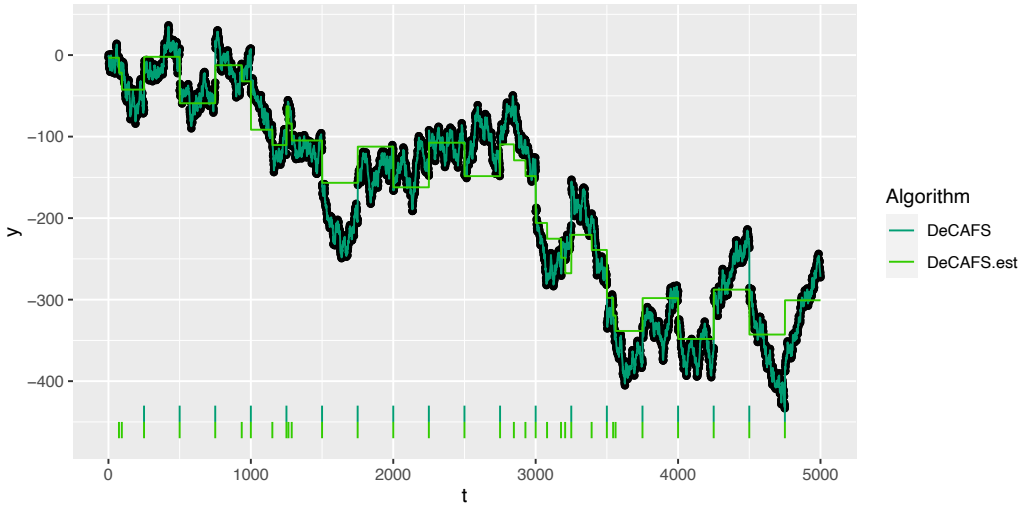


Figure 12: An example of a sequence generated with $\sigma_\eta = 4$, $\sigma_\nu = 2$, $\phi = 0.14$, with relative signal and changepoints estimates of DeCAFS with real parameter values compared to DeCAFS with estimated ones. On this particular sequence, our estimator returns values for initial parameters of $\hat{\sigma}_\eta = 0$, $\hat{\sigma}_\nu = 4.6$, $\hat{\phi} = 0.98$, resulting in a distorted signal estimation.

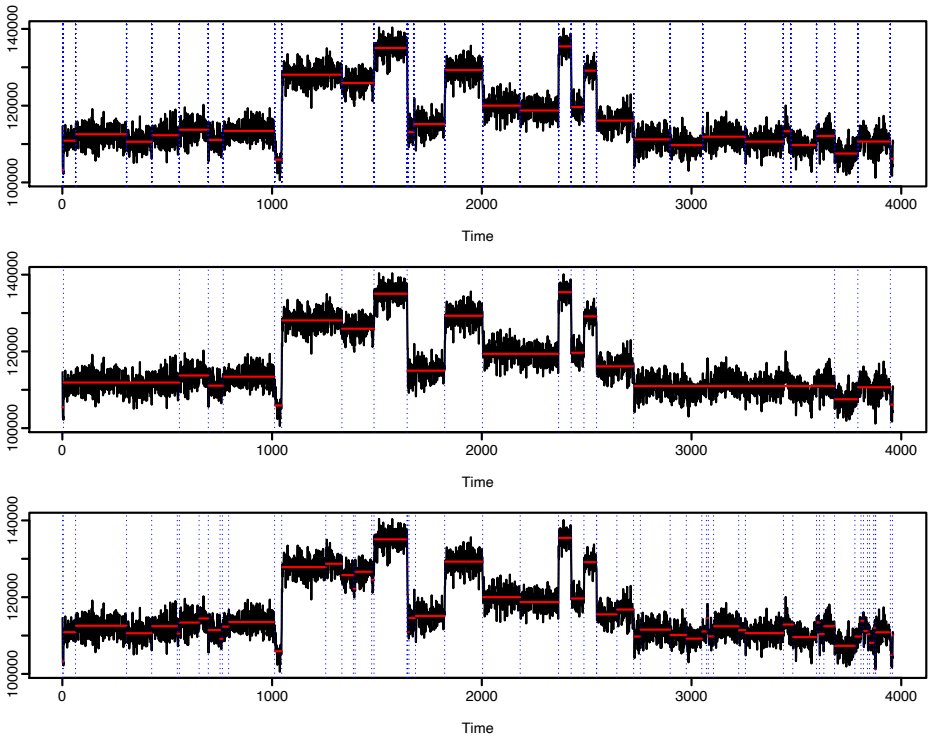


Figure 13: Segmentations of well-log data: Optimal segmentation under square error loss with the default, BIC, penalty (top); segmentation with the AR1-seg method of Chakar et al. (2017) that models the data as piecewise constant mean with AR(1) noise (middle); optimal segmentation for constant-mean model with WBS2 and the number of changes detected by the steepest drop to low levels criteria of Fryzlewicz (2018*a*) (bottom). Each plot shows the data (black line) the estimated mean (red line) and changepoint location (vertical blue dashed lines).

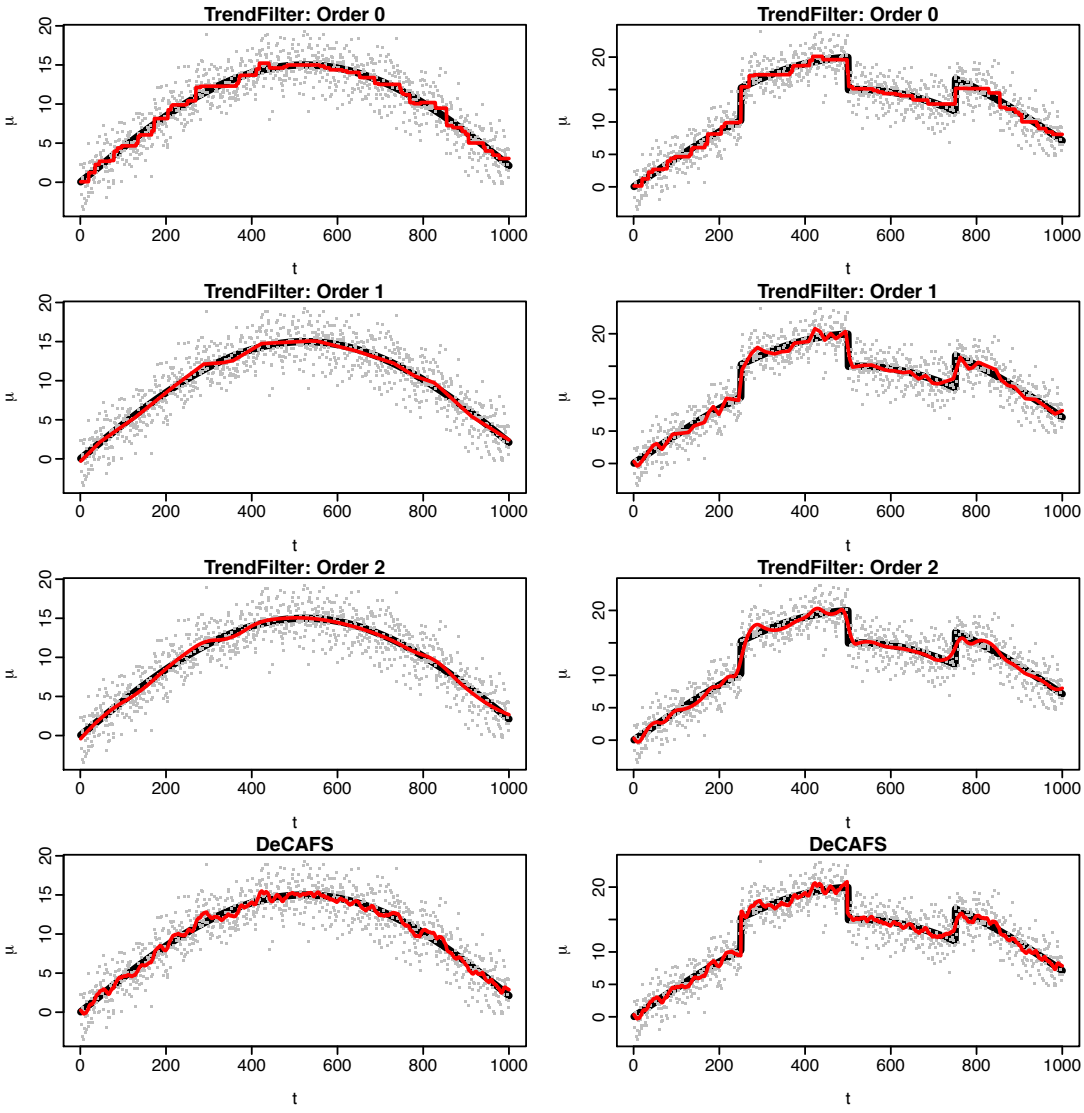


Figure 14: Comparison of segmentations for Trend Filtering and DeCAFS. Left-hand column is data with a sinusoidal mean; and the right-hand column in addition includes 3 changepoints. Columns, respectively from top, are for Trend Filtering with order 0 (the fused Lasso), 1 (fitting piecewise linear) and 2 (fitting piecewise quadratic) and DeCAFS. In each plot the true mean is the black line, the estimated mean is the red line, and the data are shown by the grey dots.

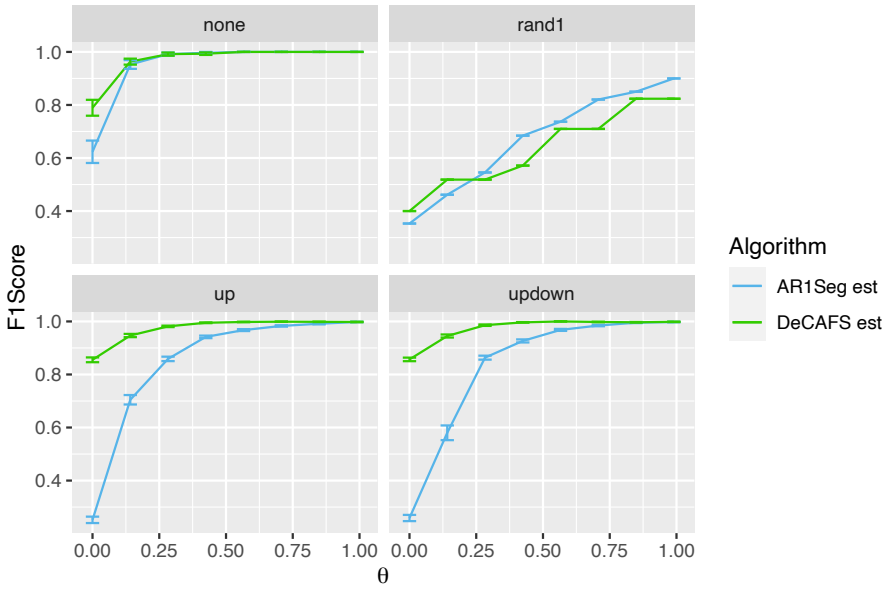


Figure 15: F1 score on different scenarios with an underlying OU process as we vary θ. Data simulated fixing $\sigma_\nu = 1$, $\sigma_\eta = 1$ and $\sigma = 1$ over a change of size 10.

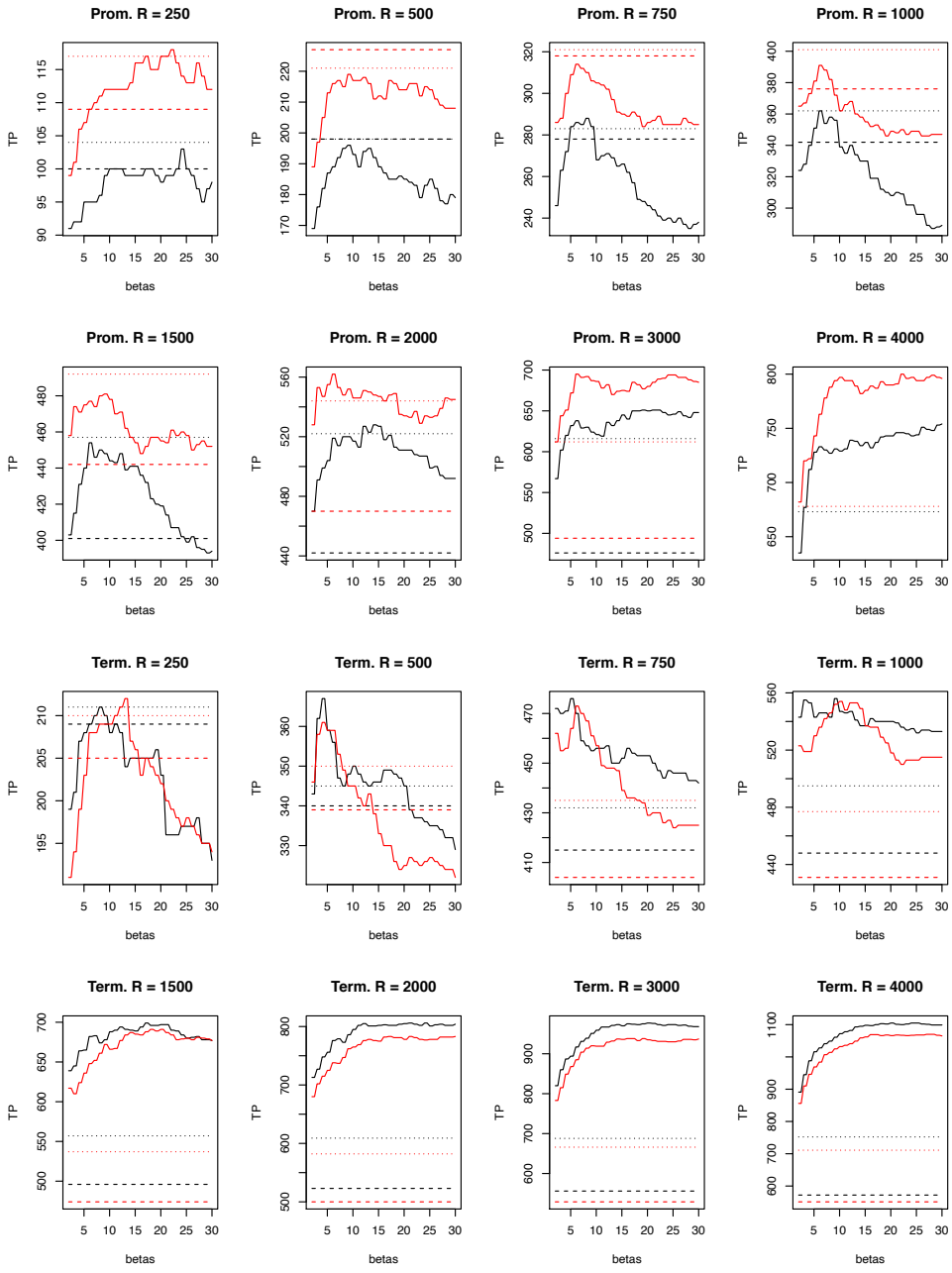


Figure 16: $M(\delta)$ as a function of β for promoters (line 1 and 2) and terminators (line 3 and 4), for various value of $R(\delta)$ (250, 500, 750, 1000, 2000, 3000 and 4000) for the plus strand in red and the minus strand in black. Results of hmmTiling.ori on the plus and minus strand are represented as horizontal dotted red and black lines. Results of hmmTiling.all (using all probes rather than only those called transitions).

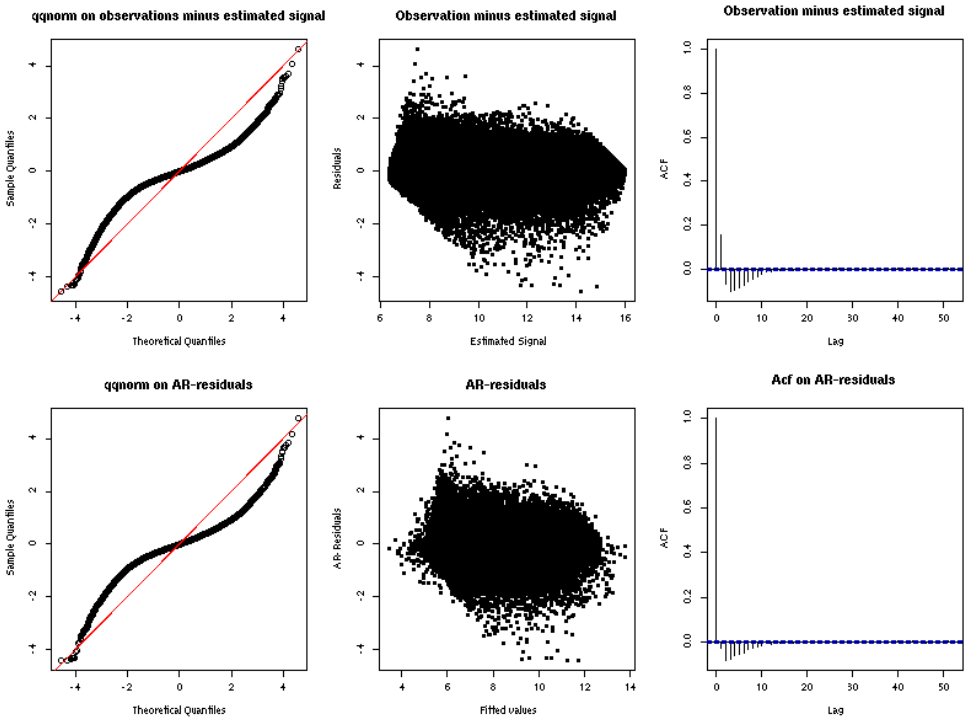


Figure 17: Model checking plots on the residuals and AR-residuals of DeCAFS on the plus strand using a penalty of $8\log(n)$ (learnt on the minus strand). The top line correspond to the residuals $(Y_i - \hat{\mu}_i)$ and the bottom line to the AR-residuals $(Y_i - \hat{\mu}_i) - \hat{\phi}(Y_{i-1} - \hat{\mu}_{i-1})$. The right column show qq-plots versus normal quantiles. In the middle column the residuals are plotted as a function of the fitted values. In the right column are the acf plots.

G Additional Simulation Results

In Figures 18 and 19 we summarize the results of the first simulation of Section 6 in terms of Precision (the proportion of detected changes which are correct) and Recall (the proportion of true changes that are detected). Similarly, Figure 20 shows Precision and Recall for the simulation with an AR(2) noise, Figure 21 shows Precision and Recall for the simulation with an underlying sinusoidal signal, and Figure 22 shows Precision and Recall for the simulations where the local fluctations in the mean are from an Ornstein-Uhlenbeck process.

As an extension on the simple AR(1) noise (Figure 6.1 A), we investigate a further case of model misspecification. Differently to what already shown, we now assume independence in the AR(1) noise across the various segments. Results for F1Score, Precision and Recall across the 3 change scenarios are summarised in Figures 23. For values of $\phi \leq 0.5$ DeCAFS has comparable performances to the ones of the model where we have dependence across segment. Throughout DeCAFS tends to perform similarly to or better than AR1Seg.

Figure 24 shows a comparison between LAVA and DeCAFS when the mean is sinusoidal with abrupt jumps.

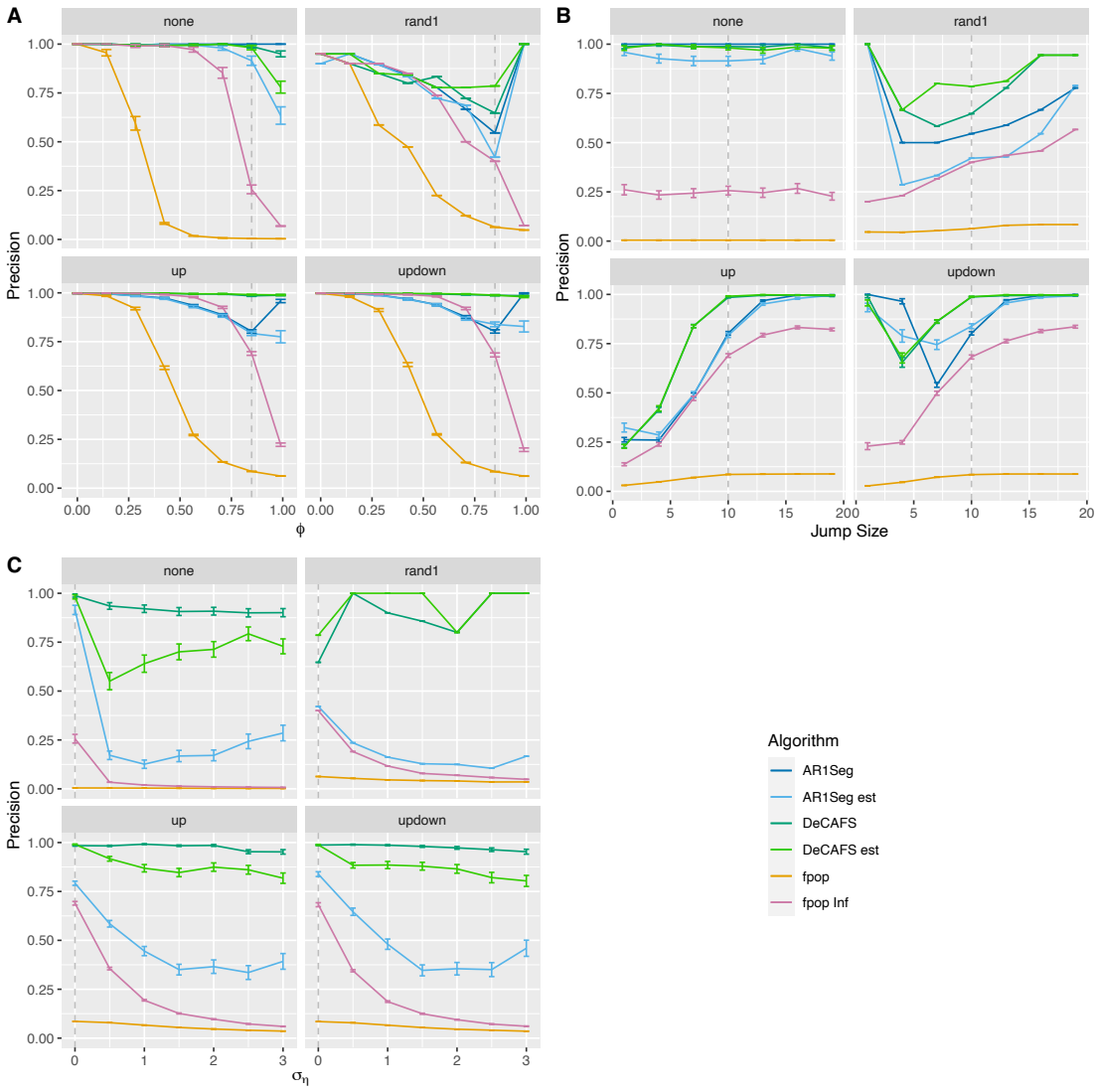


Figure 18: Precision on the 4 different scenarios from the main simulation study of Section 6. Should be read in conjunction with Figure 6.1.

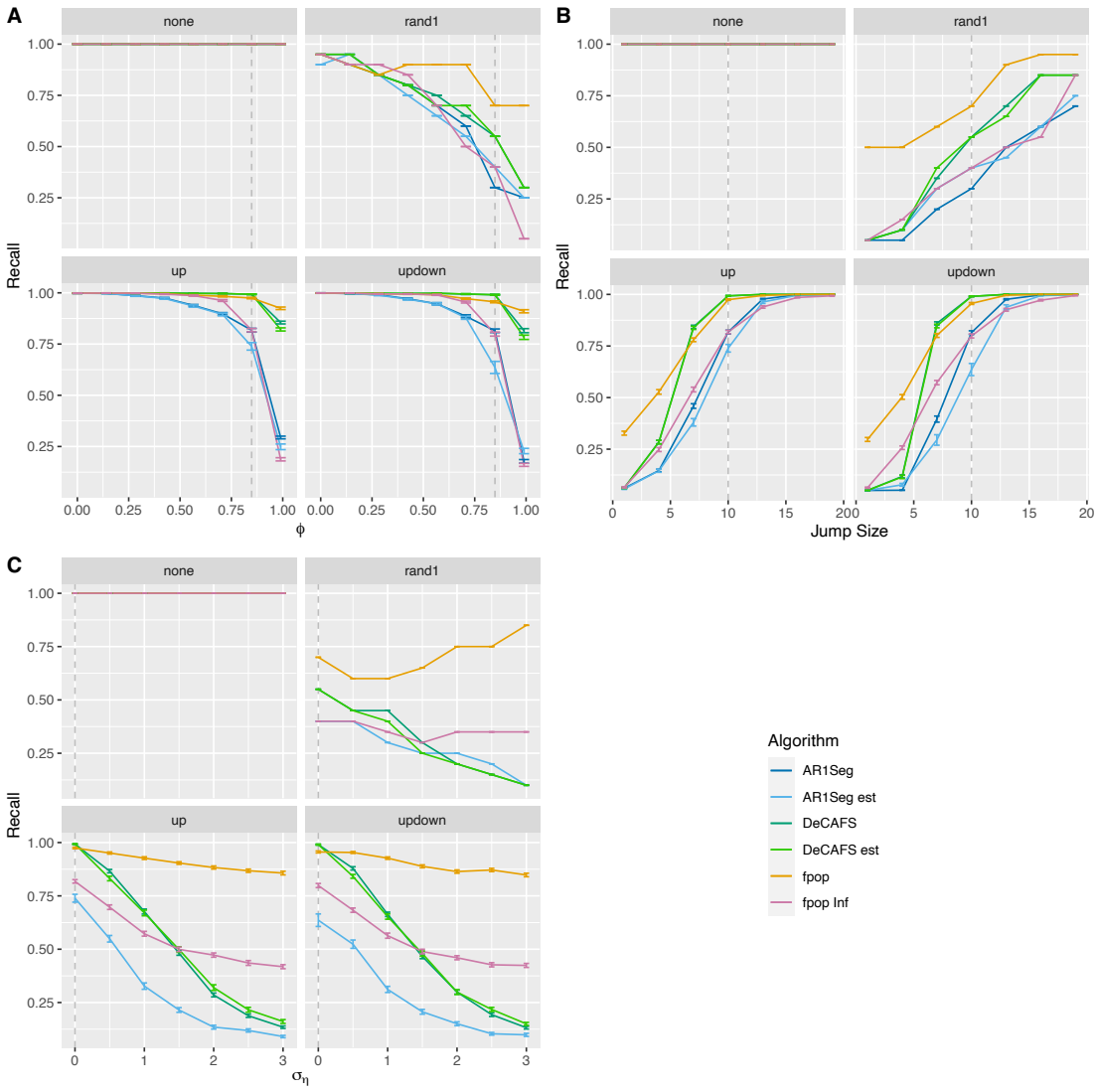
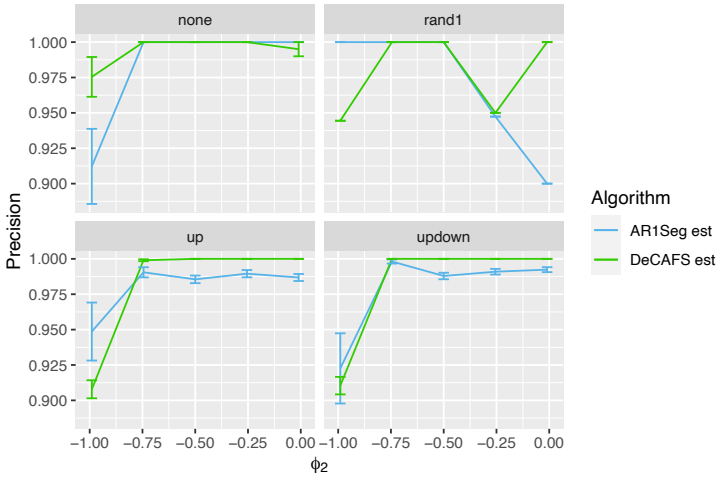
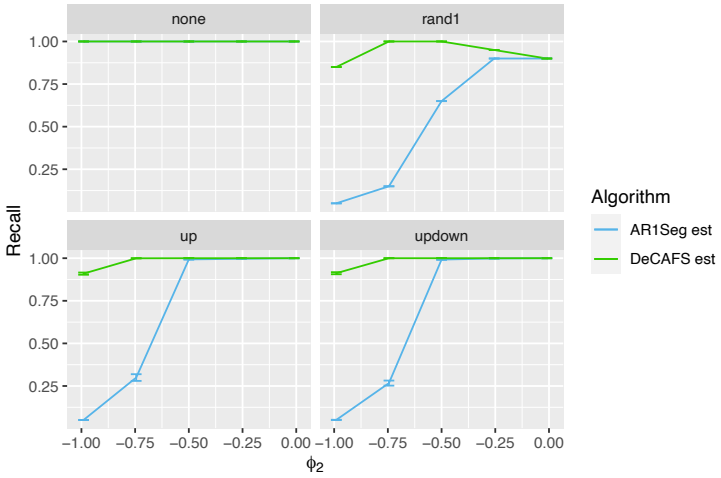


Figure 19: Recall on the 4 different scenarios from the main simulation study of Section 6. Should be read in conjunction with Figure 6.1.

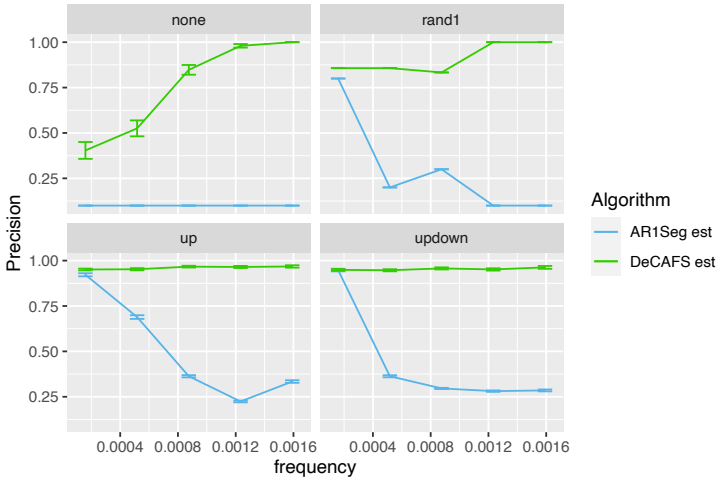


(a)

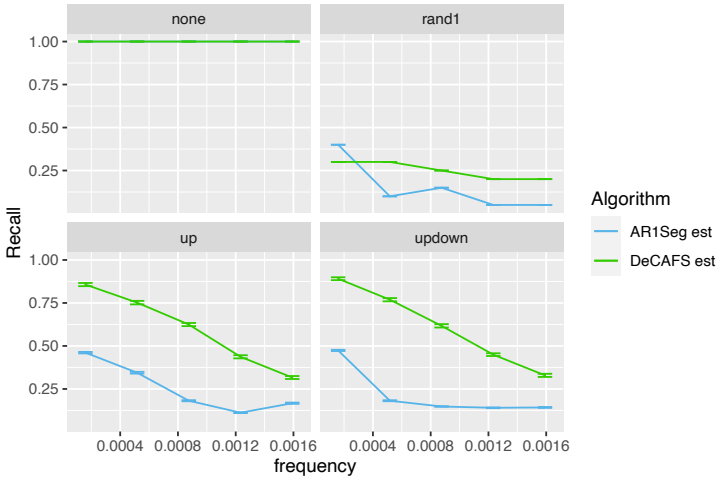


(b)

Figure 20: Precision (a) and Recall (b) on different scenarios with a AR(2) noise. Should be read in conjunction with Figure 5.

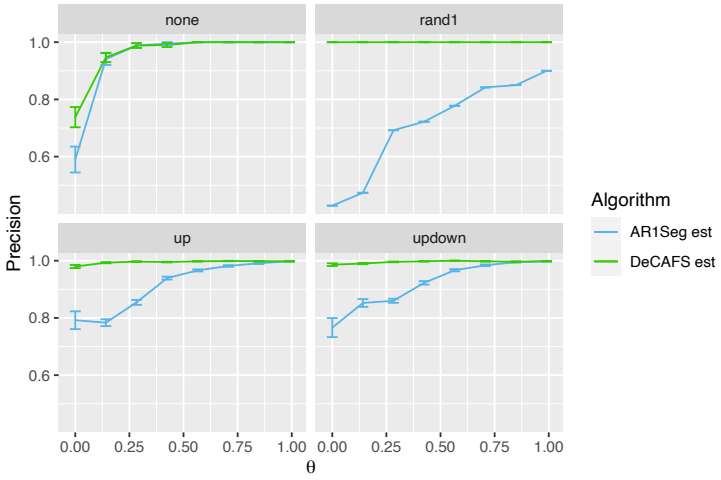


(a)

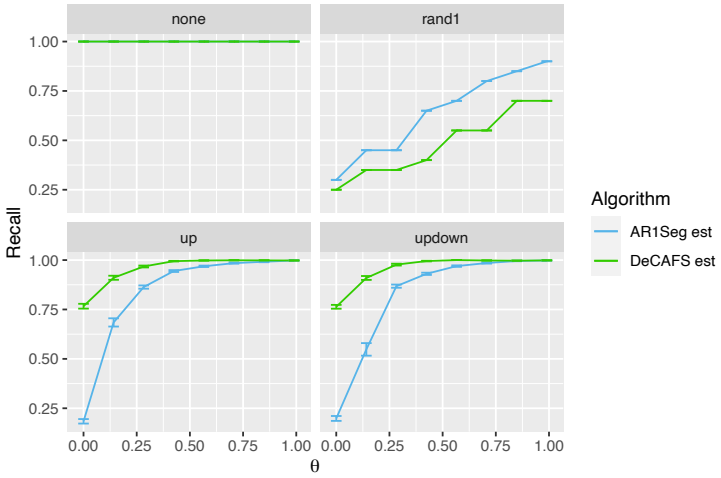


(b)

Figure 21: Precision (a) and Recall (b) on different scenarios with an underlying sinusoidal process. Should be read in conjunction with Figure 6.

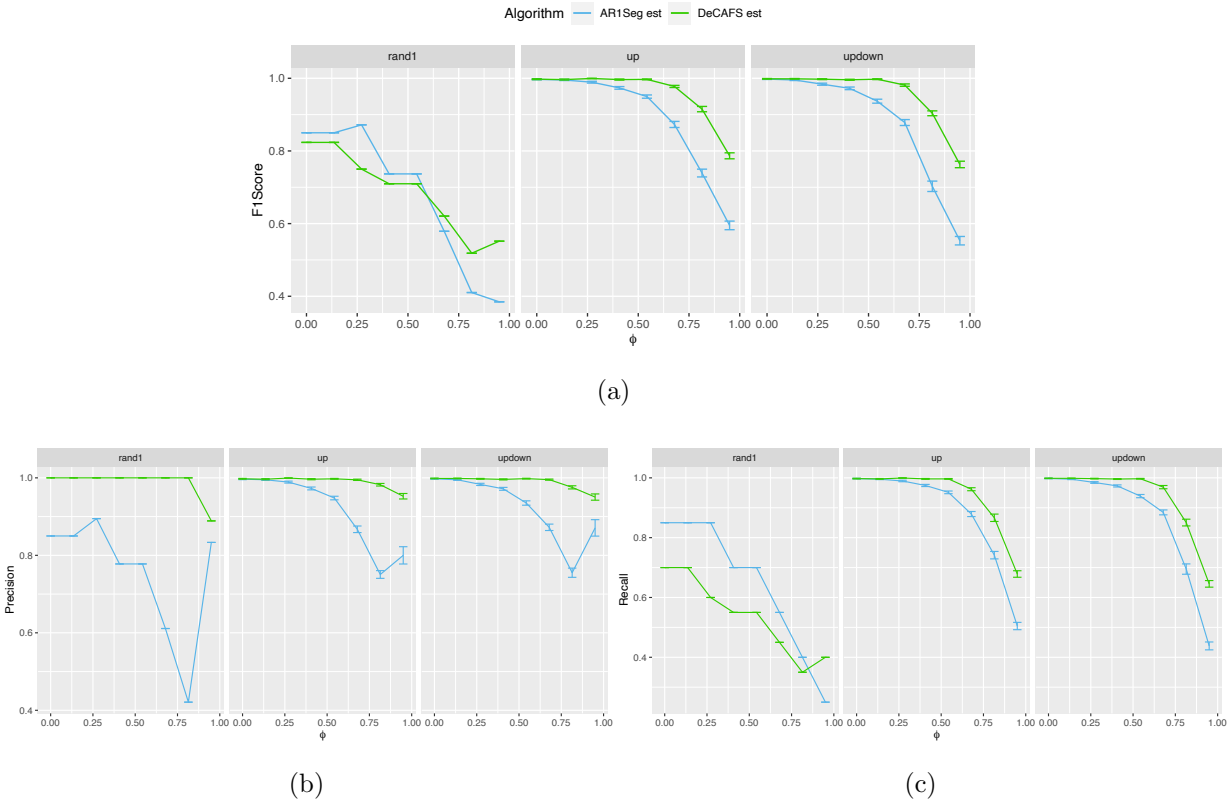


(a)



(b)

Figure 22: Precision (a) and Recall (b) on different scenarios with an underlying Ornstein-Uhlenbeck process. Should be read in conjunction with Figure 15.



(a)

(b)

(c)

Figure 23: F1 score (a), Precision (b) and Recall (c) on 3 different change scenarios with an independent between-the-changes AR(1) noise as we vary ϕ. Data simulated fixing $\sigma_\nu = 2$ over a change of size 10.

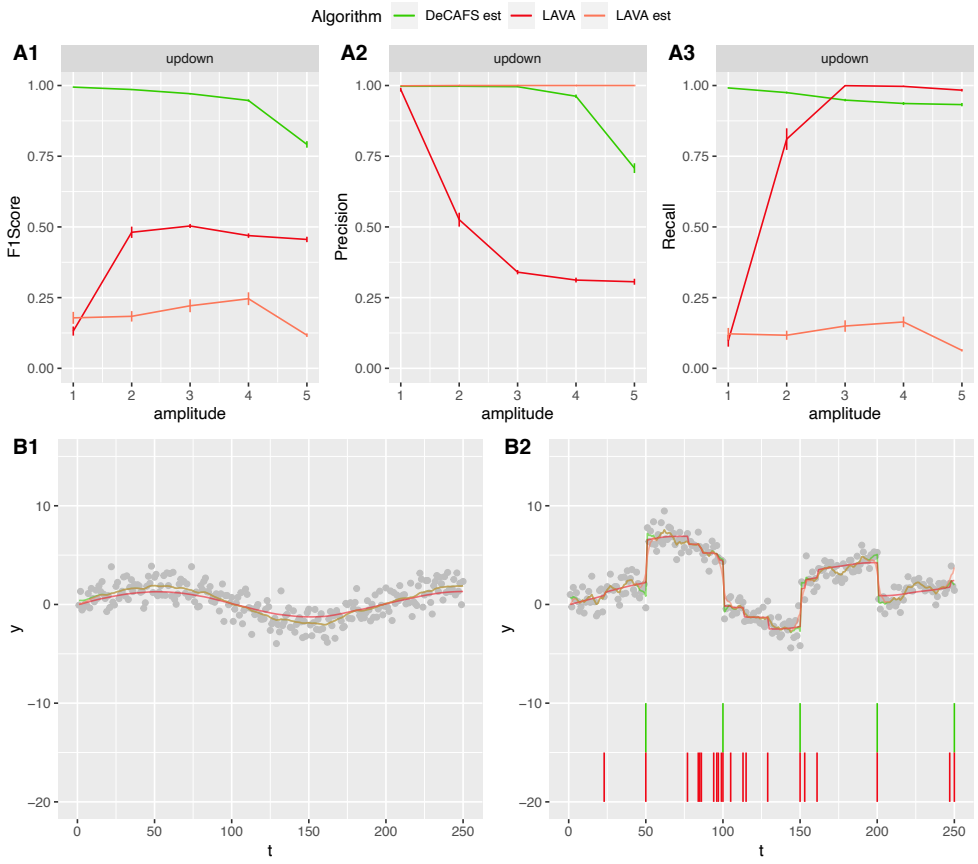


Figure 24: On top: comparison of the F1 Score, in **A1**, Precision in **A2** and Recall, in **A3**, for DeCAFS est (in light green) and LAVA (red) and LAVA est (in orange) on the updown scenario for a sinusoidal signal over a range of different amplitudes. On the bottom the first 250 observations of two realization of the experiment with an amplitude of 2, in **B1** with no changes, whilst in **B2** with 20 changes. The continuous line over the data points represent the relative signal estimations of DeCAFS est LAVA oracle, and LAVA est; the segments their changepoint locations estimates. In **B1**, in particular, LAVA est and DeCAFS est have an almost equal signal estimation.