Cross-validation for change-point regression: pitfalls and solutions

Florian Pein[1,2] **and Rajen D. Shah**[2]

[1] *Lancaster University, UK, e-mail:* f.pein@lancaster.ac.uk

[2] *University of Cambridge, UK, e-mail:* r.shah@statslab.cam.ac.uk

Abstract: Cross-validation is the standard approach for tuning parameter selection in many non-parametric regression problems. However its use is less common in change-point regression, perhaps as its prediction error-based criterion may appear to permit small spurious changes and hence be less well-suited to estimation of the number and location of change-points. We show that in fact the problems of cross-validation with squared error loss are more severe and can lead to systematic under- or over-estimation of the number of change-points, and highly suboptimal estimation of the mean function in simple settings where changes are easily detectable. We propose two simple approaches to remedy these issues, the first involving the use of absolute error rather than squared error loss, and the second involving modifying the holdout sets used. For the latter, we provide conditions that permit consistent estimation of the number of change-points for a general change-point estimation procedure. We show these conditions are satisfied for least squares estimation using new results on its performance when supplied with the incorrect number of change-points. Numerical experiments show that our new approaches are competitive with common change-point methods using classical tuning parameter choices when error distributions are well-specified, but can substantially outperform these in misspecified models. An implementation of our methodology is available in the `R` package `crossvalidationCP` on CRAN.

MSC2020 subject classifications: Primary 62G08; secondary 62G20.
Keywords and phrases: Change-point regression, Cross-validation, Segment Neighbourhood, Sample splitting, Selection consistency, Tuning parameter selection.

1. Introduction

Driven by a need to study datasets that exhibit abrupt changes in distribution, often across time, the field of change-point analysis has received a great deal of attention in recent years. Application areas where such data are common include biochemistry (Pein, Eltzner and Munk, 2021), finance (Bai and Perron, 2003; Kim, Morley and Nelson, 2005), genomics (Olshen et al., 2004), quality monitoring (D'Angelo et al., 2011) and speech processing (Harchaoui et al., 2009), to name a few. Perhaps the simplest model studied involves data $Y = (Y_1, \dots, Y_n) \in \mathbb{R}^n$ satisfying

$$Y_i = \mu_i + \varepsilon_i, \qquad i = 1, \dots, n, \tag{1}$$

where the ε_i are independent mean-zero errors, and $\mu := (\mu_1, \ldots, \mu_n)$ is piecewise constant with change-points $\tau_1 < \cdots < \tau_K$; that is $\mu_i \neq \mu_{i+1}$ if and only if $i = \tau_k$ for some k, for $i = 1, \ldots, n-1$.

A variety of methods exists for estimating μ and the unknown change-points, for instance binary segmentation (Vostrikova, 1981) and its variants (Olshen et al., 2004; Fryzlewicz, 2014, 2020; Kovács et al., 2020), (penalized) cost methods such as Segment Neighbourhood (Auger and Lawrence, 1989; Jackson et al., 2005; Zhang and Siegmund, 2007; Killick, Fearnhead and Eckley, 2012; Maidstone et al., 2017; Verzelen et al., 2020), multiscale methods (Frick, Munk and Sieling, 2014; Li, Munk and Sieling, 2016) and Bayesian approaches (Fearnhead, 2006; Du, Kao and Kou, 2016), among many others; for further detail see Niu, Hao and Zhang (2016); Truong, Oudre and Vayatis (2020); Fearnhead and Rigaill (2020) and references therein. The empirical and theoretical properties of these methods typically rely on selecting appropriate choices for tuning parameters. For instance, Segment Neighbourhood (Auger and Lawrence, 1989) requires selection of the number of change-points to estimate, and then determines the location of these to minimise the residual sum of squares. Given the correct number of change-points, the least squares estimate of the change-point locations is minimax rate-optimal and its L_2-error rate is optimal up to log-factors (see Theorem 5 in Section 3.4 and the following discussion as well as (18)). Other approaches require different tuning parameters, and for some of these methods, theoretically motivated choices of those tuning parameters can be very successful in idealised settings where the joint distribution of the ε_i is known. However, their performance can deteriorate when the error distribution is misspecified, as is likely to be the case in practice, particularly when the errors have heavy tails or are heteroscedastic.

The problem of selecting regression procedures indexed by tuning parameters is of course encountered in more general regression settings, and here cross-validation is typically the method of choice. One of the appeals of cross-validation is its inherently model-free nature which confers a certain robustness. It has been shown to be very successful empirically in non-parametric and high-dimensional regression settings and theoretical guarantees have provided additional support for its usage (Wong, 1983; Yang, 2007; Arlot and Celisse, 2010; Yu and Feng, 2014; Chetverikov, Liao and Chernozhukov, 2021).

The use of cross-validation is however less common in change-point regression, and the only theoretical contributions we are aware of are Arlot and Celisse (2011) and Zou, Wang and Li (2020). The former provides results on the quality of a cross-validation estimation of the prediction error for a least squares estimate based on a given fixed set of putative change-points. These results however do not directly tackle the problem of whether cross-validation can provide a consistent estimate of the true number of changes; this latter problem is studied in Zou, Wang and Li (2020) which we discuss further in the following.

A misgiving one may have about cross-validation is that it typically targets procedures with good prediction properties, and a method that introduces many spurious changes with small jump sizes may not suffer too much from this perspective. This may be concerning given that the goal in change-point regression

is often accurate estimation of the number and locations of the change-points.

In this work however, we show that the shortcomings of standard cross-validation with squared error loss are more serious, and can lead to both under- and over-selection of the number of change-points, and perhaps surprisingly, poor performance in terms of mean squared error. The central issue is that if one of the holdout sets includes a point immediately following or preceding a large change-point, the squared error incurred when predicting at that data point can dominate the cross-validation error criterion. We detail this in Section 2 where we provide some formal negative results for the use of cross-validation with least squares estimation. In Sections 4.1, S1.1 and S1.2 we demonstrate empirically that this can substantially affect the performance of cross-validation in practice, and moving a single change-point from an even to an odd location for example, can lead to a drastic deterioration in performance. This issue extends also to the cross-validation procedures of Arlot and Celisse (2011), which are also based on squared error loss.

One reason this problem has not (to our knowledge) been highlighted in prior literature, is that existing theoretical results on cross-validation consider asymptotic regimes which may either implicitly or explicitly assume that a change-point procedure trained on a subset of the data will make bounded expected squared errors on the remaining data as is the case in Arlot and Celisse (2011, Prop. 1). However by bypassing the issue of poor predictive performance described above, the insights of such asymptotic regimes for finite samples are thus perhaps somewhat limited.

Our second contribution is to further advance the point made in Zou, Wang and Li (2020) that the basic intuition that cross-validation encourages too many small spurious changes, is not necessarily well-founded. In Sections 3.1 and 3.2 we propose two simple approaches to avoid the problems associated with large changes. The first involves using absolute error rather than squared error in the cross-validation criterion. The second involves modifying the cross-validation score for squared error loss to avoid the problematic points. For the latter, we provide relatively mild conditions on the underlying change-point regression procedure under which the cross-validation approach is consistent for selecting the number of change-points; no additional penalisation (or need for choosing appropriate tuning parameters that would come with it) is required to achieve this consistency.

Our theory builds on the work of Zou, Wang and Li (2020), who show consistency of a cross-validation scheme in their Theorems 1 and 2. However inspection of their proofs shows that their conclusions as stated may need some caveats. Firstly, though not explicitly stated, the proofs require all change-points to occur at even locations. This simplification of the model may be justified when the noise and signal strength are bounded away from 0 and ∞ respectively such that the expected errors of the fitted regression function are bounded; however, as explained above, this asymptotic regime may then have less relevance to practice. Secondly, it is unclear to us how the arguments in their proofs may be extended to allow for the maximum number of change-points considered and the true number of change-points to diverge; see the discussion before and after (66) in

Section S5 in the supplementary material of our paper. This seems particularly relevant given that the number of change-points is the object of inference. In contrast, our result (Theorem 4) allows for the number of change-points to tend to infinity. We verify the conditions of our general result for the case of least squares estimation by employing a new result (Theorem 5 in Section 3.4) on the existence of estimated change-points in the neighbourhood of true changes even when the number of changes has been incorrectly specified, which may be of independent interest.

In Section 4, we present numerical experiments that illustrate the performance of our new cross-validation schemes in comparison with commonly used change-point procedures using classical tuning parameter choices. We see that cross-validation with absolute error loss is competitive when the error distribution has been well-specified, but substantially outperforms classical methods in settings with heteroscedastic or heavy-tailed errors, or when outliers are present. We conclude with a discussion in Section 5. The Appendix contains descriptions of generalisations of our methodology; additional numerical experiments as well as all proofs are contained in the supplementary material.

2. Pitfalls of using cross-validation with squared error loss

In this section, we give examples of simple settings where changes are easily detectable but where using cross-validation with squared error loss can lead to both systematic under- and over-estimation of the number of change-points. For these negative results, we focus on the univariate mean change-point regression problem (1), where additionally $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. Furthermore, we consider a version of two-fold cross-validation, termed the COPSS procedure in Zou, Wang and Li (2020), with least squares estimation for estimating the change-point locations[1]. However it will be clear that our constructions highlighting the undesirable properties of cross-validation, and our conclusions, can be generalised to other settings and other forms of cross-validation employing squared error loss.

2.1. Setting

In order to describe and study the COPSS procedure, we introduce some notation. Let $\tau_0 = 0$ and $\tau_{K+1} = n$. Let β_k for $k = 1, \ldots, K$ be the mean of the signal in the kth constant segment, so

$$\mu_i = \beta_k, \quad \text{if and only if } \tau_k < i \leq \tau_{k+1}, \ k = 0, \ldots, K; \ i = 1, \ldots, n. \tag{2}$$

[1] Throughout the paper we will use the terminology least squares estimation. Zou, Wang and Li (2020) called it optimal partitioning, highlighting the fact that least squares estimation is performed for various putative numbers of change-points. Elsewhere in the literature however, this is known as Segment Neighbourhood (Auger and Lawrence, 1989), with optimal partitioning referring to a penalised version. We finally remark that least squares estimation coincides with maximum likelihood estimation if the noise is Gaussian.

For an arbitrary vector of observations $Z := (Z_1, \ldots, Z_m)$ and a putative number of change-points L, least squares estimation obtains the change-points as

$$\left(\hat{\tau}^Z_{L,1}, \ldots, \hat{\tau}^Z_{L,L}\right) := \underset{0=:t_0<t_1<\cdots<t_L<t_{L+1}:=n}{\operatorname{argmin}} \sum_{l=0}^{L} \sum_{i=t_l+1}^{t_{l+1}} \left(Z_i - \overline{Z}_{t_l:t_{l+1}}\right)^2, \tag{3}$$

where $\overline{Z}_{a:b} := (b-a)^{-1} \sum_{i=a+1}^{b} Z_i$. In Section 3.4 we will give theoretical guarantees for the estimates $\hat{\tau}^Z_{L,l}$. We will assume for simplicity here and throughout that n is even; if not, the final observation may be dropped.

In order to perform COPSS with least squares estimation, we will apply the above to the odd and even indexed observations separately. To study this, we introduce

$$\begin{aligned} Y^O_i &:= Y_{2i-1}, \quad \mu^O_i := \mu_{2i-1}, \quad \varepsilon^O_i := \varepsilon_{2i-1}, \quad i = 1, \ldots, n/2, \\ Y^E_i &:= Y_{2i}, \quad \mu^E_i := \mu_{2i}, \quad \varepsilon^E_i := \varepsilon_{2i}, \quad i = 1, \ldots, n/2. \end{aligned} \tag{4}$$

We write $\mathcal{T}^O := \{\tau^O_0, \ldots, \tau^O_{K+1}\}$ for the set of true change-points among the odd observations $Y^O := (Y^O_1, \ldots, Y^O_{n/2})$, so

$$\mu^O_i = \beta_k, \quad \text{if and only if } \tau^O_k < i \leq \tau^O_{k+1}, \ k = 0, \ldots, K; \ i = 1, \ldots, n/2.$$

We denote the estimated change-point set obtained by applying least squares estimation to Y^O by

$$\hat{\mathcal{T}}^O_L := \{0 = \hat{\tau}^O_{L,0} < \hat{\tau}^O_{L,1} < \cdots < \hat{\tau}^O_{L,L} < \hat{\tau}^O_{L,L+1} = n/2\}, \tag{5}$$

and define $\mathcal{T}^E$ and $\hat{\mathcal{T}}^E_L$ analogously for the even observations.

With these, we may now define the cross-validation criterion for L change-points using squared error loss as

$$\mathrm{CV}_{(2)}(L) := \sum_{l=0}^{L} \sum_{i=\hat{\tau}^O_{L,l}+1}^{\hat{\tau}^O_{L,l+1}} \left(Y^E_i - \overline{Y}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}}\right)^2 + \sum_{l=0}^{L} \sum_{i=\hat{\tau}^E_{L,l}+1}^{\hat{\tau}^E_{L,l+1}} \left(Y^O_i - \overline{Y}^E_{\hat{\tau}^E_{L,l}:\hat{\tau}^E_{L,l+1}}\right)^2, \tag{6}$$

where the subscript (2) in $\mathrm{CV}_{(2)}(L)$ emphasises the use of squared error loss. Lastly, the estimated number of change-points K is given by

$$\hat{K} := \underset{L=0,\ldots,K_{\max}}{\operatorname{argmin}} \ \mathrm{CV}_{(2)}(L), \tag{7}$$

where $K_{\max} \geq 1$ is a pre-specified upper bound for the number of change-points. Given $\hat{K}$, we may perform least squares estimation with $L = \hat{K}$ on the full vector of observations Y to produce a final set of estimated change-points $0 =: \hat{\tau}_0 < \hat{\tau}_1 < \cdots < \hat{\tau}_{\hat{K}} < \hat{\tau}_{\hat{K}+1} := n$ and an estimate $\hat{f} : [0,1] \to \mathbb{R}, \ t \mapsto \sum_{k=0}^{\hat{K}} \overline{Y}_{\hat{\tau}_k:\hat{\tau}_{k+1}} \mathbb{1}_{(\hat{\tau}_k/n, \hat{\tau}_{k+1}/n]}(t)$ of the true mean function $f : [0,1] \to \mathbb{R}, \ t \mapsto \sum_{k=0}^{K} \beta_k \mathbb{1}_{(\tau_k/n, \tau_{k+1}/n]}(t)$.

The motivation for this approach is that in (6), Y_i^E is (almost) an unbiased proxy for μ_i^O, and similarly for the odd and even designators interchanged. The "almost" qualification is due to the fact that the unbiasedness may fail to hold immediately after a change, so for example μ_i^O and $\mathbb{E}Y_i^E = \mu_i^E$ can be very different when $i = \tau_k^O$. Whilst this will only occur at isolated points, the fact that the errors are squared in (6) can lead these discrepancies to dominate the cross-validation criterion when changes are large, and this has severe consequences for the quality of the estimate $\hat{K}$ as we show formally in the following.

2.2. Underestimation

In this section we present a scenario in which cross-validation will under-estimate the number of change-points with high probability.

Example 1 (Underestimation). Let $Y \in \mathbb{R}^n$ be of the form in (1) with mean vector μ as in (2). Suppose n is even, but $n/2$ odd. Let $K = 2$ and for odd $\underline{\lambda} < n/4$, let $\tau_1 = n/2 - \underline{\lambda}$, $\tau_2 = n/2$. We set $\Delta_1 = \beta_1$, $\Delta_2 = \beta_3$ and $\beta_2 = 0$ and suppose $\Delta_1 < \Delta_2$, so Δ_2 and Δ_1 are the sizes of the largest and smallest jumps respectively. An illustration of this construction is given in Figure 1.

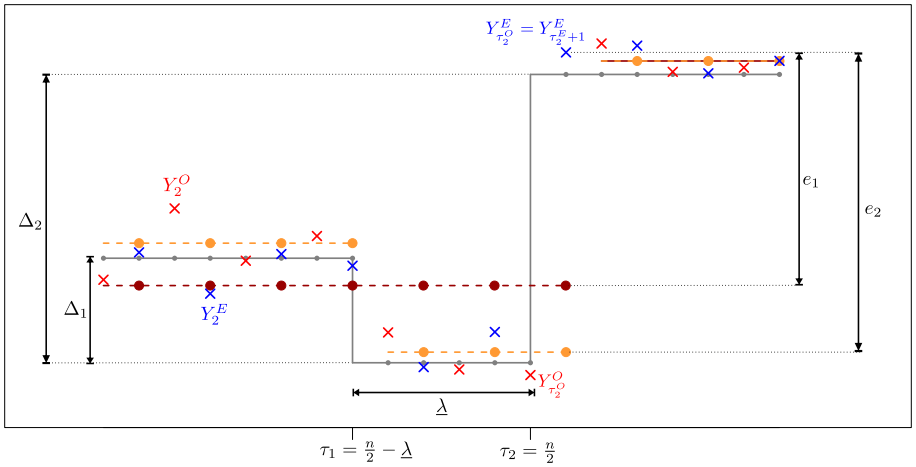

FIG 1. *Schematic of Example 1. Expectations are visualized by grey dots and observations by coloured crosses, split into the two folds given by* $\mathbf{Y^O}$ *(red) and* $\mathbf{Y^E}$ *(blue). The predictions* $(\overline{Y}^{\mathbf{O}}_{\hat{\tau}^{\mathbf{O}}_{L,l}:\hat{\tau}^{\mathbf{O}}_{L,l+1}})_{l=0}^{L}$ *from the odd observations, are shown for* $L = 2$ *(orange dots) and* $L = 1$ *(brown dots). Distances* e_2 *and* e_1 *between* $Y^E_{\tau_2^O}$ *and the corresponding predictions are much larger than those corresponding to other observations, and can dominate the cross-validation criterion. The larger size of* e_2 *results in* $\mathrm{CV}_{(2)}$ *being minimised at* $L = 1$.

In Example 1 above, we have that as $n/2$ is odd, $\tau_2^O = (n/2+1)/2$, but $\tau_2^E = (n/2-1)/2$, and so $\mu^E_{\tau_2^O} = \Delta_2 \neq \mu^O_{\tau_2^O} = 0$. Thus, if Δ_2 is large, the point following the large second change, $Y^E_{\tau_2^E+1} = Y^E_{\tau_2^O}$, contributes heavily to the cross-validation

criterion. To see why this is problematic, it is instructive to first consider a noiseless setting where $\sigma = 0$. Then with a correctly specified $L = K = 2$, we obtain

$$\begin{aligned}\mathrm{CV}_{(2)}(2) &= \left(\mu^O_{\tau_2^E+1} - \beta_2\right)^2 + \left(\mu^E_{\tau_2^O} - \beta_1\right)^2 = \left(\mu^O_{\tau_2^O} - \beta_2\right)^2 + \left(\mu^E_{\tau_2^E+1} - \beta_1\right)^2 \\ &= (\beta_1 - \beta_2)^2 + (\beta_2 - \beta_1)^2 = \Delta_2^2 + \Delta_2^2.\end{aligned}$$

On the other hand, when $L = 1$, least squares estimation recovers only the large second change-point. Hence, the first segment consists of $n/2 - \underline{\lambda}$ observations with value $\beta_0 = \Delta_1$ and $\underline{\lambda}$ observations with value $\beta_1 = 0$. Hence, $\overline{\mu}^O_{0:\hat{\tau}^O_{1,1}} = \overline{\mu}^E_{0:\hat{\tau}^E_{1,1}} \approx \frac{n-2\underline{\lambda}}{n}\Delta_1$. Thus,

$$\begin{aligned}\mathrm{CV}_{(2)}(1) &\approx \left(\frac{n}{2} - \underline{\lambda}\right)\left(\frac{2\underline{\lambda}}{n}\right)^2 \Delta_1^2 + \underline{\lambda}\left(\frac{n-2\underline{\lambda}}{n}\right)^2 \Delta_1^2 + \left(\mu^O_{\tau_2^E+1} - \beta_2\right)^2 + \left(\mu^E_{\tau_2^O} - \frac{n-2\underline{\lambda}}{n}\Delta_1\right)^2 \\ &= \left(1 - \frac{2\underline{\lambda}}{n}\right)\underline{\lambda}\Delta_1^2 + \Delta_2^2 + \left(\Delta_2 - \frac{n-2\underline{\lambda}}{n}\Delta_1\right)^2.\end{aligned}$$

We then obtain

$$\mathrm{CV}_{(2)}(2) - \mathrm{CV}_{(2)}(1) \approx \Delta_1^2\left(2\frac{\Delta_2}{\Delta_1} - \underline{\lambda}\right)\left(1 - \frac{2\underline{\lambda}}{n}\right) > \frac{1}{2}\Delta_1^2\left(2\frac{\Delta_2}{\Delta_1} - \underline{\lambda}\right). \tag{8}$$

Thus when $2\Delta_2/(\underline{\lambda}\Delta_1) > 1$ we can expect cross-validation to favour $L = 1$ change-point.

More precisely, and taking account of the presence of noise, we have the following asymptotic result. Note that here and in the sequel, all parameters (in the current case Δ_1, Δ_2, $\underline{\lambda}$ and σ) are permitted to change with n, though we suppress this in the notation. Also, although our result is stated for simplicity for the case of COPSS, a similar conclusion would hold for other squared error cross-validation schemes, possibly with the large change at an a different location depending on the details of the method employed such as the number of folds used. One example is the LooVF procedure from Arlot and Celisse (2011), where the issues described above occur when one considers Example 1 in reverse order. This is because they extrapolate to the left when predicting at unseen time points. In general, a choice has to be made to extrapolate in either direction, and depending on the underlying signal, each strategy can result in poor performance.

Theorem 1. *Let $Y \in \mathbb{R}^n$ be as in Example 1. Suppose that the following hold:*

$$\frac{\underline{\lambda}\Delta_1^2}{\sigma^2 \log(n)} \to \infty, \tag{9}$$

$$\liminf_{n\to\infty} \frac{2}{\underline{\lambda}}\frac{\Delta_2}{\Delta_1} > 1. \tag{10}$$

Then $\hat{K}$ (7) *satisfies* $\mathbb{P}(\hat{K} = 2) \to 0$. *Moreover, if additionally* $K_{\max} = K = 2$,

$$\left[\int_0^1 \left(\hat{f}(t) - f(t)\right)^2 dt\right]^{-1} = O_{\mathbb{P}}\left(\frac{n}{\underline{\lambda}\Delta_1^2}\right). \tag{11}$$

Condition (9) is the minimum requirement that ensures that both change-points are detectable; see the discussion after Theorem 3. Cross-validation however is unlikely to select the correct number of change-points provided the size Δ_2 of the largest jump is large compared to the product of the minimum gap $\underline{\lambda}$ between the change-points and the smallest jump size Δ_1 (10).

If overestimation is not permitted, i.e. $K_{\max} = K$, (11) shows that the L_2-loss of the estimated function is at least of order $\underline{\lambda}\Delta_1^2/n$. We may contrast this with the corresponding L_2-loss realised by our proposed criterion, which for general signals is $O_{\mathbb{P}}\left(n^{-1}\sigma^2 \log\log \overline{\lambda}\right)$, see (18). It follows from (9) that in the setting of Theorem 1, the quotient of these losses converges to zero, underlining the suboptimality of cross-validation based on squared error loss even with respect to L_2-loss.

2.3. Overestimation

We now introduce an example where cross-validation with least squares loss has a tendency to overestimate the number of change-points.

Example 2 (Overestimation). As in Example 1 let $Y \in \mathbb{R}^n$ be of the form in (1) with n even but $n/2$ odd. Let $K = 1$, $\tau_1 = n/2$, $\beta_0 = 0$ and $\beta_1 = \Delta_1$.

In this example, $\tau_1^O = (n/2+1)/2$ and $\mu^E_{\tau_1^O} - \mu^O_{\tau_1^O} = \Delta_1$. Thus, if Δ_1 is large, the cross-validation criterion will be heavily influenced by $Y^E_{\tau_1^O}$. More precisely, τ_1^O is estimated exactly with high probability if Δ_1 is large. Thus on this event, the criterion contains the term

$$\left(Y^E_{\tau_1^O} - \overline{Y}^O_{\hat{\tau}:\tau_1^O}\right)^2 = \left(\varepsilon^E_{\tau_1^O} + \Delta_1 - \overline{\varepsilon}^O_{\hat{\tau}:\tau_1^O}\right)^2, \tag{12}$$

where $\hat{\tau}$ denotes the last estimated change-point before τ_1^O. If $L = K = 1$, $\hat{\tau} = 0$. However, if $L > K$, $\mathbb{P}(\hat{\tau} > 0) \geq \frac{1}{2}$ and $\mathbb{P}\left(\overline{\varepsilon}^O_{\hat{\tau}:\tau_1^O} > \overline{\varepsilon}^O_{0:\tau_1^O}\right) \geq \frac{1}{2}$. Thus, if Δ_1 is large, the cross-term $\Delta_1 \overline{\varepsilon}^O_{\hat{\tau}:\tau_1^O}$ in (12) can outweigh the costs incurred by the additional change-points and cause overestimation. This is formalised in Theorem 2. Note that for this effect it is only important that τ_1 is at an odd location: all other choices are just made to simplify the analysis.

Theorem 2. *Let* Y *be as in Example 2 and* $\hat{K}$ *be as in* (7) *with* $K_{\max} > 1$. *Then if*

$$\frac{\Delta_1}{\sigma\sqrt{n \log\log n}} \to \infty, \quad \text{as } n \to \infty, \tag{13}$$

we have

$$\liminf_{n\to\infty} \mathbb{P}\left(\hat{K} > 1\right) > 0. \tag{14}$$

The simulations in Section S1.2 in the supplementary material suggest that the probability in (14) can be roughly 2/3 when the signal to noise ratio is large. In these settings we also see that the probability drops to less than 9% when the change-point is moved by a single design point to an even location, illustrating that the single point is solely responsible for the unfavourable behaviour. Similarly to Example 1, there exist procedures that correctly estimate the number of change-points to be 1 with probability converging to 1. In particular, our modified squared error loss criterion, which we describe in the next section, would achieve this; see Theorem 3.

3. New cross-validation criteria

We now introduce two new cross-validation criteria which circumvent the issues of squared error loss presented in the previous section. In Section 3.1 we motivate the use of absolute errors and in Section 3.2 we propose a modified criterion with squared errors. For the latter we prove theoretical guarantees on consistent estimation of the number of change-points in a more general multivariate model introduced in Section 3.3.

3.1. Cross-validation with absolute error loss

We have seen how when standard squared error loss is used, a single point following a large change-point can dominate the cross-validation criterion. For instance in Example 1, $Y^E_{\tau_2^E+1} = Y^E_{\tau_2^O}$ contributes e_L^2 to $\mathrm{CV}_{(2)}(L)$, where

$$e_1^2 \approx \left(\varepsilon^E_{\tau_2^O} + \Delta_2 - \frac{n-2\underline{\lambda}}{n}\Delta_1\right)^2 \qquad \text{and} \qquad e_2^2 \approx (\varepsilon^E_{\tau_2^O} + \Delta_2)^2;$$

see Figure 1. If Δ_2 is large, the difference $e_2^2 - e_1^2$ will be large, resulting in cross-validation erroneously favouring $\hat{K} = 1$ change-point.

Consider now cross-validation with absolute error loss:

$$\mathrm{CV}_{(1)}(L) := \sum_{l=0}^{L} \sum_{i=\hat{\tau}^O_{L,l}+1}^{\hat{\tau}^O_{L,l+1}} \left|Y_i^E - \overline{Y}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}}\right| + \sum_{l=0}^{L} \sum_{i=\hat{\tau}^E_{L,l}+1}^{\hat{\tau}^E_{L,l+1}} \left|Y_i^O - \overline{Y}^E_{\hat{\tau}^E_{L,l}:\hat{\tau}^E_{L,l+1}}\right|. \tag{15}$$

Then the difference of the contribution of $Y^E_{\tau_2^O}$ to the criterion will be $|e_2| - |e_1|$, which when Δ_2 is large will, with high probability, simply be

$$e_2 - e_1 \approx \frac{n-2\underline{\lambda}}{n}\Delta_1,$$

which importantly does not feature Δ_2, and is of the same order as the additional loss incurred at other points. Specifically, considering a noiseless version

of Example 1 as in (8), we have

$$\begin{aligned}\mathrm{CV}_{(1)}(1) - \mathrm{CV}_{(1)}(2) &\approx \frac{2\underline{\lambda}\Delta_1}{n}\left(\frac{n}{2} - \underline{\lambda}\right) + \underline{\lambda}\Delta_1\frac{n-2\underline{\lambda}}{n} + \Delta_2 + \left(\Delta_2 - \Delta_1\frac{n-2\underline{\lambda}}{n}\right) - 2\Delta_2 \\ &= \Delta_1\left((\underline{\lambda}-1)\frac{n-2\underline{\lambda}}{n} + \left(\frac{n}{2} - \underline{\lambda}\right)\frac{2\underline{\lambda}}{n}\right) > 0.\end{aligned}$$

We show in Section 4 through numerical experiments that a 5-fold version of $\mathrm{CV}_{(1)}$ (15) works well in general, and can substantially outperform other methods in the realistic scenario where models have been misspecified. We therefore recommend this as the default option for cross-validation in change-point regression. In Appendix A we extend (15) to multivariate settings and general V-fold cross-validation.

3.2. Modified cross-validation with squared error loss

A second approach to addressing the issues of standard cross-validation with squared error loss is to remove those observations that have the potential to dominate the criterion as follows:

$$\begin{aligned}\mathrm{CV}_{\mathrm{mod}}(L) := &\sum_{l=0}^{L}\frac{\hat{\tau}^O_{L,l+1} - \hat{\tau}^O_{L,l}}{\hat{\tau}^O_{L,l+1} - \hat{\tau}^O_{L,l} - 1}\sum_{i=\hat{\tau}^O_{L,l}+1}^{\hat{\tau}^O_{L,l+1}-1}\left(Y^E_i - \overline{Y}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}}\right)^2 \\ &+ \sum_{l=0}^{L}\frac{\hat{\tau}^E_{L,l+1} - \hat{\tau}^E_{L,l}}{\hat{\tau}^E_{L,l+1} - \hat{\tau}^E_{L,l} - 1}\sum_{i=\hat{\tau}^E_{L,l}+2}^{\hat{\tau}^E_{L,l+1}}\left(Y^O_i - \overline{Y}^E_{\hat{\tau}^E_{L,l}:\hat{\tau}^E_{L,l+1}}\right)^2.\end{aligned} \tag{16}$$

Observe that compared to (6), the squared error terms involving $Y^E_{\hat{\tau}^O_{L,l+1}}$ and $Y^O_{\hat{\tau}^E_{L,l}+1}$, $l = 1, \ldots, L$ have been removed. Out-of-sample prediction at these points necessarily involves extrapolation and thus large change-points immediately following or preceding these can be problematic. To compensate for the removal of these observations, we rescale the sums of squared error terms so that each estimated constant segment gives a contribution proportional to its length in the case where all change-points have been estimated correctly. We thus require that $\hat{\tau}^O_{L,l+1} - \hat{\tau}^O_{L,l} \geq 2$ and $\hat{\tau}^E_{L,l+1} - \hat{\tau}^E_{L,l} \geq 2 \;\forall\; l = 0, \ldots, L$.

In Theorem 3 below, we show that this criterion does indeed deliver consistent estimation of the number of change-points. We consider a generalisation of the setting of Section 2.1 where we relax the Gaussian assumption on the errors and instead require the ε_i to be independent (possibly with different distributions) and sub-Gaussian with variance proxy σ^2, so $\max_i \mathbb{E}[\exp(s\varepsilon_i)] \leq \exp(s^2\sigma^2/2)$ for all $s \in \mathbb{R}$. We assume additionally that all observations on the same segment have the same variance, i.e. $\mathrm{Var}[\varepsilon_i] = \mathrm{Var}[\varepsilon_j]$ if there exists a k such that $\tau_k < i < j \leq \tau_{k+1}$, and writing $\underline{\sigma}^2 := \min_{i=1,\ldots,n}\mathrm{Var}[\varepsilon_i]$ we have $\limsup_{n\to\infty}\sigma/\underline{\sigma} < \infty$.

In order to present our result, we introduce some further notation. We denote by $\Delta_k := |\beta_k - \beta_{k-1}|$ the size of the kth change-point and by $\Delta_{(k)}$ the kth

smallest order statistic of $\Delta_1, \ldots, \Delta_K$, i.e. $\{\Delta_1, \ldots, \Delta_K\} = \{\Delta_{(1)}, \ldots, \Delta_{(K)}\}$ and $\Delta_{(1)} \leq \cdots \leq \Delta_{(K)}$. Further, we denote by $\underline{\lambda}$ and $\overline{\lambda}$ the minimal and maximal distance between two change-points, i.e.

$$\underline{\lambda} := \min_{k=0,\ldots,K} \tau_{k+1} - \tau_k \text{ and } \overline{\lambda} := \max_{k=0,\ldots,K} \tau_{k+1} - \tau_k.$$

We require the following bounds on the speed at which the number of change-points K and its upper bound $K_{\max}$ are permitted to increase.

Assumption 1 (Number of Change-points). (i) $K_{\max} \geq K$ for all n sufficiently large, and K is non-decreasing.
(ii) $K = o(\underline{\lambda})$ and $K(\log(K \vee 1))^2 = o(\log\log\overline{\lambda})$.
(iii) $(K_{\max}\log K_{\max})^{1/2} = o(\log\log\overline{\lambda})$.

Condition (i) ensures that the true number of change-points is considered by the method. This is not too difficult to satisfy as comparing (ii) and (iii) shows that $K_{\max}$ can increase more than quadratically in K. Condition (ii) however restricts the growth of K quite substantially: we have $\overline{\lambda} \geq n/K$ and hence $K(\log(K \vee 1))^2 = o(\log\log n)$. This condition is sufficient to ensure that the costs of adding a false positive, which is of order $\sigma^2 \log\log\overline{\lambda}$, dominates the cost of miss-estimating change-point locations when $L = K$. On the other hand, (iii) ensures that the former also dominates the variance of the cross-validation error criterion for all $K \leq L \leq K_{\max}$. Relative to other results on (non-cross-validation based) change-point approaches, the condition on the growth of K is quite restrictive (Frick, Munk and Sieling, 2014; Garreau and Arlot, 2018; Verzelen et al., 2020). We remark however that (ii) and (iii) are likely artefacts of our proof strategy, which considers worst case scenarios for each estimated change-point location, rather than a fundamental limitation of cross-validation. In practice, $K_{\max}$ does not have to be fixed in advance. Calculating $\mathrm{CV}_{(1)}(L)$ for increasing L and stopping once a clear local minimum is found is possible and is the approach we use in our numerical results; further details are given in Section 3.5.

Theorem 3. *Suppose Assumption 1 holds, and in the case where $K > 0$ eventually,*

$$\liminf_{n\to\infty} \frac{\underline{\lambda}\Delta_{(1)}^2}{K\sigma^2(\log\overline{\lambda})^2} = \infty \quad \textit{and} \quad K\log\log\left((\log(K)\vee 1)\frac{\sigma^2}{\Delta_{(1)}^2} \vee e\right) = o(\log\log\overline{\lambda}). \tag{17}$$

Then we have

$$\mathbb{P}\left(\hat{K} = K\right) \to 1, \ \textit{as } n \to \infty,$$
$$\int_0^1 \left(\hat{f}(t) - f(t)\right)^2 dt = O_{\mathbb{P}}\left(n^{-1}\sigma^2\left(\log(K)\vee 1\right)\log\log\overline{\lambda}\right). \tag{18}$$

Theorem 3 shows that $\mathrm{CV}_{\mathrm{mod}}$ (19), used with least squares estimation, estimates the number of change-points consistently, under mild conditions. The

first condition in (17) is slightly stronger than the minimax rate for change-points to be detectable, which is $\liminf_{n\to\infty} \underline{\lambda}\Delta_{(1)}^2/(\sigma^2 \log \overline{\lambda}) = \infty$; see Chan and Walther (2013, Equation (2.2)) and Dümbgen and Spokoiny (2001); Frick, Munk and Sieling (2014); Fryzlewicz (2014) for related discussions. This first condition may be an artefact of our proof strategy and could perhaps be relaxed. The second condition in (17) ensures that change-point locations are estimated accurately enough such that the costs caused by miss-estimating change-point locations are smaller than the costs of adding a false positive.

To the best of our knowledge, existing L_2-error bounds for other estimators are $O_{\mathbb{P}}\left(n^{-1}\sigma^2 \log(n)\right)$ or worse (Lin et al., 2016, Remarks 4 and 10). In comparison, our bound is only a factor $K(\log(K) \vee 1) \log\log \overline{\lambda} \leq K(\log(K) \vee 1) \log\log n$ larger than the lower bound in Li, Guo and Munk (2019, Theorem 1(ii)). As a side effect we also see that the same bound holds for least squares estimation given the correct number of change-points.

We note that if the second condition in (17) does not hold, but if instead the left hand side is $o(\log(\overline{\lambda}))$, then we may conclude $\mathbb{P}\left(\hat{K} = K \text{ or } \hat{K} = K+1\right) \to 1$, as $n \to \infty$, i.e. the number change-points will be over-estimated by at most one change-point.

3.3. General multivariate model

Our result on the consistency of $\mathrm{CV}_{\mathrm{mod}}$ used with least squares estimation is based on a general result placing conditions on an arbitrary estimation procedure that yield consistency, which we now present. We consider a multivariate parametric change-point regression model with potentially non sub-Gaussian errors. We build on earlier notation but with estimated change-point sets $\hat{\mathcal{T}}_L^O$ and $\hat{\mathcal{T}}_L^E$ (5) now not necessarily estimated by least squares estimation.

We consider a multivariate version of $\mathrm{CV}_{\mathrm{mod}}$:

$$\begin{aligned}\mathrm{CV}_{\mathrm{mod}}(L) := &\sum_{l=0}^{L} \sum_{i=\hat{\tau}_{L,l}^O+1}^{\hat{\tau}_{L,l+1}^O-1} \frac{\hat{\tau}_{L,l+1}^O - \hat{\tau}_{L,l}^O}{\hat{\tau}_{L,l+1}^O - \hat{\tau}_{L,l}^O - 1} \left\| Y_i^E - \overline{Y}^O_{\hat{\tau}_{L,l}^O:\hat{\tau}_{L,l+1}^O} \right\|_2^2 \\ &+ \sum_{l=0}^{L} \sum_{i=\hat{\tau}_{L,l}^E+2}^{\hat{\tau}_{L,l+1}^E} \frac{\hat{\tau}_{L,l+1}^E - \hat{\tau}_{L,l}^E}{\hat{\tau}_{L,l+1}^E - \hat{\tau}_{L,l}^E - 1} \left\| Y_i^O - \overline{Y}^E_{\hat{\tau}_{L,l}^E:\hat{\tau}_{L,l+1}^E} \right\|_2^2,\end{aligned} \tag{19}$$

where here $Y_i^E, Y_i^0 \in \mathbb{R}^d$ for all i. We first establish some additional notation in order to state our result. Let $\Sigma_k := \mathrm{Cov}\,[Y_i]$, for $i = \tau_k + 1, \ldots, \tau_{k+1}$, $k = 0, \ldots, K$ and let $\overline{\sigma}(\Sigma_k)^2$ be the maximum eigenvalue of Σ_k. Further let $\overline{\sigma}^2 := \max_{k=0,\ldots,K} \overline{\sigma}(\Sigma_k)^2$. Finally, for any set of candidate change-points $\mathcal{U} = \{t_0 < t_1 < \cdots < t_K < t_{K+1}\}$ and any collection of vectors $X = (X_1, \ldots, X_{t_{K+1}}) \in \mathbb{R}^{d \times t_{K+1}}$ we use the notation

$$S_X(\mathcal{U}) := \sum_{k=0}^{K} \sum_{i=t_k+1}^{t_{k+1}} \left\| X_i - \overline{X}_{t_k:t_{k+1}} \right\|_2^2. \tag{20}$$

We now introduce assumptions under which consistent estimation of K holds. In addition to Assumption 1, we require the following assumptions for consistent estimation of K. These parallel assumptions required for Zou, Wang and Li (2020, Theorems 1 and 2), but are in places stronger as they guarantee consistency even when K is allowed to increase; see the discussion before and after (66) in Section S5 in the supplementary material.

We require the following uniform Bernstein condition on the errors.

Assumption 2 (Noise). The covariance matrices Σ_k are positive-definite and there exists a constant $c > 0$ such that

$$\limsup_{n\to\infty} \max_{k=0,\ldots,K} \max_{\tau_k < i \le \tau_{k+1}} \mathbb{E}\left[\left\|\Sigma_k^{-1/2}\varepsilon_i\right\|_2^q\right] \le \frac{q!}{2}c^{q-2} \quad \forall\, q \ge 3.$$

The next assumption ensures precise estimation of the change-point locations. In the special case where $K = 0$ for all n, this and Assumption 5 are not required.

Assumption 3 (Estimation precision). Let $Q := K_{\max} - K$. There exists a sequence of matrices $(\delta_{q,k})$ where $q = 0, \ldots, Q$ and $k = 1, \ldots, K$, and a sequence (C_n) with $C_n \to 0$ such that the following are satisfied:

(i)
$$\mathbb{P}\left(\forall\, L = K, \ldots, K_{\max},\ \exists\, \hat\tau^O_{L,i_1}, \ldots, \hat\tau^O_{L,i_K} \in \hat{\mathcal{T}}_L\ :\ \left|\hat\tau^O_{L,i_k} - \tau_k\right| \le \delta_{L-K,k},\ k = 1, \ldots, K\right) \to 1,$$

(ii) $\max_{0\le q\le Q, 1\le k\le K} K \log\log\left(\delta_{q,k} \vee e\right) = o(\log\log\overline{\lambda})$,

(iii) $\sum_{k=1}^K \delta_{0,k}\Delta_k^2 = o(\overline{\sigma}^2 \log\log\overline{\lambda})$,

(iv) for each $k = 1, \ldots, K$, if $\overline{\sigma}^2\log\log\overline{\lambda}/(K\Delta_k^2) \le C_n$, then $\delta_{q,k} = 0$ for all $q = 0, \ldots, Q$,

(v) writing $\mathcal{K}_n := \{k : \overline{\sigma}^2\log\log\overline{\lambda}/(K\Delta_k^2) > C_n\}$, we have that

$$\mathbb{P}\left(\forall\, L < K,\ \forall\, k \in \mathcal{K}_n,\ \sum_{i=\tau_k - \underline{\lambda}/2+1}^{\tau_k + \underline{\lambda}/2} \|\mu_i - \overline{\mu}_{L,i}\|_2^2 \ge C\underline{\lambda}\Delta_k^2 \text{ or } \exists\, \hat\tau \in \hat{\mathcal{T}}_L\ :\ \hat\tau = \tau_k\right) \to 1,$$

as $n \to \infty$, with $C > 0$ a constant and $\overline{\mu}_{L,i} := \sum_{l=0}^L \mathbb{1}_{\{\hat\tau^O_{L,l}+1 \le i \le \hat\tau^O_{L,l+1}\}}\overline{\mu}_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}}$.

Assumption (i) states that the miss-estimation is uniformly bounded by the matrices $(\delta_{q,k})$. The size of those values are limited by (ii)–(iv) to ensure that the influence of miss-estimation on the cross-validation criterion is under control. While (ii) ensures that the errors are generally small, (iii) and (iv) bound the errors of large changes when $q = 0$ and $q \ge 0$, respectively. Part (v) deals with the case where $L < K$ and states that for large changes either the change-point is estimated precisely or the mean estimate is poor. Theorem 5 shows that all of these conditions can be satisfied by least squares estimation as long as changes are large enough; see Section 3.4 and the proof of Theorem 3.

Next, we have to assume that the costs of over-fitting are at least of order $\overline{\sigma}^2\log\log\overline{\lambda}$; see the discussion after Assumption 1. One can show that this is satisfied by least squares estimation; see for instance Lemma 16 in the supplementary material which relies on Zou, Wang and Li (2020, Theorem 2).

Assumption 4 (Over-fitting). For all $\epsilon > 0$,

$$\mathbb{P}\left(\frac{\min_{L=K+1,\ldots,K_{\max}}\left\{S_{\varepsilon^O}\left(\mathcal{T}_K^O\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}_L^O \cup \mathcal{T}_K^O\right)\right\}}{\overline{\sigma}^2 \log\log\overline{\lambda}} < \epsilon\right) \to 0,$$

and as above but with all instances of O replaced by E.

Finally, we assume that all change-points are sufficiently large; see also the discussion following Theorem 3.

Assumption 5 (Minimum Signal). The minimum jump size $\Delta_{(1)}$ satisfies

$$\frac{\underline{\lambda}\Delta_{(1)}^2}{K\overline{\sigma}^2(\log\overline{\lambda})^2} \to \infty, \quad \text{as } n \to \infty.$$

We may now state our general result on consistency of our modified cross-validation $\mathrm{CV}_{\mathrm{mod}}$.

Theorem 4 (Consistency)**.** *Suppose that Assumptions 1–5 hold. Then,*

$$\mathbb{P}\left(\hat{K} = K\right) \to 1, \ \textit{as } n \to \infty.$$

3.4. Least Squares Estimation

In this section we give theoretical guarantees for the change-points estimated by least squares estimation (i.e. the Segment Neighbourhood algorithm) (Auger and Lawrence, 1989). This allows us to verify Assumption 3 for least squares estimation in the case where we have sub-Gaussian noise. More precisely, we assume the setting of Section 3.2, with the exception that the condition $\limsup_{n\to\infty} \sigma/\underline{\sigma} < \infty$ is not required. Similarly to Section 2.1, we denote the unknown set of true change-points by

$$\mathcal{T} := \{0 = \tau_0 < \tau_1 < \cdots < \tau_K < \tau_{K+1} = n\}. \tag{21}$$

Furthermore, for $L \in \mathbb{N}$, we write

$$\hat{\mathcal{T}}_L := \{0 = \hat{\tau}_{L,0} < \hat{\tau}_{L,1} < \cdots < \hat{\tau}_{L,L} < \hat{\tau}_{L,L+1} = n\}. \tag{22}$$

for the set of estimated change-points using least squares estimation, i.e. the output of (3) with Z being Y. Then, we have the following guarantees for the estimated change-point locations.

Theorem 5. *Let $K \le \overline{\lambda}$ eventually and assume that*

$$\frac{\underline{\lambda}\Delta_{(1)}^2}{K\sigma^2\log(\overline{\lambda})} \to \infty, \ \textit{as } n \to \infty. \tag{23}$$

Then, for any sequence of matrices $(\gamma_{L,k})$, $L = 0, \ldots, n-1$, $k = 1, \ldots, K$ *such that*

$$\begin{aligned} \max_{k=1,\ldots,K} \left(\gamma_{K,k}\right)^{-1} (\log(K) \vee 1) \frac{\sigma^2}{\Delta_k^2} &= o(1), \\ \max_{L \neq K} \max_{k=1,\ldots,K} \left(\gamma_{L,k}\right)^{-1} \left(\log\left(K \frac{\sigma^2}{\Delta_k^2} \right) \vee 1 \right) \frac{\sigma^2}{\Delta_k^2} &= o(1), \end{aligned} \tag{24}$$

we have

$$\lim_{n\to\infty} \mathbb{P}\left(\forall\, L \geq K \, \exists\, \hat{\tau}_{L,i_1}, \ldots, \hat{\tau}_{L,i_K} \in \hat{\mathcal{T}}_L \, : \, \left|\hat{\tau}_{L,i_k} - \tau_k\right| \leq \gamma_{L,k}, \;\; k = 1, \ldots, K \right) = 1, \tag{25}$$

and

$$\begin{aligned} \lim_{n\to\infty} \mathbb{P}\Bigg(\forall\, L < K, \, \forall\, k = 1, \ldots, K, \, \sum_{i=\tau_k - \frac{\lambda}{2}+1}^{\tau_k + \frac{\lambda}{2}} (\mu_i - \overline{\mu}_{L,i})^2 \geq \frac{\lambda \Delta_k^2}{100} \;\; \textit{or} \\ \exists\, \hat{\tau} \in \hat{\mathcal{T}}_L : |\hat{\tau} - \tau_k| \leq \gamma_{L,k} \Bigg) = 1, \end{aligned} \tag{26}$$

where $\overline{\mu}_{L,i} := \sum_{l=0}^{L} \mathbb{1}_{\{\hat{\tau}_{L,l}+1 \leq i \leq \hat{\tau}_{L,l+1}\}} \overline{\mu}_{\hat{\tau}_{L,l}:\hat{\tau}_{L,l+1}}$.

Assumption (23) ensures that each change-point is detectable; see the discussion after Theorem 3. The assumption $K \leq \overline{\lambda}$ is rather weak. Moreover, if it is not satisfied, we can obtain a similar statement by replacing $\log(\overline{\lambda})$ by $\log(K)$ in (23). The guarantees obtained for the estimated change-point locations are essential for the proof of Theorem 3. However, while (26) is rather technical, (25) may be of independent interest. Related results about least squares estimation, which is equivalent to maximum likelihood estimation in the Gaussian case, exist in the literature (Yao, 1988; Yao and Au, 1989). However, to the best of our knowledge, no existing result covers the case of $L \neq K$. Moreover, Theorem 5 permits all parameters to vary with n. For finite sample results where $L = K$ (beyond estimating changes in mean in piecewise constant signals with sub-Gaussian noise) see (Garreau and Arlot, 2018, Theorem 3.1) and following discussion. If we only want versions of (25) and (26) to hold for a specific change-point (rather than for all simultaneously), the $\log K$ factors may be dropped in (24).

The rate obtained in (24) is optimal when $L = K$ (Verzelen et al., 2020, Proposition 6). Interestingly, for $L > K$ an additional $\log(\sigma^2/\Delta_k^2)$ factor is required, which we believe to be necessary. The factor may be seen as a consequence of the fact that adding one false positive in a setting when $K = 0$ (i.e. considering $L = 1$) only decreases the least square loss (3) by $O_{\mathbb{P}}(\sigma^2 \log\log n)$, while adding two false positives results in a decrease of $O_{\mathbb{P}}(\sigma^2 \log n)$, see Lemmas 6 and 5, respectively. Hence, placing one incorrect change-point due to noise has a larger contribution when there is at least one false positive, i.e. $L > K$.

3.5. Data-driven choice of $K_{\max}$

For our theoretical results, we have considered a deterministic choice for the maximum number of potential change-points $K_{\max}$ (though this is permitted to increase with n). In practice however, it is helpful to be able to set this in a data-driven way. To do this, we can seek a local minimum of the cross-validation criterion by evaluating $\mathrm{CV}(L)$ for each $L = 0, 1, \ldots$ and stopping when $\mathrm{CV}(L) > \mathrm{CV}(L-1)$, i.e. taking $K_{\max}$ to be the first L where this occurs, and $\hat{K} = K_{\max} - 1$. In order to protect against the fact that such a local minimum may occur largely as a result of the noise, we can instead insist that $\hat{K}$ is further away from the boundary of values at which we have evaluated the criterion. We take such an approach in our numerical experiments, with precise implementation details given in Appendix A.1.

4. Numerical experiments

In this section we study the performance of different methods and tuning parameter selection procedures empirically[2]. We consider a variety of univariate change in mean settings and report for each setting, the proportion of times the number of change-points is underestimated ($\hat{K} < K$), correctly estimated ($\hat{K} = K$) and overestimated ($\hat{K} > K$) over $M = 10\,000$ simulation runs. We also report the mean integrated squared error MISE.

In Section 4.1, we consider a more complicated and perhaps more realistic signal than those described in Examples 1 and 2 in Section 2, but which still demonstrates the sensitivity of vanilla cross-validation to the locations of large change-points. In Section 4.2, we apply our procedures to a simulation setting from Arlot and Celisse (2011). In Section 4.3, we investigate the performance of our new criteria in settings with the famous 'block' (Donoho and Johnstone, 1994) and 'stairs' (Fryzlewicz, 2014) signals, and in Section 4.4 we consider settings with non-Gaussian and non-i.i.d. error distributions. Section S1 in the supplementary material contains the results of further simulations relating to Examples 1 and 2, and a systematic study of the detection power.

We compare our cross-validation approaches $\mathrm{CV}_{(1)}$ (15), $\mathrm{CV}_{\mathrm{mod}}$ (16), and a V-fold version of $\mathrm{CV}_{(1)}$ (see Appendix A and (27) in particular), all used with least squares estimation, to a number of competitors. Note that for V-fold cross-validation with $\mathrm{CV}_{(1)}$ we use our adaptive choice for $K_{\max}$, see Section 3.5 and Appendix A.1 for details. For all other procedures and all competitors that require such a choice we set $K_{\max} = 30$, in order to allow for a more direct comparison with the original `COPSS` procedure. Note that the choice of $K_{\max}$ and whether it is adaptively chosen or not has no significant influence on the results; see also Appendix A.2 for a brief simulation study. For the simulations, we use the functions `CV1`, `CVmod`, and `VfoldCV` from our `R` package `crossvalidationCP`, available on CRAN, which is coded entirely in `R` for maximum flexibility; we note

[2]Code for the simulations is available at https://github.com/FlorianPein/SimulationsCrossvalidationCP

that if a fast implementation for a specialised setting is required, compiled code will be beneficial. We do however use the function `Fpsn` from the package `fpopw`, which is a fast implementation of the Segment Neighbourhood algorithm with functional pruning (see Rigaill (2015); Maidstone et al. (2017)), for the runtime intensive calculation of least squares estimation for $L = 0, \ldots, K_{\max}$ change-points.

The list of competitors includes the COPSS procedure from Zou, Wang and Li (2020) described in Section 2.1 (we use our own implementation given by the function `COPSS` in the `R` package `crossvalidationCP`) and the two stage V-fold cross-validation procedure LooVF from Arlot and Celisse (2011)[3]; note that due to the slower run-time of the latter, we only include it in our smaller-scale simulations. We also compare to classical change-point procedures such as PELT (Killick, Fearnhead and Eckley, 2012), with the SIC-penalty implemented in the R-package `change-point`; wild binary segmentation (WBS) (Fryzlewicz, 2014), with penalty 1.3 times the SIC-penalty implemented in the `R` package `wbs`; FDRSeg (Li, Munk and Sieling, 2016) with $\alpha = 0.9$ as implemented by the `R` package `FDRSeg`; and Ms. FPOP (Liehrmann and Rigaill, 2023; Verzelen et al., 2020), with recommend parameters $\alpha = 9 + 2.25\log(n)$ and $\beta = 2.25$. We also included the robust methods from Fearnhead and Rigaill (2019), implemented in the R-package `robseg`. We use Biweight loss (which outperformed the other available option of Huber loss in all simulations) with default parameters $\lambda = 2\log(n)$ and threshold 3. These provide genuinely different estimation methods to least squares, and so a comparison with the cross-validation approaches using the latter loss function should be interpreted with care. We also experimented with HSMUCE (Pein, Sieling and Munk, 2017); however the strong error control it provides meant that it lacked sufficient power to detect change-points so we do not present these results.

4.1. Sensitivity to locations of large changes

Here we consider a setting with $n = 2048$ observations and $K = 11$ change-points. The signal, shown in Figure 2, has change-points at 204, 470, 778, 878, 883, 894, 984, 1414, 1638, 1680, 1740 and function values −2.32, 15.98, 5, 20, 0, 70, 0, −15, −7.32, 8.42, −2.93, 4.76. To form the observations, we add to the signal $\mathcal{N}(0, \sigma^2)$ errors with $\sigma = 7$. We also consider a modified signal where the change-point at 883 is shifted to 884, an even location.

We see that COPSS performs poorly compared to PELT and other change-point procedures when the large change is at an odd location, both in terms of estimation of K and MISE; in contrast, the new cross-validation approaches perform well in both settings, with 5-fold $\mathrm{CV}_{(1)}$ the best among these. This is mirrored in the additional results presented in Sections S1.1 and S1.2 in the supplementary material.

[3]Matlab implementation available at https://www.imo.universite-paris-saclay.fr/ sylvain.arlot/code/CHPTCV.htm

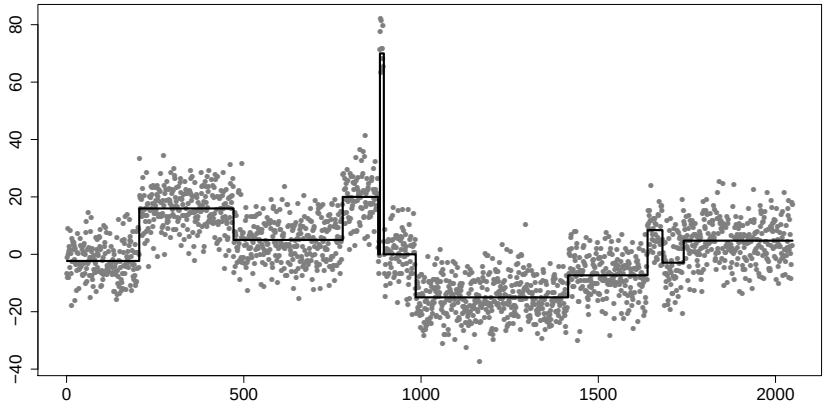

FIG 2. *Signal with* $n = 2048$ *observations and* $K = 11$ *change-points.*

	Original signal				Modified signal			
Method	$\hat{K} < K$	$\hat{K} = K$	$\hat{K} > K$	MISE	$\hat{K} < K$	$\hat{K} = K$	$\hat{K} > K$	MISE
5-fold $\mathrm{CV}_{(1)}$	3.56	81.15	15.29	0.9061	3.46	81.14	15.4	0.9011
$\mathrm{CV}_{(1)}$	16.27	73.49	10.24	1.004	13.28	75.54	11.18	0.9811
$\mathrm{CV}_{\mathrm{mod}}$	15.61	74.94	9.45	1.006	15.55	75.07	9.38	1.003
COPSS	60.7	33.5	5.8	1.493	12.84	74.78	12.38	0.9794
PELT	0.91	95.9	3.19	0.8369	0.78	95.83	3.39	0.831
WBS	4.7	40.65	54.65	1.14	4.73	39.95	55.32	1.135
FDRSeg	1.05	74.22	24.73	0.9133	0.78	75.17	24.05	0.9086
Ms. FPOP	6.62	92.83	0.55	0.8754	6.04	93.47	0.49	0.865
Biweight	2.69	94.65	2.66	0.8935	2.37	94.9	2.73	0.8888

TABLE 1

Results of simulations with the signal as in Figure 2. COPSS *systematically underestimates the number of change-points in the setting with the original signal, while the suggested alternatives perform well.*

4.2. Small sample size

In this section we consider a simulation setting of Arlot and Celisse (2011) involving their s_1 signal (see Fig. 2 in their paper). This has four change-points with each constant segment containing 20 time points and jumps of size one in alternating directions. We consider the standard deviation functions $\sigma_c = 0.25$ (constant), $\sigma_{pc,3} = 0.6\mathbb{1}_{[0,1/3)}(t) + 0.15\mathbb{1}_{[1/3,1]}(t)$ (piecewise constant), and $\sigma_s = 0.5\sin(t\pi/4)$.

We see that the proposed cross-validation criteria perform well, and in particular dominate LooVF here. PELT, WBS, Ms. FPOP and Biweight all perform well in the constant variance setting of σ_c, but their performances greatly deteriorate in the heteroscedastic settings. In contrast, our proposed $\mathrm{CV}_{\mathrm{mod}}$ is competitive in the idealised homoscedastic setting, and is the best performer in the heteroscedastic settings.

	σ_c				$\sigma_{pc,3}$			
Method	$\hat{K} < K$	$\hat{K} = K$	$\hat{K} > K$	MISE	$\hat{K} < K$	$\hat{K} = K$	$\hat{K} > K$	MISE
2-fold $\text{CV}_{(1)}$	0	83.5	16.5	0.00613	0.34	67.74	31.92	0.02539
5-fold $\text{CV}_{(1)}$	0	77.74	22.26	0.006659	0.2	70.66	29.14	0.02445
10-fold $\text{CV}_{(1)}$	0	73.41	26.59	0.007089	0.13	68.61	31.26	0.02479
20-fold $\text{CV}_{(1)}$	0	69.42	30.58	0.007407	0.16	66.76	33.08	0.02529
$\text{CV}_{(1)}$	0	84.26	15.74	0.006199	0.21	66.97	32.82	0.02725
CV_{mod}	0	90.41	9.59	0.005648	2.87	81.36	15.77	0.02342
COPSS	0	82.3	17.7	0.006193	0.76	71.6	27.64	0.02531
LooVF_2	0	74.03	25.97	0.006866	1.26	65.33	33.41	0.02598
LooVF_5	0	69.47	30.53	0.007342	0.6	68.66	30.74	0.0243
PELT	0	87.11	12.89	0.006028	0	0.36	99.64	0.08918
WBS	0	90.95	9.05	0.005569	0	0.13	99.87	0.08213
FDRSeg	0	80.51	19.49	0.006702	0	0.02	99.98	0.1032
Ms. FPOP	0	99.25	0.75	0.005122	0	12.32	87.68	0.05875
Biweight	0	89.3	10.7	0.006057	0	12.35	87.65	0.0559

TABLE 2

Results for the s_1 signal of Arlot and Celisse (2011) with various standard deviation functions.

	σ_s			
Method	$\hat{K} < K$	$\hat{K} = K$	$\hat{K} > K$	MISE
2-fold $\text{CV}_{(1)}$	0	80.41	19.59	0.005233
5-fold $\text{CV}_{(1)}$	0	76.49	23.51	0.005743
10-fold $\text{CV}_{(1)}$	0	72.9	27.1	0.006056
20-fold $\text{CV}_{(1)}$	0	70.73	29.27	0.00623
$\text{CV}_{(1)}$	0	77.57	22.43	0.005724
CV_{mod}	0	87.77	12.23	0.004541
COPSS	0	78.35	21.65	0.00549
LooVF_2	0	72.75	27.25	0.005826
LooVF_5	0	70.4	29.6	0.006228
PELT	0	8.68	91.32	0.01827
WBS	0	5.24	94.76	0.01726
FDRSeg	0	2.76	97.24	0.02421
Ms. FPOP	0	49.7	50.3	0.009256
Biweight	0	24.41	75.59	0.0137

TABLE 3

Continuation of Table 2

4.3. Blocks and stairs signal

We consider the famous block signal with $K = 11$ change-points. As in Zou, Wang and Li (2020) we choose 2048 observations, $\mathcal{N}(0, 7^2)$ errors as before, and set change-points at 205, 267, 308, 472, 512, 820, 902, 1332, 1557, 1598, 1659 with corresponding function values 0, 14.64, −3.66, 7.32, −7.32, 10.98, −4.39, 3.29, 19.03, 7.68, 15.37, 0; see Figure 3. Secondly, we consider the stairs example from Fryzlewicz (2014) with $n = 150$, errors distributed as $\mathcal{N}(0, 0.3^2)$, and a change-point with jump size 1 every ten observations, so $K = 14$.

In these settings with well-specified errors, we might expect classical change-point procedures to have a noticeable advantage. However, from Table 4 we see that cross-validation remains quite competitive.

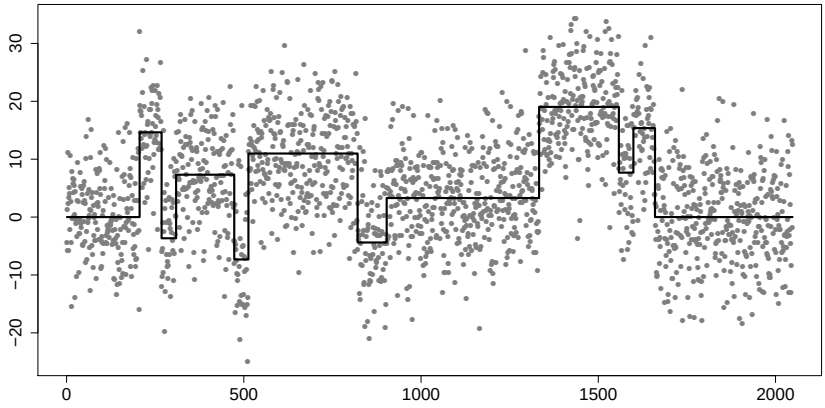

FIG 3. *Blocks signal with* $K = 11$ *change-points.*

	Blocks signal				Stairs signal			
Method	$\hat{K} < K$	$\hat{K} = K$	$\hat{K} > K$	MISE	$\hat{K} < K$	$\hat{K} = K$	$\hat{K} > K$	MISE
5-fold $\text{CV}_{(1)}$	7.78	76.46	15.76	1.047	0.27	75.57	24.16	0.02192
$\text{CV}_{(1)}$	23.13	66.59	10.28	1.109	0.53	67.64	31.83	0.02275
CV_{mod}	21.76	69.15	9.09	1.099	8.63	82.21	9.16	0.02377
COPSS	24.5	66.27	9.23	1.114	0.54	64.93	34.53	0.02296
PELT	3.76	92.9	3.34	0.9741	0.42	94	5.58	0.02066
WBS	25.66	72.44	1.9	1.205	6.14	86.05	7.81	0.02778
FDRSeg	2.51	75.16	22.33	1.046	1.43	82.28	16.29	0.02231
Ms. FPOP	10.08	89.18	0.74	0.994	9.74	90.2	0.06	0.02326
Biweight	4.39	92.7	2.91	1.006	0.47	95.5	4.03	0.021

TABLE 4

Results for the blocks, see Figure 3, and stairs signal.

4.4. Robustness

In this section, we explore the robustness of cross-validation to a misspecified models. We continue to use the blocks signal (Figure 3) but with various violations of the change-point model in Section 2 as detailed below.

Misspecified error distribution (Table 5) We consider t-distributed errors with 5 degrees of freedom and exponentially distributed errors with mean 1 which we then mean-centre. All errors are standardised such that the standard deviation is $\sigma = 7$.

Heteroscedastic errors (Table 6) We consider two settings with heteroscedastic Gaussian errors: the first where the errors have different standard deviations on each piecewise constant segment, and the second where the standard deviation instead changes after each block of 32 observations. The standard deviations are drawn independently from $U[0, 8]$.

Outliers (Table 7) We use the same setting as that of Section 4.3 but randomly sample ten observations and add a Poisson distributed random variable with intensity $\lambda \in \{20, 30\}$.

Method	$\hat{K} < K$	$\hat{K} = K$	$\hat{K} > K$	MISE	$\hat{K} < K$	$\hat{K} = K$	$\hat{K} > K$	MISE
	t_5 error distribution				exponential error distribution			
5-fold $\text{CV}_{(1)}$	17.4	58.08	24.52	1.722	27.51	50.73	21.76	1.488
$\text{CV}_{(1)}$	28.87	35.62	35.51	2.153	40.87	27.13	32	1.99
CV_{mod}	72.8	26.71	0.49	3.565	83.77	15.88	0.35	2.893
COPSS	48.56	35.54	15.9	2.152	65.09	25.3	9.61	1.945
PELT	0.02	3.24	96.74	4.261	0	0	100	11.61
WBS	0.75	6.65	92.6	3.324	0	0	100	7.304
FDRSeg	0	0.12	99.88	6.265	0	0	100	18.81
Ms. FPOP	0.84	22.98	76.18	2.873	0.01	0.03	99.96	6.978
Biweight	1.6	93.11	5.29	0.8514	0.13	75.1	24.77	1.898

TABLE 5
Results with misspecified error distributions.

Method	$\hat{K} < K$	$\hat{K} = K$	$\hat{K} > K$	MISE	$\hat{K} < K$	$\hat{K} = K$	$\hat{K} > K$	MISE
5-fold $\text{CV}_{(1)}$	0.88	80.11	19.01	0.4409	0.2	81.66	18.14	0.4148
$\text{CV}_{(1)}$	3.72	80.98	15.3	0.4408	1.95	84.44	13.61	0.4063
CV_{mod}	6.51	82.22	11.27	0.4454	4.17	85.69	10.14	0.4074
COPSS	7.56	80.79	11.65	0.4531	4.9	84.3	10.8	0.4128
PELT	0.01	3.75	96.24	4.8	0	0.01	99.99	5.034
WBS	0	5.01	94.99	2.628	0	0.01	99.99	2.764
FDRSeg	0	0.24	99.76	8.905	0	0	100	9.567
Ms. FPOP	0.01	12.82	87.17	2.881	0	1.24	98.76	2.515
Biweight	0.02	9.62	90.36	2.41	0	1.98	98.02	2.04

TABLE 6
Results with heteroscedastic errors.

Method	$\hat{K} < K$	$\hat{K} = K$	$\hat{K} > K$	MISE	$\hat{K} < K$	$\hat{K} = K$	$\hat{K} > K$	MISE
	$\lambda = 20$				$\lambda = 30$			
5-fold $\text{CV}_{(1)}$	10.21	77.51	12.28	1.101	18.86	71	10.14	1.264
$\text{CV}_{(1)}$	27.59	62.76	9.65	1.204	40.52	41.85	17.63	1.573
CV_{mod}	30.88	64.32	4.8	1.206	64.02	35.29	0.69	1.816
COPSS	31.42	62.07	6.51	1.208	54.54	39.62	5.84	1.572
PELT	3.41	76.94	19.65	1.17	0.74	14.33	84.93	2.73
WBS	22.85	61.79	15.36	1.364	7.19	18.13	74.68	2.372
FDRSeg	1.54	41.13	57.33	1.505	0.08	1.62	98.3	3.937
Ms. FPOP	10.81	87.2	1.99	1.06	7.96	59.39	32.65	1.575
Biweight	5	92.13	2.87	1.034	5.33	91.83	2.84	1.031

TABLE 7
Results with 10 *Poisson-distributed outliers.*

The results above indicate that cross-validation and change-point regression with Biweight loss are more robust to model misspecifications than classical approaches designed for homoscedastic Gaussian errors. Usage of Biweight loss outperforms cross-validation in some scenarios, but we also see that cross-validation performs reasonably well without any knowledge of the type of violation to be expected, and without sacrificing much performance in idealised settings. We

find that 5-fold $\mathrm{CV}_{(1)}$ performs better than other cross-validation approaches for misspecified error distributions, but $\mathrm{CV}_{\mathrm{mod}}$ outperforms this slightly when the errors are heteroscedastic.

5. Discussion

In sharp contrast to its ubiquity in high-dimensional and non-parametric regression, cross-validation has received little attention and use in change-point problems. There is good reason for this: as we show in this work, standard cross-validation with squared error loss may not correctly estimate the number of change-points in settings where all changes are easily detectable, and can yield an estimated regression function that has an integrated squared error that is orders of magnitude larger than achievable by other methods. On the other hand, there may be much to be gained from deeper investigation of the use of cross-validation in change-point problems. We propose two simple approaches to remedy these deficiencies and show empirically that they perform well in settings with Gaussian errors, and are relatively robust to heavy-tailed, non-symmetric and heteroscedastic errors.

We expect that cross-validation-type approaches for tuning parameter selection may be even more successful in more complex change-point settings, for example those involving piecewise smooth mean functions, in part due to the fact that no prior estimate of the noise variance is required. As well as exploring such settings, it would also be of interest to develop theory for the absolute error approach similar to that which we present for our modified squared error criterion. It may also be fruitful to develop alternatives to cross-validation that are also model agnostic, for example based on the bootstrap (Antoch, Hušková and Veraverbeke, 1995; Hušková and Kirch, 2008; Sharipov, Tewes and Wendler, 2016).

Appendix A: Generalised procedure

In this section we consider the setting of Section 3.3 and describe a more general V-fold cross-validation procedure that selects a tuning parameter ψ of an arbitrary change-point estimation procedure $\mathcal{A}(\psi, Y_1, \ldots, Y_n)$. The estimation procedure $\mathcal{A}$ requires a tuning parameter ψ and a vector of observations $(Y_1, \ldots, Y_n)$ and returns the estimated change-point locations $0 = \hat{\tau}_{\psi,0} < \hat{\tau}_{\psi,1} < \cdots < \hat{\tau}_{\psi,\hat{K}_\psi} < \hat{\tau}_{\psi,\hat{K}_\psi+1} = n$ and parameter estimates $\hat{\beta}_{\psi,0} \neq \hat{\beta}_{\psi,1} \neq \cdots \neq \hat{\beta}_{\psi,\hat{K}_\psi}$. The tuning parameter may for instance be the number of change-points as in the case of least squares estimation / Segment Neighbourhood (Auger and Lawrence, 1989).

Let Ψ be a set of potential tuning parameters and define fold $F_v := \{v+i\cdot V,\ i \in \mathbb{N} : v + i \cdot V \leq n\}$ for $v = 1, \ldots, V$. We then select $\psi \in \Psi$ as follows.

For each fold F_v and every tuning parameter $\psi \in \Psi$, the procedure $\mathcal{A}$ is applied using all observations with indices not in F_v. We denote the estimated change-points by $0 = \hat{\tau}^{-F_v}_{\psi,0} < \hat{\tau}^{-F_v}_{\psi,1} < \cdots < \hat{\tau}^{-F_v}_{\psi,\hat{K}^{-F_v}_\psi} < \hat{\tau}^{-F_v}_{\psi,\hat{K}^{-F_v}_\psi+1} = n$ and the

parameters by $\hat{\beta}^{-F_v}_{\psi,0}, \ldots, \hat{\beta}^{-F_v}_{\psi,\hat{K}^{-F_v}_{\psi}}$. Note that we ask the estimated change-points to be a subset of $\{1, \ldots, n\}$, i.e. if $F^c_v = \{i_1, \ldots, i_m\}$ and $\mathcal{A}$ returns change-points $0 = \tilde{\tau}_{\psi,0} < \tilde{\tau}_{\psi,1} < \cdots < \tilde{\tau}_{\psi,\hat{K}^{-F_v}_{\psi}} < \tilde{\tau}_{\psi,\hat{K}^{-F_v}_{\psi}+1} = m$, we set $\hat{\tau}^{-F_v}_{\psi,l} := i_{\tilde{\tau}_{\psi,l}}$, $l = 1, \ldots, \hat{K}^{-F_v}_{\psi}$.

To evaluate the quality of the estimates, we use extended versions of the criteria we have proposed in Section 3. We extend cross-validation with absolute error loss (15) to

$$\mathrm{CV}^V_{(1)}(\psi) := \sum_{v=1}^{V} \sum_{k=0}^{\hat{K}^{-F_v}_{\psi}} \sum_{\substack{i \in F_v, \\ \hat{\tau}^{-F_v}_{\psi,k} < i \leq \hat{\tau}^{-F_v}_{\psi,k+1}}} \left\| Y_i - \hat{\beta}^{-F_v}_{\psi,k} \right\|_2. \tag{27}$$

Finally, the tuning parameter $\psi \in \Psi$ that minimizes cross-validation criterion is selected, i.e.

$$\hat{\psi} := \operatorname*{argmin}_{\psi \in \Psi} \mathrm{CV}^V_{(1)}(\psi). \tag{28}$$

In the case where $\mathcal{A}$ is least squares estimation, but also for many other estimators, a natural choice for $\hat{\beta}^{-F_v}_{\psi,k}$ is

$$\overline{Y}^{-F_v}_{\psi,k} := \left| \left\{ i \notin F_v;\ \hat{\tau}^{-F_v}_{\psi,k} < i \leq \hat{\tau}^{-F_v}_{\psi,k+1} \right\} \right|^{-1} \sum_{\substack{i \notin F_v, \\ \hat{\tau}^{-F_v}_{\psi,k} < i \leq \hat{\tau}^{-F_v}_{\psi,k+1}}} Y_i. \tag{29}$$

Alternatively, a slight generalisation of (16) and (19) gives

$$\mathrm{CV}^V_{\mathrm{mod}}(\psi) := \sum_{v=1}^{V} \sum_{k=0}^{\hat{K}^{-F_v}_{\psi}} \sum_{\substack{i \in F_v, \\ \hat{\tau}^{-F_v}_{\psi,k}+1 < i \leq \hat{\tau}^{-F_v}_{\psi,k+1}}} \frac{\hat{\tau}^{-F_v}_{\psi,k+1} - \hat{\tau}^{-F_v}_{\psi,k}}{\hat{\tau}^{-F_v}_{\psi,k+1} - \hat{\tau}^{-F_v}_{\psi,k} - 1} \left\| Y_i - \hat{\beta}^{-F_v}_{\psi,k} \right\|_2^2. \tag{30}$$

Note that for this criterion we require that each segment is at least of length $2(V-1)$, i.e. $\hat{\tau}^{-F_v}_{\psi,k+1} - \hat{\tau}^{-F_v}_{\psi,k} \geq 2(V-1)$.

A.1. Adaptive choice of $K_{\max}$

As discussed in Section 3.5, in practice, rather than fixing $K_{\max}$ in advance, we can choose $K_{\max}$ in a data-driven way so as to ensure that $\hat{K}$ is not too close to $K_{\max}$. To implement this, since the function `Fpsn` from the package `fpopw`, which we use to calculate the least squares estimator, only allows one to perform calculations for $L = 1, 2, \ldots, K_{\max}$ for any given $K_{\max}$, we proceed as follows. We start with $K_{\max} = 8$. If $\hat{K} < K_{\max} - 3$, we stop and return $\hat{K}$. Otherwise, we double $K_{\max}$ and rerun the procedure until either our stopping criterium is satisfied or in extreme cases $K_{\max} = n/2$. The choices to start at $K_{\max} = 8$ and to subtract 3 do not follow any specific considerations and results does not depend much on it as we see in the following section.

A.2. Influence of $K_{\max}$ in simulations

In this simulation study we compare different starting values for $K_{\max}$ for the adaptive procedure and also use the non-adaptive procedure where we fixed $K_{\max} = 30$. We use the blocks setting from Section 4.3 with 5-fold $\mathrm{CV}_{(1)}$.

Method	$\hat{K} < K$	$\hat{K} = K$	$\hat{K} > K$	MISE
$K_{\max} = 30$, fixed	7.47	77.26	15.27	1.041
$K_{\max} = 5$, adaptive	7.47	77.26	15.27	1.041
$K_{\max} = 6$, adaptive	7.47	77.26	15.27	1.041
$K_{\max} = 7$, adaptive	7.49	77.26	15.25	1.041
$K_{\max} = 8$, adaptive	7.47	77.31	15.22	1.04
$K_{\max} = 9$, adaptive	7.47	77.26	15.27	1.041
$K_{\max} = 10$, adaptive	7.47	77.26	15.27	1.041
$K_{\max} = 11$, adaptive	7.47	77.26	15.27	1.041
$K_{\max} = 12$, adaptive	7.47	77.26	15.27	1.041
$K_{\max} = 13$, adaptive	7.47	77.26	15.27	1.041
$K_{\max} = 14$, adaptive	7.49	77.26	15.25	1.041

TABLE 8
Different choices for $K_{\max}$.

We see from Table 8 that results are nearly the same for all choices. We obtained similar results when we tried different choices for a fixed $K_{\max}$ (not displayed) as long as the number was a bit larger than the true number of change-points. This confirms our reasoning for the design of the adaptive procedure.

Acknowledgments

The authors would like to thank the associate editor and two anonymous referees for their valuable feedback that helped us to improve the manuscript.

Funding

The authors were supported by EPSRC grant EP/N031938/1.

Supplementary Material

The supplementary material contains additional simulations and all proofs.

References

ANTOCH, J., HUŠKOVÁ, M. and VERAVERBEKE, N. (1995). Change-point problem and bootstrap. *Journaltitle of Nonparametric Statistics* **5** 123–144.

ARLOT, S. and CELISSE, A. (2010). A survey of cross-validation procedures for model selection. *Stat. Surv.* **4** 40–79.

Arlot, S. and Celisse, A. (2011). Segmentation of the mean of heteroscedastic data via cross-validation. *Stat. Comput.* **21** 613–632.

Auger, I. E. and Lawrence, C. E. (1989). Algorithms for the optimal identification of segment neighborhoods. *B. Math. Biol.* **51** 39–54.

Bai, J. and Perron, P. (2003). Computation and analysis of multiple structural change models. *J. Appl. Econ.* **18** 1–22.

Chan, H. P. and Walther, G. (2013). Detection with the scan and the average likelihood ratio. *Stat. Sin.* 409–428.

Chetverikov, D., Liao, Z. and Chernozhukov, V. (2021). On cross-validated lasso in high dimensions. *Ann. Stat.* **49** 1300–1317.

D'Angelo, M., Palhares, R. M., Takahashi, R., Loschi, R. H., Baccarini, L. and Caminhas, W. (2011). Incipient fault detection in induction machine stator-winding using a fuzzy-Bayesian change point detection approach. *Appl. Soft. Comput.* **11** 179–192.

Donoho, D. L. and Johnstone, J. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455.

Du, C., Kao, C. L. M. and Kou, S. C. (2016). Stepwise signal extraction via marginal likelihood. *J. Am. Stat. Assoc.* **111** 314–330.

Dümbgen, L. and Spokoiny, V. G. (2001). Multiscale testing of qualitative hypotheses. *Ann. Stat.* 124–152.

Fearnhead, P. (2006). Exact and efficient Bayesian inference for multiple changepoint problems. *Stat. Comput.* **16** 203–213.

Fearnhead, P. and Rigaill, G. (2019). Changepoint detection in the presence of outliers. *J. Am. Stat. Assoc.* **114** 169–183.

Fearnhead, P. and Rigaill, G. (2020). Relating and comparing methods for detecting changes in mean. *Stat* **9** e291.

Frick, K., Munk, A. and Sieling, H. (2014). Multiscale change point inference. *J. R. Stat. Soc., B: Stat. Methodol.* 495–580.

Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *Ann. Stat.* **42** 2243–2281.

Fryzlewicz, P. (2020). Detecting possibly frequent change-points: Wild Binary Segmentation 2 and steepest-drop model selection. *J. Korean Stat. Soc.* **49** 1027–1070.

Garreau, D. and Arlot, S. (2018). Consistent change-point detection with kernels. *Electronic Journal of Statistics* **12** 4440 – 4486.

Harchaoui, Z., Vallet, F., Lung-Yut-Fong, A. and Cappé, O. (2009). A regularized kernel-based approach to unsupervised audio segmentation. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing* 1665–1668. IEEE.

Hušková, M. and Kirch, C. (2008). Bootstrapping confidence intervals for the change-point of time series. *Journal of Time Series Analysis* **29** 947–972.

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

Killick, R., Fearnhead, P. and Eckley, I. A. (2012). Optimal detection

of changepoints with a linear computational cost. *J. Am. Stat. Assoc.* **107** 1590–1598.

Kim, C. J., Morley, J. C. and Nelson, C. R. (2005). The structural break in the equity premium. *J. Bus. Econ. Stat.* **23** 181–191.

Kovács, S., Li, H., Bühlmann, P. and Munk, A. (2020). Seeded Binary Segmentation: A general methodology for fast and optimal change point detection. *arXiv preprint arXiv:2002.06633*.

Li, H., Guo, Q. and Munk, A. (2019). Multiscale change-point segmentation: Beyond step functions. *Electron. J. Stat.* **13** 3254–3296.

Li, H., Munk, A. and Sieling, H. (2016). FDR-control in multiscale change-point segmentation. *Electron. J. Stat.* **10** 918–959.

Li, J., Fearnhead, P., Fryzlewicz, P. and Wang, T. (2022). Automatic change-point detection in time series via deep learning. *arXiv preprint arXiv:2211.03860*.

Liehrmann, A. and Rigaill, G. (2023). Ms. FPOP: An Exact and Fast Segmentation Algorithm With a Multiscale Penalty. *arXiv preprint arXiv:2303.08723*.

Lin, K., Sharpnack, J., Rinaldo, A. and Tibshirani, R. J. (2016). Approximate Recovery in Changepoint Problems, from l_2 Estimation Error Rates. *arXiv preprint arXiv:1606.06746*.

Maidstone, R., Hocking, T., Rigaill, G. and Fearnhead, P. (2017). On optimal multiple changepoint algorithms for large data. *Stat. Comput.* **27** 519–533.

Niu, Y. S., Hao, N. and Zhang, H. (2016). Multiple change-point detection: A selective overview. *Stat. Sci.* 611–623.

Olshen, A. B., Venkatraman, E., Lucito, R. and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5** 557–572.

Pein, F., Eltzner, B. and Munk, A. (2021). Analysis of patchclamp recordings: model-free multiscale methods and software. *Eur. Biophys. J.* **50** 187–209.

Pein, F., Sieling, H. and Munk, A. (2017). Heterogeneous change point inference. *J. R. Stat. Soc., B: Stat. Methodol.* **79** 1207–1227.

Rigaill, G. (2015). A pruned dynamic programming algorithm to recover the best segmentations with 1 to $K\_\{max\}$ change-points. *Journal de la Société Française de Statistique* **156** 180–205.

Sharipov, O., Tewes, J. and Wendler, M. (2016). Sequential block bootstrap in a Hilbert space with application to change point analysis. *Canadian Journal of Statistics* **44** 300–322.

Truong, C., Oudre, L. and Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Process.* **167** 107299.

Verzelen, N., Fromont, M., Lerasle, M. and Reynaud-Bouret, P. (2020). Optimal Change-Point Detection and Localization. *arXiv preprint arXiv:2010.11470*.

Vostrikova, L. Y. (1981). Detecting "disorder" in multidimensional random processes. *Dokl. Akad. Nauk* **259** 270–274.

Wong, W. H. (1983). On the consistency of cross-validation in kernel nonparametric regression. *Ann. Stat.* **11** 1136–1141.

Yang, Y. (2007). Consistency of cross validation for comparing regression procedures. *Ann. Stat.* **35** 2450–2473.

Yao, Y. C. (1988). Estimating the number of change-points via Schwarz' criterion. *Stat. Probab. Lett.* **6** 181–189.

Yao, Y. C. and Au, S. T. (1989). Least-squares estimation of a step function. *Sankhya: Indian J. Stat.* 370–381.

Yu, Y. and Feng, Y. (2014). Modified cross-validation for penalized high-dimensional linear regression models. *J. Comput. Graph. Stat.* **23** 1009–1027.

Zhang, N. R. and Siegmund, D. O. (2007). A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics* **63** 22–32.

Zou, C., Wang, G. and Li, R. (2020). Consistent selection of the number of change-points via sample-splitting. *Ann. Stat.* **48** 413–439.

Supplementary material to Cross-validation for change-point regression: pitfalls and solutions

In the following we collect additional simulations and all of our proofs. First of all, in Section S1 we present further simulation settings. Section S2 begins with the proof of Theorem 5 as the strategy used to split the least squares objective will be helpful in other proofs as well. Theorems 1, 2 and 4 are proved in Sections S3, S4 and S5 respectively. Finally, Section S6 gives a proof of Theorem 3, which largely follows from Theorems 4 and 5 and is hence shown last.

S1. Additional simulations

In this section we include further simulation results. The methods under consideration include V-fold $\mathrm{CV}_{(1)}$ procedures as well as those mentioned in Section 4, with leave-one-out cross-validation denoted by LOOCV $\mathrm{CV}_{(1)}$.

S1.1. Underestimation example

Observations are as in Example 1. We take $n = 202$, $\Delta_1 = 10$, $\sigma = 1$, and $\lambda = 5$. We vary $\Delta_2 = D\Delta_1$ as factor of Δ_1, with $D \in \{2, 3, 5\}$. For all D the same set of seeds was used.

	$D = 5$				$D = 3$			
Method	$\hat{K} < K$	$\hat{K} = K$	$\hat{K} > K$	MISE	$\hat{K} < K$	$\hat{K} = K$	$\hat{K} > K$	MISE
2-fold $\mathrm{CV}_{(1)}$	0	86.98	13.02	0.02159	0	86.98	13.02	0.02159
5-fold $\mathrm{CV}_{(1)}$	0	79.39	20.61	0.02747	0	79.39	20.61	0.02747
10-fold $\mathrm{CV}_{(1)}$	0	73.25	26.75	0.03249	0	73.25	26.75	0.03249
20-fold $\mathrm{CV}_{(1)}$	0	67.72	32.28	0.0374	0	67.72	32.28	0.0374
$\mathrm{CV}_{(1)}$	0	87.18	12.82	0.02183	0	87.18	12.82	0.02183
$\mathrm{CV}_{\mathrm{mod}}$	0	92.03	7.97	0.0191	0	92.03	7.97	0.0191
COPSS	99.95	0	0.05	2.361	91.28	0.06	8.66	2.166
LooVF_2	100	0	0	2.362	99.87	0.08	0.05	2.359
LooVF_5	100	0	0	2.362	99.81	0.09	0.1	2.358
PELT	0	91.34	8.66	0.02118	0	91.34	8.66	0.02118
WBS	0	87.38	12.62	0.0204	0	87.38	12.62	0.0204
FDRSeg	0	82.19	17.81	0.02409	0	82.19	17.81	0.02409
Ms. FPOP	0	98.89	1.11	0.01549	0	98.89	1.11	0.01549
Biweight	0	93.21	6.79	0.05512	0	93.21	6.79	0.03418

TABLE 9

Simulation results relating to Example 1. LooVF_2 *and* LooVF_5 *were applied to the observations in reverse order.*

The results in Tables 9 and 10 support our theoretical findings from Section 2.2 that cross-validation with least squares loss (COPSS and LooVF_v) underestimates the number of change-points when $2\Delta_2 > \underline{\lambda}\Delta_1$ in Example 1 (see Theorem 1). Moreover these methods also have very large MISE. Note that the results above for LooVF were obtained by applying the method to

	$D = 2$			
Method	$\hat{K} < K$	$\hat{K} = K$	$\hat{K} > K$	MISE
2-fold $\text{CV}_{(1)}$	0	86.98	13.02	0.02159
5-fold $\text{CV}_{(1)}$	0	79.39	20.61	0.02747
10-fold $\text{CV}_{(1)}$	0	73.25	26.75	0.03249
20-fold $\text{CV}_{(1)}$	0	67.72	32.28	0.0374
$\text{CV}_{(1)}$	0	87.18	12.82	0.02183
CV_{mod}	0	92.03	7.97	0.0191
COPSS	0.47	69.69	29.84	0.04654
LooVF_2	50.14	40.25	9.61	1.196
LooVF_5	9.43	63.72	26.85	0.252
PELT	0	91.34	8.66	0.02118
WBS	0	87.38	12.62	0.0204
FDRSeg	0	82.19	17.81	0.02409
Ms. FPOP	0	98.89	1.11	0.01549
Biweight	0	93.21	6.79	0.02723

TABLE 10

Simulation results relating to Example 1. LooVF_2 *and* LooVF_5 *were applied to the observations in reverse order.*

the observations in reverse order, which as explained in Example 1, is where the issue of LooVF occurs here. Other approaches do not underestimate the number of change-points and our new cross-validation approaches (particularly CV_{mod} here) are competitive with these. With a moderate number of folds, our new cross-validation approaches are competitive with classical change-point approaches.

S1.2. Overestimation example

Consider observations as in Example 2. We choose $n = 202$, $\Delta_1 = 1$ and vary σ. Moreover, in one simulation the change-point will be at $\tau_1 = n/2 + 1$, an even location.

	$\sigma = 1$				$\sigma = 0.1$			
Method	$\hat{K} < K$	$\hat{K} = K$	$\hat{K} > K$	MISE	$\hat{K} < K$	$\hat{K} = K$	$\hat{K} > K$	MISE
2-fold $\text{CV}_{(1)}$	38.54	53.03	8.43	0.194	0	88.13	11.87	0.0001636
5-fold $\text{CV}_{(1)}$	25.21	58.53	16.26	0.1931	0	79.6	20.4	0.0002293
10-fold $\text{CV}_{(1)}$	21.37	56.51	22.12	0.2055	0	73.69	26.31	0.0002821
20-fold $\text{CV}_{(1)}$	17.45	54.51	28.04	0.2182	0	67.94	32.06	0.0003258
$\text{CV}_{(1)}$	38.62	52.82	8.56	0.1949	0	86.57	13.43	0.0001711
CV_{mod}	37.53	56.56	5.91	0.1842	0	91.33	8.67	0.0001449
COPSS	37.18	56.08	6.74	0.1865	0	74.72	25.28	0.0002274
LooVF_2	40.96	51.61	7.43	0.1938	0	76.99	23.01	0.0002136
LooVF_5	26.01	59.55	14.44	0.1869	0	73.04	26.96	0.0002593
PELT	23.32	66.57	10.11	0.1765	0	91.5	8.5	0.0001619
WBS	39.42	50.16	10.42	0.1988	0	86.32	13.68	0.0001598
FDRSeg	30.42	64.59	4.99	0.1722	0	83.93	16.07	0.0001845
Ms. FPOP	33.4	65.35	1.25	0.1625	0	98.9	1.1	0.0001059
Biweight	25.14	67.37	7.49	0.1736	0	93	7	0.0001644

TABLE 11

Simulation results relating to Example 2.

	$\sigma = 0.01$				$\sigma = 0.001$			
Method	$\hat{K} < K$	$\hat{K} = K$	$\hat{K} > K$	MISE	$\hat{K} < K$	$\hat{K} = K$	$\hat{K} > K$	MISE
2-fold $CV_{(1)}$	0	87.54	12.46	1.649e-06	0	87.66	12.34	1.648e-08
5-fold $CV_{(1)}$	0	80.75	19.25	2.227e-06	0	79.81	20.19	2.264e-08
10-fold $CV_{(1)}$	0	74.41	25.59	2.753e-06	0	73.77	26.23	2.74e-08
20-fold $CV_{(1)}$	0	68.21	31.79	3.225e-06	0	68.07	31.93	3.214e-08
$CV_{(1)}$	0	86.53	13.47	1.72e-06	0	86.51	13.49	1.712e-08
CV_{mod}	0	91.33	8.67	1.453e-06	0	91.31	8.69	1.463e-08
COPSS	0	32.34	67.66	1.227e-05	0	18.6	81.4	1.982e-07
$LooVF_2$	0	39.78	60.22	6.111e-05	0	23.11	76.89	5.024e-05
$LooVF_5$	0	40.16	59.84	0.009707	0	20.25	79.75	0.0451
PELT	0	91.71	8.29	1.603e-06	0	91.12	8.88	1.669e-08
WBS	0	86.39	13.61	1.561e-06	0	86.61	13.39	1.595e-08
FDRSeg	0	87.88	12.12	1.659e-06	0	86.97	13.03	1.739e-08
Ms. FPOP	0	98.85	1.15	1.074e-06	0	98.75	1.25	1.091e-08
Biweight	0	93.21	6.79	1.735e-05	0	93.04	6.96	2.081e-05

TABLE 12
Simulation results relating to Example 2.

	$\sigma = 0.0001$				$\sigma = 0.0001$, $\tau_1 = n/2 + 1$			
Method	$\hat{K} < K$	$\hat{K} = K$	$\hat{K} > K$	MISE	$\hat{K} < K$	$\hat{K} = K$	$\hat{K} > K$	MISE
2-fold $CV_{(1)}$	0	87.15	12.85	1.691e-10	0	86.74	13.26	1.726e-10
5-fold $CV_{(1)}$	0	79.72	20.28	2.289e-10	0	79.9	20.1	2.294e-10
10-fold $CV_{(1)}$	0	72.89	27.11	2.838e-10	0	73.92	26.08	2.81e-10
20-fold $CV_{(1)}$	0	67.37	32.63	3.288e-10	0	68.21	31.79	3.237e-10
$CV_{(1)}$	0	86.13	13.87	1.752e-10	0	88.22	11.78	1.674e-10
CV_{mod}	0	90.94	9.06	1.49e-10	0	91.17	8.83	1.48e-10
COPSS	0	18.47	81.53	2.034e-09	0	91.5	8.5	1.522e-10
$LooVF_2$	0	22.01	77.99	5e-05	0	22.64	77.36	4.951e-05
$LooVF_5$	0	18.9	81.1	0.05105	0	19.95	80.05	0.05203
PELT	0	91.29	8.71	1.659e-10	0	90.66	9.34	1.707e-10
WBS	0	86.84	13.16	1.624e-10	0	86.18	13.82	1.637e-10
FDRSeg	0	88.8	11.2	1.64e-10	0	88.63	11.37	1.651e-10
Ms. FPOP	0	98.82	1.18	1.097e-10	0	98.92	1.08	1.092e-10
Biweight	0	93	7	1.535e-05	0	92.59	7.41	8.911e-06

TABLE 13
Simulation results relating to Example 2.

Tables 11–13 show that cross-validation with least squares loss is likely to overestimate the number of change-points in the setting of Example 2 when the signal to noise ratio increases. In accordance with the explanations in Section 2, this phenomenon does not occur when the change-point is at an even location. Cross-validation with our proposed criteria does not suffer from this deficiency and is insensitive to whether the change-point is at an odd or even location. Similarly to the previous set of results, the likelihood of overestimation increases with the number of folds, and classical change-point approaches have a slightly smaller tendency to overestimate than cross-validation.

S1.3. Detection power

In this section we provide a more systematic study of detection power. To this end, we consider a signal with a single bump in the middle. We chose $n = 200$,

$K = 2$, $\tau_1 = (n - \lambda)/2$, $\tau_2 = (n + \lambda)/2$, $\beta_0 = \beta_2 = 0$ and $\beta_1 = \Delta_1 = -\Delta_2 = \delta$, and $\sigma = 1$. We vary λ and δ to change the length and size of the bump.

	$\lambda = 6,\ \delta = 2$				$\lambda = 6,\ \delta = 3$			
Method	$\hat{K} < K$	$\hat{K} = K$	$\hat{K} > K$	MISE	$\hat{K} < K$	$\hat{K} = K$	$\hat{K} > K$	MISE
2-fold $\text{CV}_{(1)}$	57.91	32.98	9.11	0.09507	11.29	75.34	13.37	0.06415
5-fold $\text{CV}_{(1)}$	31.82	50.91	17.27	0.07822	2.09	76.7	21.21	0.04884
10-fold $\text{CV}_{(1)}$	26.91	50.92	22.17	0.07852	1.62	71.4	26.98	0.05183
20-fold $\text{CV}_{(1)}$	24.35	48.35	27.3	0.08044	1.49	66.44	32.07	0.05528
$\text{CV}_{(1)}$	67.33	25.7	6.97	0.1033	27.37	62.29	10.34	0.103
CV_{mod}	53.08	37.46	9.46	0.08971	6.86	80.83	12.31	0.05266
COPSS	67.59	26.38	6.03	0.1029	26.68	62.75	10.57	0.1017
LooVF_2	61.92	29	9.08	0.09794	18.61	66.11	15.28	0.08379
LooVF_5	33.03	48.53	18.44	0.07943	2.26	73.44	24.3	0.05007
PELT	30.06	62.79	7.15	0.07085	0.56	91.1	8.34	0.03865
WBS	22.9	65.67	11.43	0.06609	0.28	87.44	12.28	0.03777
FDRSeg	56.55	35.87	7.58	0.09902	4.87	84.75	10.38	0.05035
Ms. FPOP	60.7	38.56	0.74	0.09193	3.97	94.97	1.06	0.04143
Biweight	38.47	56.35	5.18	0.07644	1.78	91.56	6.66	0.04118

TABLE 14
Detection of a single bump of length λ and size δ.

	$\lambda = 6,\ \delta = 4$				$\lambda = 8,\ \delta = 2$			
Method	$\hat{K} < K$	$\hat{K} = K$	$\hat{K} > K$	MISE	$\hat{K} < K$	$\hat{K} = K$	$\hat{K} > K$	MISE
2-fold $\text{CV}_{(1)}$	1.11	85.19	13.7	0.03563	36.39	51.07	12.54	0.09098
5-fold $\text{CV}_{(1)}$	0.05	78.73	21.22	0.0366	14.44	65.34	20.22	0.07011
10-fold $\text{CV}_{(1)}$	0.05	73.53	26.42	0.04118	11.18	63.67	25.15	0.07032
20-fold $\text{CV}_{(1)}$	0.02	67.78	32.2	0.0454	10	59.21	30.79	0.07334
$\text{CV}_{(1)}$	9.21	77.75	13.04	0.07238	30.77	53.74	15.49	0.08562
CV_{mod}	0.33	89.91	9.76	0.03021	31.48	56.9	11.62	0.08361
COPSS	9.56	75.71	14.73	0.07406	28.7	57.08	14.22	0.08195
LooVF_2	3.96	78.78	17.26	0.05039	33.15	53.26	13.59	0.08703
LooVF_5	0.06	74.79	25.15	0.03826	15.48	64.13	20.39	0.07086
PELT	0	91.43	8.57	0.02975	12.12	79.95	7.93	0.06073
WBS	0	88.42	11.58	0.02889	8.78	79.34	11.88	0.05714
FDRSeg	0.01	87.06	12.93	0.03281	35.91	56.87	7.22	0.09241
Ms. FPOP	0	99.01	0.99	0.0244	34.08	65.01	0.91	0.08182
Biweight	0.02	93.1	6.88	0.02935	16.75	77.1	6.15	0.06554

TABLE 15
Detection of a single bump of length λ and size δ.

We see from Tables 14–16 that the newly proposed criteria have only a slightly smaller detection power than the one based on least square loss. Note that the issues of least square loss do not have a significance influence in this setting. Detection power increases with the number of folds, but this also increases the likelihood of false positives. We find that overall, 5-fold cross-validation with absolute error offers a good balance between the risks of under- or overestimate, and also performs well with respect to MISE. Its performance is also quite competitive with classical change-point regression approaches.

	$\lambda = 12,\ \delta = 2$				$\lambda = 20,\ \delta = 2$			
Method	$\hat{K} < K$	$\hat{K} = K$	$\hat{K} > K$	MISE	$\hat{K} < K$	$\hat{K} = K$	$\hat{K} > K$	MISE
2-fold $\text{CV}_{(1)}$	13.29	71.2	15.51	0.07345	1.23	83.32	15.45	0.05142
5-fold $\text{CV}_{(1)}$	2.68	76.03	21.29	0.05779	0.06	78.33	21.61	0.05305
10-fold $\text{CV}_{(1)}$	1.89	71.16	26.95	0.06026	0.05	73.14	26.81	0.05731
20-fold $\text{CV}_{(1)}$	1.73	66.36	31.91	0.06397	0.02	67.75	32.23	0.06125
$\text{CV}_{(1)}$	10.61	72.83	16.56	0.0693	0.97	83.14	15.89	0.05101
CV_{mod}	9.75	77.59	12.66	0.06499	0.75	86.53	12.72	0.0479
COPSS	8.92	75.85	15.23	0.06492	0.66	84.38	14.96	0.04919
LooVF_2	10.92	71.37	17.71	0.06956	1.01	80.68	18.31	0.05152
LooVF_5	2.61	74.58	22.81	0.05815	0.04	77.58	22.38	0.0525
PELT	1.38	90.1	8.52	0.04915	0.01	90.75	9.24	0.04632
WBS	1.01	86.82	12.17	0.04803	0.01	87.84	12.15	0.0455
FDRSeg	6.03	81.59	12.38	0.06085	0.07	86.51	13.42	0.04913
Ms. FPOP	6.59	92.32	1.09	0.05332	0.05	98.94	1.01	0.04055
Biweight	2.01	91.1	6.89	0.05034	0.01	92.64	7.35	0.04613

TABLE 16
Detection of a single bump of length λ and size δ.

S2. Proof of Theorem 5

We first show that with high probability, for each true change-point, there is an estimated change-point that is $\underline{\lambda}/4$ close, see (32). We then working on this event and show that any change-point set that violates the condition in Theorem 5 has larger costs than the true change-point set. To this end, we use Lemma 4 to split the difference of these costs into several events, which we use the following large deviations bounds to control. Throughout the proofs, we write $k \in [a, b]$ instead of $k \in \{\lceil a \rceil, \ldots, \lfloor b \rfloor\}$ when it is clear from the context that k is an integer.

Lemma 1 (Lemma 4 on p. 33 in Verzelen et al. (2020))**.** *Let $\varepsilon_1, \varepsilon_2, \ldots$ be independent centred sub-Gaussian random variables with variance proxy 1. Then, for any integer $d \geq 1$, any $\alpha > 0$ and any $x > 0$,*

$$\mathbb{P}\left(\max_{k \in [d, (1+\alpha)d]} \frac{\sum_{i=1}^{k} \varepsilon_i}{\sqrt{k}} \geq x \right) \leq \exp\left(-\frac{x^2}{2(1+\alpha)} \right).$$

Lemma 2. *Let $\varepsilon_1, \varepsilon_2, \ldots$ be independent centred sub-Gaussian random variables with variance proxy 1. Then, for any integer $c \geq 1$ and any $x > 0$ such that $cx^2 > 4$,*

$$\mathbb{P}\left(\sup_{k \geq c} \frac{\sum_{i=1}^{k} \varepsilon_i}{k} \geq x \right) \leq 2 \exp\left(-\frac{1}{4} c x^2 \right).$$

Proof. It follows from a union bound and Lemma 1 that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{k\geq c}\frac{\sum_{i=1}^{k}\varepsilon_i}{k}\geq x\right)=\mathbb{P}\left(\sup_{s\in\mathbb{Z}:s\geq 0}\max_{k\in[c2^s,c2^{s+1}]}\frac{\sum_{i=1}^{k}\varepsilon_i}{k}\geq x\right)\\
&\leq\sum_{s=0}^{\infty}\mathbb{P}\left(\max_{k\in[c2^s,c2^{s+1}]}\frac{\sum_{i=1}^{k}\varepsilon_i}{\sqrt{k}}\geq x\sqrt{c2^s}\right)\\
&\leq\sum_{s=0}^{\infty}\exp\left(-\frac{1}{4}cx^2 2^s\right)\leq\sum_{s=1}^{\infty}\exp\left(-\frac{1}{4}cx^2 s\right)=\frac{\exp\left(-\frac{1}{4}cx^2\right)}{1-\exp\left(-\frac{1}{4}cx^2\right)}\leq 2\exp\left(-\frac{1}{4}cx^2\right).
\end{aligned}$$

□

Lemma 3. *Let $\varepsilon_1,\varepsilon_2,\ldots$ be independent centred sub-Gaussian random variables with variance proxy 1. Then, for any integer $c\geq 1$ and any $x>0$ such that $cx^2\geq 4$,*

$$\mathbb{P}\left(\sup_{k\geq c,\,l\geq 1}\frac{\sum_{i=k+1}^{k+l}\varepsilon_i}{\sqrt{l}\sqrt{k+l}}\geq x\right)\leq 2c^2\exp\left(-\frac{1}{2}cx^2\right).$$

Proof. We have that

$$\mathbb{P}\left(\sup_{k\geq c,\,l\geq 1}\frac{\sum_{i=k+1}^{k+l}\varepsilon_i}{\sqrt{l}\sqrt{k+l}}\geq x\right)\leq\mathbb{P}\left(\sup_{k\geq c,\,l\in[1,c-1]}\frac{\sum_{i=k+1}^{k+l}\varepsilon_i}{\sqrt{l}\sqrt{k+l}}\geq x\right)+\mathbb{P}\left(\sup_{k\geq c,\,l\geq c}\frac{\sum_{i=k+1}^{k+l}\varepsilon_i}{\sqrt{l}\sqrt{k+l}}\geq x\right).$$

From a union bound, sub-Gaussianity and $cx^2\geq 4$, we have that

$$\begin{aligned}
\mathbb{P}\left(\sup_{k\geq c,\,l\in[1,c-1]}\frac{\sum_{i=k+1}^{k+l}\varepsilon_i}{\sqrt{l}\sqrt{k+l}}\geq x\right)&\leq\sum_{l=1}^{c-1}\sum_{s=1}^{\infty}\sum_{k\in[cs,c(s+1))}\mathbb{P}\left(\frac{\sum_{i=k+1}^{k+l}\varepsilon_i}{\sqrt{l}}\geq x\sqrt{cs+l}\right)\\
&\leq c\sum_{s=1}^{\infty}\sum_{l=1}^{c-1}\exp\left(-\frac{1}{2}x^2(cs+l)\right)\\
&\leq c^2\frac{\exp\left(-\frac{1}{2}cx^2\right)}{1-\exp\left(-\frac{1}{2}cx^2\right)}\\
&= c^2\frac{\exp\left(-\frac{1}{4}cx^2\right)}{1-\exp\left(-\frac{1}{2}cx^2\right)}\exp\left(-\frac{1}{4}cx^2\right)\leq c^2\exp\left(-\frac{1}{4}cx^2\right).
\end{aligned}$$

It follows from a union bound, Lemma 1 and the fact that $cx^2 \geq 4$, that

$$\begin{aligned}
\mathbb{P}\left(\sup_{k\geq c, l\geq c} \frac{\sum_{i=k+1}^{k+l}\varepsilon_i}{\sqrt{l}\sqrt{k+l}} \geq x\right) &\leq \sum_{s=1}^{\infty}\sum_{t=0}^{\infty}\sum_{k\in[sc,(s+1)c)} \mathbb{P}\left(\max_{l\in[c2^t, c2^{t+1}]} \frac{\sum_{i=k+1}^{k+l}\varepsilon_i}{\sqrt{l}} \geq x\sqrt{sc+2^t c}\right) \\
&\leq \sum_{s=1}^{\infty}\sum_{t=0}^{\infty} c\exp\left(-\frac{1}{4}cx^2(s+2^t)\right) \\
&\leq c\left(\sum_{s=1}^{\infty}\exp\left(-\frac{1}{4}cx^2 s\right)\right)\left(\sum_{t=1}^{\infty}\exp\left(-\frac{1}{4}cx^2 t\right)\right) \\
&\leq c\left(\frac{\exp\left(-\frac{1}{4}cx^2\right)}{1-\exp\left(-\frac{1}{4}cx^2\right)}\right)^2 \leq c\exp\left(-\frac{1}{4}cx^2\right),
\end{aligned}$$

in the final line using the fact that $\{u/(1-u)\}^2 \leq u$ for $0 \leq u \leq (3-\sqrt{(}5))/2$ and that $e^{-1} \leq (3-\sqrt{(}5))/2$. Combing all inequalities shows the statement. □

For a set of putative change-points $\mathcal{U} = \{t_0 < t_1 < \cdots < t_L < t_{L+1}\}$ we will write $\mathcal{U} \setminus t_1$ and $\mathcal{U} \cup t_2$ instead of $\mathcal{U} \setminus \{t_1\}$ and $\mathcal{U} \cup \{t_2\}$ throughout the paper, for ease of presentation.

Lemma 4. *Let $Y_i = \mu_i + \varepsilon_i$ and let $\mathcal{U} = \{0 = t_0 < t_1 < \cdots < t_L < t_{L+1} = n\}$ be an arbitrary set of candidate change-points. Suppose integer t is such that $t_{l-1} < t < t_l$ and $\overline{\mu}_{t:t_l} = 0$. Then,*

$$\begin{aligned}
&S_Y(\mathcal{U}) - S_Y((\mathcal{U}\setminus t_l)\cup t) \\
&= (t_l - t)\left[\frac{t-t_{l-1}}{t_l - t_{l-1}}\overline{\mu}^2_{t_{l-1}:t} - \frac{t_{l+1}-t_l}{t_{l+1}-t}\overline{\mu}^2_{t_l:t_{l+1}}\right] \\
&\quad + 2(t_l - t)\left[\frac{t-t_{l-1}}{t_l - t_{l-1}}\overline{\mu}_{t_{l-1}:t}\left(\overline{\varepsilon}_{t_{l-1}:t} - \overline{\varepsilon}_{t:t_l}\right) + \frac{t_{l+1}-t_l}{t_{l+1}-t}\overline{\mu}_{t_l:t_{l+1}}\left(\overline{\varepsilon}_{t:t_l} - \overline{\varepsilon}_{t_l:t_{l+1}}\right)\right] \\
&\quad + (t_l - t)\left[\frac{(t_l - t)\,\overline{\varepsilon}^2_{t:t_l} + 2(t_{l+1}-t_l)\,\overline{\varepsilon}_{t_l:t_{l+1}}\overline{\varepsilon}_{t:t_l} - (t_{l+1}-t_l)\,\overline{\varepsilon}^2_{t_l:t_{l+1}}}{t_{l+1}-t}\right. \\
&\qquad \left. + \frac{(t-t_{l-1})\,\overline{\varepsilon}^2_{t_{l-1}:t} - 2(t-t_{l-1})\,\overline{\varepsilon}_{t_{l-1}:t}\overline{\varepsilon}_{t:t_l} - (t_l - t)\,\overline{\varepsilon}^2_{t:t_l}}{t_l - t_{l-1}}\right].
\end{aligned}$$

The assumption $\overline{\mu}_{t:t_l} = 0$ may appear restrictive at first glance. But note that we can replace Y_i with $Y_i - \overline{\mu}_{t:t_l}$, since it does not changes the costs. Alternatively, one can replace every μ_i with $\mu_i - \overline{\mu}_{t:t_l}$ in the right hand side to obtain a more general lemma that does not require this assumption.

Proof of Lemma 4. We have that

$$\begin{aligned}
&\sum_{i=t_l+1}^{t_{l+1}} \left(Y_i - \overline{Y}_{t_l:t_{l+1}}\right)^2 \\
&= \sum_{i=t_l+1}^{t_{l+1}} \left(\mu_i + \varepsilon_i - \overline{\mu}_{t_l:t_{l+1}} - \overline{\varepsilon}_{t_l:t_{l+1}}\right)^2 \\
&= \sum_{i=t_l+1}^{t_{l+1}} \left(\mu_i - \overline{\mu}_{t_l:t_{l+1}}\right)^2 + 2\sum_{i=t_l+1}^{t_{l+1}} \left(\mu_i - \overline{\mu}_{t_l:t_{l+1}}\right)\left(\varepsilon_i - \overline{\varepsilon}_{t_l:t_{l+1}}\right) + \sum_{i=t_l+1}^{t_{l+1}} \left(\varepsilon_i - \overline{\varepsilon}_{t_l:t_{l+1}}\right)^2 \\
&= \sum_{i=t_l+1}^{t_{l+1}} \mu_i \left(\mu_i - \overline{\mu}_{t_l:t_{l+1}}\right) + 2\sum_{i=t_l+1}^{t_{l+1}} \mu_i \left(\varepsilon_i - \overline{\varepsilon}_{t_l:t_{l+1}}\right) - (t_{l+1} - t_l)\,\overline{\varepsilon}^2_{t_l:t_{l+1}} + \sum_{i=t_l+1}^{t_{l+1}} \varepsilon_i^2,
\end{aligned} \tag{31}$$

since

$$\sum_{i=t_l+1}^{t_{l+1}} \overline{\mu}_{t_l:t_{l+1}} \left(\mu_i - \overline{\mu}_{t_l:t_{l+1}}\right) = 0 = \sum_{i=t_l+1}^{t_{l+1}} \overline{\mu}_{t_l:t_{l+1}} \left(\varepsilon_i - \overline{\varepsilon}_{t_l:t_{l+1}}\right).$$

Thus,

$$\begin{aligned}
&S_Y(\mathcal{U}) - S_Y((\mathcal{U} \setminus t_l) \cup t) \\
&= -\sum_{i=t_{l-1}+1}^{t_l} \mu_i \overline{\mu}_{t_{l-1}:t_l} - \sum_{i=t_l+1}^{t_{l+1}} \mu_i \overline{\mu}_{t_l:t_{l+1}} + \sum_{i=t_{l-1}+1}^{t} \mu_i \overline{\mu}_{t_{l-1}:t} + \sum_{i=t+1}^{t_{l+1}} \mu_i \overline{\mu}_{t:t_{l+1}} \\
&\quad - 2\sum_{i=t_{l-1}+1}^{t_l} \mu_i \overline{\varepsilon}_{t_{l-1}:t_l} - 2\sum_{i=t_l+1}^{t_{l+1}} \mu_i \overline{\varepsilon}_{t_l:t_{l+1}} + 2\sum_{i=t_{l-1}+1}^{t} \mu_i \overline{\varepsilon}_{t_{l-1}:t} + 2\sum_{i=t+1}^{t_{l+1}} \mu_i \overline{\varepsilon}_{t:t_{l+1}} \\
&\quad + (t - t_{l-1})\,\overline{\varepsilon}^2_{t_{l-1}:t} + (t_{l+1} - t)\,\overline{\varepsilon}^2_{t:t_{l+1}} - (t_l - t_{l-1})\,\overline{\varepsilon}^2_{t_{l-1}:t_l} - (t_{l+1} - t_l)\,\overline{\varepsilon}^2_{t_l:t_{l+1}}.
\end{aligned}$$

Since $\overline{\mu}_{t:t_l} = 0$, we have that $(t_l - t_{l-1})\,\overline{\mu}_{t_{l-1}:t_l} = (t - t_{l-1})\,\overline{\mu}_{t_{l-1}:t}$ and $(t_{l+1} - t)\,\overline{\mu}_{t:t_{l+1}} = (t_{l+1} - t_l)\,\overline{\mu}_{t_l:t_{l+1}}$. Thus,

$$\begin{aligned}
&-\sum_{i=t_{l-1}+1}^{t_l} \mu_i \overline{\mu}_{t_{l-1}:t_l} - \sum_{i=t_l+1}^{t_{l+1}} \mu_i \overline{\mu}_{t_l:t_{l+1}} + \sum_{i=t_{l-1}+1}^{t} \mu_i \overline{\mu}_{t_{l-1}:t} + \sum_{i=t+1}^{t_{l+1}} \mu_i \overline{\mu}_{t:t_{l+1}} \\
&= -\sum_{i=t_{l-1}+1}^{t_l} \mu_i \frac{(t - t_{l-1})\,\overline{\mu}_{t_{l-1}:t}}{t_l - t_{l-1}} + \sum_{i=t_{l-1}+1}^{t} \mu_i \overline{\mu}_{t_{l-1}:t} \\
&\quad + \sum_{i=t+1}^{t_{l+1}} \mu_i \frac{(t_{l+1} - t_l)\,\overline{\mu}_{t_l:t_{l+1}}}{t_{l+1} - t} - \sum_{i=t_l+1}^{t_{l+1}} \mu_i \overline{\mu}_{t_l:t_{l+1}} \\
&= (t_l - t)\left[\frac{t - t_{l-1}}{t_l - t_{l-1}} \overline{\mu}^2_{t_{l-1}:t} - \frac{t_{l+1} - t_l}{t_{l+1} - t} \overline{\mu}^2_{t_l:t_{l+1}}\right].
\end{aligned}$$

Since $(t_l - t_{l-1})\,\overline{\varepsilon}_{t_{l-1}:t_l} = (t - t_{l-1})\,\overline{\varepsilon}_{t_{l-1}:t} + (t_l - t)\,\overline{\varepsilon}_{t:t_l}$ and

$(t_{l+1}-t)\,\overline{\varepsilon}_{t:t_{l+1}} = (t_{l+1}-t_l)\,\overline{\varepsilon}_{t_l:t_{l+1}} + (t_l-t)\,\overline{\varepsilon}_{t:t_l}$, we have that

$$\begin{aligned}
&-\sum_{i=t_{l-1}+1}^{t_l}\mu_i\overline{\varepsilon}_{t_{l-1}:t_l} - \sum_{i=t_l+1}^{t_{l+1}}\mu_i\overline{\varepsilon}_{t_l:t_{l+1}} + \sum_{i=t_{l-1}+1}^{t}\mu_i\overline{\varepsilon}_{t_{l-1}:t} + \sum_{i=t+1}^{t_{l+1}}\mu_i\overline{\varepsilon}_{t:t_{l+1}}\\
&= -\sum_{i=t_{l-1}+1}^{t_l}\mu_i\frac{(t-t_{l-1})\,\overline{\varepsilon}_{t_{l-1}:t} + (t_l-t)\,\overline{\varepsilon}_{t:t_l}}{t_l-t_{l-1}} + \sum_{i=t_{l-1}+1}^{t}\mu_i\overline{\varepsilon}_{t_{l-1}:t}\\
&\quad + \sum_{i=t+1}^{t_{l+1}}\mu_i\frac{(t_{l+1}-t_l)\,\overline{\varepsilon}_{t_l:t_{l+1}} + (t_l-t)\,\overline{\varepsilon}_{t:t_l}}{t_{l+1}-t} - \sum_{i=t_l+1}^{t_{l+1}}\mu_i\overline{\varepsilon}_{t_l:t_{l+1}}\\
&= (t_l-t)\left[\frac{t-t_{l-1}}{t_l-t_{l-1}}\overline{\mu}_{t_{l-1}:t}\left(\overline{\varepsilon}_{t_{l-1}:t}-\overline{\varepsilon}_{t:t_l}\right) + \frac{t_{l+1}-t_l}{t_{l+1}-t}\overline{\mu}_{t_l:t_{l+1}}\left(\overline{\varepsilon}_{t:t_l}-\overline{\varepsilon}_{t_l:t_{l+1}}\right)\right],
\end{aligned}$$

where we have used the assumption $\overline{\mu}_{t:t_l} = 0$. Moreover, it follows from the same splitting of the errors that

$$\begin{aligned}
&(t-t_{l-1})\,\overline{\varepsilon}^2_{t_{l-1}:t} + (t_{l+1}-t)\,\overline{\varepsilon}^2_{t:t_{l+1}} - (t_l-t_{l-1})\,\overline{\varepsilon}^2_{t_{l-1}:t_l} - (t_{l+1}-t_l)\,\overline{\varepsilon}^2_{t_l:t_{l+1}}\\
&= \frac{\left[(t_{l+1}-t_l)\,\overline{\varepsilon}_{t_l:t_{l+1}} + (t_l-t)\,\overline{\varepsilon}_{t:t_l}\right]^2}{t_{l+1}-t} - (t_{l+1}-t_l)\,\overline{\varepsilon}^2_{t_l:t_{l+1}}\\
&\quad - \frac{\left[(t-t_{l-1})\,\overline{\varepsilon}_{t_{l-1}:t} + (t_l-t)\,\overline{\varepsilon}_{t:t_l}\right]^2}{t_l-t_{l-1}} + (t-t_{l-1})\,\overline{\varepsilon}^2_{t_{l-1}:t}\\
&= (t_l-t)\left[\frac{(t_l-t)\,\overline{\varepsilon}^2_{t:t_l} + 2\,(t_{l+1}-t_l)\,\overline{\varepsilon}_{t_l:t_{l+1}}\overline{\varepsilon}_{t:t_l} - (t_{l+1}-t_l)\,\overline{\varepsilon}^2_{t_l:t_{l+1}}}{t_{l+1}-t}\right.\\
&\qquad \left. + \frac{(t-t_{l-1})\,\overline{\varepsilon}^2_{t_{l-1}:t} - 2\,(t-t_{l-1})\,\overline{\varepsilon}_{t_{l-1}:t}\overline{\varepsilon}_{t:t_l} - (t_l-t)\,\overline{\varepsilon}^2_{t:t_l}}{t_l-t_{l-1}}\right].
\end{aligned}$$

Combining all equalities completes the proof. □

Proof of Theorem 5. We define $\hat{\mathcal{T}}_L^{\varepsilon}$ as the set of L change-points that minimises S_{ε}, i.e. $\hat{\mathcal{T}}_L^{\varepsilon} = \operatorname{argmin}_{\mathcal{U}:|\mathcal{U}|=L} S_{\varepsilon}(\mathcal{U})$.

Proof of (25) We begin by showing that with high probability, there is an estimated change-point in each $\underline{\lambda}/4$-neighbourhood of a true change, i.e. that the sequence of events

$$\Omega_{1n} := \left\{\forall\, L \geq K\ \exists\, \hat{\tau}_{L,i_1},\ldots,\hat{\tau}_{L,i_K} \in \hat{\mathcal{T}}_L \ :\ \max_{1\leq k\leq K}\left|\hat{\tau}_{L,i_k}-\tau_k\right| \leq \frac{\underline{\lambda}}{4}\right\} \tag{32}$$

has $\mathbb{P}(\Omega_{1n}) \to 1$.

We denote by $\hat{\mathcal{T}}_L([a,b])$ the set of $L \in \mathbb{N}$ change-points estimated by least squares estimation, with the restriction that no change-point is in $[a,b]$, so

$$\hat{\mathcal{T}}_L([a,b]) := \operatorname*{argmin}_{\substack{0=:t_0<t_1<\cdots<t_L<t_{L+1}:=n,\\ t_1,\ldots,t_l\notin[a,b]}} \sum_{l=0}^{L}\sum_{i=t_l+1}^{t_{l+1}}\left(Y_i - \overline{Y}_{t_l:t_{l+1}}\right)^2.$$

We also use the notation $\hat{\mathcal{T}}_L^{-k} := \hat{\mathcal{T}}_L\left(\left[\tau_k - \frac{\lambda}{4}, \tau_k + \frac{\lambda}{4}\right]\right)$. Note that the event

$$\left\{S_Y\left(\hat{\mathcal{T}}_L^{-k}\right) > S_Y\left(\hat{\mathcal{T}}_L\right) \ \forall\ L \geq K,\ \forall\ k = 1, \ldots, K\right\}, \tag{33}$$

is contained in Ω_{1n} (32). Now from (31), we have that for any $l = 0, \ldots, K$,

$$\begin{aligned}
&\sum_{i=\tau_l+1}^{\tau_{l+1}} \left(Y_i - \overline{Y}_{\tau_l:\tau_{l+1}}\right)^2 \\
&= \sum_{i=\tau_l+1}^{\tau_{l+1}} \left(\mu_i - \overline{\mu}_{\tau_l:\tau_{l+1}}\right)^2 + 2\sum_{i=\tau_l+1}^{\tau_{l+1}} \left(\mu_i - \overline{\mu}_{\tau_l:\tau_{l+1}}\right)\left(\varepsilon_i - \overline{\varepsilon}_{\tau_l:\tau_{l+1}}\right) + \sum_{i=\tau_l+1}^{\tau_{l+1}} \left(\varepsilon_i - \overline{\varepsilon}_{\tau_l:\tau_{l+1}}\right)^2 \\
&= \sum_{i=\tau_l+1}^{\tau_{l+1}} \left(\mu_i - \overline{\mu}_{\tau_l:\tau_{l+1}}\right)^2 + 2\sum_{i=\tau_l+1}^{\tau_{l+1}} \left(\mu_i - \overline{\mu}_{\tau_l:\tau_{l+1}}\right)\varepsilon_i + \sum_{i=\tau_l+1}^{\tau_{l+1}} \left(\varepsilon_i - \overline{\varepsilon}_{\tau_l:\tau_{l+1}}\right)^2,
\end{aligned}$$

since $\sum_{i=\tau_l+1}^{\tau_{l+1}} \left(\mu_i - \overline{\mu}_{\tau_l:\tau_{l+1}}\right)\overline{\varepsilon}_{\tau_l:\tau_{l+1}} = 0$. As adding change-points can never increase the cost,

$$\begin{aligned}
S_Y\left(\hat{\mathcal{T}}_L^{-k}\right) \geq & S_Y\left(\hat{\mathcal{T}}_L^{-k} \cup \bigcup_{\substack{j=1,\ldots,K \\ j\neq k}} \tau_j \cup \left\{\tau_k - \frac{\lambda}{4}, \tau_k + \frac{\lambda}{4}\right\}\right) \\
= & \sum_{i=\tau_k-\frac{\lambda}{4}+1}^{\tau_k+\frac{\lambda}{4}} \left[\left(\mu_i - \overline{\mu}_{\tau_k-\frac{\lambda}{4}:\tau_k+\frac{\lambda}{4}}\right)^2 + 2\left(\mu_i - \overline{\mu}_{\tau_k-\frac{\lambda}{4}:\tau_k+\frac{\lambda}{4}}\right)\varepsilon_i\right] \\
& + S_\varepsilon\left(\hat{\mathcal{T}}_L^{-k} \cup \bigcup_{\substack{j=1,\ldots,K \\ j\neq k}} \tau_j \cup \left\{\tau_k - \frac{\lambda}{4}, \tau_k + \frac{\lambda}{4}\right\}\right) \\
= & \frac{\lambda}{2}\left(\frac{\Delta_k}{2}\right)^2 - 2\,\mathrm{sgn}(\beta_k - \beta_{k-1})\frac{\lambda}{4}\frac{\Delta_k}{2}\left(\overline{\varepsilon}_{\tau_k-\frac{\lambda}{4}:\tau_k} - \overline{\varepsilon}_{\tau_k:\tau_k+\frac{\lambda}{4}}\right) \\
& + S_\varepsilon\left(\hat{\mathcal{T}}_L^{-k} \cup \bigcup_{\substack{j=1,\ldots,K \\ j\neq k}} \tau_j \cup \left\{\tau_k - \frac{\lambda}{4}, \tau_k + \frac{\lambda}{4}\right\}\right).
\end{aligned} \tag{34}$$

As a consequence of the sub-Gaussianity of ε,

$$\max_{k=1,\ldots,K}\left|\overline{\varepsilon}_{\tau_k-\frac{\lambda}{4}:\tau_k} - \overline{\varepsilon}_{\tau_k:\tau_k+\frac{\lambda}{4}}\right| = O_{\mathbb{P}}\left(\frac{\sigma\sqrt{\log(eK)}}{\sqrt{\lambda}}\right). \tag{35}$$

Since removing restrictions and adding change-points can never increase the cost,

$$S_\varepsilon\left(\hat{\mathcal{T}}_L^{-k} \cup \bigcup_{\substack{j=1,\ldots,K \\ j\neq k}} \tau_j \cup \left\{\tau_k - \frac{\lambda}{4}, \tau_k + \frac{\lambda}{4}\right\}\right) \geq S_\varepsilon\left(\mathcal{T} \cup \hat{\mathcal{T}}_{L+2}^{\varepsilon}\right). \tag{36}$$

Next it follows from the fact that $\sum_{i=\tau_l+1}^{\tau_{l+1}} \left(\varepsilon_i - \overline{\varepsilon}_{\tau_l:\tau_{l+1}}\right)^2 = \sum_{i=\tau_l+1}^{\tau_{l+1}} \varepsilon_i^2 - (\tau_{l+1} - \tau_l)\,\overline{\varepsilon}^2_{\tau_l:\tau_{l+1}}$, sub-Gaussianity, and $K < \overline{\lambda}$ that

$$\begin{aligned}
&\max_{L \geq K} \left\{ S_\varepsilon\left(\mathcal{T} \cup \hat{\mathcal{T}}^{\varepsilon}_{L-K}\right) - S_\varepsilon\left(\mathcal{T} \cup \hat{\mathcal{T}}^{\varepsilon}_{L+2}\right) \right\} \\
\leq & (K+2) \max_{1 \leq \tau_1 < \tau_2 \leq n} (\tau_2 - \tau_1)\,\overline{\varepsilon}^2_{\tau_1:\tau_2} \\
= & O_{\mathbb{P}}\left(K\sigma^2 \log(n)\right) = O_{\mathbb{P}}\left(K\sigma^2 \log(K\overline{\lambda})\right) = O_{\mathbb{P}}\left(K\sigma^2 \log(\overline{\lambda})\right). \qquad (37)
\end{aligned}$$

Also,

$$S_Y\left(\hat{\mathcal{T}}_L\right) \leq S_Y\left(\mathcal{T} \cup \hat{\mathcal{T}}^{\varepsilon}_{L-K}\right) = S_\varepsilon\left(\mathcal{T} \cup \hat{\mathcal{T}}^{\varepsilon}_{L-K}\right). \qquad (38)$$

Hence combining (34), (35), (36), (37) and (38), we have that

$$\begin{aligned}
&\min_{L>K} \min_{k=1,\ldots,K} \left\{ S_Y\left(\hat{\mathcal{T}}^{-k}_L\right) - S_Y\left(\hat{\mathcal{T}}_L\right) \right\} \\
\geq & \frac{\underline{\lambda}}{2}\left(\frac{\Delta_{(1)}}{2}\right)^2 + O_{\mathbb{P}}\left(\sqrt{\underline{\lambda}\Delta^2_{(1)} \log(eK)}\sigma\right) + O_{\mathbb{P}}\left(K\sigma^2 \log(\overline{\lambda})\right),
\end{aligned}$$

where the second inequality holds for large n because of (23). Moreover, the right-hand side is positive with probability converging to 1 due to (23). Thus $\mathbb{P}(\Omega_{1n}) \to 1$ as claimed.

We will now introduce classes of change-point sets $\mathfrak{T}_L(\cdot)$ that violate the statement for $\hat{\mathcal{T}}_L$ in Theorem 5. We then compute the difference of the costs of such change-point sets and the cost of $\hat{\mathcal{T}}_L$. To complete the proof, it will remain to show that this difference is positive simultaneously for all L and all change-point sets with probability converging to 1.

Let

$$c_{L,k} := \begin{cases} \log(eK)\frac{\sigma^2}{\Delta_k^2}, & \text{if } L = K, \\ \log\left(eK\frac{\sigma^2}{\Delta_k^2}\right)\frac{\sigma^2}{\Delta_k^2}, & \text{if } L \neq K. \end{cases}$$

Then, by definition $\max_{L,k} c_{L,k}\gamma^{-1}_{L,k} = o(1)$. Now let C be a sequence with $C \to \infty$, but

$$C\frac{\sigma^2 \log(\overline{\lambda})}{\underline{\lambda}\Delta^2_{(1)}} \to 0, \quad \text{as } n \to \infty, \qquad (39)$$

which is possible because of (23). Then, for fixed n and $L \geq K$, for any $\mathcal{I} \subseteq \{1, \ldots, K\}$, let $\mathfrak{T}_L(\mathcal{I})$ be the class of change-point sets $\mathcal{U}$ of size L such that the follow two properties hold:

$$\begin{aligned}
&\text{for each } k = 1, \ldots, K,\ \exists\ \tau \in \mathcal{U} : |\tau - \tau_k| \leq \underline{\lambda}/4, \\
&\{k : |\tau_k - \tau| > Cc_{L,k}\ \forall \tau \in \mathcal{U}\} = \mathcal{I}. \qquad (40)
\end{aligned}$$

We note that $\mathcal{I}$ denotes the indices of those change-points which are not well estimated. Conversely, $\{1, \ldots, K\} \setminus \mathcal{I}$ are the indices that satisfy the result in

Theorem 5. Hence defining

$$\Omega_{2n} := \left\{ S_Y\left(\mathcal{U}\right) > S_Y\left(\hat{\mathcal{T}}_L\right) \ \forall\, \mathcal{U} \in \mathfrak{T}_L(\mathcal{I}),\ \forall\, \mathcal{I} \subseteq \{1,\ldots,K\},\ \mathcal{I} \neq \emptyset,\ \text{and}\ \forall\, L \geq K \right\}, \tag{41}$$

we have that $\Omega_{1n} \cap \Omega_{2n}$ implies the event in Theorem 5. Indeed, suppose that $\hat{\mathcal{I}}_L := \{k : |\tau_k - \hat{\tau}| > Cc_{L,k}\ \forall \hat{\tau} \in \hat{\mathcal{T}}_L\}$ is non-empty for some $L \geq K$. Then on Ω_{1n} we have $\hat{\mathcal{T}}_L \in \mathfrak{T}_L(\hat{\mathcal{I}}_L)$, implying that Ω_{2n} has not occurred. Thus it suffices to show $\mathbb{P}(\Omega_{2n}) \to 1$, as $n \to \infty$. To this end, we will bound $S_Y\left(\mathcal{U}\right) - S_Y\left(\hat{\mathcal{T}}_L\right)$ next.

Let n and $L \geq K$ be fixed. Take a non-empty $\mathcal{I} = \{i_1,\ldots,i_{\tilde{K}}\} \subseteq \{1,\ldots,K\}$, and $i_1,\ldots,i_{\tilde{K}}$ ordered such that $\Delta_{i_k} \geq \Delta_{i_l} \Leftrightarrow i_k < i_l$. Further let $\mathcal{U} = \{0 = \tilde{\tau}_0 < \tilde{\tau}_1 < \cdots \tilde{\tau}_L < \tilde{\tau}_{L+1} = n\} \in \mathfrak{T}_L(\mathcal{I})$ be fixed For each $k \in \mathcal{I}$, let j_k be such that $\tilde{\tau}_{j_k} \in \mathcal{U}$ is the closest change-point in $\mathcal{U}$ to τ_k. In the case of a tie we pick the change-point on the right hand side. Thus, by definition of $\hat{\mathcal{T}}_L$,

$$\begin{aligned}
&S_Y\left(\mathcal{U}\right) - S_Y\left(\hat{\mathcal{T}}_L\right) \\
\geq &S_Y\left(\mathcal{U}\right) - S_Y\left(\left[\mathcal{U} \setminus \bigcup_{l=1}^{\tilde{K}} \tilde{\tau}_{j_k}\right] \cup \bigcup_{l=1}^{\tilde{K}} \tau_k\right) \\
= &\sum_{k=1}^{\tilde{K}} \left[S_Y\left(\left[\mathcal{U} \setminus \bigcup_{l=1}^{k-1} \tilde{\tau}_l\right] \cup \bigcup_{l=1}^{k-1} \tau_k\right) - S_Y\left(\left[\mathcal{U} \setminus \bigcup_{l=1}^{k} \tilde{\tau}_{j_l}\right] \cup \bigcup_{l=1}^{k} \tau_l\right)\right].
\end{aligned}$$

We now show that with probability converging to 1, each summand, and hence the sum, is positive, simultaneously for all L, $\mathcal{I}$ and $\mathcal{U} \in \mathfrak{T}_L(\mathcal{I})$.

For any fixed k, write

$$\tilde{\mathcal{U}} := \left(\mathcal{U} \setminus \bigcup_{l=1}^{k-1} \tilde{\tau}_{j_l}\right) \cup \bigcup_{l=1}^{k-1} \tau_l. \tag{42}$$

Further write $l := j_k$. Without loss of generality we may assume $\tau_k < \tilde{\tau}_l$, as otherwise we can instead consider the observations in reverse order, which does not change the costs. Also, since replacing every μ_i by $\mu_i - c$, with c a global constant leaves the costs unchanged, we can assume $\overline{\mu}_{\tau_k:\tilde{\tau}_l} = 0$ for the calculations below without loss of generality. Indeed, each subsequent instance of μ_i can be interpreted as the original mean with $\overline{\mu}_{\tau_k:\tilde{\tau}_l}$ subtracted. Hence, it follows

from Lemma 4 that

$$
\begin{aligned}
&S_Y\left(\tilde{\mathcal{U}}\right) - S_Y\left((\tilde{\mathcal{U}} \setminus \tilde{\tau}_l) \cup \tau_k\right) \\
&= (\tilde{\tau}_l - \tau_k)\left[\frac{\tau_k - \tilde{\tau}_{l-1}}{\tilde{\tau}_l - \tilde{\tau}_{l-1}}\overline{\mu}^2_{\tilde{\tau}_{l-1}:\tau_k} - \frac{\tilde{\tau}_{l+1} - \tilde{\tau}_l}{\tilde{\tau}_{l+1} - \tau_k}\overline{\mu}^2_{\tilde{\tau}_l:\tilde{\tau}_{l+1}}\right] \\
&\quad + 2(\tilde{\tau}_l - \tau_k)\left[\frac{\tau_k - \tilde{\tau}_{l-1}}{\tilde{\tau}_l - \tilde{\tau}_{l-1}}\overline{\mu}_{\tilde{\tau}_{l-1}:\tau_k}\left(\overline{\varepsilon}_{\tilde{\tau}_{l-1}:\tau_k} - \overline{\varepsilon}_{\tau_k:\tilde{\tau}_l}\right)\right. \\
&\qquad\qquad \left. + \frac{\tilde{\tau}_{l+1} - \tilde{\tau}_l}{\tilde{\tau}_{l+1} - \tau_k}\overline{\mu}_{\tilde{\tau}_l:\tilde{\tau}_{l+1}}\left(\overline{\varepsilon}_{\tau_k:\tilde{\tau}_l} - \overline{\varepsilon}_{\tilde{\tau}_l:\tilde{\tau}_{l+1}}\right)\right] \\
&\quad + (\tilde{\tau}_l - \tau_k)\left[\frac{(\tilde{\tau}_l - \tau_k)\,\overline{\varepsilon}^2_{\tau_k:\tilde{\tau}_l} + 2(\tilde{\tau}_{l+1} - \tilde{\tau}_l)\,\overline{\varepsilon}_{\tilde{\tau}_l:\tilde{\tau}_{l+1}}\overline{\varepsilon}_{\tau_k:\tilde{\tau}_l} - (\tilde{\tau}_{l+1} - \tilde{\tau}_l)\,\overline{\varepsilon}^2_{\tilde{\tau}_l:\tilde{\tau}_{l+1}}}{\tilde{\tau}_{l+1} - \tau_k}\right. \\
&\qquad\qquad \left. + \frac{(\tau_k - \tilde{\tau}_{l-1})\,\overline{\varepsilon}^2_{\tilde{\tau}_{l-1}:\tau_k} - 2(\tau_k - \tilde{\tau}_{l-1})\,\overline{\varepsilon}_{\tilde{\tau}_{l-1}:\tau_k}\overline{\varepsilon}_{\tau_k:\tilde{\tau}_l} - (\tilde{\tau}_l - \tau_k)\,\overline{\varepsilon}^2_{\tau_k:\tilde{\tau}_l}}{\tilde{\tau}_l - \tilde{\tau}_{l-1}}\right].
\end{aligned} \tag{43}
$$

We now bound this cost difference. Without loss of generality we assume that the kth jump at τ_k is downwards; otherwise, we can consider $-Y_i$ instead of Y_i which does not change the costs. Thus, since $\overline{\mu}_{\tau_k:\tilde{\tau}_l} = 0 = \beta_k$, we have that $\beta_{k-1} = \Delta_k$, $\beta_{k-2} = \Delta_k \pm \Delta_{k-1}$ and $\beta_{k+1} = \pm\Delta_{k-1}$. Recall that we have assumed $\tau_k < \tilde{\tau}_l$. Thus, if $\tilde{\tau}_{l+1} \le \tau_{k+1}$, then $\overline{\mu}_{\tilde{\tau}_l:\tilde{\tau}_{l+1}} = 0$. If $\tilde{\tau}_{l+1} > \tau_{k+1}$, then

$$
\overline{\mu}_{\tilde{\tau}_l:\tilde{\tau}_{l+1}} \le \frac{\tilde{\tau}_{l+1} - \tau_{k+1}}{\tilde{\tau}_{l+1} - \tilde{\tau}_l}\Delta_{k-1} \le \frac{1}{4}\Delta_k,
$$

since $\tilde{\tau}_{l+1} - \tilde{\tau}_l \ge \tilde{\tau}_{l+1} - \tau_{k+1} + \frac{3}{4}\underline{\lambda}$ and $\tilde{\tau}_{l+1} - \tau_{k+1} \le \frac{1}{4}\underline{\lambda}$. If $\Delta_{k-1} > \Delta_k$, we have also that $\tilde{\tau}_{l+1} - \tau_{k+1} \le C c_{L,k+1}$, since we have ordered the changes in $\mathcal{I}$ according to their size. Hence, it follows from (23) and (39) that

$$
\begin{aligned}
\frac{\overline{\mu}_{\tilde{\tau}_l:\tilde{\tau}_{l+1}}}{\Delta_k} &\le \frac{\tilde{\tau}_{l+1} - \tau_{k+1}}{\tilde{\tau}_{l+1} - \tilde{\tau}_l}\frac{\Delta_{k-1}}{\Delta_k} \le \frac{C c_{L,k+1}\Delta_{k-1}}{\frac{3}{4}\underline{\lambda}\Delta_k} \\
&\le C\frac{(\log(K) \vee 1)\sigma^2}{\frac{3}{4}\underline{\lambda}\Delta^2_{(1)}} + C\frac{\sigma^2\log(\overline{\lambda})}{\frac{3}{4}\underline{\lambda}\Delta^2_{(1)}}\frac{\log\left(\frac{\sigma^2}{\Delta^2_{(1)}} \vee 1\right)}{\log(\overline{\lambda})} \to 0,
\end{aligned}
$$

as $n \to \infty$. Thus for all n sufficiently large, we have $\overline{\mu}_{\tilde{\tau}_l:\tilde{\tau}_{l+1}} \le \frac{1}{4}\Delta_k$ uniformly in k and $\mathcal{U}$.

Using similar arguments, we also have $\overline{\mu}_{\tilde{\tau}_{l-1}:\tau_k} \ge \frac{4}{5}\Delta_k$ for n sufficiently large, uniformly in k and $\mathcal{U}$. Hence for such n,

$$
\frac{\tau_k - \tilde{\tau}_{l-1}}{\tilde{\tau}_l - \tilde{\tau}_{l-1}}\overline{\mu}^2_{\tilde{\tau}_{l-1}:\tau_k} - \frac{\tilde{\tau}_{l+1} - \tilde{\tau}_l}{\tilde{\tau}_{l+1} - \tau_k}\overline{\mu}^2_{\tilde{\tau}_l:\tilde{\tau}_{l+1}} \ge \left(\frac{1}{2}\frac{16}{25} - 1\frac{1}{16}\right)\Delta_k^2 > \frac{1}{4}\Delta_k^2, \tag{44}
$$

since $\tau_k - \tilde{\tau}_{l-1} \ge \tilde{\tau}_l - \tau_k$ as we have chosen l such that $\tilde{\tau}_l \in \mathcal{U}$ is the closest change-point in $\mathcal{U}$ to τ_k.

Now if, for a fixed $\tilde{C} > 0$, the event

$$\Omega(\mathcal{U}, k) := \left\{ |\overline{\varepsilon}_{\tau_k:\tilde{\tau}_l}| \leq \tilde{C}\Delta_k,\ |\overline{\varepsilon}_{\tilde{\tau}_{l-1}:\tau_k}| \leq \tilde{C}\Delta_k \text{ and } \left| \frac{\tilde{\tau}_{l+1} - \tilde{\tau}_l}{\tilde{\tau}_{l+1} - \tau_k} \overline{\varepsilon}^2_{\tilde{\tau}_l:\tilde{\tau}_{l+1}} \right| \leq \tilde{C}^2\Delta_k^2 \right\} \tag{45}$$

occurs, then it follows from similar arguments that

$$\left| \frac{\tau_k - \tilde{\tau}_{l-1}}{\tilde{\tau}_l - \tilde{\tau}_{l-1}} \overline{\mu}_{\tilde{\tau}_{l-1}:\tau_k} \left(\overline{\varepsilon}_{\tilde{\tau}_{l-1}:\tau_k} - \overline{\varepsilon}_{\tau_k:\tilde{\tau}_l} \right) \right| \leq \frac{6}{5}\tilde{C}\Delta_k^2, \tag{46}$$

$$\left| \frac{\tilde{\tau}_{l+1} - \tilde{\tau}_l}{\tilde{\tau}_{l+1} - \tau_k} \overline{\mu}_{\tilde{\tau}_l:\tilde{\tau}_{l+1}} \left(\overline{\varepsilon}_{\tau_k:\tilde{\tau}_l} - \overline{\varepsilon}_{\tilde{\tau}_l:\tilde{\tau}_{l+1}} \right) \right| \leq \frac{1}{4}\tilde{C}\Delta_k^2, \tag{47}$$

and that

$$\begin{aligned} &\left| \frac{(\tilde{\tau}_l - \tau_k)\,\overline{\varepsilon}^2_{\tau_k:\tilde{\tau}_l} + 2\,(\tilde{\tau}_{l+1} - \tilde{\tau}_l)\,\overline{\varepsilon}_{\tilde{\tau}_l:\tilde{\tau}_{l+1}}\overline{\varepsilon}_{\tau_k:\tilde{\tau}_l} - (\tilde{\tau}_{l+1} - \tilde{\tau}_l)\,\overline{\varepsilon}^2_{\tilde{\tau}_l:\tilde{\tau}_{l+1}}}{\tilde{\tau}_{l+1} - \tau_k} \right. \\ &\left. + \frac{(\tau_k - \tilde{\tau}_{l-1})\,\overline{\varepsilon}^2_{\tilde{\tau}_{l-1}:\tau_k} - 2\,(\tau_k - \tilde{\tau}_{l-1})\,\overline{\varepsilon}_{\tilde{\tau}_{l-1}:\tau_k}\overline{\varepsilon}_{\tau_k:\tilde{\tau}_l} - (\tilde{\tau}_l - \tau_k)\,\overline{\varepsilon}^2_{\tau_k:\tilde{\tau}_l}}{\tilde{\tau}_l - \tilde{\tau}_{l-1}} \right| \\ &\leq 8\tilde{C}^2\Delta_k^2. \end{aligned} \tag{48}$$

Suppose then that $\Omega(\mathcal{U}, k)$ (45) holds for $\tilde{C} := \frac{1}{24}$. Then it follows from (43), (44), (46), (47), and (48) that

$$S_Y\left(\tilde{\mathcal{U}}\right) - S_Y\left((\tilde{\mathcal{U}} \setminus \tilde{\tau}_l) \cup \tau_k\right) > (\tilde{\tau}_l - \tau_k)\,\Delta_k^2 \left(\frac{1}{4} - \frac{12}{5}\tilde{C} - \frac{1}{2}\tilde{C} - 8\tilde{C}^2 \right) > 0, \tag{49}$$

with $\tilde{\mathcal{U}}$ as in (42).

Now let

$$\Omega_{3n} := \bigcap_{\substack{\forall\ L \geq K,\\ \forall\ I \subseteq \{1,\ldots,K\},\ I \neq \emptyset,\\ \forall k \in I,\ \forall\ \mathcal{U} \in \mathfrak{T}_L(I)}} \Omega(\mathcal{U}, k)$$

We have that $\Omega_{3n} \subseteq \Omega_{2n}$, so it suffices to show $\mathbb{P}(\Omega_{3n}) \to 1$. To this end, let $\Omega_{3n,+}$ and $\Omega_{3n,-}$ be as Ω_{3n}, but with the absolute value in the definition of $\Omega(\mathcal{U}, k)$ (45) replaced by the positive and negative part. Then, $\mathbb{P}(\Omega_{3n}) \geq 1 - (1 - \mathbb{P}(\Omega_{3n,+})) - (1 - \mathbb{P}(\Omega_{3n,-}))$. Furthermore, it follows from a union bound,

the definition of $\mathfrak{T}_L(\mathcal{I})$ (40) that

$$\begin{aligned}
\mathbb{P}(\Omega_{3n,+}) \geq & 1 - \mathbb{P}\Bigg(\max_{\substack{\mathcal{I}\subseteq\{1,\dots,K\},\\ \mathcal{I}\neq\emptyset}} \max_{k\in\mathcal{I}} \max_{L\geq K} \max_{t-\tau_k > Cc_{L,k}} (\overline{\varepsilon}_{\tau_k:t} - \tilde{C}\Delta_k) > 0 \Bigg) \\
& - \mathbb{P}\Bigg(\max_{\substack{\mathcal{I}\subseteq\{1,\dots,K\},\\ \mathcal{I}\neq\emptyset}} \max_{k\in\mathcal{I}} \max_{L\geq K} \max_{\tau_k - t > Cc_{L,k}} (\overline{\varepsilon}_{t:\tau_k} - \tilde{C}\Delta_k) > 0 \Bigg) \\
& - \mathbb{P}\Bigg(\max_{\substack{\mathcal{I}\subseteq\{1,\dots,K\},\\ \mathcal{I}\neq\emptyset}} \max_{k\in\mathcal{I}} \max_{L> K} \max_{t_2>t_1>\tau_k+Cc_{L,k}} \Big(\frac{t_2-t_1}{t_2-\tau_k} \overline{\varepsilon}^2_{t_1:t_2} - \tilde{C}^2\Delta_k^2 \Big) > 0 \Bigg) \\
& - \mathbb{P}\Bigg(\max_{\substack{\mathcal{I}\subseteq\{1,\dots,K\},\\ \mathcal{I}\neq\emptyset}} \max_{k\in\mathcal{I}} \max_{\substack{t_2>t_1>\tau_k+Cc_{K,k}\\ |\tau_{k+1}-t_2|\leq \underline{\lambda}/4}} \Big(\frac{t_2-t_1}{t_2-\tau_k} \overline{\varepsilon}^2_{t_1:t_2} - \tilde{C}^2\Delta_k^2 \Big) > 0 \Bigg) \\
\geq & 1 - \sum_{k=1}^K \mathbb{P}\Bigg(\max_{L\geq K} \max_{t-\tau_k>Cc_{L,k}} \overline{\varepsilon}_{\tau_k:t} > \tilde{C}\Delta_k \Bigg) - \sum_{k=1}^K \mathbb{P}\Bigg(\max_{L\geq K} \max_{\tau_k - t > Cc_{L,k}} \overline{\varepsilon}_{t:\tau_k} > \tilde{C}\Delta_k \Bigg) \\
& - \sum_{k=1}^K \mathbb{P}\Bigg(\max_{L>K} \max_{t_2>t_1>\tau_k+Cc_{L,k}} \frac{t_2-t_1}{t_2-\tau_k} \overline{\varepsilon}^2_{t_1:t_2} > \tilde{C}^2\Delta_k^2 \Bigg) \\
& - \mathbb{P}\Bigg(\max_{k=1,\dots,K} \max_{\substack{t_2>t_1>\tau_k+Cc_{K,k}\\ |\tau_{k+1}-t_2|\leq \underline{\lambda}/4}} \frac{t_2-t_1}{t_2-\tau_k} \overline{\varepsilon}^2_{t_1:t_2} > \tilde{C}^2\Delta_{(1)}^2 \Bigg).
\end{aligned}$$

The same bound applies for $\mathbb{P}(\Omega_{3n,-})$.

Then, since $c_{L,k} \geq (\log(K)\vee 1)\frac{\sigma^2}{\Delta_k^2}$ $\forall$ $L \geq K$, it follows from Lemma 2 that

$$\begin{aligned}
& \mathbb{P}\Bigg(\max_{L\geq K} \max_{t-\tau_k>Cc_{L,k}} \overline{\varepsilon}_{\tau_k:t} > \tilde{C}\Delta_k \Bigg) + \mathbb{P}\Bigg(\max_{L\geq K} \max_{\tau_k - t > Cc_{L,k}} \overline{\varepsilon}_{t:\tau_k} > \tilde{C}\Delta_k \Bigg) \\
\leq & 2\mathbb{P}\Bigg(\max_{r\geq C\frac{(\log(K)\vee 1)\sigma^2}{\Delta_k^2}} \frac{\sum_{i=1}^r \varepsilon_i/\sigma}{r} \geq \tilde{C}\frac{\Delta_k}{\sigma} \Bigg) \leq 4\exp\Big(-\frac{1}{4} C\tilde{C}^2 (\log(K)\vee 1) \Big).
\end{aligned}$$

Next, since $c_{L,k} = c_{K+1,k} \geq \log\Big(eK\frac{\sigma^2}{\Delta_k^2}\Big)\frac{\sigma^2}{\Delta_k^2}$ $\forall$ $L > K$, it follows from Lemma 3 that for n large enough

$$\begin{aligned}
& \mathbb{P}\Bigg(\max_{L>K} \max_{t_2>t_1>\tau_k+Cc_{L,k}} \frac{t_2-t_1}{t_2-\tau_k} \overline{\varepsilon}^2_{t_1:t_2} > \tilde{C}^2\Delta_k^2 \Bigg) \\
\leq & \mathbb{P}\Bigg(\max_{r\geq Cc_{K+1,k},\, s\geq 1} \frac{\sum_{j=r+1}^{r+s} \varepsilon_j/\sigma}{\sqrt{s}\sqrt{r+s}} \geq \tilde{C}\frac{\Delta_k}{\sigma} \Bigg) \\
\leq & 2C^2 c_{K+1,k}^2 \exp\Bigg(-\frac{1}{2} Cc_{K+1,k}\tilde{C}^2 \frac{\Delta_k^2}{\sigma^2} \Bigg) \\
\leq & 2\exp\Big(-\frac{1}{4} C\tilde{C}^2 (\log(K)\vee 1) \Big).
\end{aligned}$$

Because of sub-Gaussianity and since $|\tau_{k+1} - t_2| \leq \underline{\lambda}/4$ implies $\tau_k + \frac{3}{4}\underline{\lambda} \leq t_2 \leq \tau_k + \overline{\lambda} + \frac{1}{4}\underline{\lambda}$, we have that

$$\max_{k=1,\ldots,K} \max_{\substack{t_2 > t_1 > \tau_k + Cc_{K,k} \\ |\tau_{k+1} - t_2| \leq \underline{\lambda}/4}} \frac{t_2 - t_1}{t_2 - \tau_k} \overline{\varepsilon}^2_{t_1:t_2}$$

$$\leq \max_{k=1,\ldots,K} \max_{\tau_k \leq t_1 \leq t_2 \leq \tau_k + \overline{\lambda} + \frac{1}{4}\underline{\lambda}} \frac{4}{3} \underline{\lambda}^{-1} (t_2 - t_1) \overline{\varepsilon}^2_{t_1:t_2} \leq O_{\mathbb{P}}\left(\underline{\lambda}^{-1} \log(K\overline{\lambda}) \sigma^2 \right).$$

Thus, it follows from (23) that

$$\mathbb{P}\left(\max_{k=1,\ldots,K} \max_{\substack{t_2 > t_1 > \tau_k + Cc_{K,k} \\ |\tau_{k+1} - t_2| \leq \underline{\lambda}/4}} \frac{t_2 - t_1}{t_2 - \tau_k} \overline{\varepsilon}^2_{t_1:t_2} > \tilde{C}^2 \Delta^2_{(1)} \right) \to 0, \text{ as } n \to \infty.$$

Thus,

$$\begin{aligned} \mathbb{P}(\Omega_{3n}) \geq & 1 - 12K \exp\left(-\frac{1}{4} C\tilde{C}^2 (\log(K) \vee 1) \right) + o(1) \\ \geq & 1 - 12 \exp\left(-\frac{1}{4} C\tilde{C}^2 \right) + o(1) \to 1, \end{aligned}$$

since $C \to \infty$, as $n \to \infty$. This completes the proof of (25).

Proof of (26) If $|\hat{\tau} - \tau_k| > \frac{\underline{\lambda}}{4} \, \forall \, \hat{\tau} \in \hat{\mathcal{T}}_L$, then

$$\sum_{i=\tau_k - \frac{\underline{\lambda}}{2}+1}^{\tau_k + \frac{\underline{\lambda}}{2}} (\mu_i - \overline{\mu}_{L,i})^2 \geq \sum_{i=\tau_k - \frac{\underline{\lambda}}{4}+1}^{\tau_k + \frac{\underline{\lambda}}{4}} (\mu_i - \overline{\mu}_{L,i})^2 \geq \frac{\underline{\lambda}}{4} \frac{\Delta_k^2}{4}.$$

Otherwise, let l be such that $\hat{\tau}_{L,l}$ is the closest change-point to τ_k. In case of a tie, we pick the change-point on the right hand side. Without loss of generality let $\hat{\tau}_{L,l} \geq \tau_k$. If this is not true, we can instead consider the observations in reverse order. Additionally, since replacing every μ_i by $\mu_i - c$, with c a global constant, does not change the costs, we can assume $\beta_k = \overline{\mu}_{\tau_k:\hat{\tau}_{L,l}} = 0$ without loss of generality as in our previous argument. Finally, without loss of generality we may assume that the kth jump at τ_k is downwards, as otherwise we can consider $-Y_i$ instead of Y_i, which does not change the costs.

If $\overline{\mu}_{\hat{\tau}_{L,l-1}:\tau_k} > \frac{6}{5}\Delta_k$ or $\overline{\mu}_{\hat{\tau}_{L,l-1}:\tau_k} < \frac{4}{5}\Delta_k$, then $\hat{\tau}_{L,l-1} < \tau_{k-1}$, since otherwise $\overline{\mu}_{\hat{\tau}_{L,l-1}:\tau_k} = \Delta_k$. Hence,

$$\sum_{i=\tau_k - \frac{\underline{\lambda}}{2}+1}^{\tau_k + \frac{\underline{\lambda}}{2}} (\mu_i - \overline{\mu}_{L,i})^2 \geq \sum_{i=\tau_k - \frac{\underline{\lambda}}{4}+1}^{\tau_k} (\mu_i - \overline{\mu}_{L,i})^2 \geq \frac{\underline{\lambda}}{4} \frac{\Delta_k^2}{25}.$$

If $|\overline{\mu}_{\hat{\tau}_{L,l}:\hat{\tau}_{L,l+1}}| > \frac{1}{4}\Delta_k$, then $\hat{\tau}_{L,l+1} > \tau_{k+1}$, since otherwise $\overline{\mu}_{\hat{\tau}_{L,l}:\hat{\tau}_{L,l+1}} = \beta_k = 0$. Hence,

$$\sum_{i=\tau_k - \frac{\underline{\lambda}}{2}+1}^{\tau_k + \frac{\underline{\lambda}}{2}} (\mu_i - \overline{\mu}_{L,i})^2 \geq \sum_{i=\tau_k + \frac{\underline{\lambda}}{4}+1}^{\tau_k + \frac{\underline{\lambda}}{2}} (\mu_i - \overline{\mu}_{L,i})^2 \geq \frac{\underline{\lambda}}{4} \frac{\Delta_k^2}{16}.$$

But, if $\overline{\mu}_{\hat{\tau}_{L,l-1}:\tau_k} \in \left[\frac{4}{5}\Delta_k, \frac{6}{5}\Delta_k\right]$ and $|\overline{\mu}_{\hat{\tau}_{L,l}:\hat{\tau}_{L,l+1}}| \leq \frac{1}{4}\Delta_k$, then the statement in (26) follows from the same arguments as used to show (25).

□

S3. Proofs of Theorem 1

Proof of Theorem 1. Recall that $\tau_1^O = \tau_1^E = (n/2 - \underline{\lambda})/2$, $\tau_2^O = (n/2+1)/2$ and $\tau_2^E = (n/2-1)/2$. We start by showing that

$$\mathbb{P}\left(\hat{\tau}_{1,1}^O = \tau_2^O, \hat{\tau}_{1,1}^E = \tau_2^E\right) \to 1. \tag{50}$$

We have that

$$\left\{\hat{\tau}_{1,1}^O = \tau_2^O\right\} = \left\{S_Y\left(\left\{\tau_0^O, \tau, \tau_3^O\right\}\right) > S_Y\left(\left\{\tau_0^O, \tau_2^O, \tau_3^O\right\}\right) \ \forall\ \tau \neq \tau_2^O\right\}.$$

We distinguish the cases $\tau > \tau_2^O$, $\tau \in [\tau_1^O, \tau_2^O)$ and $\tau < \tau_1^O$. Let $\tau > \tau_2^O$. In order to use Lemma 4, we consider $-(Y_i - \Delta_2)$ instead of Y_i. Note that this does not change the costs. Hence, $\beta_0 = \Delta_2 - \Delta_1$, $\beta_1 = \Delta_2$ and $\beta_2 = 0$. Thus, $\overline{\mu}_{\tau_2^O:\tau} = 0$, $\overline{\mu}_{\tau:\tau_3^O} = 0$ and $\overline{\mu}_{0:\tau_2^O} = \Delta_2 - \tau_1^O\Delta_1/\tau_2^O = c\Delta_2$ for a $c \in (0,1)$, since $\Delta_1 < \Delta_2$. Consequently, Lemma 4 gives us

$$\begin{aligned}
&\left(\tau - \tau_2^O\right)^{-1}\left[S_Y\left(\left\{\tau_0^O, \tau, \tau_3^O\right\}\right) - S_Y\left(\left\{\tau_0^O, \tau_2^O, \tau_3^O\right\}\right)\right] \\
=&\frac{\tau_2^O}{\tau}\overline{\mu}_{0:\tau_2^O}^2 + 2\frac{\tau_2^O}{\tau}\overline{\mu}_{0:\tau_2^O}\left(\overline{\varepsilon}_{0:\tau_2^O} - \overline{\varepsilon}_{\tau_2^O:\tau}\right) \\
&+ \frac{\left(\tau - \tau_2^O\right)\overline{\varepsilon}_{\tau_2^O:\tau}^2 + 2\left(\tau_3^O - \tau\right)\overline{\varepsilon}_{\tau:\tau_3^O}\overline{\varepsilon}_{\tau_2^O:\tau} - \left(\tau_3^O - \tau\right)\overline{\varepsilon}_{\tau:\tau_3^O}^2}{\tau_3^O - \tau_2^O} \\
&+ \frac{\tau_2^O\overline{\varepsilon}_{0:\tau_2^O}^2 - 2\tau_2^O\overline{\varepsilon}_{0:\tau_2^O}\overline{\varepsilon}_{\tau_2^O:\tau} - \left(\tau - \tau_2^O\right)\overline{\varepsilon}_{\tau_2^O:\tau}^2}{\tau} \\
\geq&\frac{1}{2}c^2\Delta_2^2 + O_{\mathbb{P}}\left(\Delta_2\sigma\right) + O_{\mathbb{P}}\left(\sigma^2\right),
\end{aligned}$$

since $\overline{\varepsilon}_{0:\tau_2^O} = O_{\mathbb{P}}(\sigma)$, $\overline{\varepsilon}_{\tau_2^O:\tau} = O_{\mathbb{P}}(\sigma)$, and $\overline{\varepsilon}_{\tau_2^O:\tau} = O_{\mathbb{P}}(\sigma)$ for all $\tau > \tau_2^O$ simultaneously. Then as (9) and (10) imply $\Delta_2^2/\sigma^2 \to \infty$, we have that

$$\frac{1}{\sigma^2}\left(\tau - \tau_2^O\right)^{-1}\left[S_Y\left(\left\{\tau_0^O, \tau, \tau_3^O\right\}\right) - S_Y\left(\left\{\tau_0^O, \tau_2^O, \tau_3^O\right\}\right)\right] \xrightarrow{p} \infty.$$

Next consider $\tau \in [\tau_1^O, \tau_2^O)$. To use Lemma 4, we consider $Y_1^O, \ldots, Y_{n/2}^O$ in reverse order. Note that this does not change the costs. Hence, $\beta_0 = \Delta_2$, $\beta_1 = 0$ and $\beta_2 = \Delta_1$. It follows $\overline{\mu}_{0:\tau_1^O} = \Delta_2$, $\overline{\mu}_{\tau_1^O:\tau} = 0$ and $\overline{\mu}_{\tau:\tau_3^O} = \frac{\tau_3^O - \tau_2^O}{\tau_3^O - \tau}\Delta_1$. Consequently,

Lemma 4 gives us

$$\begin{aligned}
&\left(\tau-\tau_1^O\right)^{-1}\left[S_Y\left(\{\tau_0^O,\tau,\tau_3^O\}\right)-S_Y\left(\{\tau_0^O,\tau_1^O,\tau_3^O\}\right)\right]\\
=&\frac{\tau_1^O}{\tau}\overline{\mu}^2_{0:\tau_1^O}-\frac{\tau_3^O-\tau}{\tau_3^O-\tau_1^O}\overline{\mu}^2_{\tau:\tau_3^O}\\
&+2\frac{\tau_1^O}{\tau}\overline{\mu}_{0:\tau_1^O}\left(\overline{\varepsilon}_{0:\tau_1^O}-\overline{\varepsilon}_{\tau_1^O:\tau}\right)+\frac{\tau_3^O-\tau}{\tau_3^O-\tau_1^O}\overline{\mu}_{\tau:\tau_3^O}\left(\overline{\varepsilon}_{\tau_1^O:\tau}-\overline{\varepsilon}_{\tau:\tau_3^O}\right)\\
&+\frac{\left(\tau-\tau_1^O\right)\overline{\varepsilon}^2_{\tau_1^O:\tau}+2\left(\tau_3^O-\tau\right)\overline{\varepsilon}_{\tau:\tau_3^O}\overline{\varepsilon}_{\tau_1^O:\tau}-\left(\tau_3^O-\tau\right)\overline{\varepsilon}^2_{\tau:\tau_3^O}}{\tau_3^O-\tau_1^O}\\
&+\frac{\tau_1^O\overline{\varepsilon}^2_{0:\tau_1^O}-2\tau_1^O\overline{\varepsilon}_{0:\tau_1^O}\overline{\varepsilon}_{\tau_1^O:\tau}-\left(\tau-\tau_1^O\right)\overline{\varepsilon}^2_{\tau_1^O:\tau}}{\tau}.
\end{aligned}\tag{51}$$

We now return to our original notation, i.e. we consider $Y_1^O,\ldots,Y_{n/2}^O$ in their original order. Then, using similar arguments to earlier, we see there exists a $c\in(0,1)$ such that

$$\begin{aligned}
&\frac{1}{\sigma^2}\left(\tau-\tau_1^O\right)^{-1}\left[S_Y\left(\{\tau_0^O,\tau,\tau_3^O\}\right)-S_Y\left(\{\tau_0^O,\tau_2^O,\tau_3^O\}\right)\right]\\
\geq& c\frac{\Delta_2^2}{\sigma^2}+O_{\mathbb{P}}\left(\frac{\Delta_2}{\sigma}\right)+O_{\mathbb{P}}(1)\xrightarrow{p}\infty.
\end{aligned}$$

Thus,

$$\mathbb{P}\left(S_Y\left(\{\tau_0^O,\tau,\tau_3^O\}\right)>S_Y\left(\{\tau_0^O,\tau_2^O,\tau_3^O\}\right)\ \forall\ \tau\in[\tau_1^O,\tau_2^O)\right)\to 1.$$

The case $\tau<\tau_1^O$ follows from similar arguments. Consequently, $\mathbb{P}\left(\hat{\tau}_{1,1}^O=\tau_2^O\right)\to 1$, as $n\to\infty$. Similarly we may show that $\mathbb{P}\left(\hat{\tau}_{1,1}^E=\tau_2^E\right)\to 1$, as $n\to\infty$, giving (50) as desired.

Next, since $\Delta_2/\sigma\to\infty$, $\underline{\lambda}\Delta_1^2/(\sigma^2\log n)\to\infty$ and $\sqrt{\underline{\lambda}}\Delta_1^2/\sigma^2\to\infty$, as $n\to\infty$, it follows from Theorem 5 that writing $\delta_0:=\lfloor\sqrt{\underline{\lambda}}\rfloor$,

$$\mathbb{P}\big(|\hat{\tau}_{2,1}^O-\tau_1^O|\leq\delta_0,\ |\hat{\tau}_{2,1}^E-\tau_1^E|\leq\delta_0,\ \hat{\tau}_{2,2}^O=\tau_2^O,\ \hat{\tau}_{2,2}^E=\tau_2^E\big)\to 1.\tag{52}$$

We denote by Ω_n the intersection of the events in (50) and (52). We have shown $\mathbb{P}(\Omega_n)\to 1$. In the following we work on the sequence Ω_n.

Observe that $\mu_i^E-\mu_i^O=\Delta_2$ if $i=(n/2+1)/2$, but $\mu_i^E-\mu_i^O=0$ otherwise.

Thus,

$$\mathrm{CV}_{(2)}(L) = \sum_{l=0}^{L} \sum_{i=\hat{\tau}^O_{L,l}+1}^{\hat{\tau}^O_{L,l+1}} \left(Y_i^E - \overline{Y}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}}\right)^2 + \sum_{l=0}^{L} \sum_{i=\hat{\tau}^E_{L,l}+1}^{\hat{\tau}^E_{L,l+1}} \left(Y_i^O - \overline{Y}^E_{\hat{\tau}^E_{L,l}:\hat{\tau}^E_{L,l+1}}\right)^2$$

$$= \sum_{l=0}^{L} \sum_{i=\hat{\tau}^O_{L,l}+1}^{\hat{\tau}^O_{L,l+1}} \left(\varepsilon_i^E + \mu_i^O - \overline{Y}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}}\right)^2 + \Delta_2^2 + 2\Delta_2 \left(\varepsilon^E_{\frac{n}{2}+1} + \mu^O_{\frac{n}{2}+1} - \overline{Y}^O_{\hat{\tau}^O_{L,L-1}:\hat{\tau}^O_{L,L}}\right)$$

$$+ \sum_{l=0}^{L} \sum_{i=\hat{\tau}^E_{L,l}+1}^{\hat{\tau}^E_{L,l+1}} \left(\varepsilon_i^O + \mu_i^E - \overline{Y}^E_{\hat{\tau}^E_{L,l}:\hat{\tau}^E_{L,l+1}}\right)^2 + \Delta_2^2 - 2\Delta_2 \left(\varepsilon^O_{\frac{n}{2}+1} + \mu^E_{\frac{n}{2}+1} - \overline{Y}^E_{\hat{\tau}^E_{L,L}:\hat{\tau}^E_{L,L+1}}\right).$$

Consequently,

$$\mathrm{CV}_{(2)}(2) - \mathrm{CV}_{(2)}(1)$$

$$= \sum_{i=1}^{\hat{\tau}^O_{2,1}} \left(\varepsilon_i^E + \mu_i^O - \overline{Y}^O_{0:\hat{\tau}^O_{2,1}}\right)^2 + \sum_{i=\hat{\tau}^O_{2,1}+1}^{\tau_2^O} \left(\varepsilon_i^E + \mu_i^O - \overline{Y}^O_{\hat{\tau}^O_{2,1}:\tau_2^O}\right)^2 - \sum_{i=1}^{\tau_2^O} \left(\varepsilon_i^E + \mu_i^O - \overline{Y}^O_{0:\tau_2^O}\right)^2$$

$$+ \sum_{i=1}^{\hat{\tau}^E_{2,1}} \left(\varepsilon_i^O + \mu_i^E - \overline{Y}^E_{0:\hat{\tau}^E_{2,1}}\right)^2 + \sum_{i=\hat{\tau}^E_{2,1}+1}^{\tau_2^E} \left(\varepsilon_i^O + \mu_i^E - \overline{Y}^E_{\hat{\tau}^E_{2,1}:\tau_2^E}\right)^2 - \sum_{i=1}^{\tau_2^E} \left(\varepsilon_i^O + \mu_i^E - \overline{Y}^E_{0:\tau_2^E}\right)^2$$

$$+ 2\Delta_2 \left(\overline{Y}^O_{0:\tau_2^O} - \overline{Y}^O_{\hat{\tau}^O_{2,1}:\tau_2^O}\right)$$

$$= 2\sum_{i=1}^{\hat{\tau}^O_{2,1}} \varepsilon_i^E \left(\overline{Y}^O_{0:\tau_2^O} - \overline{Y}^O_{0:\hat{\tau}^O_{2,1}}\right) + 2\sum_{i=\hat{\tau}^O_{2,1}+1}^{\tau_2^O} \varepsilon_i^E \left(\overline{Y}^O_{0:\tau_2^O} - \overline{Y}^O_{\hat{\tau}^O_{2,1}:\tau_2^O}\right)$$

$$+ 2\sum_{i=1}^{\hat{\tau}^E_{2,1}} \varepsilon_i^O \left(\overline{Y}^E_{0:\tau_2^E} - \overline{Y}^E_{0:\hat{\tau}^E_{2,1}}\right) + 2\sum_{i=\hat{\tau}^E_{2,1}+1}^{\tau_2^E} \varepsilon_i^O \left(\overline{Y}^E_{0:\tau_2^E} - \overline{Y}^E_{\hat{\tau}^E_{2,1}:\tau_2^E}\right)$$

$$+ \sum_{i=1}^{\hat{\tau}^O_{2,1}} \left(\mu_i^O - \overline{Y}^O_{0:\hat{\tau}^O_{2,1}}\right)^2 + \sum_{i=\hat{\tau}^O_{2,1}+1}^{\tau_2^O} \left(\mu_i^O - \overline{Y}^O_{\hat{\tau}^O_{2,1}:\tau_2^O}\right)^2 - \sum_{i=1}^{\tau_2^O} \left(\mu_i^O - \overline{Y}^O_{0:\tau_2^O}\right)^2$$

$$+ \sum_{i=1}^{\hat{\tau}^E_{2,1}} \left(\mu_i^E - \overline{Y}^E_{0:\hat{\tau}^E_{2,1}}\right)^2 + \sum_{i=\hat{\tau}^E_{2,1}+1}^{\tau_2^E} \left(\mu_i^E - \overline{Y}^E_{\hat{\tau}^E_{2,1}:\tau_2^E}\right)^2 - \sum_{i=1}^{\tau_2^E} \left(\mu_i^E - \overline{Y}^E_{0:\tau_2^E}\right)^2$$

$$+ 2\Delta_2 \left(\overline{Y}^O_{0:\tau_2^O} - \overline{Y}^O_{\hat{\tau}^O_{2,1}:\tau_2^O}\right) =: A_n.$$

We have shown $\{\mathrm{CV}_{(2)}(2) - \mathrm{CV}_{(2)}(1)\}\mathbb{1}_{\Omega_n} = A_n\mathbb{1}_{\Omega_n}$. Note that to complete the proof it suffices to show that $\mathbb{P}(A_n\mathbb{1}_{\Omega_n} > 0) \to 1$ as $n \to \infty$.

Now let $A_n = A_{1,n} + A_{2,n}$, with

$$\begin{aligned} A_{1,n} :=& 2\sum_{i=1}^{\hat{\tau}^O_{2,1}} \varepsilon_i^E \left(\overline{Y}^O_{0:\tau_2^O} - \overline{Y}^O_{0:\hat{\tau}^O_{2,1}}\right) + 2\sum_{i=\hat{\tau}^O_{2,1}+1}^{\tau_2^O} \varepsilon_i^E \left(\overline{Y}^O_{0:\tau_2^O} - \overline{Y}^O_{\hat{\tau}^O_{2,1}:\tau_2^O}\right) \\ &+ 2\sum_{i=1}^{\hat{\tau}^E_{2,1}} \varepsilon_i^O \left(\overline{Y}^E_{0:\tau_2^E} - \overline{Y}^E_{0:\hat{\tau}^E_{2,1}}\right) + 2\sum_{i=\hat{\tau}^E_{2,1}+1}^{\tau_2^E} \varepsilon_i^O \left(\overline{Y}^E_{0:\tau_2^E} - \overline{Y}^E_{\hat{\tau}^E_{2,1}:\tau_2^E}\right) \end{aligned}$$

and $A_{2,n} := A_n - A_{1,n}$. We start by bounding $A_{1,n}\mathbb{1}_{\Omega_n}$. Let us consider the first two terms of $A_{1,n}\mathbb{1}_{\Omega_n}$. It follows from the total law of probability that

$$\begin{aligned} &\left(\sum_{i=1}^{\hat{\tau}^O_{2,1}} \varepsilon_i^E \left(\overline{Y}^O_{0:\tau_2^O} - \overline{Y}^O_{0:\hat{\tau}^O_{2,1}}\right) + \sum_{i=\hat{\tau}^O_{2,1}+1}^{\tau_2^O} \varepsilon_i^E \left(\overline{Y}^O_{0:\tau_2^O} - \overline{Y}^O_{\hat{\tau}^O_{2,1}:\tau_2^O}\right)\right)\mathbb{1}_{\Omega_n} \\ =& O_{\mathbb{P}}\left(\sigma\left[\hat{\tau}^O_{2,1}\left(\overline{Y}^O_{0:\tau_2^O} - \overline{Y}^O_{0:\hat{\tau}^O_{2,1}}\right)^2 + \left(\tau_2^O - \hat{\tau}^O_{2,1}\right)\left(\overline{Y}^O_{0:\tau_2^O} - \overline{Y}^O_{\hat{\tau}^O_{2,1}:\tau_2^O}\right)^2\right]^{1/2}\mathbb{1}_{\Omega_n}\right), \end{aligned}$$

since conditional on ε^O, $\varepsilon_i^E \left(\overline{Y}^O_{0:\tau_2^O} - \overline{Y}^O_{0:\hat{\tau}^O_{2,1}}\right)$ is centred Gaussian with variance $\sigma^2\left(\overline{Y}^O_{0:\tau_2^O} - \overline{Y}^O_{0:\hat{\tau}^O_{2,1}}\right)^2$. Next,

$$\begin{aligned} &\left[\hat{\tau}^O_{2,1}\left(\overline{Y}^O_{0:\tau_2^O} - \overline{Y}^O_{0:\hat{\tau}^O_{2,1}}\right)^2 + \left(\tau_2^O - \hat{\tau}^O_{2,1}\right)\left(\overline{Y}^O_{0:\tau_2^O} - \overline{Y}^O_{\hat{\tau}^O_{2,1}:\tau_2^O}\right)^2\right]\mathbb{1}_{\Omega_n} \\ \leq& 2\left[\hat{\tau}^O_{2,1}\left(\overline{\mu}^O_{0:\tau_2^O} - \overline{\mu}^O_{0:\hat{\tau}^O_{2,1}}\right)^2 + \left(\tau_2^O - \hat{\tau}^O_{2,1}\right)\left(\overline{\mu}^O_{0:\tau_2^O} - \overline{\mu}^O_{\hat{\tau}^O_{2,1}:\tau_2^O}\right)^2\right]\mathbb{1}_{\Omega_n} \\ &+ 2\left[\hat{\tau}^O_{2,1}\left(\overline{\varepsilon}^O_{0:\tau_2^O} - \overline{\varepsilon}^O_{0:\hat{\tau}^O_{2,1}}\right)^2 + \left(\tau_2^O - \hat{\tau}^O_{2,1}\right)\left(\overline{\varepsilon}^O_{0:\tau_2^O} - \overline{\varepsilon}^O_{\hat{\tau}^O_{2,1}:\tau_2^O}\right)^2\right]\mathbb{1}_{\Omega_n} \\ \leq& 2\left[\max_{\tau\in\{1,\ldots,\tau_2^O-1\}}\left\{\tau\left(\overline{\mu}^O_{0:\tau_2^O} - \overline{\mu}^O_{0:\tau}\right)^2 + \left(\tau_2^O - \tau\right)\left(\overline{\mu}^O_{0:\tau_2^O} - \overline{\mu}^O_{\tau:\tau_2^O}\right)^2\right\}\right] \\ &+ 4\left[\hat{\tau}^O_{2,1}\left(\overline{\varepsilon}^O_{0:\hat{\tau}^O_{2,1}}\right)^2 + \tau_2^O\left(\overline{\varepsilon}^O_{0:\tau_2^O}\right)^2 + \left(\tau_2^O - \hat{\tau}^O_{2,1}\right)\left(\overline{\varepsilon}^O_{\hat{\tau}^O_{2,1}:\tau_2^O}\right)^2\right]\mathbb{1}_{\Omega_n} \\ \leq& 2\left[\tau_1^O\left(\overline{\mu}^O_{0:\tau_2^O} - \overline{\mu}^O_{0:\tau_1^O}\right)^2 + \left(\tau_2^O - \tau_1^O\right)\left(\overline{\mu}^O_{0:\tau_2^O} - \overline{\mu}^O_{\tau_1^O:\tau_2^O}\right)^2\right] \\ &+ 4\left[\hat{\tau}^O_{2,1}\left(\overline{\varepsilon}^O_{0:\hat{\tau}^O_{2,1}}\right)^2 + \tau_2^O\left(\overline{\varepsilon}^O_{0:\tau_2^O}\right)^2 + \left(\tau_2^O - \hat{\tau}^O_{2,1}\right)\left(\overline{\varepsilon}^O_{\hat{\tau}^O_{2,1}:\tau_2^O}\right)^2\right]\mathbb{1}_{\Omega_n}. \end{aligned}$$

We now bound the last two terms separately:

$$\begin{aligned}
&\tau_1^O \left(\overline{\mu}^O_{0:\tau_2^O} - \overline{\mu}^O_{0:\tau_1^O}\right)^2 + \left(\tau_2^O - \tau_1^O\right)\left(\overline{\mu}^O_{0:\tau_2^O} - \overline{\mu}^O_{\tau_1^O:\tau_2^O}\right)^2 \\
=&\tau_1^O \left(\frac{\tau_1^O}{\tau_2^O}\Delta_1 - \Delta_1\right)^2 + \left(\tau_2^O - \tau_1^O\right)\left(\frac{\tau_1^O}{\tau_2^O}\Delta_1\right)^2 \\
=&\Delta_1^2 \frac{\tau_1^O\left(\tau_2^O - \tau_1^O\right)}{\tau_2^O} = O\left(\Delta_1^2 \underline{\lambda}(1 - 2\underline{\lambda}/n)\right).
\end{aligned}$$

Next, since $\tau_1^O > \tau_2^O - \tau_1^O \geq (\underline{\lambda} - 1)/2$,

$$\begin{aligned}
&\left(\hat{\tau}^O_{2,1}\right)^{\frac{1}{2}} \left|\overline{\varepsilon}^O_{0:\hat{\tau}^O_{2,1}}\right| \mathbb{1}_{\Omega_n} \leq \max_{\tau \in \{\tau_1^O - \delta_0, \ldots, \tau_1^O + \delta_0\}} \tau^{-\frac{1}{2}} \left|\sum_{i=1}^{\tau} \varepsilon_i^O\right| \\
\leq &\left(\tau_1^O - \delta_0\right)^{-\frac{1}{2}} \left(\left|\sum_{i=1}^{\tau_1^O} \varepsilon_i^O\right| + \max_{\tau \in \{\tau_1^O - \delta_0, \ldots, \tau_1^O - 1\}} \left|\sum_{i=\tau}^{\tau_1^O - 1} \varepsilon_i^O\right| + \max_{\tau \in \{\tau_1^O + 1, \ldots, \tau_1^O + \delta_0\}} \left|\sum_{i=\tau_1^O + 1}^{\tau} \varepsilon_i^O\right|\right) \\
\leq &O_{\mathbb{P}}\left(\sigma\left(\tau_1^O - \sqrt{\underline{\lambda}}\right)^{-\frac{1}{2}} \left(\sqrt{\tau_1^O} + \sqrt{\sqrt{\underline{\lambda}}\log\underline{\lambda}}\right)\right) \leq O_{\mathbb{P}}(\sigma).
\end{aligned}$$

Similarly,

$$\left(\tau_2^O - \hat{\tau}^O_{2,1}\right)^{\frac{1}{2}} \left|\overline{\varepsilon}^O_{\hat{\tau}^O_{2,1}:\tau_2^O}\right| \mathbb{1}_{\Omega_n} \leq O_{\mathbb{P}}(\sigma). \tag{53}$$

Finally,

$$\tau_2^O \left(\overline{\varepsilon}^O_{0:\tau_2^O}\right)^2 \mathbb{1}_{\Omega_n} \leq \tau_2^O \left(\overline{\varepsilon}^O_{0:\tau_2^O}\right)^2 = O_{\mathbb{P}}\left(\sigma^2\right).$$

The same bounds can be achieved for the last two terms in $A_{1,n}\mathbb{1}_{\Omega_n}$ by interchanging O's and E's. Thus, we obtain

$$A_{1,n}\mathbb{1}_{\Omega_n} = O_{\mathbb{P}}\left(\sigma\sqrt{\Delta_1^2\underline{\lambda}(1 - 2\underline{\lambda}/n) + \sigma^2}\right) = O_{\mathbb{P}}\left(\sigma\sqrt{\Delta_1^2\underline{\lambda}(1 - 2\underline{\lambda}/n)}\right), \tag{54}$$

the last equality following from (9).

We now consider $A_{2,n}$:

$$\begin{aligned}
A_{2,n} =& \sum_{i=1}^{\hat{\tau}^O_{2,1}} \left(\mu^O_i - \overline{Y}^O_{0:\hat{\tau}^O_{2,1}}\right)^2 + \sum_{i=\hat{\tau}^O_{2,1}+1}^{\tau^O_2} \left(\mu^O_i - \overline{Y}^O_{\hat{\tau}^O_{2,1}:\tau^O_2}\right)^2 - \sum_{i=1}^{\tau^O_2} \left(\mu^O_i - \overline{Y}^O_{0:\tau^O_2}\right)^2 \\
&+ \sum_{i=1}^{\hat{\tau}^E_{2,1}} \left(\mu^E_i - \overline{Y}^E_{0:\hat{\tau}^E_{2,1}}\right)^2 + \sum_{i=\hat{\tau}^E_{2,1}+1}^{\tau^E_2} \left(\mu^E_i - \overline{Y}^E_{\hat{\tau}^E_{2,1}:\tau^E_2}\right)^2 - \sum_{i=1}^{\tau^E_2} \left(\mu^E_i - \overline{Y}^E_{0:\tau^E_2}\right)^2 \\
&+ 2\Delta_2 \left(\overline{Y}^O_{0:\tau^O_2} - \overline{Y}^O_{\hat{\tau}^O_{2,1}:\tau^O_2}\right) \\
\geq& 2\Delta_2 \left(\overline{Y}^O_{0:\tau^O_2} - \overline{Y}^O_{\hat{\tau}^O_{2,1}:\tau^O_2}\right) - \sum_{i=1}^{\tau^O_2} \left(\mu^O_i - \overline{Y}^O_{0:\tau^O_2}\right)^2 - \sum_{i=1}^{\tau^E_2} \left(\mu^E_i - \overline{Y}^E_{0:\tau^E_2}\right)^2 \\
\geq& 2\Delta_2 \left(\overline{\mu}^O_{0:\tau^O_2} - \overline{\mu}^O_{\hat{\tau}^O_{2,1}:\tau^O_2} - \left|\overline{\varepsilon}^O_{0:\tau^O_2}\right| - \left|\overline{\varepsilon}^O_{\hat{\tau}^O_{2,1}:\tau^O_2}\right|\right) \\
&- \sum_{i=1}^{\tau^O_2} \left(\mu^O_i - \overline{\mu}^O_{0:\tau^O_2}\right)^2 - 2\sqrt{\sum_{i=1}^{\tau^O_2} \left(\mu^O_i - \overline{\mu}^O_{0:\tau^O_2}\right)^2} \sqrt{\tau^O_2 \left(\overline{\varepsilon}^O_{0:\tau^O_2}\right)^2} - \tau^O_2 \left(\overline{\varepsilon}^O_{0:\tau^O_2}\right)^2 \\
&- \sum_{i=1}^{\tau^E_2} \left(\mu^E_i - \overline{\mu}^E_{0:\tau^E_2}\right)^2 - 2\sqrt{\sum_{i=1}^{\tau^E_2} \left(\mu^E_i - \overline{\mu}^E_{0:\tau^E_2}\right)^2} \sqrt{\tau^E_2 \left(\overline{\varepsilon}^E_{0:\tau^E_2}\right)^2} - \tau^E_2 \left(\overline{\varepsilon}^E_{0:\tau^E_2}\right)^2 .
\end{aligned}$$

We bound the terms separately. Using (53) gives us

$$\begin{aligned}
& 2\Delta_2 \left(\overline{\mu}^O_{0:\tau^O_2} - \overline{\mu}^O_{\hat{\tau}^O_{2,1}:\tau^O_2} - \left|\overline{\varepsilon}^O_{0:\tau^O_2}\right| - \left|\overline{\varepsilon}^O_{\hat{\tau}^O_{2,1}:\tau^O_2}\right|\right) \mathbb{1}_{\Omega_n} \\
\geq& 2\Delta_2 \left(\overline{\mu}^O_{0:\tau^O_2} - \overline{\mu}^O_{\tau^O_1 - \sqrt{\underline{\lambda}}:\tau^O_2} - \left|\overline{\varepsilon}^O_{0:\tau^O_2}\right| - \left(\tau^O_2 - \tau^O_1 - \delta_0\right)^{-\frac{1}{2}} \left(\tau^O_2 - \hat{\tau}^O_{2,1}\right)^{\frac{1}{2}} \left|\overline{\varepsilon}^O_{\hat{\tau}^O_{2,1}:\tau^O_2}\right| \mathbb{1}_{\Omega_n}\right) \\
\geq& 2\Delta_2 \left[\frac{\tau^O_1}{\tau^O_2}\Delta_1 - \frac{\sqrt{\underline{\lambda}}}{\tau^O_2 - \tau^O_1 + \sqrt{\underline{\lambda}}}\Delta_1 + O_{\mathbb{P}}\left(\sigma\left(\tau^O_2\right)^{-\frac{1}{2}}\right) + O_{\mathbb{P}}\left(\sigma\left(\tau^O_2 - \tau^O_1 - \sqrt{\underline{\lambda}}\right)^{-\frac{1}{2}}\right)\right].
\end{aligned}$$

Next we have,

$$\sum_{i=1}^{\tau^O_2} \left(\mu^O_i - \overline{\mu}^O_{0:\tau^O_2}\right)^2 \mathbb{1}_{\Omega_n} \leq \sum_{i=1}^{\tau^O_2} \left(\mu^O_i - \overline{\mu}^O_{0:\tau^O_2}\right)^2 = \frac{\tau^O_1\left(\tau^O_2 - \tau^O_1\right)}{\tau^O_2}\Delta_1^2 = O\left(\Delta_1^2 \underline{\lambda}(1 - 2\underline{\lambda}/n)\right)$$

and

$$\tau^O_2 \left(\overline{\varepsilon}^O_{0:\tau^O_2}\right)^2 \mathbb{1}_{\Omega_n} \leq \tau^O_2 \left(\overline{\varepsilon}^O_{0:\tau^O_2}\right)^2 = O_{\mathbb{P}}\left(\sigma^2\right).$$

The same bounds can be achieved when interchanging O's and E's. Now as

$$\frac{\tau^O_1}{\tau^O_2} = \frac{n/2 - \underline{\lambda}}{n/2 + 1} = \left(1 - \frac{2\underline{\lambda}}{n}\right)\left(1 - \frac{1}{n/2+1}\right) = \left(1 - \frac{2\underline{\lambda}}{n}\right)(1 + o(1)),$$

$\sqrt{\underline{\lambda}} = O\left(\sqrt{\tau_2^O - \tau_1^O}\right)$, $\tau_1^O > \tau_2^O - \tau_1^O$, and by assumption $\underline{\lambda}\Delta_1^2/\sigma^2 \to \infty$, it follows that

$$\begin{aligned} A_{2,n}\mathbb{1}_{\Omega_n} \geq & 2\Delta_2 \left[\frac{\tau_1^O}{\tau_2^O}\Delta_1 - \frac{\sqrt{\underline{\lambda}}}{\tau_2^O - \tau_1^O + \sqrt{\underline{\lambda}}}\Delta_1 + O_{\mathbb{P}}\left(\sigma\left(\tau_2^O\right)^{-\frac{1}{2}}\right) + O_{\mathbb{P}}\left(\sigma\left(\tau_2^O - \tau_1^O - \sqrt{\underline{\lambda}}\right)^{-\frac{1}{2}}\right)\right] \\ & + O\left(\Delta_1^2\underline{\lambda}(1 - 2\underline{\lambda}/n)\right) + O_{\mathbb{P}}\left(\Delta_1^2\sqrt{\underline{\lambda}(1 - 2\underline{\lambda}/n)}\sigma\right) + O_{\mathbb{P}}\left(\sigma^2\right) \\ = & \Delta_1^2\left(1 - \frac{2\underline{\lambda}}{n}\right)\left[2\frac{\Delta_2}{\Delta_1} - \underline{\lambda}\right](1 + o_{\mathbb{P}}(1)). \end{aligned}$$

This and (54) then imply

$$\lim_{n\to\infty} \mathbb{P}\left(A_{1,n}\mathbb{1}_{\Omega_n} + A_{2,n}\mathbb{1}_{\Omega_n} > 0\right) = 1,$$

since the assumptions of the theorem imply

$$\frac{\Delta_1^2\left(1 - 2\underline{\lambda}/n\right)\left[2\Delta_2/\Delta_1 - \underline{\lambda}\right]}{\left(\Delta_1^2\sigma^2\underline{\lambda}\left(1 - 2\underline{\lambda}/n\right)\right)^{\frac{1}{2}}} = \frac{\Delta_1}{\sigma}\sqrt{\underline{\lambda}}\sqrt{1 - \frac{2\underline{\lambda}}{n}}\left(\frac{2}{\underline{\lambda}}\frac{\Delta_2}{\Delta_1} - 1\right) \to \infty.$$

This completes the proof of $\mathbb{P}(\hat{K} = 2) \to 0$.

It remains to show (11). We define $\hat{f}_L : [0,1] \to \mathbb{R}$, $t \mapsto \sum_{l=0}^{L} \overline{Y}_{\hat{\tau}_{L,l}^O:\hat{\tau}_{L,l+1}^O} \mathbb{1}_{(\hat{\tau}_{L,l}^O/n, \hat{\tau}_{L,l+1}^O/n]}(t)$. Given $K_{\max} = 2$, it is straightforward to see that $\mathbb{P}(\hat{K} = K) \to 1$, as $n \to \infty$. Secondly, with the same arguments as before, we see that $\mathbb{P}\left(\hat{\tau}_{1,1} = \tau_2\right) \to 1$. Moreover, a straightforward calculation shows that the L_2-loss minimizer and hence $\hat{f}_1$ satisfies

$$\int_0^1 \left(\hat{f}_1(t) - f(t)\right)^2 dt \mathbb{1}_{\{\hat{\tau}_{1,1}=\tau_2\}} \geq \left(1 - \frac{\underline{\lambda}}{n}\right)\frac{\underline{\lambda}\Delta_1^2}{n}\mathbb{1}_{\hat{\tau}_{1,1}=\tau_2} \geq \frac{\underline{\lambda}\Delta_1^2}{2n}\mathbb{1}_{\{\hat{\tau}_{1,1}=\tau_2\}} \quad \text{a.s.}$$

Thus, (11) follows from the following calculation

$$\begin{aligned} & \mathbb{P}\left(\left[\int_0^1 \left(\hat{f}_{\hat{K}}(t) - f(t)\right)^2 dt\right]^{-1} \leq \frac{2n}{\underline{\lambda}\Delta_1^2}\right) \\ = & \mathbb{P}\left(\int_0^1 \left(\hat{f}_{\hat{K}}(t) - f(t)\right)^2 dt \geq \frac{\underline{\lambda}\Delta_1^2}{2n}\right) \\ = & \mathbb{P}\left(\int_0^1 \left(\hat{f}_1(t) - f(t)\right)^2 dt \geq \frac{\underline{\lambda}\Delta_1^2}{2n}, \hat{\tau}_{1,1} = \tau_2\right) - \mathbb{P}(\hat{K} \neq 1) - \mathbb{P}(\hat{\tau}_{1,1} \neq \tau_2) \to 1, \end{aligned}$$

as $n \to \infty$. □

S4. Proof of Theorem 2

Proof of Theorem 2. We focus to begin on the first term of the cross-validation criterion:

$$\mathrm{CV}_{(2)}^O(L) := \sum_{l=0}^{L} \sum_{i=\hat{\tau}_{L,l}^O+1}^{\hat{\tau}_{L,l+1}^O} \left(Y_i^E - \overline{Y}_{\hat{\tau}_{L,l}^O:\hat{\tau}_{L,l+1}^O}^O\right)^2.$$

We now briefly outline how our arguments proceed. We begin by showing $\{\mathrm{CV}^O_{(2)}(1) - \mathrm{CV}^O_{(2)}(2)\}\mathbb{1}_{\Omega^O_1 \cap \Omega^O_2} = A^O_n \mathbb{1}_{\Omega^O_1 \cap \Omega^O_2}$ for some events Ω^O_1 and Ω^O_2 and a random variable A^O_n. We then lower bound the probability of the event $\{A^O_n > 0\} \cap \Omega^O_1 \cap \Omega^O_2$. To this end, we split the event $\{A^O_n > 0\}$ into further events $\Omega^O_3, \ldots, \Omega^O_{10}$. The proof concludes by symmetry arguments to include the second term of the cross-validation criterion, which we denote with a superscript E.

First observe that $\tau^O_1 = (n/2+1)/2$ and $\mu^E_i - \mu^O_i = \Delta_1$ if $i = (n/2+1)/2$, but $\mu^E_i - \mu^O_i = 0$ otherwise. We now define the events

$$\Omega^O_1 := \{\hat{\tau}^O_{1,1} = \tau^O_1\} \text{ and } \Omega^O_2 := \{\hat{\tau}^O_{2,2} = \tau^O_1\}.$$

In the following, we work on $\Omega^O_1 \cap \Omega^O_2$. We have,

$$\begin{aligned}
\mathrm{CV}^O_{(2)}(1) &= \sum_{i=1}^{\tau^O_1} \left(Y^E_i - \overline{Y}^O_{0:\tau^O_1}\right)^2 + \sum_{i=\tau^O_1+1}^{\tau^O_2} \left(Y^E_i - \overline{Y}^O_{\tau^O_1:\tau^O_2}\right)^2 \\
&= \sum_{i=1}^{\tau^O_1} \left(\varepsilon^E_i - \overline{\varepsilon}^O_{0:\tau^O_1}\right)^2 + \Delta_1^2 + 2\Delta_1 \left(\varepsilon^E_{\tau^O_1} - \overline{\varepsilon}^O_{0:\tau^O_1}\right) + \sum_{i=\tau^O_1+1}^{\tau^O_2} \left(Y^E_i - \overline{Y}^O_{\tau^O_1:\tau^O_2}\right)^2
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{CV}^O_{(2)}(2) &= \sum_{i=1}^{\tilde{\tau}^O_{2,1}} \left(Y^E_i - \overline{Y}^O_{0:\tilde{\tau}^O_{2,1}}\right)^2 + \sum_{i=\tilde{\tau}^O_{2,1}+1}^{\tau^O_1} \left(Y^E_i - \overline{Y}^O_{\tilde{\tau}^O_{2,1}:\tau^O_1}\right)^2 + \sum_{i=\tau^O_1+1}^{\tau^O_2} \left(Y^E_i - \overline{Y}^O_{\tau^O_1:\tau^O_2}\right)^2 \\
&= \sum_{i=1}^{\tilde{\tau}^O_{2,1}} \left(\varepsilon^E_i - \overline{\varepsilon}^O_{0:\tilde{\tau}^O_{2,1}}\right)^2 + \sum_{i=\tilde{\tau}^O_{2,1}+1}^{\tau^O_1} \left(\varepsilon^E_i - \overline{\varepsilon}^O_{\tilde{\tau}^O_{2,1}:\tau^O_1}\right)^2 \\
&\quad + \Delta_1^2 + 2\Delta_1 \left(\varepsilon^E_{\tau^O_1} - \overline{\varepsilon}^O_{\tilde{\tau}^O_{2,1}:\tau^O_1}\right) + \sum_{i=\tau^O_1+1}^{\tau^O_2} \left(Y^E_i - \overline{Y}^O_{\tau^O_1:\tau^O_2}\right)^2,
\end{aligned}$$

with

$$\begin{aligned}
\tilde{\tau}^O_{2,1} &:= \operatorname*{argmin}_{\tau=1,\ldots,\tau^O_1-1} \left\{ \sum_{i=1}^{\tau} \left(\varepsilon^O_i - \overline{\varepsilon}^O_{0:\tau}\right)^2 + \sum_{\tau+1}^{\tau^O_1} \left(\varepsilon^O_i - \overline{\varepsilon}^O_{\tau:\tau^O_1}\right)^2 \right\} \\
&= \operatorname*{argmin}_{\tau=1,\ldots,\tau^O_1-1} \left\{ \sum_{i=1}^{\tau^O_1} \left(\varepsilon^O_i\right)^2 - \tau \left(\overline{\varepsilon}^O_{0:\tau}\right)^2 - \left(\tau^O_1 - \tau\right) \left(\overline{\varepsilon}^O_{\tau:\tau^O_1}\right)^2 \right\} \qquad (55) \\
&= \operatorname*{argmax}_{\tau=1,\ldots,\tau^O_1-1} \left\{ \tau \left(\overline{\varepsilon}^O_{0:\tau}\right)^2 + \left(\tau^O_1 - \tau\right) \left(\overline{\varepsilon}^O_{\tau:\tau^O_1}\right)^2 \right\}.
\end{aligned}$$

Hence,

$$\mathrm{CV}^O_{(2)}(1) - \mathrm{CV}^O_{(2)}(2) = \sum_{i=1}^{\tau_1^O} \left(\varepsilon_i^E - \overline{\varepsilon}^O_{0:\tau_1^O}\right)^2 - \sum_{i=1}^{\tilde{\tau}_{2,1}^O} \left(\varepsilon_i^E - \overline{\varepsilon}^O_{0:\tilde{\tau}_{2,1}^O}\right)^2 - \sum_{i=\tilde{\tau}_{2,1}^O+1}^{\tau_1^O} \left(\varepsilon_i^E - \overline{\varepsilon}^O_{\tilde{\tau}_{2,1}^O:\tau_1^O}\right)^2 + 2\Delta_1 \left(\overline{\varepsilon}^O_{\tilde{\tau}_{2,1}^O:\tau_1^O} - \overline{\varepsilon}^O_{0:\tau_1^O}\right) =: A_n^O.$$

To summarise, we have shown

$$\{\mathrm{CV}^O_{(2)}(1) - \mathrm{CV}^O_{(2)}(2)\}\mathbb{1}_{\Omega_1^O \cap \Omega_2^O} = A_n^O \mathbb{1}_{\Omega_1^O \cap \Omega_2^O}. \tag{56}$$

From (13) we have that

$$r_n := \frac{\Delta_1}{\sigma\sqrt{n \log\log n}} \to \infty, \quad \text{as } n \to \infty.$$

Then, let

$$\Omega_3^O := \left\{ \overline{\varepsilon}^O_{\tilde{\tau}_{2,1}^O:\tau_1^O} > s_1\sigma\sqrt{\frac{\log\log n}{n}} \right\},$$

$$\Omega_4^O := \left\{ \left(\tau_1^O\right)^{1/2} \overline{\varepsilon}^O_{0:\tau_1^O} \le s_2\sigma \right\},$$

$$\Omega_5^O := \left\{ \tilde{\tau}_{2,1}^O \left(\overline{\varepsilon}^O_{0:\tilde{\tau}_{2,1}^O}\right)^2 + \left(\tau_1^O - \tilde{\tau}_{2,1}^O\right)\left(\overline{\varepsilon}^O_{\tilde{\tau}_{2,1}^O:\tau_1^O}\right)^2 \le s_3\sigma^2 \log\log n \right\},$$

$$\Omega_6^O := \left\{ \left(\tau_1^O\right)^{1/2} \overline{\varepsilon}^E_{0:\tau_1^O} \le s_4\sigma,\ \left(\tilde{\tau}_{2,1}^O\right)^{1/2} \overline{\varepsilon}^E_{0:\tilde{\tau}_{2,1}^O} \le s_4\sigma,\ \text{and}\ \left(\tau_1^O - \tilde{\tau}_{2,1}^O\right)^{1/2} \overline{\varepsilon}^E_{\tilde{\tau}_{2,1}^O:\tau_1^O} \le s_4\sigma \right\},$$

with $s_1 := \max(r_n^{-1/2}, (\log\log n)^{-1/4})$, $s_2 := (\log\log n)^{1/8}$, $s_3 := (r_n s_1)^{1/2}$ and $s_4 := (\log\log n)^{1/4}$. Hence, $s_1 \to 0$, $s_2 \to \infty$, $s_3 \to \infty$, $s_4 \to \infty$, as $n \to \infty$. Moreover, we have that

$$\begin{aligned}
&\sum_{i=1}^{\tau_1^O} \left(\varepsilon_i^E - \overline{\varepsilon}^O_{0:\tau_1^O}\right)^2 - \sum_{i=1}^{\tilde{\tau}_{2,1}^O} \left(\varepsilon_i^E - \overline{\varepsilon}^O_{0:\tilde{\tau}_{2,1}^O}\right)^2 - \sum_{i=\tilde{\tau}_{2,1}^O+1}^{\tau_1^O} \left(\varepsilon_i^E - \overline{\varepsilon}^O_{\tilde{\tau}_{2,1}^O:\tau_1^O}\right)^2 \\
=&\tau_1^O \left(\overline{\varepsilon}^O_{0:\tau_1^O}\right)^2 - \tilde{\tau}_{2,1}^O \left(\overline{\varepsilon}^O_{0:\tilde{\tau}_{2,1}^O}\right)^2 - \left(\tau_1^O - \tilde{\tau}_{2,1}^O\right)\left(\overline{\varepsilon}^O_{\tilde{\tau}_{2,1}^O:\tau_1^O}\right)^2 \\
&- 2\left[\tau_1^O \overline{\varepsilon}^E_{0:\tau_1^O} \overline{\varepsilon}^O_{0:\tau_1^O} - \tilde{\tau}_{2,1}^O \overline{\varepsilon}^E_{0:\tilde{\tau}_{2,1}^O} \overline{\varepsilon}^O_{0:\tilde{\tau}_{2,1}^O} - \left(\tau_1^O - \tilde{\tau}_{2,1}^O\right) \overline{\varepsilon}^E_{\tilde{\tau}_{2,1}^O:\tau_1^O} \overline{\varepsilon}^O_{\tilde{\tau}_{2,1}^O:\tau_1^O}\right].
\end{aligned}$$

Thus, for n large enough,

$$\bigcap_{j=3,4,5,6} \Omega_j^O \subseteq \{A_n^O > 0\}. \tag{57}$$

Next, let

$$\Omega_7^O := \left\{ \max_{t<\tau_1^O} \left\{ t\left(\overline{\varepsilon}_{0:t}^O\right)^2 + (\tau_1^O - t)\left(\overline{\varepsilon}_{t:\tau_1^O}^O\right)^2 \right\} > \max_{\tau_1^O<t<n/2} \left\{ (t-\tau_1^O)\left(\overline{\varepsilon}_{\tau_1^O:t}^O\right)^2 + (n/2-t)\left(\overline{\varepsilon}_{t:n/2}^O\right)^2 \right\} \right\}$$

and

$$\Omega_8^O := \{\hat{\tau}_{2,1}^O = \tau_1^O \text{ or } \hat{\tau}_{2,2}^O = \tau_1^O\}.$$

Then, it follows from the fact that $\sum_{i=a}^b \left(\varepsilon_i^O - \overline{\varepsilon}_{a:b}^O\right)^2 = \sum_{i=a}^b \left(\varepsilon_i^O\right)^2 - (b-a+1)\left(\overline{\varepsilon}_{a:b}^O\right)^2$ that $\Omega_2^O = \Omega_7^O \cap \Omega_8^O$. We define $\mathrm{CV}_{(2)}^E(L)$ and $\Omega_1^E, \ldots, \Omega_8^E$ in the same way but with O and E interchanged and the observations considered in reverse order. Note that considering the observations in reverse order does not change the value of $\mathrm{CV}_{(2)}^E(L)$. Hence, we conclude that the same statements hold as for their counterparts denoted by O.

Note that $\mathrm{CV}_{(2)}(L) = \mathrm{CV}_{(2)}^O(L) + \mathrm{CV}_{(2)}^E(L)$. Hence,

$$\left\{\mathrm{CV}_{(2)}(1) > \mathrm{CV}_{(2)}(2)\right\} \supseteq \left\{\mathrm{CV}_{(2)}^O(1) > \mathrm{CV}_{(2)}^O(2), \mathrm{CV}_{(2)}^E(1) > \mathrm{CV}_{(2)}^E(2)\right\}.$$

Then, it follows from (56), $\Omega_2^O = \Omega_7^O \cap \Omega_8^O$ and (57), that if n is large enough for (57) to hold, then

$$\begin{aligned}
&\mathbb{P}\left(\mathrm{CV}_{(2)}^O(1) > \mathrm{CV}_{(2)}^O(2), \mathrm{CV}_{(2)}^E(1) > \mathrm{CV}_{(2)}^E(2)\right) \\
\geq &\mathbb{P}\left(A_n^O > 0,\ \Omega_1^O \cap \Omega_2^O,\ A_n^E > 0,\ \Omega_1^E \cap \Omega_2^E\right) \\
= &\mathbb{P}\left(A_n^O > 0,\ \Omega_1^O \cap \Omega_7^O \cap \Omega_8^O,\ A_n^E > 0,\ \Omega_1^E \cap \Omega_7^E \cap \Omega_8^E\right) \\
\geq &\mathbb{P}\left(\Omega_1^O \cap \bigcap_{j=3}^8 \Omega_j^O,\ \Omega_1^E \cap \bigcap_{j=3}^8 \Omega_j^E\right).
\end{aligned}$$

Moreover, it follows from the symmetry of events and from the fact that $\Omega_3^O \cap \Omega_7^O$ and $\Omega_3^E \cap \Omega_7^E$ are independent since they depend on different ε_i's, that

$$\begin{aligned}
&\mathbb{P}\left(\Omega_1^O \cap \bigcap_{j=3}^8 \Omega_j^O,\ \Omega_1^E \cap \bigcap_{j=3}^8 \Omega_j^E\right) \\
\geq &\mathbb{P}\left(\Omega_3^O \cap \Omega_7^O,\ \Omega_3^E \cap \Omega_7^E\right) - \mathbb{P}\left(\left(\Omega_1^O\right)^C \cup \left(\Omega_8^O\right)^C\right) - \mathbb{P}\left(\left(\Omega_1^E\right)^C \cup \left(\Omega_8^E\right)^C\right) \\
&- \sum_{j=4}^6 \mathbb{P}\left(\left(\Omega_j^O\right)^C\right) - \sum_{j=4}^6 \mathbb{P}\left(\left(\Omega_j^E\right)^C\right) \\
= &\mathbb{P}\left(\Omega_3^O \cap \Omega_7^O\right)^2 - 2\mathbb{P}\left(\left(\Omega_1^O\right)^C \cup \left(\Omega_2^O\right)^C\right) - 2\sum_{j=4}^6 \mathbb{P}\left(\left(\Omega_j^O\right)^C\right).
\end{aligned}$$

By combining all the inequalities, we obtain, for n sufficiently large that (57) holds,

$$\begin{aligned}&\mathbb{P}\left(\mathrm{CV}_{(2)}(1) > \mathrm{CV}_{(2)}(2)\right)\\ \geq&\mathbb{P}\left(\Omega_3^O \cap \Omega_7^O\right)^2 - 2\mathbb{P}\left((\Omega_1^O)^C \cup (\Omega_2^O)^C\right) - 2\sum_{j=4}^{6}\mathbb{P}\left((\Omega_j^O)^C\right).\end{aligned} \tag{58}$$

As $\Delta_1/\sigma \to \infty$ due to (13), it follows from Theorem 5 that $\mathbb{P}(\Omega_1^O \cap \Omega_2^O) \to 1$, as $n \to \infty$. Since $\overline{\varepsilon}_{a:b}^O = O_{\mathbb{P}}\left(\sigma(b-a)^{-\frac{1}{2}}\right)$ and $\overline{\varepsilon}_{a:b}^E = O_{\mathbb{P}}\left(\sigma(b-a)^{-\frac{1}{2}}\right)$ for all $a \leq b$ we have that $\mathbb{P}\left((\Omega_4^O)^C\right) \to 0$, as $n \to \infty$. Additionally, $\mathbb{P}\left((\Omega_6^O)^C | \tilde{\tau}_{2,1}^O\right) \to 0$ almost surely, and so by dominated convergence, $\mathbb{P}\left((\Omega_6^O)^C\right) \to 0$ as well. Furthermore, it follows from standard extreme value theory that

$$\begin{aligned}&\tilde{\tau}_{2,1}^O\left(\overline{\varepsilon}_{0:\tilde{\tau}_{2,1}^O}^O\right)^2 + \left(\tau_1^O - \tilde{\tau}_{2,1}^O\right)\left(\overline{\varepsilon}_{\tilde{\tau}_{2,1}^O:\tau_1^O}^O\right)^2 = \max_{\tau<\tau_1^O}\left\{\tau\left(\varepsilon_{0:\tau}^O\right)^2 + (\tau_1^O - \tau)\left(\varepsilon_{\tau:\tau_1^O}^O\right)^2\right\}\\ \leq&\max_{\tau=1,\ldots,\tau_1^O-1}\left\{\tau\left(\overline{\varepsilon}_{0:\tau}^O\right)^2\right\} + \max_{\tau=1,\ldots,\tau_1^O-1}\left\{\left(\tau_1^O - \tau\right)\left(\overline{\varepsilon}_{\tau:\tau_1^O}^O\right)^2\right\} = O_{\mathbb{P}}\left(\sigma^2\log\log n\right)\end{aligned}$$

and hence $\mathbb{P}\left((\Omega_5^O)^C\right) \to 0$, as $n \to \infty$. Finally, let

$$\begin{aligned}\Omega_9^O :=& \left\{\tilde{\tau}_{2,1}^O\left(\overline{\varepsilon}_{0:\tilde{\tau}_{2,1}^O}^O\right)^2 + \left(\tau_1^O - \tilde{\tau}_{2,1}^O\right)\left(\overline{\varepsilon}_{\tilde{\tau}_{2,1}^O:\tau_1^O}^O\right)^2 > 2s_1^2\sigma^2\log\log n\right\} \text{ and}\\ \Omega_{10}^O :=& \left\{\left(\tau_1^O - \tilde{\tau}_{2,1}^O\right)\left(\overline{\varepsilon}_{\tilde{\tau}_{2,1}^O:\tau_1^O}^O\right)^2 > \tilde{\tau}_{2,1}^O\left(\overline{\varepsilon}_{0:\tilde{\tau}_{2,1}^O}^O\right)^2\right\}.\end{aligned}$$

Note that on $\Omega_9^O \cap \Omega_{10}^O$, $\left(\tau_1^O - \tilde{\tau}_{2,1}^O\right)\left(\overline{\varepsilon}_{\tilde{\tau}_{2,1}^O:\tau_1^O}^O\right)^2 > s_1^2\sigma^2\log\log n$, so

$$\begin{aligned}\mathbb{P}\left(\Omega_3^O \cap \Omega_7^O\right) =&\mathbb{P}\left(\Omega_7^O\right)\mathbb{P}\left(\overline{\varepsilon}_{\tilde{\tau}_{2,1}^O:\tau_1^O}^O > s_1\sigma\sqrt{\frac{\log\log n}{n}}\,\middle|\,\Omega_7^O\right)\\ \geq&\mathbb{P}\left(\Omega_7^O\right)\mathbb{P}\left(\overline{\varepsilon}_{\tilde{\tau}_{2,1}^O:\tau_1^O}^O > 0\,\middle|\,\Omega_7^O \cap \Omega_9^O \cap \Omega_{10}^O\right)\mathbb{P}\left(\Omega_{10}^O\,\middle|\,\Omega_7^O \cap \Omega_9^O\right)\mathbb{P}(\Omega_9^O\,|\Omega_7^O).\end{aligned}$$

Because of symmetry and since the left segment is one observations longer, it follows that $\mathbb{P}\left(\Omega_7^O\right) \geq \frac{1}{2}$. Additionally, since the ε_i^O's are independent and symmetric around zero as centred Gaussian distributed errors we have that

$$\mathbb{P}\left(\overline{\varepsilon}_{\tilde{\tau}_{2,1}^O:\tau_1^O}^O > 0\,\middle|\,\Omega_7^O \cap \Omega_9^O \cap \Omega_{10}^O\right) = \mathbb{P}\left(\overline{\varepsilon}_{\tilde{\tau}_{2,1}^O:\tau_1^O}^O > 0\right) = \frac{1}{2}.$$

Moreover, $\tilde{\tau}_{2,1}^O\left(\overline{\varepsilon}_{0:\tilde{\tau}_{2,1}^O}^O\right)^2$ and $\left(\tau_1^O - \tilde{\tau}_{2,1}^O\right)\left(\overline{\varepsilon}_{\tilde{\tau}_{2,1}^O:\tau_1^O}^O\right)^2$ have the same distribution. This still holds when we condition on $\Omega_7^O \cap \Omega_9^O$. Thus, $\mathbb{P}(\Omega_{10}^O\,|\Omega_7^O \cap \Omega_9^O) \geq \frac{1}{2}$.

Finally, from the definition of $\tilde{\tau}^O_{2,1}$ (55) and standard extreme value theory, it follows that

$$\tilde{\tau}^O_{2,1}\left(\overline{\varepsilon}^O_{0:\tilde{\tau}^O_{2,1}}\right)^2 + \left(\tau^O_1 - \tilde{\tau}^O_{2,1}\right)\left(\overline{\varepsilon}^O_{\tilde{\tau}^O_{2,1}:\tau^O_1}\right)^2 \geq \max_{\tau=1,\ldots,\tau^O_1-1} \tau\left(\overline{\varepsilon}^O_{0:\tau}\right)^2$$

and

$$\left\{\max_{\tau=1,\ldots,\tau^O_1-1} \tau\left(\overline{\varepsilon}^O_{0:\tau}\right)^2\right\}^{-1} = O_{\mathbb{P}}\{(\sigma^2 \log\log n)^{-1}\}. \tag{59}$$

Moreover, the conditional distribution of the maximum on the l.h.s. of (59) given Ω^O_7 is stochastically larger than its unconditional distribution. Hence we may conclude that $\liminf_{n\to\infty} \mathbb{P}\left(\Omega^O_3 \cap \Omega^O_7\right) \geq 1/8$. Then, combining (58) and the probabilities above gives us

$$\liminf_{n\to\infty} \mathbb{P}(\mathrm{CV}_{(2)}(1) > \mathrm{CV}_{(2)}(2)) > 1/64.$$

The proof concludes by noting that $\mathbb{P}(\mathrm{CV}_{(2)}(0) > \mathrm{CV}_{(2)}(1)) \to 1$, as $n \to \infty$, and hence

$$\liminf_{n\to\infty} \mathbb{P}\left(\hat{K} > 1\right) \geq \liminf_{n\to\infty} \mathbb{P}\left(\mathrm{CV}_{(2)}(1) > \mathrm{CV}_{(2)}(2)\right) \geq 1/64 > 0.$$

□

S5. Proof of Theorem 4

We will show that

$$\mathbb{P}\left(\hat{K} = K\right) \geq \mathbb{P}\left(\min_{\substack{L=0,\ldots,K_{\max}\\ L\neq K}} \mathrm{CV}_{\mathrm{mod}}(L) - \mathrm{CV}_{\mathrm{mod}}(K) > 0\right) \to 1, \text{ as } n \to \infty. \tag{60}$$

Note that the two terms in $\mathrm{CV}_{\mathrm{mod}}(L)$ are symmetric. Hence, we can focus most of the time on the first term only.

To this end, we use the notations $E_i := \varepsilon^E_i + \mu^O_i,\ i = 1, \ldots, n/2$,

$$\mathrm{CV}^O_{\mathrm{mod}}(L) := \sum_{l=0}^{L} \sum_{i=\hat{\tau}^O_{L,l}+1}^{\hat{\tau}^O_{L,l+1}-1} \frac{\hat{n}^O_l}{\hat{n}^O_l - 1}\left(Y^E_i - \overline{Y}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}}\right)^2, \tag{61}$$

and

$$\widetilde{\mathrm{CV}}^O_{\mathrm{mod}}(L) := \sum_{l=0}^{L} \sum_{i=\hat{\tau}^O_{L,l}+1}^{\hat{\tau}^O_{L,l+1}} \left(E_i - \overline{Y}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}}\right)^2. \tag{62}$$

We may later also use the same definitions with all instances of O's and E's interchanged to define O_i, $\mathrm{CV}^E_{\mathrm{mod}}$ and $\widetilde{\mathrm{CV}}^E_{\mathrm{mod}}(L)$.

Then, our main argument proceeds as follows. In Lemma 14 we lower bound

$$\min_{L \neq K} \widetilde{\mathrm{CV}}^{O}_{\mathrm{mod}}(L) - \widetilde{\mathrm{CV}}^{O}_{\mathrm{mod}}(K). \tag{63}$$

Next, in Lemma 15 we upper bound

$$\max_{L=0,\ldots,K_{\max}} |\,\mathrm{CV}^{O}_{\mathrm{mod}}(L) - \widetilde{\mathrm{CV}}^{O}_{\mathrm{mod}}(L)|.$$

Putting these together, along with corresponding results for $\mathrm{CV}^{E}_{\mathrm{mod}}$ and $\widetilde{\mathrm{CV}}^{E}_{\mathrm{mod}}(L)$, which follow from identical arguments, gives the final result. To show Lemma 14, we follow the strategy in Zou, Wang and Li (2020) and split $\widetilde{\mathrm{CV}}^{O}_{\mathrm{mod}}(L)$ as follows.

For any set of time points $\mathcal{U} = \{t_0 < t_1 < \cdots < t_K < t_{K+1}\}$ and any collection of vectors $X = (X_1, \ldots, X_{t_{K+1}}) \in \mathbb{R}^{d \times t_{K+1}}$ and $Y = (Y_1, \ldots, Y_{t_{K+1}}) \in \mathbb{R}^{d \times t_{K+1}}$, define

$$S_{X,Y}(\mathcal{U}) := \sum_{k=0}^{K} \sum_{i=t_k+1}^{t_{k+1}} \left(X_i - \overline{X}_{t_k:t_{k+1}}\right)\left(Y_i - \overline{Y}_{t_k:t_{k+1}}\right). \tag{64}$$

Then for any $L = 0, \ldots, K_{\max}$, it may be shown that

$$\widetilde{\mathrm{CV}}^{O}_{\mathrm{mod}}(L) = S_E\left(\hat{\mathcal{T}}^{O}_{L}\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}^{O}_{L}\right) - S_{\varepsilon^E}\left(\hat{\mathcal{T}}^{O}_{L}\right) + 2S_{\varepsilon^O,\varepsilon^E}\left(\hat{\mathcal{T}}^{O}_{L}\right) + \sum_{i=1}^{n/2}\left(\varepsilon^{O}_{i} - \varepsilon^{E}_{i}\right)^2,$$

with $S_X\left(\hat{\mathcal{T}}^{O}_{L}\right)$ defined as in (20).

Consequently, for any $L \neq K$,

$$\begin{aligned}
&\widetilde{\mathrm{CV}}^{O}_{\mathrm{mod}}(L) - \widetilde{\mathrm{CV}}^{O}_{\mathrm{mod}}(K) \\
&= \left\{S_E\left(\hat{\mathcal{T}}^{O}_{L}\right) - S_E\left(\hat{\mathcal{T}}^{O}_{K}\right)\right\} - \left\{S_{\varepsilon^O}\left(\hat{\mathcal{T}}^{O}_{L}\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}^{O}_{K}\right)\right\} - \left\{S_{\varepsilon^E}\left(\hat{\mathcal{T}}^{O}_{L}\right) - S_{\varepsilon^E}\left(\hat{\mathcal{T}}^{O}_{K}\right)\right\} \\
&\quad + 2\left\{S_{\varepsilon^O,\varepsilon^E}\left(\hat{\mathcal{T}}^{O}_{L}\right) - S_{\varepsilon^O,\varepsilon^E}\left(\hat{\mathcal{T}}^{O}_{K}\right)\right\}.
\end{aligned} \tag{65}$$

The main argument needed to bound the terms in (65) uniformly will be derived in Lemmas 11–13. Preliminary results needed for this are given in Lemmas 5–8, which bound the maximum of rescaled local sums, and in Lemmas 9 and 10, where we bound $S_X(\mathcal{U}) - S_X(\mathcal{V})$ and $S_{X,Y}(\mathcal{U}) - S_{X,Y}(\mathcal{V})$, respectively, when $\mathcal{U} \subset \mathcal{V}$. Lemma 15 is shown by splitting up terms in a similar fashion.

A lower bound of $\widetilde{\mathrm{CV}}^{O}_{\mathrm{mod}}(L) - \widetilde{\mathrm{CV}}^{O}_{\mathrm{mod}}(K)$ is developed in Zou, Wang and Li (2020) (see (66) below) where a simplifying assumption is made that change-points do not occur at odd locations. This additional assumption ensures that the first term in $\mathrm{CV}_{(2)}(L)$ is identical to $\widetilde{\mathrm{CV}}^{O}_{\mathrm{mod}}(L)$ as claimed by their second displayed formula on pg 433, where it is assumed that O_i and E_i have the same expectation. As discussed at the beginning of the proof of Theorem 1 on pg. 8 in

the supplementary material of Wang, Zou and Qiu (2021), such an assumption may be justified when the noise variance is bounded away from zero and the signal magnitude is bounded; we however do not make these assumptions in our results in order to maintain greater fidelity to phenomena observed in finite sample settings.

In Zou, Wang and Li (2020), results of the form

$$\mathbb{P}\left(\mathrm{CV}_{(2)}(L) - \mathrm{CV}_{(2)}(K) > a_n\right) \to 1, \text{ as } n \to \infty \tag{66}$$

are shown, for each fixed $L = 0, \ldots, K_{\max}$, $L \neq K$, where a_n is a positive sequence. However while these immediately give consistency when K and $K_{\max}$ are finite, it is not clear to us how the arguments may be extended directly to allow for diverging K and $K_{\max}$ as this would require delicate control of the rates at which the probabilities above approach 1. Several of the lemmas below, which we require in our proof of Lemma 14, are therefore uniform versions of Lemmas 1–5 in Zou, Wang and Li (2020).

For notational convenience we use the following convention: whenever u and v are vectors of the same dimension, uv should be understood to mean $u^t v$, and similarly $|u|$ and u^2 should be taken to mean $\|u\|_2$ and $\|u\|_2^2$ respectively. Furthermore, we use the notation $\hat{n}_l^O := \hat{\tau}_{L,l+1}^O - \hat{\tau}_{L,l}^O$.

The following two lemmas are uniform versions of Lemmas 1 and 2 in Zou, Wang and Li (2020).

Lemma 5. *Let $(Y_{nkj})_{n=1,2,\ldots,k=1,\ldots,K_n,j=1,\ldots,n}$ be independent (potentially multivariate) mean-zero random variables with $\mathbb{E}[\|Y_{nkj}\|_2^2] \leq 1$ and $\mathbb{E}\|Y_{nkj}\|_2^q \leq \frac{q!}{2}c^{q-2}$ for all $q \geq 3$ and a constant $c > 0$, $n \in \mathbb{N}$, $k = 1, \ldots, K_n$, $j = 1, \ldots, n$. Then,*

$$\max_{k=1,\ldots,K_n} \max_{0\leq i<j\leq n} \frac{1}{j-i} \left\| \sum_{l=i+1}^{j} Y_{nkl} \right\|_2^2 = O_{\mathbb{P}}\big((\log n)^2 + (\log K_n)^2 \big).$$

Proof. Without loss of generality, we may assume that Y_{nkj} is one-dimensional. If it is multidimensional, we can bound each coordinate individually. To this end, note that the conditions of Bernstein's inequality (see Theorem 2.10 in Boucheron, Lugosi and Massart (2013)) are still satisfied. Hence, we observe that the bound only increases by a constant depending solely on the dimension.

It follows from union bounds and Bernstein's inequality that for $\xi > 0$ sufficiently large,

$$\begin{aligned}
&\mathbb{P}\left(\max_{k=1,\ldots,K_n} \max_{0\leq i<j\leq n} \frac{1}{j-i} \left(\sum_{l=i+1}^{j} Y_{nkl} \right)^2 > \left((\sqrt{2}+c)\xi(\log n + \log K_n) \right)^2 \right) \\
&\leq \sum_{k=1}^{K_n} \sum_{0\leq i<j\leq n} \mathbb{P}\left(\left| \sum_{l=i+1}^{j} Y_{nkl} \right| > (\sqrt{2}+c)\xi(\log n + \log K_n)\sqrt{j-i} \right) \\
&\leq 2K_n \sum_{0\leq i<j\leq n} \exp\left(-\xi(\log n + \log K_n)\right) \\
&\leq 2K_n^{-\xi+1} n^{-\xi+2},
\end{aligned}$$

Hence, the r.h.s. can be made arbitrarily small by choosing ξ large enough. □

Lemma 6. *Consider the setup of Lemma 5. Then,*

$$\max_{k=1,\dots,K_n} \max_{1\le j\le n} \frac{1}{j} \left\| \sum_{i=1}^{j} Y_{nki} \right\|_2^2 = O_{\mathbb{P}}\Big(\log\log n + (\log K_n)^2 \Big).$$

Proof. Without loss of generality, we may assume $K_n \to \infty$ as $n \to \infty$. Similarly to the proof of Lemma 5, we may assume without loss of generality that Y_{nkj} is one-dimensional.

Let n and $k \in \{1,\dots,K_n\}$ be fixed. Furthermore, let $M_t := \sum_{i=1}^{t} Y_{nki}$, $\xi_t := M_t - M_{t-1} = Y_{nki}$, and $V_t := \sum_{i=1}^{t} \mathbb{E}\left[\xi_i^2\right] = \sum_{i=1}^{t} \mathbb{E}\left[Y_{nki}^2\right] \le t$. Then M_t is a martingale and the Bernstein condition in Lemma 23 in Balsubramani (2014) is satisfied. Hence, it follows from Theorem 5 in Balsubramani (2014) (see the discussion after their Theorem 5 which explains that the Bernstein condition may replace the interval condition in Theorem 5) that for all $\delta \in (0,1)$,

$$\mathbb{P}\left(|M_t| \le \sqrt{6(\sqrt{2c}-2)t\left(2\log\log\left(\frac{3(\sqrt{2c}-2)t}{|M_t|}\right) + \log\left(\frac{2}{\delta}\right)\right)} \quad \forall\, t \ge \tau_0 \right) \ge 1-\delta,$$

with $\tau_0 := \lceil 173\log(4/\delta)/\{2(\sqrt{2c}-2)\}\rceil$. (Note that we may assume without loss of generality that $c > 2$.)

Therefore for all $\delta \in (0,1)$ and all n sufficiently large, we have

$$\mathbb{P}\left(\max_{t=\tau_0,\dots,n} M_t^2/t > 6(\sqrt{2c}-2)\big\{2\log\log(3(\sqrt{2c}-2)n) + \log(2/\delta)\big\} \right) < \delta.$$

Note here we have used the fact that for those t for which $|M_t| < 1$, the inequality is trivially satisfied for n sufficiently large.

Now let $\delta := 2\exp\left(-\frac{x}{6(\sqrt{2c}-2)}\right)$. Then,

$$\mathbb{P}\left(\max_{t=\tau_0,\dots,n} M_t^2/t > 12(\sqrt{2c}-2)\log\log\Big(3(\sqrt{2c}-2)n\Big) + x \right) < 2\exp\left(-\frac{x}{6(\sqrt{2c}-2)}\right)$$

and

$$\tau_0 = \left\lceil \frac{173}{12(\sqrt{2c}-2)^2}x + \frac{173\log(2)}{2(\sqrt{2c}-2)} \right\rceil.$$

Let us consider $x = A\log K_n$, with $A > 0$ a constant. Then a union bound yields

$$\begin{aligned}
&\mathbb{P}\left(\max_{k=1,\dots,K_n} \max_{t=\tau_0,\dots,n} \left(\sum_{l=1}^{t} Y_{nkt} \right)^2 /t > 12(\sqrt{2c}-2)\log\log\Big(3(\sqrt{2c}-2)n\Big) + A\log K_n \right) \\
&< 2K_n \exp\left(-\frac{A\log K_n}{6(\sqrt{2c}-2)}\right) = 2K_n^{-A/(6(\sqrt{2c}-2))+1},
\end{aligned} \tag{67}$$

with $\tau_0 = B \log K_n$, where $B > 0$ is a constant only depending on c and A. The r.h.s. in (67) can be made arbitrarily small by choosing A large enough.

It remains to consider $t < \tau_0 = B \log K_n$. It follows from a union bound and Bernstein's inequality, see Corollary 2.11 in Boucheron, Lugosi and Massart (2013), that for any constant $C > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\max_{k=1,\dots,K_n} \max_{t=1,\dots,\tau_0-1} \frac{1}{t} \left(\sum_{l=1}^{t} Y_{nkt} \right)^2 > C(\log K_n)^2 \right) \\
&\leq \sum_{k=1}^{K_n} \sum_{t=1}^{\tau_0-1} \left\{ \mathbb{P}\left(\sum_{l=1}^{t} Y_{nkt} > \sqrt{Ct} \log K_n \right) + \mathbb{P}\left(\sum_{l=1}^{t} Y_{nkt} < -\sqrt{Ct} \log K_n \right) \right\} \\
&\leq 2K_n B \log K_n \exp\left(-\sqrt{C}/(\sqrt{2}+c) \log K_n \right) = 2BK_n^{-\sqrt{C}/(\sqrt{2}+c)-1} \log K_n.
\end{aligned}$$

The r.h.s. can be made arbitrarily small by choosing C large enough, which completes the proof. □

Lemma 7. *Consider the setup of Lemma 5 but where $K_n \geq 2$ eventually. Then, for any constant $C > 2$ and any sequence $(\delta_n)_{n=1}^{\infty}$, with $1 < \delta_n < n/C$, we have that*

$$\max_{k=1,\dots,K_n} \delta_n \max_{\lceil C\delta_n \rceil \leq j \leq n} \left\| \frac{1}{j} \sum_{i=1}^{j} Y_{nki} \right\|_2^2 = O_{\mathbb{P}}\big((\log K_n)^2\big).$$

Proof. A union bound and Bernstein's inequality, see Corollary 2.11 in Boucheron, Lugosi and Massart (2013), yield that for any $A > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\max_{k=1,\dots,K_n} \delta_n \max_{\lceil C\delta_n \rceil \leq j \leq n} \left(\frac{\sum_{i=1}^{j} Y_{nki}}{j} \right)^2 > A^2(\log K_n)^2 \right) \\
&\leq \sum_{k=1}^{K_n} \sum_{j=\lceil C\delta_n \rceil,\dots,n} \left\{ \mathbb{P}\left(\sum_{i=1}^{j} Y_{nki} > \frac{j}{\sqrt{\delta_n}} A \log K_n \right) + \mathbb{P}\left(\sum_{i=1}^{j} Y_{nki} < -\frac{j}{\sqrt{\delta_n}} A \log K_n \right) \right\} \\
&\leq 2K_n \sum_{j=\lceil C\delta_n \rceil,\dots,n} \exp\left(-\frac{j^2 A^2 (\log K_n)^2/\delta_n}{2\big(j + cA \log(K_n) j/\sqrt{\delta_n}\big)} \right).
\end{aligned}$$

Moreover, there exist constants $B > 0$ and $D' > 0$ not depending on A, such

that for $D := D'A$,

$$K_n \sum_{j=\lceil C\delta_n\rceil,\dots,n} \exp\left(-\frac{j^2A^2(\log K_n)^2/\delta_n}{2\big(j + cA\log(K_n)j/\sqrt{\delta_n}\big)}\right) \le K_n \sum_{j=\lceil C\delta_n\rceil,\dots,n} \exp\left(-\frac{jA\log K_n}{B\sqrt{\delta_n}}\right)$$
$$\le K_n \sum_{j=\lceil C\delta_n\rceil,\dots,n} \exp\left(-\sqrt{j}A\log(K_n)/B\right) = \sum_{j=\lceil C\delta_n\rceil,\dots,n} K_n^{-\sqrt{j}A/B+1}$$
$$\le \int_1^\infty \left(K_n^{-D}\right)^{\sqrt{x}}\,dx = \left[\frac{2\left(K_n^{-D}\right)^{\sqrt{x}}\left(-\sqrt{x}D\log K_n - 1\right)}{D^2(\log K_n)^2}\right]_1^\infty$$
$$= \frac{2K_n^{-D}(D\log K_n + 1)}{D^2(\log K_n)^2}.$$

The r.h.s. can be made arbitrarily small by choosing A and hence D large enough. □

Lemma 8. *Let $(Y_{nj})_{n=1,2,\dots,j=1,\dots,n}$ be independent (potentially multivariate) mean-zero random variables with $\mathbb{E}[\|Y_{nj}\|_2^2] \le 1$ and $\mathbb{E}\|Y_{nj}\|_2^q \le \frac{q!}{2}c^{q-2}$ for all $q \ge 3$ and a constant $c > 0$, $n \in \mathbb{N}$, $j = 1,\dots,n$. For any $L = 1,\dots,L_{\max}$, where $L_{\max} \ge 2$, let $0 \le \tau_{L,1,s} < \tau_{L,1,e} \le \tau_{L,2,s} < \tau_{L,2,e} \le \cdots \tau_{L,L_{\max},s} < \tau_{L,L_{\max},e} \le n$ be sequences of (random) time points that are independent of $(Y_{nj})_{n=1,2,\dots,j=1,\dots,n}$. Then,*

$$\max_{L=1,\dots,L_{\max}} \sum_{l=1}^{L_{\max}} \left(\tau_{L,l,e} - \tau_{L,l,s}\right)^{-1} \left\| \sum_{j=\tau_{L,l,s}+1}^{\tau_{L,l,e}} Y_{nj} \right\|_2^2 = O_{\mathbb{P}}\big((L_{\max}\log L_{\max})^{1/2}\big) \quad (68)$$

and

$$\max_{L=1,\dots,L_{\max}} \left| \sum_{l=1}^{L_{\max}} \left\{ \left(\tau_{L,l,e} - \tau_{L,l,s}\right)^{-1} \sum_{j=\tau_{L,l,s}+1}^{\tau_{L,l,e}} \left\|Y_{nj}\right\|_2^2 - \mathbb{E}\left[\left(\tau_{L,l,e} - \tau_{L,l,s}\right)^{-1} \sum_{j=\tau_{L,l,s}+1}^{\tau_{L,l,e}} \left\|Y_{nj}\right\|_2^2 \right] \right\} \right|$$
$$= O_{\mathbb{P}}\big((L_{\max}\log L_{\max})^{1/2}\big).$$

Proof. Similarly to the proof of Lemma 5, we may assume without loss of generality that Y_{nj} is one-dimensional.

We consider the first bound to begin with. For any $L = 1,\dots,L_{\max}$, let $\mathcal{T}_L$ denote the σ-algebra generated by $\tau_{L,1,s}$, $\tau_{L,1,e}$, $\tau_{L,2,s}$, $\tau_{L,2,e},\dots,\tau_{L,L_{\max},s}$, $\tau_{L,L_{\max},e}$. Then, it follows from a union bound and the law of total probability

that for any $\xi > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\max_{L=1,\ldots,L_{\max}} \sum_{l=1}^{L_{\max}} \left(\tau_{L,l,e} - \tau_{L,l,s}\right)^{-1} \left(\sum_{j=\tau_{L,l,s}+1}^{\tau_{L,l,e}} Y_{nj}\right)^2 > \xi\right) \\
&\leq \sum_{L=1}^{L_{\max}} \mathbb{P}\left(\sum_{l=1}^{L_{\max}} \left(\tau_{L,l,e} - \tau_{L,l,s}\right)^{-1} \left(\sum_{j=\tau_{L,l,s}+1}^{\tau_{L,l,e}} Y_{nj}\right)^2 > \xi\right) \\
&= \sum_{L=1}^{L_{\max}} \mathbb{E}\left[\mathbb{P}\left(\sum_{l=1}^{L_{\max}} \left(\tau_{L,l,e} - \tau_{L,l,s}\right)^{-1} \left(\sum_{j=\tau_{L,l,s}+1}^{\tau_{L,l,e}} Y_{nj}\right)^2 > \xi \middle| \mathcal{T}_L\right)\right].
\end{aligned}$$

In the following we focus on the conditional probability inside the expectation. Conditional on $\mathcal{T}_L$, $\left(\tau_{L,l,e} - \tau_{L,l,s}\right)^{-1} \left(\sum_{j=\tau_{L,l,s}+1}^{\tau_{L,l,e}} Y_{nj}\right)^2$, $l = 1, \ldots, L_{\max}$ are independent. Bernstein's inequality, see Theorem 2.10 in Boucheron, Lugosi and Massart (2013), yields that $\left(\tau_{L,l,e} - \tau_{L,l,s}\right)^{-1/2} \sum_{j=\tau_{L,l,s}+1}^{\tau_{L,l,e}} Y_{nj}$ is sub-exponentially distributed with Orlicz norm bounded uniformly over $l = 1, \ldots, L_{\max}$, $L = 1, \ldots, L_{\max}$, i.e. there exist a constant $\psi^{(1)} < \infty$ such that

$$\inf\left\{C \in (0,\infty) \,:\, \mathbb{E}\left[\exp\left(C^{-1}\left(\tau_{L,l,e} - \tau_{L,l,s}\right)^{-1}\left(\sum_{j=\tau_{L,l,s}+1}^{\tau_{L,l,e}} Y_{nj}\right)^2\right)\right] \leq 2\right\} \leq \psi^{(1)}$$

for all $l = 1, \ldots, L_{\max}$, $L = 1, \ldots, L_{\max}$. Consequently, it follows from Corollary 3 in Zhang and Wei (2021) that $\left(\tau_{L,l,e} - \tau_{L,l,s}\right)^{-1} \left(\sum_{j=\tau_{L,l,s}+1}^{\tau_{L,l,e}} Y_{nj}\right)^2$ are sub-Weibull distributed with parameter $\theta = 1/2$ and uniformly bounded sub-Weibull norms, i.e. there exist a constant $\psi^{(1/2)} < \infty$ such that

$$\inf\left\{C \in (0,\infty) \,:\, \mathbb{E}\left[\exp\left(C^{-1/2}\left(\left(\tau_{L,l,e} - \tau_{L,l,s}\right)^{-1}\left(\sum_{j=\tau_{L,l,s}+1}^{\tau_{L,l,e}} Y_{nj}\right)^2\right)^{1/2}\right)\right] \leq 2\right\} \leq \psi^{(1/2)}$$

for all $l = 1, \ldots, L_{\max}$, $L = L1, \ldots, L_{\max}$. Hence, it follows from Corollary 6.4 in Zhang and Chen (2020) (see also Theorem 1 in Zhang and Wei (2021)) that there exist constants $C_1, C_2 > 0$ (not depending on L) such that for any $t > 0$,

$$\mathbb{P}\left(\sum_{l=1}^{L_{\max}} \left(\tau_{L,l,e} - \tau_{L,l,s}\right)^{-1} \left(\sum_{j=\tau_{L,l,s}+1}^{\tau_{L,l,e}} Y_{nj}\right)^2 \geq C_1\sqrt{L_{\max} t} + C_2 t^2 \middle| \mathcal{T}_L\right) \leq 2e^{-t} \qquad (69)$$

for all $L = 1, \ldots, L_{\max}$. Thus, for any $a > 1$,

$$\begin{aligned}
&\mathbb{P}\left(\max_{L=1,\ldots,L_{\max}} \sum_{l=1}^{L_{\max}} \left(\tau_{L,l,e} - \tau_{L,l,s}\right)^{-1} \left(\sum_{j=\tau_{L,l,s}+1}^{\tau_{L,l,e}} Y_{nj}\right)^2 > a\left[C_1 \left(L_{\max}\log L_{\max}\right)^{1/2} + C_2 \left(\log L_{\max}\right)^2\right]\right) \\
&\leq \sum_{L=1}^{L_{\max}} 2\exp\left(-\sqrt{a}\log L_{\max}\right) = 2L_{\max}^{-\sqrt{a}+1}.
\end{aligned}$$

The r.h.s. can be made arbitrarily small by choosing a large enough, thus showing (68).

We will now show the second bound. From following the same steps as before we see that it suffices to show that

$$\sum_{l=1}^{L_{\max}} \left\{ \left(\tau_{L,l,e} - \tau_{L,l,s}\right)^{-1} \sum_{j=\tau_{L,l,s}+1}^{\tau_{L,l,e}} \left\|Y_{nj}\right\|_2^2 - \mathbb{E}\left[\left(\tau_{L,l,e} - \tau_{L,l,s}\right)^{-1} \sum_{j=\tau_{L,l,s}+1}^{\tau_{L,l,e}} \left\|Y_{nj}\right\|_2^2 \right] \right\}$$

conditional on $\mathcal{T}_L$ is sub-Weibull distributed with parameter $\theta = 1/2$ and uniformly bounded sub-Weibull norms.

It follows from the assumptions that the $\left\|Y_{nj}\right\|_2$'s are independent and sub-exponentially distributed with uniformly bounded Orlicz norms. Consequently, it follows from Corollary 3 in Zhang and Wei (2021) that the $\left\|Y_{nj}\right\|_2^2$'s are sub-Weibull distributed with parameter $\theta = 1/2$ and uniformly bounded sub-Weibull norms. The same applies for the centred random variables and the mean of those centred random variables. Hence, the bound follows from the same arguments as used to derive the bound for the first term. □

Lemma 9. *For any $L \in \mathbb{N}$ and any sequence of time points $1 \leq a =: t_0 < t_1 < \cdots < t_L < t_{L+1} := b \leq n$ and any vectors $X_1, \ldots, X_n \in \mathbb{R}^d$ we have that*

$$\begin{aligned} S_X(\{a,b\}) - S_X(\{t_0, t_1, \ldots, t_L, t_{L+1}\}) \leq & 2\sum_{l=0}^{L} \frac{(b-a) - (t_{l+1} - t_l)}{b-a} (t_{l+1} - t_l) \left(\overline{X}_{t_l:t_{l+1}}\right)^2 \\ \leq & 2\sum_{l=0}^{L} (t_{l+1} - t_l) \left(\overline{X}_{t_l:t_{l+1}}\right)^2. \end{aligned}$$

Proof. From (31), we have that

$$\begin{aligned} & S_X(\{a,b\}) - S_X(\{t_0, t_1, \ldots, t_L, t_{L+1}\}) \\ = & \sum_{l=0}^{L} (t_{l+1} - t_l) \left(\overline{X}_{t_l:t_{l+1}}\right)^2 - (b-a) \left(\overline{X}_{a:b}\right)^2 \\ = & \sum_{l=0}^{L} (t_{l+1} - t_l) \left(\overline{X}_{t_l:t_{l+1}}\right)^2 - \sum_{l_1=0}^{L} \sum_{l_2=0}^{L} \frac{(t_{l_1+1} - t_{l_1})(t_{l_2+1} - t_{l_2})}{b-a} \overline{X}_{t_{l_1}:t_{l_1+1}} \overline{X}_{t_{l_2}:t_{l_2+1}} \\ = & \sum_{l=0}^{L} \left(1 - \frac{t_{l+1} - t_l}{b-a}\right) (t_{l+1} - t_l) \left(\overline{X}_{t_l:t_{l+1}}\right)^2 \\ & - \sum_{l=0}^{L} (b-a)^{-1} \sum_{l'=0,\ldots,L;\ l' \neq l} (t_{l'+1} - t_{l'})(t_{l+1} - t_l) \overline{X}_{t_l:t_{l+1}} \overline{X}_{t_{l'}:t_{l'+1}}. \end{aligned}$$

Using the inequality $2xy \leq x^2 + y^2$ yields

$$\begin{aligned}
&\left|\sum_{l=0}^{L}(b-a)^{-1}\sum_{l'=0,\ldots,L;\ l'\neq l}(t_{l'+1}-t_{l'})(t_{l+1}-t_l)\overline{X}_{t_l:t_{l+1}}\overline{X}_{t_{l'}:t_{l'+1}}\right| \\
\leq&\frac{1}{2}\sum_{l=0}^{L}(b-a)^{-1}\sum_{l'=0,\ldots,L;\ l'\neq l}(t_{l'+1}-t_{l'})(t_{l+1}-t_l)\left\{\left(\overline{X}_{t_l:t_{l+1}}\right)^2+\left(\overline{X}_{t_{l'}:t_{l'+1}}\right)^2\right\} \\
=&\sum_{l=0}^{L}(b-a)^{-1}(t_{l+1}-t_l)\left(\overline{X}_{t_l:t_{l+1}}\right)^2\sum_{l'=0,\ldots,L;\ l'\neq l}(t_{l'+1}-t_{l'}).
\end{aligned}$$

The proof is completed by noting that $\sum_{l'=0,\ldots,L;\ l'\neq l}(t_{l'+1}-t_{l'}) = (b-a)-(t_{l+1}-t_l) < b-a$. □

Lemma 10. *For any $L \in \mathbb{N}$ and any sequence of time points $1 \leq a =: t_0 < t_1 < \cdots < t_L < t_{L+1} := b \leq n$ and any vectors $X_1, \ldots, X_n \in \mathbb{R}^d$ and $Y_1, \ldots, Y_n \in \mathbb{R}^d$ we have that*

$$\begin{aligned}
&2\left|S_{X,Y}\left(\{t_0,t_1,\ldots,t_L,t_{L+1}\}\right)-S_{X,Y}\left(\{a,b\}\right)\right| \\
\leq&\left\{S_X\left(\{a,b\}\right)-S_X\left(\{t_0,t_1,\ldots,t_L,t_{L+1}\}\right)\right\}+\left\{S_Y\left(\{a,b\}\right)-S_Y\left(\{t_0,t_1,\ldots,t_L,t_{L+1}\}\right)\right\}.
\end{aligned}$$

Proof. Let $P_0 \in \mathbb{R}^{n\times n}$ and $P_1 \in \mathbb{R}^{n\times n}$ be the orthogonal projection matrices that project vectors onto the constant segments given by the change-point sets $\{a,b\}$ and $\{t_0,t_1,\ldots,t_L,t_{L+1}\}$ respectively. Then, $X^t(I-P_0)Y = S_{X,Y}(\{a,b\})$ and $X^t(I-P_1)Y = S_{X,Y}(\{t_0,t_1,\ldots,t_L,t_{L+1}\})$. Hence,

$$\begin{aligned}
&2\left|S_{X,Y}\left(\{a,b\}-S_{X,Y}\left(\{t_0,t_1,\ldots,t_L,t_{L+1}\}\right)\right)\right| \\
=&2\left|X^t(P_1-P_0)Y\right| \leq X^t(P_1-P_0)X+Y^t(P_1-P_0)Y \\
=&\left\{S_X\left(\{a,b\}\right)-S_X\left(\{t_0,t_1,\ldots,t_L,t_{L+1}\}\right)\right\}+\left\{S_Y\left(\{a,b\}\right)-S_Y\left(\{t_0,t_1,\ldots,t_L,t_{L+1}\}\right)\right\}.
\end{aligned}$$

□

Lemma 11. *Suppose that Assumptions 1–5 hold in the case where $K \geq 1$ eventually, or only Assumptions 1, 2 and 4 in the case $K = 0\ \forall\ n$. Then,*

(i) $\max_{L=0,\ldots,K_{\max}}\left\{S_{\varepsilon^E}\left(\hat{\mathcal{T}}_L^O\right)-S_{\varepsilon^E}\left(\hat{\mathcal{T}}_L^O\cup\mathcal{T}_K^O\right)\right\} = o_{\mathbb{P}}(\overline{\sigma}^2\log\log\overline{\lambda})$.

(ii) $\max_{L=0,\ldots,K_{\max}}\left\{S_{\varepsilon^E}\left(\mathcal{T}_K^O\right)-S_{\varepsilon^E}\left(\hat{\mathcal{T}}_L^O\cup\mathcal{T}_K^O\right)\right\} = o_{\mathbb{P}}(\overline{\sigma}^2\log\log\overline{\lambda})$.

Proof. To show (i), first let $L \in \{0,\ldots,K_{\max}\}$ be fixed. Let $\hat{\tau}_{L,l}^O =: \hat{\tau}_{L,l,0}^O < \hat{\tau}_{L,l,1}^O < \cdots < \hat{\tau}_{L,l,\hat{K}_l}^O < \hat{\tau}_{L,l,\hat{K}_l+1}^O := \hat{\tau}_{L,l+1}^O$ be the true change-points between $\hat{\tau}_{L,l}^O$ and $\hat{\tau}_{L,l+1}^O$, so $\bigcup_{l=0}^{L}\bigcup_{k=0}^{\hat{K}_l}\hat{\tau}_{L,l,k}^O = \bigcup_{l=0}^{L}\hat{\tau}_{L,l}^O \cup \bigcup_{k=0}^{K}\tau_k^O$. It follows from Lemma 9 that

$$0 \leq S_{\varepsilon^E}\left(\hat{\mathcal{T}}_L^O\right)-S_{\varepsilon^E}\left(\hat{\mathcal{T}}_L^O\cup\mathcal{T}_K^O\right) \leq 2\sum_{\substack{l=0,\ldots,L\\ \hat{K}_l>0}}\sum_{k=0}^{\hat{K}_l}\left(\hat{\tau}_{L,l,k+1}^O-\hat{\tau}_{L,l,k}^O\right)\left(\overline{\varepsilon}^E_{\hat{\tau}_{L,l,k}^O:\hat{\tau}_{L,l,k+1}^O}\right)^2.$$

Then, the fact that $\sum_{\substack{l=0,\dots,L\\ \hat{K}_l>0}}(\hat{K}_l+1) \le 2K$ and Lemma 8 gives us that

$$\max_{L=0,\dots,K_{\max}} \sum_{\substack{l=0,\dots,L\\ \hat{K}_l>0}} \sum_{k=0}^{\hat{K}_l} \left(\hat{\tau}^O_{L,l,k+1} - \hat{\tau}^O_{L,l,k}\right) \left(\overline{\varepsilon}^E_{\hat{\tau}^O_{L,l,k}:\hat{\tau}^O_{L,l,k+1}}\right)^2 = O_{\mathbb{P}}\left(\overline{\sigma}^2 (K_{\max}\log K_{\max})^{1/2}\right).$$

Finally, Assumption 1(iii) yields $(K_{\max}\log K_{\max})^{1/2} = o(\log\log\overline{\lambda})$.

We now show (ii). Let $L \in \{0,\dots,K_{\max}\}$ be fixed. Let $\tau^O_k =: \tilde{\tau}^O_{L,k,0} \le \tilde{\tau}^O_{L,k,1} < \dots < \tilde{\tau}^O_{L,k,\tilde{L}_k} \le \tilde{\tau}^O_{L,k,\tilde{L}_k+1} := \tau^O_{k+1}$ be the estimated change-points in $\hat{\mathcal{T}}^O_L$ between τ^O_k and τ^O_{k+1}, so $\bigcup_{k=0}^K \bigcup_{l=0}^{\tilde{L}_k} \tilde{\tau}^O_{L,k,l} = \bigcup_{k=0}^K \tau^O_k \cup \bigcup_{l=0}^L \hat{\tau}^O_{L,l}$ and $\sum_{k=0}^K \tilde{L}_k = L$. It follows from Lemma 9 that

$$0 \le S_{\mathcal{E}^E}\left(\mathcal{T}^O_K\right) - S_{\mathcal{E}^E}\left(\hat{\mathcal{T}}^O_L \cup \mathcal{T}^O_K\right) \le 2 \sum_{\substack{k=0,\dots,K\\ \tilde{L}_k>0}} \sum_{l=0}^{\tilde{L}_k} \left(\tilde{\tau}^O_{L,k,l+1} - \tilde{\tau}^O_{L,k,l}\right) \left(\overline{\varepsilon}^E_{\tilde{\tau}^O_{L,k,l}:\tilde{\tau}^O_{L,k,l+1}}\right)^2.$$

Then, the fact that $\sum_{\substack{k=0,\dots,K\\ \tilde{L}_k>0}}(\tilde{L}_k+1) \le 2L$ and Lemma 8 gives us that

$$\max_{L=0,\dots,K_{\max}} \left\{ \sum_{\substack{k=0,\dots,K\\ \tilde{L}_k>0}} \sum_{l=0}^{\tilde{L}_k} \left(\tilde{\tau}^O_{L,k,l+1} - \tilde{\tau}^O_{L,k,l}\right) \left(\overline{\varepsilon}^E_{\tilde{\tau}^O_{L,k,l}:\tilde{\tau}^O_{L,k,l+1}}\right)^2 \right\} = O_{\mathbb{P}}\left(\overline{\sigma}^2 (K_{\max}\log K_{\max})^{1/2}\right),$$

and so the result follows similarly to (i). □

The following lemma is a uniform version of Lemma 4 in Zou, Wang and Li (2020). Note that (i) below extends Zou, Wang and Li (2020, Lemma 4 (i)) to allow for sequences of index sets $\mathcal{I}_L$ of missing change-points rather than a fixed single change-point.

Lemma 12. *Let $E_i := \mu^O_i + \varepsilon^E_i$, $i = 1,\dots,n/2$. Suppose that Assumptions 1–5 hold in the case where $K \ge 1$ eventually, or only Assumptions 1, 2 and 4 in the case $K = 0\ \forall\ n$. Then we have the following.*

(i) Suppose that, in addition, for a constant $A > 0$, there exist sequences of non-empty sets $\mathcal{I}_L \subseteq \{1,\dots,K\}$ such that

$$\mathbb{P}\left(\forall\ L < K,\ \forall\ k \in \mathcal{I}_L,\ \sum_{i=\tau^O_k-\frac{\underline{\lambda}}{4}+1}^{\tau^O_k+\frac{\underline{\lambda}}{4}} (\mu^O_i - \overline{\mu}^O_{L,i})^2 \ge A\underline{\lambda}\Delta^2_k\right) \to 1, \tag{70}$$

with $\overline{\mu}^O_{L,i} := \sum_{l=0}^L \mathbb{1}_{\{\hat{\tau}^O_{L,l}+1 \le i \le \hat{\tau}^O_{L,l+1}\}} \overline{\mu}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}}$. Then

$$\min_{L=0,\dots,K-1} \left(\sum_{k\in\mathcal{I}_L} \Delta^2_k\right)^{-1} \left\{S_E\left(\hat{\mathcal{T}}^O_L\right) - S_E\left(\hat{\mathcal{T}}^O_K\right)\right\} \ge \underline{\lambda}\left(A + o_{\mathbb{P}}(1)\right).$$

(ii) $\max_{L=0,\dots,K_{\max}} \left\{ S_E\left(\mathcal{T}_K^O\right) - S_E\left(\hat{\mathcal{T}}_L^O \cup \mathcal{T}_K^O\right)\right\} = o_{\mathbb{P}}(\overline{\sigma}^2 \log\log \overline{\lambda})$.

(iii) $S_E\left(\hat{\mathcal{T}}_K^O\right) - S_E\left(\mathcal{T}_K^O\right) = o_{\mathbb{P}}(\overline{\sigma}^2 \log\log \overline{\lambda})$.

Proof. We will show (i) after (ii) and (iii), since (i) is more complex and uses other results.

Turning to (ii), note that

$$S_E\left(\mathcal{T}_K^O\right) - S_E\left(\hat{\mathcal{T}}_L^O \cup \mathcal{T}_K^O\right) = S_{\varepsilon^E}\left(\mathcal{T}_K^O\right) - S_{\varepsilon^E}\left(\hat{\mathcal{T}}_L^O \cup \mathcal{T}_K^O\right).$$

Then, (ii) follows from Lemma 11(i).

We now show (iii). It follows from (31) that

$$\begin{aligned}
&S_E\left(\hat{\mathcal{T}}_K^O\right) - S_E\left(\mathcal{T}_K^O\right) \\
&= \sum_{k=0}^{K} \sum_{i=\hat{\tau}_{K,k}^O+1}^{\hat{\tau}_{K,k+1}^O} \left(\mu_i^O - \overline{\mu}^O_{\hat{\tau}_{K,k}^O:\hat{\tau}_{K,k+1}^O}\right)^2 + 2\sum_{k=0}^{K} \sum_{i=\hat{\tau}_{K,k}^O+1}^{\hat{\tau}_{K,k+1}^O} \left(\mu_i^O - \overline{\mu}^O_{\hat{\tau}_{K,k}^O:\hat{\tau}_{K,k+1}^O}\right)\varepsilon_i^E \\
&\quad - \sum_{k=0}^{K} \left(\hat{\tau}_{K,k+1}^O - \hat{\tau}_{K,k}^O\right)\left(\overline{\varepsilon}^E_{\hat{\tau}_{K,k}^O:\hat{\tau}_{K,k+1}^O}\right)^2 + \sum_{k=0}^{K} \left(\tau_{k+1}^O - \tau_k^O\right)\left(\overline{\varepsilon}^E_{\tau_k^O:\tau_{k+1}^O}\right)^2 \\
&=: A_1 + 2A_2 + A_3 + A_4.
\end{aligned}$$

By Markov's inequality, $A_3 = O_{\mathbb{P}}\left(K\overline{\sigma}^2\right)$ and $A_4 = O_{\mathbb{P}}\left(K\overline{\sigma}^2\right)$.

In the following we will bound A_1 and A_2. To this end, let $\hat{\tau}_k^{\min} := \min\{\hat{\tau}_{K,k}^O, \tau_k^O\}$ and $\hat{\tau}_k^{\max} := \max\{\hat{\tau}_{K,k}^O, \tau_k^O\}$, $k = 0\dots, K+1$. Furthermore, we define $\mathbf{\Omega}_n$ to be the event in Assumption 3(i). Note that $\mathbb{P}(\mathbf{\Omega}_n) \to 1$. In the following we work on $\mathbf{\Omega}_n$.

From Assumptions 3(iii) and 5, we have that $\hat{\tau}_k^{\max} - \hat{\tau}_k^{\min} \le \delta_{0,k} \le \underline{\lambda}/4$. Consequently, $\hat{\tau}_{K,k+1}^O - \hat{\tau}_{K,k}^O \ge \underline{\lambda}/2$, $k = 0, \dots, K$.

Now for $i \in [\hat{\tau}_k^{\max}, \hat{\tau}_{k+1}^{\min}]$ we know that $\mu_i^O = \beta_k$, and

$$\begin{aligned}
\left|\overline{\mu}^O_{\hat{\tau}_{K,k}^O:\hat{\tau}_{K,k+1}^O} - \beta_k\right| &\le \frac{(\hat{\tau}_k^{\max} - \hat{\tau}_k^{\min})\Delta_k + (\hat{\tau}_{k+1}^{\max} - \hat{\tau}_{k+1}^{\min})\Delta_{k+1}}{\hat{\tau}_{K,k+1}^O - \hat{\tau}_{K,k}^O} \\
&\le \frac{\delta_{0,k}\Delta_k + \delta_{0,k+1}\Delta_{k+1}}{\max\left\{\underline{\lambda}/2, \hat{\tau}_{k+1}^{\min} - \hat{\tau}_k^{\max}\right\}},
\end{aligned}$$

for every $k = 0, \dots, K$, where we have used the notation $\Delta_0 = \Delta_{K+1} = 0$. Thus it

follows from the fact that $(x+y)^2 \le 2x^2+2y^2$ and $\delta_{0,k} \le \underline{\lambda}/2$, $k = 1, \dots, K$, that

$$\begin{aligned} A_1 \le & \sum_{k=0}^{K} \left(\hat{\tau}_{k+1}^{\min} - \hat{\tau}_k^{\max}\right) \left(\beta_k - \overline{\mu}^O_{\hat{\tau}^O_{K,k}:\hat{\tau}^O_{K,k+1}}\right)^2 \\ & + \sum_{k=1}^{K} \left(\hat{\tau}_k^{\max} - \hat{\tau}_k^{\min}\right) \max\left\{ \left(\beta_k - \overline{\mu}^O_{\hat{\tau}^O_{K,k-1}:\hat{\tau}^O_{K,k}}\right)^2, \left(\beta_{k-1} - \overline{\mu}^O_{\hat{\tau}^O_{K,k}:\hat{\tau}^O_{K,k+1}}\right)^2 \right\} \\ = & O\left(\sum_{k=1}^{K} \delta_{0,k}\Delta_k^2\right). \end{aligned}$$

Turning to A_2, for the following equality we use that, by definition of $\hat{\tau}_k^{\min}$ and $\hat{\tau}_k^{\max}$, $\mu_i^O - \overline{\mu}^O_{\hat{\tau}^O_{K,k}:\hat{\tau}^O_{K,k+1}}$ is constant between $\hat{\tau}_k^{\max}+1$ and $\hat{\tau}_{k+1}^{\min}$ as well as between $\hat{\tau}_k^{\min}$ and $\hat{\tau}_k^{\max}$. Next, the Cauchy–Schwarz inequality and $2|xy| \le x^2+y^2$ yield

$$\begin{aligned} |A_2| = & \left| \sum_{k=0}^{K} \sum_{i=\hat{\tau}_k^{\max}+1}^{\hat{\tau}_{k+1}^{\min}} \left(\mu_i^O - \overline{\mu}^O_{\hat{\tau}^O_{K,k}:\hat{\tau}^O_{K,k+1}}\right) \overline{\varepsilon}^E_{\hat{\tau}_k^{\max}:\hat{\tau}_{k+1}^{\min}} \right. \\ & \left. + \sum_{k=1}^{K} \sum_{i=\hat{\tau}_k^{\min}}^{\hat{\tau}_k^{\max}} \left(\mu_i^O - \overline{\mu}^O_{\hat{\tau}^O_{K,k}:\hat{\tau}^O_{K,k+1}}\right) \overline{\varepsilon}^E_{\hat{\tau}_k^{\min}:\hat{\tau}_k^{\max}} \right| \\ \le & \sum_{k=0}^{K} \sum_{i=\hat{\tau}^O_{K,k}+1}^{\hat{\tau}^O_{K,k+1}} \left(\mu_i^O - \overline{\mu}^O_{\hat{\tau}^O_{K,k}:\hat{\tau}^O_{K,k+1}}\right)^2 \\ & + \sum_{k=0}^{K} \left(\hat{\tau}_{k+1}^{\min} - \hat{\tau}_k^{\max}\right) \left(\overline{\varepsilon}^E_{\hat{\tau}_k^{\max}:\hat{\tau}_{k+1}^{\min}}\right)^2 + \sum_{k=1}^{K} \left(\hat{\tau}_k^{\max} - \hat{\tau}_k^{\min}\right) \left(\overline{\varepsilon}^E_{\hat{\tau}_k^{\min}:\hat{\tau}_k^{\max}}\right)^2. \end{aligned}$$

Finally, since $\varepsilon_1^O, \dots, \varepsilon_{n/2}^O$ and $\varepsilon_1^E, \dots, \varepsilon_{n/2}^E$ are independent, we obtain from Markov's inequality that

$$\begin{aligned} \sum_{k=0}^{K} \left(\hat{\tau}_{k+1}^{\min} - \hat{\tau}_k^{\max}\right) \left(\overline{\varepsilon}^E_{\hat{\tau}_k^{\max}:\hat{\tau}_{k+1}^{\min}}\right)^2 &= O_{\mathbb{P}}\left(K\overline{\sigma}^2\right), \\ \sum_{k=1}^{K} \left(\hat{\tau}_k^{\max} - \hat{\tau}_k^{\min}\right) \left(\overline{\varepsilon}^E_{\hat{\tau}_k^{\min}:\hat{\tau}_k^{\max}}\right)^2 &= O_{\mathbb{P}}\left(K\overline{\sigma}^2\right). \end{aligned}$$

Recall that $\mathbb{P}(\Omega_n) \to 1$. Thus, in total we obtain

$$S_E\left(\hat{\mathcal{T}}_K^O\right) - S_E\left(\mathcal{T}_K^O\right) = A_1 + A_2 + A_3 + A_4 = O_{\mathbb{P}}\left(\sum_{k=1}^{K} \delta_{0,k}\Delta_k^2\right) + O_{\mathbb{P}}\left(K\overline{\sigma}^2\right).$$

Moreover, Assumption 1(ii) yields $K = o(\log\log\overline{\lambda})$. Hence, it follows from Assumption 3(iv) that

$$S_E\left(\hat{\mathcal{T}}_K^O\right) - S_E\left(\mathcal{T}_K^O\right) = o_{\mathbb{P}}\left(\overline{\sigma}^2 \log\log\overline{\lambda}\right).$$

We now show (i). We have that

$$S_E\left(\hat{\mathcal{T}}_L^O\right) - S_E\left(\hat{\mathcal{T}}_K^O\right) = \left\{S_E\left(\hat{\mathcal{T}}_L^O\right) - S_E\left(\mathcal{T}_K^O\right)\right\} - \left\{S_E\left(\hat{\mathcal{T}}_K^O\right) - S_E\left(\mathcal{T}_K^O\right)\right\}.$$

From (iii) we have that $S_E\left(\hat{\mathcal{T}}_K^O\right) - S_E\left(\mathcal{T}_K^O\right) = o_{\mathbb{P}}\left(\overline{\sigma}^2 \log\log\overline{\lambda}\right)$.
Furthermore, it follows from (31) that

$$\begin{aligned}
&S_E\left(\hat{\mathcal{T}}_L^O\right) - S_E\left(\mathcal{T}_K^O\right) \\
&= \sum_{l=0}^{L} \sum_{i=\hat{\tau}_{L,l}^O+1}^{\hat{\tau}_{L,l+1}^O} \left(\mu_i^O - \overline{\mu}_{\hat{\tau}_{L,l}^O:\hat{\tau}_{L,l+1}^O}^O\right)^2 + 2\sum_{l=0}^{L} \sum_{i=\hat{\tau}_{L,l}^O+1}^{\hat{\tau}_{L,l+1}^O} \left(\mu_i^O - \overline{\mu}_{\hat{\tau}_{L,l}^O:\hat{\tau}_{L,l+1}^O}^O\right)\varepsilon_i^E + S_{\varepsilon^E}\left(\hat{\mathcal{T}}_L^O\right) - S_{\varepsilon^E}\left(\mathcal{T}_K^O\right).
\end{aligned}$$

In addition, Lemma 11 yields

$$\begin{aligned}
&\max_{L=0,\ldots,K-1} \left|S_{\varepsilon^E}\left(\hat{\mathcal{T}}_L^O\right) - S_{\varepsilon^E}\left(\hat{\mathcal{T}}_K^O\right)\right| \\
&\leq \max_{L=0,\ldots,K-1} \left\{S_{\varepsilon^E}\left(\hat{\mathcal{T}}_L^O\right) - S_{\varepsilon^E}\left(\hat{\mathcal{T}}_L^O \cup \mathcal{T}_K^O\right)\right\} + \max_{L=0,\ldots,K-1} \left\{S_{\varepsilon^E}\left(\mathcal{T}_K^O\right) - S_{\varepsilon^E}\left(\hat{\mathcal{T}}_L^O \cup \mathcal{T}_K^O\right)\right\} \\
&\quad + \left\{S_{\varepsilon^E}\left(\mathcal{T}_K^O\right) - S_{\varepsilon^E}\left(\hat{\mathcal{T}}_K^O \cup \mathcal{T}_K^O\right)\right\} + \left\{S_{\varepsilon^E}\left(\hat{\mathcal{T}}_K^O\right) - S_{\varepsilon^E}\left(\hat{\mathcal{T}}_K^O \cup \mathcal{T}_K^O\right)\right\} \\
&= o_{\mathbb{P}}(\overline{\sigma}^2 \log\log\overline{\lambda}).
\end{aligned} \tag{71}$$

Let $\hat{\tau}_{L,l,k}^O$, $l = 0,\ldots,L,\ k = 0,\ldots,\hat{K}_l$, be defined as in the proof of Lemma 11(i). Then,

$$\begin{aligned}
&\min_{L=0,\ldots,K-1} \left\{S_E\left(\hat{\mathcal{T}}_L^O\right) - S_E\left(\hat{\mathcal{T}}_K^O\right)\right\} \\
&= \min_{L=0,\ldots,K-1} \sum_{l=0}^{L} \sum_{k=0}^{\hat{K}_l} \left(\hat{\tau}_{L,l,k+1}^O - \hat{\tau}_{L,l,k}^O\right) \left[\left(\mu_{\hat{\tau}_{L,l,k}^O+1}^O - \overline{\mu}_{\hat{\tau}_{L,l}^O:\hat{\tau}_{L,l+1}^O}^O\right)^2 \right. \\
&\qquad\qquad \left. + 2\left(\mu_{\hat{\tau}_{L,l,k}^O+1}^O - \overline{\mu}_{\hat{\tau}_{L,l}^O:\hat{\tau}_{L,l+1}^O}^O\right)\overline{\varepsilon}_{\hat{\tau}_{L,l,k}^O:\hat{\tau}_{L,l,k+1}^O}^E\right] \\
&\quad + o_{\mathbb{P}}\left(\overline{\sigma}^2 \log\log(\overline{\lambda})\right).
\end{aligned}$$

Next, it follows from the Cauchy–Schwarz inequality that

$$\begin{aligned}
&\sum_{l=0}^{L} \sum_{k=0}^{\hat{K}_l} \left(\hat{\tau}_{L,l,k+1}^O - \hat{\tau}_{L,l,k}^O\right)\left(\mu_{\hat{\tau}_{L,l,k}^O+1}^O - \overline{\mu}_{\hat{\tau}_{L,l}^O:\hat{\tau}_{L,l+1}^O}^O\right)\overline{\varepsilon}_{\hat{\tau}_{L,l,k}^O:\hat{\tau}_{L,l,k+1}^O}^E \\
&\geq -\left(\sum_{l=0}^{L} \sum_{k=0}^{\hat{K}_l} \left(\hat{\tau}_{L,l,k+1}^O - \hat{\tau}_{L,l,k}^O\right)\left(\mu_{\hat{\tau}_{L,l,k}^O+1}^O - \overline{\mu}_{\hat{\tau}_{L,l}^O:\hat{\tau}_{L,l+1}^O}^O\right)^2\right)^{1/2} \\
&\quad \left(\sum_{l=0}^{L} \sum_{k=0}^{\hat{K}_l} \left(\hat{\tau}_{L,l,k+1}^O - \hat{\tau}_{L,l,k}^O\right)\left(\overline{\varepsilon}_{\hat{\tau}_{L,l,k}^O:\hat{\tau}_{L,l,k+1}^O}^E\right)^2\right)^{1/2}.
\end{aligned}$$

Then, Lemma 8 yields

$$\max_{L=0,\dots,K-1} \sum_{l=0}^{L} \sum_{k=0}^{\hat{K}_l} \left(\hat{\tau}^O_{L,l,k+1} - \hat{\tau}^O_{L,l,k}\right) \left(\overline{\varepsilon}^E_{\hat{\tau}^O_{L,l,k}:\hat{\tau}^O_{L,l,k+1}}\right)^2 = O_{\mathbb{P}}\left(\overline{\sigma}^2 (K \log K)^{1/2}\right),$$

since $\sum_{l=0}^{L}(\hat{K}_l + 1) \leq K + L \leq 2K$. In addition, Assumption 1(iii) yields that $(K \log K)^{1/2} \leq (K_{\max} \log K_{\max})^{1/2} = o(\log\log \overline{\lambda})$. Moreover, $|\mathcal{I}_L| \geq 1$ and Assumption 5 yield

$$\min_{L=0,\dots,K-1} \left(\underline{\lambda} \sum_{k \in \mathcal{I}_L} \Delta_k^2\right)^{-1} \overline{\sigma}^2 \log\log \overline{\lambda} \leq \left(\underline{\lambda} \Delta^2_{(1)}\right)^{-1} \overline{\sigma}^2 \log\log \overline{\lambda} \to 0.$$

In addition, it follows from the definition of $\overline{\mu}^O_{L,i}$ and (70) that

$$\begin{aligned}
&\min_{L=0,\dots,K-1} \left\{ \left(\underline{\lambda} \sum_{k \in \mathcal{I}_L} \Delta_k^2\right)^{-1} \sum_{l=0}^{L} \sum_{k=0}^{\hat{K}_l} \left(\hat{\tau}^O_{L,l,k+1} - \hat{\tau}^O_{L,l,k}\right) \left(\mu^O_{\hat{\tau}^O_{L,l,k}+1} - \overline{\mu}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}}\right)^2 \right\} \\
&= \min_{L=0,\dots,K-1} \left\{ \left(\underline{\lambda} \sum_{k \in \mathcal{I}_L} \Delta_k^2\right)^{-1} \sum_{k=0}^{K} \sum_{i=\tau^O_k+1}^{\tau^O_{k+1}} (\mu^O_i - \overline{\mu}^O_{L,i})^2 \right\} \\
&\geq \min_{L=0,\dots,K-1} \left\{ \left(\underline{\lambda} \sum_{k \in \mathcal{I}_L} \Delta_k^2\right)^{-1} \sum_{k \in \mathcal{I}_L} \sum_{i=\tau^O_k - \frac{\lambda}{4}+1}^{\tau^O_k + \frac{\lambda}{4}} (\mu^O_i - \overline{\mu}^O_{L,i})^2 \right\} \geq A,
\end{aligned}$$

with probability approaching 1. Thus,

$$\min_{L=0,\dots,K-1} \left\{ \left(\underline{\lambda} \sum_{k \in \mathcal{I}_L} \Delta_k^2\right)^{-1} \left(S_E\left(\hat{\mathcal{T}}^O_L\right) - S_E\left(\hat{\mathcal{T}}^O_K\right)\right) \right\} \geq A + o_{\mathbb{P}}(1).$$

□

The following lemma is a uniform version of Lemma 5 in Zou, Wang and Li (2020).

Lemma 13. *Suppose that Assumptions 1–5 hold in the case where $K \geq 1$ eventually, or only Assumptions 1, 2 and 4 in the case $K = 0$ $\forall$ n. Then,*

(i) $\max_{L=0,\dots,K-1} \left\{ S_{\varepsilon^O}\left(\mathcal{T}^O_K\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}^O_L \cup \mathcal{T}^O_K\right) \right\} = O_{\mathbb{P}}(K\overline{\sigma}^2 (\log \overline{\lambda})^2)$ *and*
$\max_{L=0,\dots,K-1} \left\{ S_{\varepsilon^O}\left(\hat{\mathcal{T}}^O_L\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}^O_L \cup \mathcal{T}^O_K\right) \right\} = O_{\mathbb{P}}(K\overline{\sigma}^2 (\log \overline{\lambda})^2)$,

(ii) $S_{\varepsilon^O}\left(\mathcal{T}^O_K\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}^O_K \cup \mathcal{T}^O_K\right) = o_{\mathbb{P}}(\overline{\sigma}^2 \log\log \overline{\lambda})$,

(iii) $\max_{L=K,\dots,K_{\max}} \left\{ S_{\varepsilon^O}\left(\hat{\mathcal{T}}^O_L\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}^O_L \cup \mathcal{T}^O_K\right) \right\} = o_{\mathbb{P}}(\overline{\sigma}^2 \log\log \overline{\lambda})$.

Proof. Without loss of generality we may assume $K > 0$, since the statements are trivial for $K = 0$.

(i) Let $L \in \{0, \dots, K-1\}$ be fixed. Recall that $\tau_k^O =: \tilde{\tau}_{L,k,0}^O \leq \tilde{\tau}_{L,k,1}^O < \cdots < \tilde{\tau}_{L,k,\tilde{L}_k}^O \leq \tilde{\tau}_{L,k,\tilde{L}_k+1}^O := \tau_{k+1}^O$ denote the estimated change-points in $\hat{\mathcal{T}}_L^O$ between τ_k^O and τ_{k+1}^O, so $\bigcup_{k=0}^K \bigcup_{l=0}^{\tilde{L}_k} \tilde{\tau}_{L,k,l}^O = \bigcup_{k=0}^K \tau_k^O \cup \bigcup_{l=0}^L \hat{\tau}_{L,l}^O$ and $\sum_{k=0}^K L_k = L$.

Then, Lemma 9 yields

$$0 \leq S_{\varepsilon^O}\left(\mathcal{T}_K^O\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}_L^O \cup \mathcal{T}_K^O\right) \leq 2 \sum_{\substack{k=0,\dots,K \\ \tilde{L}_k > 0}} \sum_{l=0}^{\tilde{L}_k} \left(\tilde{\tau}_{L,k,l+1}^O - \tilde{\tau}_{L,k,l}^O\right) \left(\overline{\varepsilon}_{\tilde{\tau}_{L,k,l}^O : \tilde{\tau}_{L,k,l+1}^O}^O\right)^2.$$

It follows from $\sum_{\substack{k=0,\dots,K \\ \tilde{L}_k > 0}} (\tilde{L}_k + 1) \leq 2L$ and Lemma 5 that

$$\begin{aligned}
&\max_{L=0,\dots,K-1} \sum_{\substack{k=0,\dots,K \\ \tilde{L}_k > 0}} \sum_{l=0}^{\tilde{L}_k} \left(\tilde{\tau}_{L,k,l+1}^O - \tilde{\tau}_{L,k,l}^O\right) \left(\overline{\varepsilon}_{\tilde{\tau}_{L,k,l}^O : \tilde{\tau}_{L,k,l+1}^O}^O\right)^2 \\
&\leq \max_{L=0,\dots,K-1} \sum_{\substack{k=0,\dots,K \\ \tilde{L}_k > 0}} (\tilde{L}_k + 1) \max_{\tau_k^O \leq i < j \leq \tau_{k+1}^O} (j-i) \left(\overline{\varepsilon}_{i:j}^O\right)^2 \\
&\leq \max_{L=0,\dots,K-1} 2L \max_{k=1,\dots,K} \max_{\tau_k^O \leq i < j \leq \tau_{k+1}^O} (j-i) \left(\overline{\varepsilon}_{i:j}^O\right)^2 \\
&= \mathcal{O}_{\mathbb{P}}\left(K\overline{\sigma}^2 \left((\log \overline{\lambda})^2 + (\log K)^2\right)\right).
\end{aligned}$$

Finally, Assumption 1(ii) yields that $\log K \leq \log \overline{\lambda}$ and the first statement follows.

For the second statement, let $L \in \{0, \dots, K-1\}$ be fixed. Recall that $\hat{\tau}_{L,l}^O =: \hat{\tau}_{L,l,0}^O < \hat{\tau}_{L,l,1}^O < \cdots < \hat{\tau}_{L,l,\hat{K}_l}^O < \hat{\tau}_{L,l,\hat{K}_l+1}^O := \hat{\tau}_{L,l+1}^O$ denote the true change-points between $\hat{\tau}_{L,l}^O$ and $\hat{\tau}_{L,l+1}^O$, so $\bigcup_{l=0}^L \bigcup_{k=0}^{\hat{K}_l} \hat{\tau}_{L,l,k}^O = \bigcup_{l=0}^L \hat{\tau}_{L,l}^O \cup \bigcup_{k=0}^K \tau_k^O$.

Then, Lemma 9 yields

$$0 \leq S_{\varepsilon^O}\left(\hat{\mathcal{T}}_L^O\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}_L^O \cup \mathcal{T}_K^O\right) \leq 2 \sum_{\substack{l=0,\dots,L \\ \hat{K}_l > 0}} \sum_{k=0}^{\hat{K}_l} \left(\hat{\tau}_{L,l,k+1}^O - \hat{\tau}_{L,l,k}^O\right) \left(\overline{\varepsilon}_{\hat{\tau}_{L,l,k}^O : \hat{\tau}_{L,l,k+1}^O}^O\right)^2.$$

It follows from Lemma 6 that

$$\begin{aligned}
&\max_{L=0,\ldots,K-1} \sum_{\substack{l=0,\ldots,L \\ \hat{K}_l>0}} \sum_{k=0}^{\hat{K}_l} \left(\hat{\tau}^O_{L,l,k+1} - \hat{\tau}^O_{L,l,k}\right) \left(\overline{\varepsilon}^O_{\hat{\tau}^O_{L,l,k}:\hat{\tau}^O_{L,l,k+1}}\right)^2 \\
&\leq \sum_{k=0}^{K} (\tau^O_{k+1} - \tau^O_k) \left(\overline{\varepsilon}^O_{\tau^O_k:\tau^O_{k+1}}\right)^2 \\
&\quad + \sum_{k=0}^{K} \max_{i=\tau^O_k+1,\ldots,\tau^O_{k+1}-1} (\tau^O_{k+1} - i) \left(\overline{\varepsilon}^O_{i:\tau^O_{k+1}}\right)^2 + \sum_{k=0}^{K} \max_{i=\tau^O_k+1,\ldots,\tau^O_{k+1}-1} (i - \tau^O_k) \left(\overline{\varepsilon}^O_{\tau^O_k:i}\right)^2 \\
&\leq O_{\mathbb{P}}\left(\sqrt{K}\overline{\sigma}^2\right) + (K+1) \max_{k=0,\ldots,K} \max_{i=\tau^O_k+1,\ldots,\tau^O_{k+1}-1} \left\{(\tau^O_{k+1} - i) \left(\overline{\varepsilon}^O_{i:\tau^O_{k+1}}\right)^2 + (i - \tau^O_k) \left(\overline{\varepsilon}^O_{\tau^O_k:i}\right)^2\right\} \\
&= O_{\mathbb{P}}\left(K\overline{\sigma}^2 \left(\log\log\overline{\lambda} + (\log K)^2\right)\right).
\end{aligned} \tag{72}$$

Finally, Assumption 1(ii) yields that $\log K \leq \log\overline{\lambda}$ and the statement follows.

(ii) Let $\mathcal{T}^\delta_L := \{\tau^O_k \pm \delta_{L-K,k},\ k = 1,\ldots,K\}$, $L = K,\ldots,K_{\max}$. Then,

$$\begin{aligned}
&S_{\varepsilon^O}\left(\mathcal{T}^O_K\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}^O_K \cup \mathcal{T}^O_K\right) \\
&= \left\{S_{\varepsilon^O}\left(\mathcal{T}^O_K\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}^O_K \cup \mathcal{T}^O_K \cup \mathcal{T}^\delta_K\right)\right\} - \left\{S_{\varepsilon^O}\left(\hat{\mathcal{T}}^O_K \cup \mathcal{T}^O_K\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}^O_K \cup \mathcal{T}^O_K \cup \mathcal{T}^\delta_K\right)\right\}.
\end{aligned}$$

We are working on the event in Assumption 3(i), but Assumption 3(i) assumes that the probability of that event converges to one. Using similar arguments as in (i) gives us

$$\begin{aligned}
&\left\{S_{\varepsilon^O}\left(\mathcal{T}^O_K\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}^O_K \cup \mathcal{T}^O_K \cup \mathcal{T}^\delta_K\right)\right\} - \left\{S_{\varepsilon^O}\left(\hat{\mathcal{T}}^O_K \cup \mathcal{T}^O_K\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}^O_K \cup \mathcal{T}^O_K \cup \mathcal{T}^\delta_K\right)\right\} \\
&= O_{\mathbb{P}}\left(K\overline{\sigma}^2 \left(\log\log\left(\max_{k=1,\ldots,K} \delta_{0,k} \vee e\right) + (\log K)^2\right)\right).
\end{aligned}$$

Finally, Assumption 3(ii) gives $K\log\log\left(\max_{k=1,\ldots,K} \delta_{0,k} \vee e\right) = o(\log\log\overline{\lambda})$ and Assumption 1(ii) yields $K(\log K)^2 = o(\log\log\overline{\lambda})$.

(iii) For $k = 1,\ldots,K$ let $\delta_k := \max_{q=0,\ldots,K_{\max}-K} \delta_{q,k}$. Moreover, let $\mathcal{T}^\delta_L := \{\tau^O_k \pm \delta_k,\ k = 1,\ldots,K\}$, $L = K,\ldots,K_{\max}$. Then,

$$\begin{aligned}
&S_{\varepsilon^O}\left(\hat{\mathcal{T}}^O_L\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}^O_L \cup \mathcal{T}^O_K\right) \\
&= \left\{S_{\varepsilon^O}\left(\hat{\mathcal{T}}^O_L\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}^O_L \cup \mathcal{T}^O_K \cup \mathcal{T}^\delta_L\right)\right\} - \left\{S_{\varepsilon^O}\left(\hat{\mathcal{T}}^O_L \cup \mathcal{T}^O_K\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}^O_L \cup \mathcal{T}^O_K \cup \mathcal{T}^\delta_L\right)\right\}.
\end{aligned}$$

We will focus on how to bound $\max_{L=K,\ldots,K_{\max}} \left\{S_{\varepsilon^O}\left(\hat{\mathcal{T}}^O_L\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}^O_L \cup \mathcal{T}^O_K \cup \mathcal{T}^\delta_L\right)\right\}$, but the same bound can be obtained for $\max_{L=K,\ldots,K_{\max}} \left\{S_{\varepsilon^O}\left(\hat{\mathcal{T}}^O_L \cup \mathcal{T}^O_K\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}^O_L \cup \mathcal{T}^O_K \cup \mathcal{T}^\delta_L\right)\right\}$ by using identical arguments.

Let $L \in \{K, \dots, K_{\max}\}$ be fixed. Let $\hat{\tau}^O_{L,l} =: \check{\tau}^O_{L,l,0} < \check{\tau}^O_{L,l,1} < \cdots < \check{\tau}^O_{L,l,\check{K}_l} < \check{\tau}^O_{L,l,\check{K}_l+1} := \hat{\tau}^O_{L,l+1}$ denote all elements of the set $\mathcal{T}^O_K \cup \mathcal{T}^\delta_L$ that are between $\hat{\tau}^O_{L,l}$ and $\hat{\tau}^O_{L,l+1}$, so $\bigcup_{l=0}^{L} \bigcup_{k=0}^{\check{K}_l} \check{\tau}^O_{L,l,k} = \bigcup_{l=0}^{L} \hat{\tau}^O_{L,l} \cup \mathcal{T}^O_K \cup \mathcal{T}^\delta_L$. Then, Lemma 9 yields

$$
\begin{aligned}
0 \leq & S_{\varepsilon^O}\left(\hat{\mathcal{T}}^O_L\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}^O_L \cup \mathcal{T}^O_K \cup \mathcal{T}^\delta_L\right) \\
\leq & 2 \sum_{\substack{l=0,\dots,L \\ \check{K}_l > 0}} \sum_{k=0}^{\check{K}_l} \frac{\left(\hat{\tau}^O_{L,l+1} - \hat{\tau}^O_{L,l}\right) - \left(\check{\tau}^O_{L,l,k+1} - \check{\tau}^O_{L,l,k}\right)}{\hat{\tau}^O_{L,l+1} - \hat{\tau}^O_{L,l}} \left(\check{\tau}^O_{L,l,k+1} - \check{\tau}^O_{L,l,k}\right) \left(\overline{\varepsilon}^O_{\check{\tau}^O_{L,l,k}:\check{\tau}^O_{L,l,k+1}}\right)^2 \\
\leq & 2 \sum_{\substack{l=0,\dots,L \\ \check{K}_l > 0}} \sum_{k=0}^{\check{K}_l} \left(\check{\tau}^O_{L,l,k+1} - \check{\tau}^O_{L,l,k}\right) \left(\overline{\varepsilon}^O_{\check{\tau}^O_{L,l,k}:\check{\tau}^O_{L,l,k+1}}\right)^2 .
\end{aligned}
$$

We are working on the event in Assumption 3(i), but Assumption 3(i) assumes that the probability of that event converges to one. Hence, for every $k = 1, \dots, K$ there exists an $l = 1, \dots, L$ (not necessarily unique) such that $\tau^O_k - \delta_k \leq \hat{\tau}^O_{L,l} \leq \tau^O_k$ or $\tau^O_k \leq \hat{\tau}^O_{L,l} \leq \tau^O_k + \delta_k$. Thus,

$$
\begin{aligned}
& \frac{1}{2} \max_{L=K,\dots,K_{\max}} \left\{ S_{\varepsilon^O}\left(\hat{\mathcal{T}}^O_L\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}^O_L \cup \mathcal{T}^O_K \cup \mathcal{T}^\delta_L\right) \right\} \\
& \leq \sum_{k=0}^{K} (\tau^O_{k+1} - \delta_{k+1} - \tau^O_k - \delta_k) \left(\overline{\varepsilon}^O_{\tau^O_k + \delta_k : \tau^O_{k+1} - \delta_{k+1}}\right)^2 + \sum_{k=1}^{K} \delta_k \left(\overline{\varepsilon}^O_{\tau^O_k - \delta_k : \tau^O_k}\right)^2 + \sum_{k=1}^{K} \delta_k \left(\overline{\varepsilon}^O_{\tau^O_k : \tau^O_k + \delta_k}\right)^2 \\
& \quad + \sum_{k=1}^{K} \max_{t = \tau^O_k - \delta_k + 1, \dots, \tau^O_k - 1} (\tau^O_k - t) \left(\overline{\varepsilon}^O_{\tau^O_k - \delta_k : t}\right)^2 + \sum_{k=1}^{K} \max_{t = \tau^O_k - \delta_k + 1, \dots, \tau^O_k - 1} (t - \tau^O_k + \delta_k) \left(\overline{\varepsilon}^O_{t : \tau^O_k}\right)^2 \\
& \quad + \sum_{k=1}^{K} \max_{t = \tau^O_k + 1, \dots, \tau^O_k + \delta_k - 1} (t - \tau^O_k) \left(\overline{\varepsilon}^O_{\tau^O_k : t}\right)^2 + \sum_{k=1}^{K} \max_{t = \tau^O_k + 1, \dots, \tau^O_k + \delta_k - 1} (\tau^O_k + \delta_k - t) \left(\overline{\varepsilon}^O_{t : \tau^O_k + \delta_k}\right)^2 \\
& \quad + \sum_{k=1}^{K} \max_{t = \tau^O_k + \delta_k + 1, \dots, \tau^O_{k+1} - \delta_{k+1} - 1} \frac{(t - \tau^O_k + \delta_k) - (t - \tau^O_k - \delta_k)}{t - \tau^O_k + \delta_k} (t - \tau^O_k - \delta_k) \left(\overline{\varepsilon}^O_{\tau^O_k + \delta_k : t}\right)^2 \\
& \quad + \sum_{k=1}^{K} \max_{t = \tau^O_{k-1} + \delta_{k-1} + 1, \dots, \tau^O_k - \delta_k - 1} \frac{(\tau^O_k + \delta_k - t) - (\tau^O_k - \delta_k - t)}{\tau^O_k + \delta_k - t} (\tau^O_k - \delta_k - t) \left(\overline{\varepsilon}^O_{t : \tau^O_k - \delta_k}\right)^2 ,
\end{aligned}
$$

where we have used the notation $\delta_0 := \delta_{K+1} := 0$.

The following calculation follows from a central limit theorem for triangular arrays. Note that Lyapunov's condition is met, since higher moments are uniformly bounded because of Assumption 2. We obtain

$$
\begin{aligned}
& \sum_{k=0}^{K} (\tau^O_{k+1} - \delta_{k+1} - \tau^O_k - \delta_k) \left(\overline{\varepsilon}^O_{\tau^O_k + \delta_k : \tau^O_{k+1} - \delta_{k+1}}\right)^2 + \sum_{k=1}^{K} \delta_k \left(\overline{\varepsilon}^O_{\tau^O_k - \delta_k : \tau^O_k}\right)^2 + \sum_{k=1}^{K} \delta_k \left(\overline{\varepsilon}^O_{\tau^O_k : \tau^O_k + \delta_k}\right)^2 \\
& = O_{\mathbb{P}}\left(\sqrt{K} \overline{\sigma}^2\right) .
\end{aligned}
$$

Furthermore, Assumption 1(ii) yields $\sqrt{K} = o(\log\log\overline{\lambda})$.

It follows from Lemma 6 that

$$\begin{aligned}
&\sum_{k=1}^{K} \max_{t=\tau_k^O-\delta_k+1,\ldots,\tau_k^O-1} (\tau_k^O - t)\left(\overline{\varepsilon}^O_{\tau_k^O-\delta_k:t}\right)^2 \\
\leq& K \max_{k=1,\ldots,K} \max_{t=\tau_k^O-\max_{k=1,\ldots,K}\delta_k+1,\ldots,\tau_k^O-1} (\tau_k^O - t)\left(\overline{\varepsilon}^O_{\tau_k^O-\max_{k=1,\ldots,K}\delta_k:t}\right)^2 \\
=& O_{\mathbb{P}}\left(K\overline{\sigma}^2\left(\log\log\left(\max_{k=1,\ldots,K}\delta_k \vee e\right) + (\log K)^2\right)\right).
\end{aligned}$$

Furthermore, Assumption 1(ii) yields $K(\log K)^2 = o(\log\log\overline{\lambda})$ and Assumption 3(ii) yields $K\log\log\left(\max_{k=1,\ldots,K}\delta_k \vee e\right) = o(\log\log\overline{\lambda})$. The same bound applies to the other terms and hence

$$\begin{aligned}
&\sum_{k=1}^{K} \max_{t=\tau_k^O-\delta_k+1,\ldots,\tau_k^O-1} (\tau_k^O - t)\left(\overline{\varepsilon}^O_{\tau_k^O-\delta_k:t}\right)^2 + \sum_{k=1}^{K} \max_{t=\tau_k^O-\delta_k+1,\ldots,\tau_k^O-1} (t - \tau_k^O + \delta_k)\left(\overline{\varepsilon}^O_{t:\tau_k^O}\right)^2 \\
&+ \sum_{k=1}^{K} \max_{t=\tau_k^O+1,\ldots,\tau_k^O+\delta_k-1} (t - \tau_k^O)\left(\overline{\varepsilon}^O_{\tau_k^O:t}\right)^2 + \sum_{k=1}^{K} \max_{t=\tau_k^O+1,\ldots,\tau_k^O+\delta_k-1} (\tau_k^O + \delta_k - t)\left(\overline{\varepsilon}^O_{t:\tau_k^O+\delta_k}\right)^2 \\
=& o_{\mathbb{P}}\left(\overline{\sigma}^2\log\log\overline{\lambda}\right).
\end{aligned}$$

Finally, let $C > 2$ be a constant. Then,

$$\begin{aligned}
&\sum_{k=1}^{K} \max_{t=\tau_k^O+\delta_k+1,\ldots,\tau_{k+1}^O-\delta_{k+1}-1} \frac{(t - \tau_k^O + \delta_k) - (t - \tau_k^O - \delta_k)}{t - \tau_k^O + \delta_k}(t - \tau_k^O - \delta_k)\left(\overline{\varepsilon}^O_{\tau_k^O+\delta_k:t}\right)^2 \\
\leq& \sum_{k=1}^{K} \max_{t=\tau_k^O+\delta_k+1,\ldots,\tau_{k+1}^O-\tau_k^O+C\{\max_{l=1,\ldots,K}\delta_l\}} (t - \tau_k^O - \delta_k)\left(\overline{\varepsilon}^O_{\tau_k^O+\delta_k:t}\right)^2 \\
&+ \sum_{k=1}^{K} \max_{t=\tau_k^O+C\{\max_{l=1,\ldots,K}\delta_l\}+1,\ldots,\tau_{k+1}^O-\delta_{k+1}-1} 2\delta_k\left(\overline{\varepsilon}^O_{\tau_k^O+\delta_k:t}\right)^2.
\end{aligned}$$

It follows from Lemma 6 that

$$\begin{aligned}
&\sum_{k=1}^{K} \max_{t=\tau_k^O+\delta_k+1,\ldots,\tau_{k+1}^O-\tau_k^O+C\{\max_{l=1,\ldots,K}\delta_l\}} (t - \tau_k^O - \delta_k)\left(\overline{\varepsilon}^O_{\tau_k^O+\delta_k:t}\right)^2 \\
\leq& K \max_{k=1,\ldots,K} \max_{t=\tau_k^O+\delta_k+1,\ldots,\tau_{k+1}^O-\tau_k^O+C\{\max_{l=1,\ldots,K}\delta_l\}} (t - \tau_k^O - \delta_k)\left(\overline{\varepsilon}^O_{\tau_k^O+\delta_k:t}\right)^2 \\
=& O_{\mathbb{P}}\left(K\overline{\sigma}^2\left(\log\log\left(\max_{k=1,\ldots,K}\delta_k \vee e\right) + (\log K)^2\right)\right).
\end{aligned}$$

Moreover, it follows from Lemma 7 that

$$\begin{aligned}
&\sum_{k=1}^{K} \max_{t=\tau_k^O+C\{\max_{l=1,\dots,K}\delta_l\}+1,\dots,\tau_{k+1}^O-\delta_{k+1}-1} 2\delta_k \left(\overline{\varepsilon}^O_{\tau_k^O+\delta_k:t}\right)^2 \\
\le& 2\sum_{k=1}^{K}\Big\{\max_{l=1,\dots,K}\delta_l\Big\}\delta_k \max_{t=\tau_k^O+C\{\max_{l=1,\dots,K}\delta_l\}+1,\dots,\tau_{k+1}^O-\delta_{k+1}-1} \left(\overline{\varepsilon}^O_{\tau_k^O+\delta_k:t}\right)^2 \\
\le& O_{\mathbb{P}}\left(K\overline{\sigma}^2(\log K)^2\right).
\end{aligned}$$

Thus, similarly to our earlier argument, it follows from Assumptions 3(ii) and Assumption 1(ii) that

$$\begin{aligned}
&\sum_{k=1}^{K} \max_{t=\tau_k^O+\delta_k+1,\dots,\tau_{k+1}^O-\delta_{k+1}-1} \frac{(t-\tau_k^O+\delta_k)-(t-\tau_k^O-\delta_k)}{t-\tau_k^O+\delta_k}(t-\tau_k^O-\delta_k)\left(\overline{\varepsilon}^O_{\tau_k^O+\delta_k:t}\right)^2 \\
&+\sum_{k=1}^{K} \max_{t=\tau_{k-1}^O+\delta_{k-1}+1,\dots,\tau_k^O-\delta_k-1} \frac{(\tau_k^O+\delta_k-t)-(\tau_k^O-\delta_k-t)}{\tau_k^O+\delta_k-t}(\tau_k^O-\delta_k-t)\left(\overline{\varepsilon}^O_{t:\tau_k^O-\delta_k}\right)^2 \\
=& o_{\mathbb{P}}\left(\overline{\sigma}^2\log\log\overline{\lambda}\right).
\end{aligned}$$

This completes the proof. □

Lemma 14. *Recall $E_i := \mu_i^O + \varepsilon_i^E$, $i = 1,\dots,n/2$ and $\widetilde{\mathrm{CV}}^O_{\mathrm{mod}}(L)$ as defined in (62). Suppose that Assumptions 1–5 hold in the case where $K \ge 1$ eventually, or only Assumptions 1, 2 and 4 in the case $K = 0\ \forall\ n$. Then we have the following.*

(i) Suppose that, in addition, $K \ge 1$ and for a constant $A > 0$ there exist sequences of non-empty sets $\mathcal{I}_L \subseteq \{1,\dots,K\}$ such that

$$\mathbb{P}\left(\forall\ L < K,\ \forall\ k \in \mathcal{I}_L,\ \sum_{i=\tau_k^O-\frac{\lambda}{4}+1}^{\tau_k^O+\frac{\lambda}{4}} (\mu_i^O - \overline{\mu}^O_{L,i})^2 \ge A\underline{\lambda}\Delta_k^2\right) \to 1,$$

with $\overline{\mu}^O_{L,i} := \sum_{l=0}^{L} \mathbb{1}_{\{\hat{\tau}^O_{L,l}+1 \le i \le \hat{\tau}^O_{L,l+1}\}} \overline{\mu}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}}$. Then,

$$\min_{L=0,\dots,K-1}\left(\sum_{k\in\mathcal{I}_L}\Delta_k^2\right)^{-1}\left\{\widetilde{\mathrm{CV}}^O_{\mathrm{mod}}(L) - \widetilde{\mathrm{CV}}^O_{\mathrm{mod}}(K)\right\} \ge \underline{\lambda}\,(A + o_{\mathbb{P}}(1)).$$

(ii)

$$\begin{aligned}
&\min_{L=K+1,\dots,K_{\max}}\left\{\widetilde{\mathrm{CV}}^O_{\mathrm{mod}}(L) - \widetilde{\mathrm{CV}}^O_{\mathrm{mod}}(K)\right\} \\
=&\min_{L=K+1,\dots,K_{\max}}\left\{S_{\varepsilon^O}\left(\mathcal{T}_K^O\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}_L^O \cup \mathcal{T}_K^O\right)\right\}(1 + o_{\mathbb{P}}(1)).
\end{aligned}$$

Proof. As seen in (65), we have for any $L \neq K$,

$$\begin{aligned}
&\widetilde{\mathrm{CV}}^O_{\mathrm{mod}}(L) - \widetilde{\mathrm{CV}}^O_{\mathrm{mod}}(K) \\
&= \left\{ S_E\left(\hat{\mathcal{T}}^O_L\right) - S_E\left(\hat{\mathcal{T}}^O_K\right) \right\} - \left\{ S_{\varepsilon^O}\left(\hat{\mathcal{T}}^O_L\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}^O_K\right) \right\} - \left\{ S_{\varepsilon^E}\left(\hat{\mathcal{T}}^O_L\right) - S_{\varepsilon^E}\left(\hat{\mathcal{T}}^O_K\right) \right\} \\
&\quad + 2\left\{ S_{\varepsilon^O,\varepsilon^E}\left(\hat{\mathcal{T}}^O_L\right) - S_{\varepsilon^O,\varepsilon^E}\left(\hat{\mathcal{T}}^O_K\right) \right\}.
\end{aligned}$$

In the following we will derive uniform bounds for each of the four terms separately.

We firstly show (i). Let $K \geq 1$. We have that

$$S_E\left(\hat{\mathcal{T}}^O_L\right) - S_E\left(\hat{\mathcal{T}}^O_K\right) = \left\{ S_E\left(\hat{\mathcal{T}}^O_L\right) - S_E\left(\mathcal{T}^O_K\right) \right\} - \left\{ S_E\left(\hat{\mathcal{T}}^O_K\right) - S_E\left(\mathcal{T}^O_K\right) \right\}.$$

For the first term, it follows from Lemma 12, (i), (iii), the fact that $|\mathcal{I}_L| \geq 1$ and Assumption 5, that

$$\min_{L=0,\ldots,K-1} \left(\sum_{k \in \mathcal{I}_L} \Delta_k^2 \right)^{-1} \left\{ S_E\left(\hat{\mathcal{T}}^O_L\right) - S_E\left(\hat{\mathcal{T}}^O_K\right) \right\} \geq \underline{\lambda}\left(A + o_{\mathbb{P}}(1)\right). \tag{73}$$

For the second term, it follows from Lemma 13 that

$$\begin{aligned}
&\max_{L=0,\ldots,K-1} \left| S_{\varepsilon^O}\left(\hat{\mathcal{T}}^O_L\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}^O_K\right) \right| \\
&\leq \max_{L=0,\ldots,K-1} \left\{ S_{\varepsilon^O}\left(\hat{\mathcal{T}}^O_L\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}^O_L \cup \mathcal{T}^O_K\right) \right\} + \max_{L=0,\ldots,K-1} \left\{ S_{\varepsilon^O}\left(\mathcal{T}^O_K\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}^O_L \cup \mathcal{T}^O_K\right) \right\} \\
&\quad + \left\{ S_{\varepsilon^O}\left(\mathcal{T}^O_K\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}^O_K \cup \mathcal{T}^O_K\right) \right\} + \left\{ S_{\varepsilon^O}\left(\hat{\mathcal{T}}^O_K\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}^O_K \cup \mathcal{T}^O_K\right) \right\} \\
&= O_{\mathbb{P}}(K\overline{\sigma}^2(\log\overline{\lambda})^2).
\end{aligned} \tag{74}$$

Abound for the third term is given in (71). We have that

$$\max_{L=0,\ldots,K-1} \left\{ S_{\varepsilon^E}\left(\hat{\mathcal{T}}^O_L\right) - S_{\varepsilon^E}\left(\hat{\mathcal{T}}^O_K\right) \right\} = o_{\mathbb{P}}(\overline{\sigma}^2 \log\log\overline{\lambda}). \tag{75}$$

For the last term, it follows from Lemma 10 that

$$
\begin{aligned}
&2 \max_{L=0,\ldots,K-1} \left| S_{\varepsilon^O,\varepsilon^E}\left(\hat{\mathcal{T}}_L^O\right) - S_{\varepsilon^O,\varepsilon^E}\left(\hat{\mathcal{T}}_K^O\right)\right| \\
\leq & 2 \max_{L=0,\ldots,K-1} \left| S_{\varepsilon^O,\varepsilon^E}\left(\hat{\mathcal{T}}_L^O\right) - S_{\varepsilon^O,\varepsilon^E}\left(\hat{\mathcal{T}}_L^O \cup \mathcal{T}_K^O\right)\right| + 2 \max_{L=0,\ldots,K-1} \left| S_{\varepsilon^O,\varepsilon^E}\left(\hat{\mathcal{T}}_L^O \cup \mathcal{T}_K^O\right) - S_{\varepsilon^O,\varepsilon^E}\left(\mathcal{T}_K^O\right)\right| \\
&+ 2\left| S_{\varepsilon^O,\varepsilon^E}\left(\mathcal{T}_K^O\right) - S_{\varepsilon^O,\varepsilon^E}\left(\hat{\mathcal{T}}_K^O \cup \mathcal{T}_K^O\right)\right| + 2\left| S_{\varepsilon^O,\varepsilon^E}\left(\hat{\mathcal{T}}_K^O \cup \mathcal{T}_K^O\right) - S_{\varepsilon^O,\varepsilon^E}\left(\hat{\mathcal{T}}_K^O\right)\right| \\
\leq & \max_{L=0,\ldots,K-1} \left\{ S_{\varepsilon^O}\left(\hat{\mathcal{T}}_L^O\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}_L^O \cup \mathcal{T}_K^O\right)\right\} + \max_{L=0,\ldots,K-1} \left\{ S_{\varepsilon^E}\left(\hat{\mathcal{T}}_L^O\right) - S_{\varepsilon^E}\left(\hat{\mathcal{T}}_L^O \cup \mathcal{T}_K^O\right)\right\} \\
&+ \max_{L=0,\ldots,K-1} \left\{ S_{\varepsilon^O}\left(\mathcal{T}_K^O\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}_L^O \cup \mathcal{T}_K^O\right)\right\} + \max_{L=0,\ldots,K-1} \left\{ S_{\varepsilon^E}\left(\mathcal{T}_K^O\right) - S_{\varepsilon^E}\left(\hat{\mathcal{T}}_L^O \cup \mathcal{T}_K^O\right)\right\} \\
&+ \left\{ S_{\varepsilon^O}\left(\mathcal{T}_K^O\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}_K^O \cup \mathcal{T}_K^O\right)\right\} + \left\{ S_{\varepsilon^E}\left(\mathcal{T}_K^O\right) - S_{\varepsilon^E}\left(\hat{\mathcal{T}}_K^O \cup \mathcal{T}_K^O\right)\right\} \\
&+ \left\{ S_{\varepsilon^O}\left(\hat{\mathcal{T}}_K^O\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}_K^O \cup \mathcal{T}_K^O\right)\right\} + \left\{ S_{\varepsilon^E}\left(\hat{\mathcal{T}}_K^O\right) - S_{\varepsilon^E}\left(\hat{\mathcal{T}}_K^O \cup \mathcal{T}_K^O\right)\right\} \\
= & O_{\mathbb{P}}\left(K\overline{\sigma}^2(\log\overline{\lambda})^2\right).
\end{aligned}
\tag{76}
$$

where the last bound is a consequence of Lemmas 9, 11 and 13 as seen before.

Thus, by combining (73)–(76) we have that

$$
\begin{aligned}
&\min_{L=0,\ldots,K-1} \left(\underline{\lambda} \sum_{k\in\mathcal{I}_L} \Delta_k^2\right)^{-1} \left\{ \widetilde{\mathrm{CV}}_{\mathrm{mod}}^O(L) - \widetilde{\mathrm{CV}}_{\mathrm{mod}}^O(K)\right\} \\
\geq & A + o_{\mathbb{P}}(1) + O_{\mathbb{P}}\left(\min_{L=0,\ldots,K-1} \left(\underline{\lambda} \sum_{k\in\mathcal{I}_L} \Delta_k^2\right)^{-1} K\overline{\sigma}^2(\log\overline{\lambda})^2\right).
\end{aligned}
$$

Hence, (i) follows from the fact that $|\mathcal{I}_L| \geq 1$ and Assumption 5.

We now show (ii). For the first term, it follows from Lemma 12(ii) and (iii) that

$$
\begin{aligned}
&\min_{L=K+1,\ldots,K_{\max}} \left\{ S_E\left(\hat{\mathcal{T}}_L^O\right) - S_E\left(\hat{\mathcal{T}}_K^O\right)\right\} \\
\geq & \min_{L=K+1,\ldots,K_{\max}} \left\{ S_E\left(\hat{\mathcal{T}}_L^O\right) - S_E\left(\hat{\mathcal{T}}_L^O \cup \mathcal{T}_K^O\right)\right\} \\
&- \max_{L=K+1,\ldots,K_{\max}} \left\{ S_E\left(\mathcal{T}_K^O\right) - S_E\left(\hat{\mathcal{T}}_L^O \cup \mathcal{T}_K^O\right)\right\} - \left| S_E\left(\hat{\mathcal{T}}_K^O\right) - S_E\left(\mathcal{T}_K^O\right)\right| \\
\geq & - \max_{L=K+1,\ldots,K_{\max}} \left\{ S_E\left(\mathcal{T}_K^O\right) - S_E\left(\hat{\mathcal{T}}_L^O \cup \mathcal{T}_K^O\right)\right\} - \left| S_E\left(\hat{\mathcal{T}}_K^O\right) - S_E\left(\mathcal{T}_K^O\right)\right| \\
= & - \left| o_{\mathbb{P}}(\overline{\sigma}^2 \log\log\overline{\lambda})\right|.
\end{aligned}
\tag{77}
$$

For the second term, it follows from Lemma 13 that

$$\begin{aligned}
&\min_{L=K+1,\ldots,K_{\max}} \left\{ S_{\varepsilon^O}\left(\hat{\mathcal{T}}_K^O\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}_L^O\right) \right\} \\
\geq & - \max_{L=K+1,\ldots,K_{\max}} \left\{ S_{\varepsilon^O}\left(\hat{\mathcal{T}}_L^O\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}_L^O \cup \mathcal{T}_K^O\right) \right\} + \min_{L=K+1,\ldots,K_{\max}} \left\{ S_{\varepsilon^O}\left(\mathcal{T}_K^O\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}_L^O \cup \mathcal{T}_K^O\right) \right\} \\
& - \left\{ S_{\varepsilon^O}\left(\mathcal{T}_K^O\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}_K^O \cup \mathcal{T}_K^O\right) \right\} - \left\{ S_{\varepsilon^O}\left(\hat{\mathcal{T}}_K^O\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}_K^O \cup \mathcal{T}_K^O\right) \right\} \\
= & \min_{L=K+1,\ldots,K_{\max}} \left\{ S_{\varepsilon^O}\left(\mathcal{T}_K^O\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}_L^O \cup \mathcal{T}_K^O\right) \right\} + o_{\mathbb{P}}(\overline{\sigma}^2 \log\log \overline{\lambda}).
\end{aligned} \tag{78}$$

For the third term, it follows from Lemma 11 that

$$\begin{aligned}
& - \min_{L=K+1,\ldots,K_{\max}} \left\{ S_{\varepsilon^E}\left(\hat{\mathcal{T}}_L^O\right) - S_{\varepsilon^E}\left(\hat{\mathcal{T}}_K^O\right) \right\} = \max_{L=K+1,\ldots,K_{\max}} \left\{ S_{\varepsilon^E}\left(\hat{\mathcal{T}}_L^O\right) - S_{\varepsilon^E}\left(\hat{\mathcal{T}}_K^O\right) \right\} \\
\leq & \max_{L=K+1,\ldots,K_{\max}} \left\{ S_{\varepsilon^E}\left(\hat{\mathcal{T}}_L^O\right) - S_{\varepsilon^E}\left(\hat{\mathcal{T}}_L^O \cup \mathcal{T}_K^O\right) \right\} + \max_{L=K+1,\ldots,K_{\max}} \left\{ S_{\varepsilon^E}\left(\hat{\mathcal{T}}_L^O \cup \mathcal{T}_K^O\right) - S_{\varepsilon^E}\left(\mathcal{T}_K^O\right) \right\} \\
& + \left\{ S_{\varepsilon^E}\left(\mathcal{T}_K^O\right) - S_{\varepsilon^E}\left(\hat{\mathcal{T}}_K^O \cup \mathcal{T}_K^O\right) \right\} + \left\{ S_{\varepsilon^E}\left(\hat{\mathcal{T}}_K^O \cup \mathcal{T}_K^O\right) - S_{\varepsilon^E}\left(\hat{\mathcal{T}}_K^O\right) \right\} \\
= & o_{\mathbb{P}}(\overline{\sigma}^2 \log\log \overline{\lambda}).
\end{aligned} \tag{79}$$

For the last term, we have that

$$\begin{aligned}
& 2\left| S_{\varepsilon^O,\varepsilon^E}\left(\hat{\mathcal{T}}_L^O\right) - S_{\varepsilon^O,\varepsilon^E}\left(\hat{\mathcal{T}}_K^O\right) \right| \\
\leq & 2\left| S_{\varepsilon^O,\varepsilon^E}\left(\hat{\mathcal{T}}_L^O\right) - S_{\varepsilon^O,\varepsilon^E}\left(\hat{\mathcal{T}}_L^O \cup \mathcal{T}_K^O\right) \right| + 2\left| S_{\varepsilon^O,\varepsilon^E}\left(\hat{\mathcal{T}}_L^O \cup \mathcal{T}_K^O\right) - S_{\varepsilon^O,\varepsilon^E}\left(\mathcal{T}_K^O\right) \right| \\
& + 2\left| S_{\varepsilon^O,\varepsilon^E}\left(\mathcal{T}_K^O\right) - S_{\varepsilon^O,\varepsilon^E}\left(\hat{\mathcal{T}}_K^O \cup \mathcal{T}_K^O\right) \right| + 2\left| S_{\varepsilon^O,\varepsilon^E}\left(\hat{\mathcal{T}}_K^O \cup \mathcal{T}_K^O\right) - S_{\varepsilon^O,\varepsilon^E}\left(\hat{\mathcal{T}}_K^O\right) \right|.
\end{aligned}$$

Moreover, it follows from the Cauchy–Schwarz inequality that

$$\begin{aligned}
& \left| S_{\varepsilon^O,\varepsilon^E}\left(\hat{\mathcal{T}}_L^O \cup \mathcal{T}_K^O\right) - S_{\varepsilon^O,\varepsilon^E}\left(\mathcal{T}_K^O\right) \right| \\
\leq & \left(S_{\varepsilon^O}\left(\mathcal{T}_K^O\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}_L^O \cup \mathcal{T}_K^O\right) \right)^{1/2} \left(S_{\varepsilon^E}\left(\mathcal{T}_K^O\right) - S_{\varepsilon^E}\left(\hat{\mathcal{T}}_L^O \cup \mathcal{T}_K^O\right) \right)^{1/2}
\end{aligned}$$

Thus, it follows from Lemmas 10, 9, 11 and 13 that

$$\begin{aligned}
& \min_{L=K+1,\ldots,K_{\max}} \left| S_{\varepsilon^O,\varepsilon^E}\left(\hat{\mathcal{T}}_L^O\right) - S_{\varepsilon^O,\varepsilon^E}\left(\hat{\mathcal{T}}_K^O\right) \right| \\
= & o_{\mathbb{P}}\left(\min_{L=K+1,\ldots,K_{\max}} \left\{ S_{\varepsilon^O}\left(\mathcal{T}_K^O\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}_L^O \cup \mathcal{T}_K^O\right) \right\} \right) + o_{\mathbb{P}}(\overline{\sigma}^2 \log\log \overline{\lambda}).
\end{aligned} \tag{80}$$

Thus, by combining (77)–(80) we have that

$$\begin{aligned}
& \min_{L=K+1,\ldots,K_{\max}} \left\{ \widetilde{\mathrm{CV}}_{\mathrm{mod}}^O(L) - \widetilde{\mathrm{CV}}_{\mathrm{mod}}^O(K) \right\} \\
= & \min_{L=K+1,\ldots,K_{\max}} \left\{ S_{\varepsilon^O}\left(\mathcal{T}_K^O\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}_L^O \cup \mathcal{T}_K^O\right) \right\} \\
& + o_{\mathbb{P}}\left(\min_{L=K+1,\ldots,K_{\max}} \left\{ S_{\varepsilon^O}\left(\mathcal{T}_K^O\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}_L^O \cup \mathcal{T}_K^O\right) \right\} \right) + o_{\mathbb{P}}(\overline{\sigma}^2 \log\log \overline{\lambda}).
\end{aligned}$$

Hence, (ii) follows from Assumption 4. □

Lemma 15. *Recall* $\mathrm{CV}^O_{\mathrm{mod}}(L)$ *and* $\widetilde{\mathrm{CV}}^O_{\mathrm{mod}}(L)$ *from* (61) *and* (62)*, respectively. Suppose that Assumptions 1–5 hold in the case where* $K \geq 1$ *eventually, or only Assumptions 1, 2 and 4 in the case* $K = 0\ \forall\ n$*. Let* $A_L^{(n)} := \left|\mathrm{CV}^O_{\mathrm{mod}}(L) - \widetilde{\mathrm{CV}}^O_{\mathrm{mod}}(L)\right|$*. Then,*

$$\max_{L=K+1,\ldots,K_{\max}} \left(S_{\varepsilon^O}\left(\mathcal{T}_K^O\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}_L^O \cup \mathcal{T}_K^O\right)\right)^{-1} \max\left(A_L^{(n)},\, A_K^{(n)}\right) = o_{\mathbb{P}}(1).$$

Moreover when $K \geq 1$ *eventually, there exist sequences of stochastic non-empty sets* $\mathcal{I}_L \subseteq \{1,\ldots,K\}$ *and a constant* $A > 0$ *such that*

$$\max_{L=0\ldots,K-1} \left(\sum_{k\in\mathcal{I}_L} \Delta_k^2\right)^{-1} \max\left(A_L^{(n)},\, A_K^{(n)}\right) = o_{\mathbb{P}}\left(\underline{\lambda}\right) \tag{81}$$

and

$$\mathbb{P}\left(\forall\, L < K,\ \sum_{i=\tau_k^O-\frac{\underline{\lambda}}{4}+1}^{\tau_k^O+\frac{\underline{\lambda}}{4}} (\mu_i^O - \overline{\mu}^O_{L,i})^2 \geq A\underline{\lambda}\Delta_k^2\ \forall\ k \in \mathcal{I}_L\right) \to 1,$$

where $\overline{\mu}^O_{L,i} := \sum_{l=0}^{L} \mathbb{1}_{\{\hat{\tau}^O_{L,l}+1 \leq i \leq \hat{\tau}^O_{L,l+1}\}} \overline{\mu}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}}$.

Proof. We have

$$\begin{aligned}
A_L^{(n)} \leq{}& \left|\sum_{l=0}^{L}\sum_{i=\hat{\tau}^O_{L,l}+1}^{\hat{\tau}^O_{L,l+1}-1} \frac{\hat{n}_l^O}{\hat{n}_l^O - 1}\left(\varepsilon_i^E - \overline{\varepsilon}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}}\right)^2 - \sum_{l=0}^{L}\sum_{i=\hat{\tau}^O_{L,l}+1}^{\hat{\tau}^O_{L,l+1}}\left(\varepsilon_i^E - \overline{\varepsilon}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}}\right)^2\right| \\
&+ \left|\sum_{l=0}^{L}\sum_{i=\hat{\tau}^O_{L,l}+1}^{\hat{\tau}^O_{L,l+1}-1} \frac{\hat{n}_l^O}{\hat{n}_l^O - 1}\left[2\left(\varepsilon_i^E - \overline{\varepsilon}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}}\right)\left(\mu_i^E - \overline{\mu}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}}\right) + \left(\mu_i^E - \overline{\mu}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}}\right)^2\right]\right. \\
&\left. - \sum_{l=0}^{L}\sum_{i=\hat{\tau}^O_{L,l}+1}^{\hat{\tau}^O_{L,l+1}}\left[2\left(\varepsilon_i^E - \overline{\varepsilon}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}}\right)\left(\mu_i^O - \overline{\mu}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}}\right) + \left(\mu_i^O - \overline{\mu}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}}\right)^2\right]\right|.
\end{aligned}$$

Also,

$$\sum_{l=0}^{L}\sum_{i=\hat\tau^O_{L,l}+1}^{\hat\tau^O_{L,l+1}-1}\frac{\hat n^O_l}{\hat n^O_l-1}\left[2\left(\varepsilon^E_i-\overline{\varepsilon}^O_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}}\right)\left(\mu^E_i-\overline{\mu}^O_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}}\right)+\left(\mu^E_i-\overline{\mu}^O_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}}\right)^2\right]$$

$$=\sum_{l=0}^{L}\sum_{i=\hat\tau^O_{L,l}+1}^{\hat\tau^O_{L,l+1}-1}\frac{\hat n^O_l}{\hat n^O_l-1}\left[2\left(\varepsilon^E_i-\overline{\varepsilon}^O_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}}\right)\left(\mu^E_i-\overline{\mu}^E_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}-1}\right)+\left(\mu^E_i-\overline{\mu}^E_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}-1}\right)^2\right]$$

$$+2\sum_{l=0}^{L}\sum_{i=\hat\tau^O_{L,l}+1}^{\hat\tau^O_{L,l+1}-1}\frac{\hat n^O_l}{\hat n^O_l-1}\left(\varepsilon^E_i-\overline{\varepsilon}^O_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}}\right)\left(\overline{\mu}^E_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}-1}-\overline{\mu}^O_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}}\right)$$

$$+\sum_{l=0}^{L}\sum_{i=\hat\tau^O_{L,l}+1}^{\hat\tau^O_{L,l+1}-1}\frac{\hat n^O_l}{\hat n^O_l-1}\left(\overline{\mu}^E_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}-1}-\overline{\mu}^O_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}}\right)^2$$

$$=\sum_{l=0}^{L}\sum_{i=\hat\tau^O_{L,l}+1}^{\hat\tau^O_{L,l+1}-1}\frac{\hat n^O_l}{\hat n^O_l-1}\left[2\left(\varepsilon^E_i-\overline{\varepsilon}^E_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}-1}\right)\left(\mu^E_i-\overline{\mu}^E_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}-1}\right)+\left(\mu^E_i-\overline{\mu}^E_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}-1}\right)^2\right]$$

$$+2\sum_{l=0}^{L}\sum_{i=\hat\tau^O_{L,l}+1}^{\hat\tau^O_{L,l+1}-1}\frac{\hat n^O_l}{\hat n^O_l-1}\left(\varepsilon^E_i-\overline{\varepsilon}^O_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}}\right)\left(\overline{\mu}^E_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}-1}-\overline{\mu}^O_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}}\right)$$

$$+\sum_{l=0}^{L}\hat n^O_l\left(\overline{\mu}^E_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}-1}-\overline{\mu}^O_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}}\right)^2.$$

For the last equation, note that

$$\sum_{i=\hat\tau^O_{L,l}+1}^{\hat\tau^O_{L,l+1}-1}\overline{\varepsilon}^O_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}}\left(\mu^E_i-\overline{\mu}^E_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}-1}\right)=0=\sum_{i=\hat\tau^O_{L,l}+1}^{\hat\tau^O_{L,l+1}-1}\overline{\varepsilon}^E_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}-1}\left(\mu^E_i-\overline{\mu}^E_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}-1}\right).$$

Hence,

$$
\begin{aligned}
A_L^{(n)} \leq & \left| \sum_{l=0}^{L} \sum_{i=\hat{\tau}_{L,l}^O+1}^{\hat{\tau}_{L,l+1}^O-1} \frac{\hat{n}_l^O}{\hat{n}_l^O-1} \left(\varepsilon_i^E - \overline{\varepsilon}^O_{\hat{\tau}_{L,l}^O:\hat{\tau}_{L,l+1}^O} \right)^2 - \sum_{l=0}^{L} \sum_{i=\hat{\tau}_{L,l}^O+1}^{\hat{\tau}_{L,l+1}^O} \left(\varepsilon_i^E - \overline{\varepsilon}^O_{\hat{\tau}_{L,l}^O:\hat{\tau}_{L,l+1}^O} \right)^2 \right| \\
& + \left| \sum_{l=0}^{L} \sum_{i=\hat{\tau}_{L,l}^O+1}^{\hat{\tau}_{L,l+1}^O-1} \frac{\hat{n}_l^O}{\hat{n}_l^O-1} \left(\mu_i^E - \overline{\mu}^E_{\hat{\tau}_{L,l}^O:\hat{\tau}_{L,l+1}^O-1} \right)^2 - \sum_{l=0}^{L} \sum_{i=\hat{\tau}_{L,l}^O+1}^{\hat{\tau}_{L,l+1}^O} \left(\mu_i^O - \overline{\mu}^O_{\hat{\tau}_{L,l}^O:\hat{\tau}_{L,l+1}^O} \right)^2 \right| \\
& + 2 \left| \sum_{l=0}^{L} \sum_{i=\hat{\tau}_{L,l}^O+1}^{\hat{\tau}_{L,l+1}^O-1} \frac{\hat{n}_l^O}{\hat{n}_l^O-1} \left(\varepsilon_i^E - \overline{\varepsilon}^E_{\hat{\tau}_{L,l}^O:\hat{\tau}_{L,l+1}^O-1} \right) \left(\mu_i^E - \overline{\mu}^E_{\hat{\tau}_{L,l}^O:\hat{\tau}_{L,l+1}^O-1} \right) \right. \\
& \left. - \sum_{l=0}^{L} \sum_{i=\hat{\tau}_{L,l}^O+1}^{\hat{\tau}_{L,l+1}^O} \left(\varepsilon_i^E - \overline{\varepsilon}^E_{\hat{\tau}_{L,l}^O:\hat{\tau}_{L,l+1}^O} \right) \left(\mu_i^O - \overline{\mu}^O_{\hat{\tau}_{L,l}^O:\hat{\tau}_{L,l+1}^O} \right) \right| \\
& + 2 \left| \sum_{l=0}^{L} \sum_{i=\hat{\tau}_{L,l}^O+1}^{\hat{\tau}_{L,l+1}^O-1} \frac{\hat{n}_l^O}{\hat{n}_l^O-1} \left(\varepsilon_i^E - \overline{\varepsilon}^O_{\hat{\tau}_{L,l}^O:\hat{\tau}_{L,l+1}^O} \right) \left(\overline{\mu}^E_{\hat{\tau}_{L,l}^O:\hat{\tau}_{L,l+1}^O-1} - \overline{\mu}^O_{\hat{\tau}_{L,l}^O:\hat{\tau}_{L,l+1}^O} \right) \right| \\
& + \left| \sum_{l=0}^{L} \hat{n}_l^O \left(\overline{\mu}^E_{\hat{\tau}_{L,l}^O:\hat{\tau}_{L,l+1}^O-1} - \overline{\mu}^O_{\hat{\tau}_{L,l}^O:\hat{\tau}_{L,l+1}^O} \right)^2 \right| \\
=: & A_{1,L}^{(n)} + A_{2,L}^{(n)} + 2A_{3,L}^{(n)} + 2A_{4,L}^{(n)} + A_{5,L}^{(n)},
\end{aligned}
$$

where we have used

$$
\sum_{i=\hat{\tau}_{L,l}^O+1}^{\hat{\tau}_{L,l+1}^O-1} \overline{\varepsilon}^E_{\hat{\tau}_{L,l}^O:\hat{\tau}_{L,l+1}^O} \left(\mu_i^O - \overline{\mu}^O_{\hat{\tau}_{L,l}^O:\hat{\tau}_{L,l+1}^O} \right) = 0 = \sum_{i=\hat{\tau}_{L,l}^O+1}^{\hat{\tau}_{L,l+1}^O} \overline{\varepsilon}^O_{\hat{\tau}_{L,l}^O:\hat{\tau}_{L,l+1}^O} \left(\mu_i^O - \overline{\mu}^O_{\hat{\tau}_{L,l}^O:\hat{\tau}_{L,l+1}^O} \right).
$$

We will bound each of the five terms in the following.

Bounding $A_{1,L}^{(n)}$

$$
\begin{aligned}
A_{1,L}^{(n)} \leq & \left| \sum_{l=0}^{L} \left\{ \frac{1}{\hat{n}_l^O - 1} \sum_{i=\hat{\tau}_{L,l}^O+1}^{\hat{\tau}_{L,l+1}^O-1} \left(\varepsilon_i^E \right)^2 - \left(\varepsilon^E_{\hat{\tau}_{L,l+1}^O} \right)^2 \right\} \right| \\
& + 2 \left| \sum_{l=0}^{L} \overline{\varepsilon}^O_{\hat{\tau}_{L,l}^O:\hat{\tau}_{L,l+1}^O} \left(\frac{1}{\hat{n}_l^O - 1} \sum_{i=\hat{\tau}_{L,l}^O+1}^{\hat{\tau}_{L,l+1}^O-1} \varepsilon_i^E - \varepsilon^E_{\hat{\tau}_{L,l+1}^O} \right) \right|.
\end{aligned} \tag{82}
$$

Let $\mathcal{E}^O$ be the sigma algebra generated by $\varepsilon_1^O, \ldots, \varepsilon_{n/2}^O$. Then,

$$\begin{aligned}
&\left| \sum_{l=0}^{L} \left\{ \frac{1}{\hat{n}_l^O - 1} \sum_{i=\hat{\tau}_{L,l}^O+1}^{\hat{\tau}_{L,l+1}^O-1} \left(\varepsilon_i^E\right)^2 - \left(\varepsilon_{\hat{\tau}_{L,l+1}^O}^E\right)^2 \right\} \right| \\
&\leq \left| \sum_{l=0}^{L} \left\{ \frac{1}{\hat{n}_l^O - 1} \sum_{i=\hat{\tau}_{L,l}^O+1}^{\hat{\tau}_{L,l+1}^O-1} \left(\varepsilon_i^E\right)^2 - \mathbb{E}\left[\frac{1}{\hat{n}_l^O - 1} \sum_{i=\hat{\tau}_{L,l}^O+1}^{\hat{\tau}_{L,l+1}^O-1} \left(\varepsilon_i^E\right)^2 \middle| \mathcal{E}^O \right] \right\} \right| \\
&\quad + \left| \sum_{l=0}^{L} \left\{ \left(\varepsilon_{\hat{\tau}_{L,l+1}^O}^E\right)^2 - \mathbb{E}\left[\left(\varepsilon_{\hat{\tau}_{L,l+1}^O}^E\right)^2 \middle| \mathcal{E}^O \right] \right\} \right| \\
&\quad + \sum_{l=0}^{L} \left| \mathbb{E}\left[\left(\varepsilon_{\hat{\tau}_{L,l+1}^O}^E\right)^2 \middle| \mathcal{E}^O \right] - \mathbb{E}\left[\frac{1}{\hat{n}_l^O - 1} \sum_{i=\hat{\tau}_{L,l}^O+1}^{\hat{\tau}_{L,l+1}^O-1} \left(\varepsilon_i^E\right)^2 \middle| \mathcal{E}^O \right] \right|.
\end{aligned} \tag{83}$$

In the following we bound the r.h.s. in (83). It follows from Lemma 8 that

$$\max_{L=0,\ldots,K_{\max}} \left| \sum_{l=0}^{L} \left\{ \frac{1}{\hat{n}_l^O - 1} \sum_{i=\hat{\tau}_{L,l}^O+1}^{\hat{\tau}_{L,l+1}^O-1} \left(\varepsilon_i^E\right)^2 - \mathbb{E}\left[\frac{1}{\hat{n}_l^O - 1} \sum_{i=\hat{\tau}_{L,l}^O+1}^{\hat{\tau}_{L,l+1}^O-1} \left(\varepsilon_i^E\right)^2 \middle| \mathcal{E}^O \right] \right\} \right| = O_{\mathbb{P}}\left((K_{\max} \log K_{\max})^{1/2}\, \overline{\sigma}^2 \right)$$

and

$$\max_{L=0,\ldots,K_{\max}} \left| \sum_{l=0}^{L} \left\{ \left(\varepsilon_{\hat{\tau}_{L,l+1}^O}^E\right)^2 - \mathbb{E}\left[\left(\varepsilon_{\hat{\tau}_{L,l+1}^O}^E\right)^2 \middle| \mathcal{E}^O \right] \right\} \right| = O_{\mathbb{P}}\left((K_{\max} \log K_{\max})^{1/2}\, \overline{\sigma}^2 \right).$$

Consider the last term on the r.h.s. in (83). Each conditional expectation is bounded by $\overline{\sigma}^2$ almost surely. Moreover, the conditional expectations can only differ if there is a true change-point between $\hat{\tau}_{L,l}^O$ and $\hat{\tau}_{L,l+1}^O$ as observations on the same segment have the same variance-covariance matrix. Hence, at most K terms can be non-zero. Thus almost surely

$$\max_{L=0,\ldots,K_{\max}} \sum_{l=0}^{L} \left| \mathbb{E}\left[\left(\varepsilon_{\hat{\tau}_{L,l+1}^O}^E\right)^2 \middle| \mathcal{E}^O \right] - \mathbb{E}\left[\frac{1}{\hat{n}_l^O - 1} \sum_{i=\hat{\tau}_{L,l}^O+1}^{\hat{\tau}_{L,l+1}^O-1} \left(\varepsilon_i^E\right)^2 \middle| \mathcal{E}^O \right] \right| \leq K\overline{\sigma}^2.$$

Hence,

$$\max_{L=0,\ldots,K_{\max}} \left| \sum_{l=0}^{L} \left\{ \frac{1}{\hat{n}_l^O - 1} \sum_{i=\hat{\tau}_{L,l}^O+1}^{\hat{\tau}_{L,l+1}^O-1} \left(\varepsilon_i^E\right)^2 - \left(\varepsilon_{\hat{\tau}_{L,l+1}^O}^E\right)^2 \right\} \right| = O_{\mathbb{P}}\left((K_{\max} \log K_{\max})^{1/2}\, \overline{\sigma}^2 \right).$$

Recall that we have assumed $(K_{\max} \log K_{\max})^{1/2} = o(\log \log \overline{\lambda})$, see Assumption 1(iii). Thus putting things together, we see that the first term in (82)

satisfies

$$\max_{L=0,\ldots,K_{\max}} \left| \sum_{l=0}^{L} \left\{ \frac{1}{\hat{n}_l^O - 1} \sum_{i=\hat{\tau}_{L,l}^O+1}^{\hat{\tau}_{L,l+1}^O-1} \left(\varepsilon_i^E\right)^2 - \left(\varepsilon_{\hat{\tau}_{L,l+1}^O}^E\right)^2 \right\} \right| = o_{\mathbb{P}}(\overline{\sigma}^2 \log\log \overline{\lambda}). \tag{84}$$

We now bound the second term in the r.h.s. of (82):

$$\begin{aligned} &\left| \sum_{l=0}^{L} \overline{\varepsilon}_{\hat{\tau}_{L,l}^O:\hat{\tau}_{L,l+1}^O}^O \left(\frac{1}{\hat{n}_l^O - 1} \sum_{i=\hat{\tau}_{L,l}^O+1}^{\hat{\tau}_{L,l+1}^O-1} \varepsilon_i^E - \varepsilon_{\hat{\tau}_{L,l+1}^O}^E \right) \right| \\ &\leq \left| \sum_{l=0}^{L} \overline{\varepsilon}_{\hat{\tau}_{L,l}^O:\hat{\tau}_{L,l+1}^O}^O \frac{1}{\hat{n}_l^O - 1} \sum_{i=\hat{\tau}_{L,l}^O+1}^{\hat{\tau}_{L,l+1}^O-1} \varepsilon_i^E \right| + \left| \sum_{l=0}^{L} \overline{\varepsilon}_{\hat{\tau}_{L,l}^O:\hat{\tau}_{L,l+1}^O}^O \varepsilon_{\hat{\tau}_{L,l+1}^O}^E \right|. \end{aligned}$$

We focus on the second term in the display above, but the same steps lead to the same bound for the first term. It follows from the law of total probability that for any random variable $x > 0$ almost surely

$$\begin{aligned} &\mathbb{P}\left(\left| \sum_{l=0}^{L} \overline{\varepsilon}_{\hat{\tau}_{L,l}^O:\hat{\tau}_{L,l+1}^O}^O \varepsilon_{\hat{\tau}_{L,l+1}^O}^E \right| \geq x \right) \\ =&\mathbb{P}\left(\sum_{l=0}^{L} \overline{\varepsilon}_{\hat{\tau}_{L,l}^O:\hat{\tau}_{L,l+1}^O}^O \varepsilon_{\hat{\tau}_{L,l+1}^O}^E \geq x \right) + \mathbb{P}\left(\sum_{l=0}^{L} \overline{\varepsilon}_{\hat{\tau}_{L,l}^O:\hat{\tau}_{L,l+1}^O}^O \varepsilon_{\hat{\tau}_{L,l+1}^O}^E \leq -x \right) \\ =&\mathbb{E}\left[\mathbb{P}\left(\sum_{l=0}^{L} \overline{\varepsilon}_{\hat{\tau}_{L,l}^O:\hat{\tau}_{L,l+1}^O}^O \varepsilon_{\hat{\tau}_{L,l+1}^O}^E \geq x \middle| \mathcal{E}^O \right) \right] + \mathbb{E}\left[\mathbb{P}\left(\sum_{l=0}^{L} \overline{\varepsilon}_{\hat{\tau}_{L,l}^O:\hat{\tau}_{L,l+1}^O}^O \varepsilon_{\hat{\tau}_{L,l+1}^O}^E \leq -x \middle| \mathcal{E}^O \right) \right]. \end{aligned}$$

We focus on the first term and the probability inside the expectation. The second term can be bounded in the same way. Let $L = 0, \ldots, K_{\max}$ be fixed. We will use Bernstein's inequality conditional on $\mathcal{E}^O$. To this end, note that $\overline{\varepsilon}_{\hat{\tau}_{L,l}^O:\hat{\tau}_{L,l+1}^O}^O \varepsilon_{\hat{\tau}_{L,l+1}^O}^E$, $l = 0, \ldots, L$, are independent conditional on $\mathcal{E}^O$ with

$$\mathbb{E}\left[\left(\overline{\varepsilon}_{\hat{\tau}_{L,l}^O:\hat{\tau}_{L,l+1}^O}^O \right)^t \varepsilon_{\hat{\tau}_{L,l+1}^O}^E \middle| \mathcal{E}^O \right] = 0, \quad \forall\, l = 0, \ldots, L.$$

Moreover, it follows from the Cauchy–Schwarz inequality and Assumption 2

that

$$\sum_{l=0}^{L}\mathbb{E}\left[\left(\left(\overline{\varepsilon}^{O}_{\hat{\tau}^{O}_{L,l}:\hat{\tau}^{O}_{L,l+1}}\right)^{t}\varepsilon^{E}_{\hat{\tau}^{O}_{L,l+1}}\right)^{2}\middle|\mathcal{E}^{O}\right] \leq \sum_{l=0}^{L}\mathbb{E}\left[\left\|\overline{\varepsilon}^{O}_{\hat{\tau}^{O}_{L,l}:\hat{\tau}^{O}_{L,l+1}}\right\|_2^2\left\|\varepsilon^{E}_{\hat{\tau}^{O}_{L,l+1}}\right\|_2^2\middle|\mathcal{E}^{O}\right]$$
$$\leq \overline{\sigma}^2\sum_{l=0}^{L}\left\|\overline{\varepsilon}^{O}_{\hat{\tau}^{O}_{L,l}:\hat{\tau}^{O}_{L,l+1}}\right\|_2^2,$$

$$\sum_{l=0}^{L}\mathbb{E}\left[\left(\left(\overline{\varepsilon}^{O}_{\hat{\tau}^{O}_{L,l}:\hat{\tau}^{O}_{L,l+1}}\right)^{t}\varepsilon^{E}_{\hat{\tau}^{O}_{L,l+1}}\right)^{q}\middle|\mathcal{E}^{O}\right] \leq \sum_{l=0}^{L}\mathbb{E}\left[\left\|\overline{\varepsilon}^{O}_{\hat{\tau}^{O}_{L,l}:\hat{\tau}^{O}_{L,l+1}}\right\|_2^q\left\|\varepsilon^{E}_{\hat{\tau}^{O}_{L,l+1}}\right\|_2^q\middle|\mathcal{E}^{O}\right]$$
$$\leq \frac{q!}{2}c^{q-2}\overline{\sigma}^{q}\sum_{l=0}^{L}\left\|\overline{\varepsilon}^{O}_{\hat{\tau}^{O}_{L,l}:\hat{\tau}^{O}_{L,l+1}}\right\|_2^q \leq \frac{q!}{2}\overline{\sigma}^2\sum_{l=0}^{L}\left\|\overline{\varepsilon}^{O}_{\hat{\tau}^{O}_{L,l}:\hat{\tau}^{O}_{L,l+1}}\right\|_2^2\left(c\,\overline{\sigma}\left(\sum_{l=0}^{L}\left\|\overline{\varepsilon}^{O}_{\hat{\tau}^{O}_{L,l}:\hat{\tau}^{O}_{L,l+1}}\right\|_2^2\right)^{1/2}\right)^{q-2},$$

for $q \geq 3$. Thus, it follows from Bernstein's inequality (Boucheron, Lugosi and Massart, 2013, Cor. 2.11) that for any $a \geq 1$

$$\mathbb{P}\left(\sum_{l=0}^{L}\overline{\varepsilon}^{O}_{\hat{\tau}^{O}_{L,l}:\hat{\tau}^{O}_{L,l+1}}\varepsilon^{E}_{\hat{\tau}^{O}_{L,l+1}} \geq a\log\left(K_{\max}\right)\overline{\sigma}\left(\sum_{l=1}^{L}\left(\overline{\varepsilon}^{O}_{\hat{\tau}^{O}_{L,l}:\hat{\tau}^{O}_{L,l+1}}\right)^2\right)^{1/2}(\sqrt{2}+c)\middle|\mathcal{E}^{O}\right)$$
$$\leq \exp(-a\log\left(K_{\max}\right)) = K_{\max}^{-a}.$$

Hence,

$$\mathbb{P}\left(\left|\sum_{l=0}^{L}\overline{\varepsilon}^{O}_{\hat{\tau}^{O}_{L,l}:\hat{\tau}^{O}_{L,l+1}}\varepsilon^{E}_{\hat{\tau}^{O}_{L,l+1}}\right| \geq a\log\left(K_{\max}\right)\overline{\sigma}\left(\sum_{l=1}^{L}\left(\overline{\varepsilon}^{O}_{\hat{\tau}^{O}_{L,l}:\hat{\tau}^{O}_{L,l+1}}\right)^2\right)^{1/2}(\sqrt{2}+c)\right) \leq 2K_{\max}^{-a}$$

and

$$\max_{L=0,\ldots,K_{\max}}\left(\sum_{l=1}^{L}\left(\overline{\varepsilon}^{O}_{\hat{\tau}^{O}_{L,l}:\hat{\tau}^{O}_{L,l+1}}\right)^2\right)^{-1/2}\left|\sum_{l=0}^{L}\overline{\varepsilon}^{O}_{\hat{\tau}^{O}_{L,l}:\hat{\tau}^{O}_{L,l+1}}\varepsilon^{E}_{\hat{\tau}^{O}_{L,l+1}}\right| = O_{\mathbb{P}}\left(\log\left(K_{\max}\right)\overline{\sigma}\right).$$

We now simplify this term. To this end, we use the notation

$$\left\{\tilde{\tau}^{O}_{L,0},\ldots,\tilde{\tau}^{O}_{L,\tilde{L}+1}\right\} := \left\{\tau^{O}_{0},\ldots,\tau^{O}_{K+1}\right\} \cup \left\{\hat{\tau}^{O}_{L,1},\ldots,\hat{\tau}^{O}_{L,L}\right\},$$

with $\tilde{L} \leq K + L$. Then,

$$\begin{aligned}
&\sum_{l=0}^{L} \left(\overline{\varepsilon}^{O}_{\hat{\tau}^{O}_{L,l}:\hat{\tau}^{O}_{L,l+1}} \right)^2 \leq \sum_{l=0}^{\tilde{L}} \left(\tilde{\tau}^{O}_{L,l+1} - \tilde{\tau}^{O}_{L,l} \right) \left(\overline{\varepsilon}^{O}_{\tilde{\tau}^{O}_{L,l}:\tilde{\tau}^{O}_{L,l+1}} \right)^2 \\
=&\sum_{k=0}^{K} \sum_{i=\tau^{O}_{k}+1}^{\tau^{O}_{k+1}} \left(\varepsilon^{O}_{i} - \overline{\varepsilon}^{O}_{\tau^{O}_{k}:\tau^{O}_{k+1}} \right)^2 - \sum_{l=0}^{\tilde{L}} \sum_{i=\tilde{\tau}^{O}_{L,l}+1}^{\tilde{\tau}^{O}_{L,l+1}} \left(\varepsilon^{O}_{i} - \overline{\varepsilon}^{O}_{\tilde{\tau}^{O}_{L,l}:\tilde{\tau}^{O}_{L,l+1}} \right)^2 \\
&+ \sum_{k=0}^{K} \left(\tau^{O}_{k+1} - \tau^{O}_{k} \right) \left(\overline{\varepsilon}^{O}_{\tau^{O}_{k}:\tau^{O}_{k+1}} \right)^2 \\
=&S_{\varepsilon^{O}} \left(\mathcal{T}^{O}_{K} \right) - S_{\varepsilon^{O}} \left(\hat{\mathcal{T}}^{O}_{L} \cup \mathcal{T}^{O}_{K} \right) + \sum_{k=0}^{K} \left(\tau^{O}_{k+1} - \tau^{O}_{k} \right) \left(\overline{\varepsilon}^{O}_{\tau^{O}_{k}:\tau^{O}_{k+1}} \right)^2 .
\end{aligned} \tag{85}$$

Note that by Markov's inequality,

$$\sum_{k=0}^{K} \left(\tau^{O}_{k+1} - \tau^{O}_{k} \right) \left(\overline{\varepsilon}^{O}_{\tau^{O}_{k}:\tau^{O}_{k+1}} \right)^2 = O_{\mathbb{P}} \left((K \vee 1) \overline{\sigma}^2 \right).$$

Let us now consider $L = K$. From Lemma 13(ii) we have that

$$S_{\varepsilon^{O}} \left(\mathcal{T}^{O}_{K} \right) - S_{\varepsilon^{O}} \left(\hat{\mathcal{T}}^{O}_{K} \cup \mathcal{T}^{O}_{K} \right) = o_{\mathbb{P}} \left(\overline{\sigma}^2 \log\log \overline{\lambda} \right).$$

Moreover, Assumption 1(ii) yields $K = o(\log\log\overline{\lambda})$. Thus,

$$\sum_{k=0}^{K} \left(\overline{\varepsilon}^{O}_{\hat{\tau}^{O}_{K,k}:\hat{\tau}^{O}_{K,k+1}} \right)^2 = o_{\mathbb{P}}(\overline{\sigma}^2 \log\log\overline{\lambda}), \tag{86}$$

and this also holds when $K = 0$. Furthermore, it follows from Assumption 1(iii) that $\log K_{\max} = o\left((\log\log\overline{\lambda})^{1/2} \right)$. To this end, note that it is trivially satisfied if $K_{\max} = O(1)$ and otherwise since $(\log K_{\max})^3 / K_{\max} \to 0$. Hence,

$$\left| \sum_{k=0}^{K} \overline{\varepsilon}^{O}_{\hat{\tau}^{O}_{K,k}:\hat{\tau}^{O}_{K,k+1}} \varepsilon^{E}_{\hat{\tau}^{O}_{K,k+1}} \right| = o_{\mathbb{P}}(\overline{\sigma}^2 \log\log\overline{\lambda})$$

and thus using (84) we have

$$A^{(n)}_{1,K} = o_{\mathbb{P}}(\overline{\sigma}^2 \log\log\overline{\lambda}).$$

Next, we consider $L > K$. Using the same arguments and Assumption 4 gives us

$$\max_{L=K+1,\ldots,K_{\max}} \left(S_{\varepsilon^{O}} \left(\mathcal{T}^{O}_{K} \right) - S_{\varepsilon^{O}} \left(\hat{\mathcal{T}}^{O}_{L} \cup \mathcal{T}^{O}_{K} \right) \right)^{-1} \sum_{l=0}^{L} \left(\overline{\varepsilon}^{O}_{\hat{\tau}^{O}_{L,l}:\hat{\tau}^{O}_{L,l+1}} \right)^2 = O_{\mathbb{P}}(1) \tag{87}$$

and hence additionally using (84) we have

$$\max_{L=K+1,\ldots,K_{\max}} \left(S_{\varepsilon^O}\left(\mathcal{T}_K^O\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}_L^O \cup \mathcal{T}_K^O\right)\right)^{-1} \max\left(A_{1,L}^{(n)}, A_{1,K}^{(n)}\right) = o_{\mathbb{P}}(1).$$

Note that this completes the proof in the case where $K = 0\ \forall\ n$ as we only have to consider $L \geq K$ and since all μ_i's are the same, we have that $A_L^{(n)} \leq A_{1,L}^{(n)}$. Moreover, note that the arguments above in the case where $K = 0$ do not use Assumptions 3 and 5. In the following we may therefore assume that $K \geq 1$.

We now consider the final case where $L < K$. From Lemma 13(i) we have that

$$\max_{L=0\ldots,K-1} \left\{S_{\varepsilon^O}\left(\mathcal{T}_K^O\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}_L^O \cup \mathcal{T}_K^O\right)\right\} = O_{\mathbb{P}}\big(K\overline{\sigma}^2(\log\overline{\lambda})^2\big).$$

Thus,

$$\max_{L=0\ldots,K-1} \sum_{l=0}^{L} \left(\overline{\varepsilon}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}}\right)^2 = O_{\mathbb{P}}\big(K\overline{\sigma}^2(\log\overline{\lambda})^2\big). \tag{88}$$

Consequently,

$$\max_{L=0\ldots,K-1} \left|\sum_{l=0}^{L} \overline{\varepsilon}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}} \varepsilon^E_{\hat{\tau}^O_{L,l+1}}\right| = O_{\mathbb{P}}\big(K\overline{\sigma}^2(\log\overline{\lambda})^2\big)$$

and hence from (84) we have

$$\max_{L=0\ldots,K-1} \max\left(A_{1,L}^{(n)}, A_{1,K}^{(n)}\right) = O_{\mathbb{P}}\big(K\overline{\sigma}^2(\log\overline{\lambda})^2\big).$$

Finally, we show below that for some (stochastic) sequences of sets $\mathcal{I}_L$ satisfying the assumptions of the lemma, we have $K\overline{\sigma}^2(\log\overline{\lambda})^2 = o_{\mathbb{P}}\left(\min_L \underline{\lambda} \sum_{k\in\mathcal{I}_L} \Delta_k^2\right)$, see (90). In summary, we will then have a version of (81), and hence of Lemma 15, with $A_L^{(n)}$ replaced by $A_{1,L}^{(n)}$.

Preliminary calculations First observe that $\mu_i^E - \mu_i^O = \mu_{2i} - \mu_{2i-1} = \beta_k - \beta_{k-1}$ if $\tau_k = 2i - 1$ and zero if such a k does not exist. Hence, $|\mu_i^E - \mu_i^O| \leq \Delta_k$ if $\tau_k^O = i$ and zero if such a k does not exist.

Now let the event Ω_n be the intersection of the events in (i) and (v) of Assumption 3. We note that Ω_n must have probability converging to 1. In the following we work on Ω_n.

Let us fix $L \geq K$. We have that for each $l = 0, \ldots, L$, there exists a unique $k = k(l) \in \{1, \ldots, K\}$ such that exactly one of the following scenarios occur:

(1) $\tau_k^O \leq \hat{\tau}_{L,l}^O < \hat{\tau}_{L,l+1}^O \leq \tau_{k+1}^O$,
(2) $\tau_{k-1}^O \leq \hat{\tau}_{L,l}^O < \tau_k^O < \hat{\tau}_{L,l+1}^O \leq \tau_{k+1}^O$,
(3) $\tau_{k-1}^O \leq \hat{\tau}_{L,l}^O < \tau_k^O < \tau_{k+1}^O < \hat{\tau}_{L,l+1}^O \leq \tau_{k+2}^O$.

Observe that if $k(l_1) = k(l_2)$ for any $l_1 \neq l_2$, then scenario (1) must occur for at least one of l_1 and l_2.

Furthermore, we have that if (1) occurs, then

$$\left|\overline{\mu}^E_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}-1} - \overline{\mu}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}}\right| = 0,$$
$$\sum_{i=\hat{\tau}^O_{L,l}+1}^{\hat{\tau}^O_{L,l+1}-1} \left|\mu^E_i - \mu^O_i\right| = 0,$$
$$\max_{i=\hat{\tau}^O_{L,l}+1,\ldots,\hat{\tau}^O_{L,l+1}-1} \left|\mu^E_i - \overline{\mu}^E_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}-1}\right| = 0,$$
$$\max_{i=\hat{\tau}^O_{L,l}+1,\ldots,\hat{\tau}^O_{L,l+1}} \left|\mu^O_i - \overline{\mu}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}}\right| = 0;$$

if (2) occurs, then

$$\hat{n}^O_l \left|\overline{\mu}^E_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}-1} - \overline{\mu}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}}\right| \le \Delta_{k(l)},$$
$$\sum_{i=\hat{\tau}^O_{L,l}+1}^{\hat{\tau}^O_{L,l+1}-1} \left|\mu^E_i - \mu^O_i\right| \le \Delta_{k(l)},$$
$$\max_{i=\hat{\tau}^O_{L,l}+1,\ldots,\hat{\tau}^O_{L,l+1}-1} \left|\mu^E_i - \overline{\mu}^E_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}-1}\right| \le \Delta_{k(l)},$$
$$\max_{i=\hat{\tau}^O_{L,l}+1,\ldots,\hat{\tau}^O_{L,l+1}} \left|\mu^O_i - \overline{\mu}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}}\right| \le \Delta_{k(l)};$$

if (3) occurs, then

$$\hat{n}^O_l \left|\overline{\mu}^E_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}-1} - \overline{\mu}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}}\right| \le \Delta_{k(l)} + \Delta_{k(l)+1},$$
$$\sum_{i=\hat{\tau}^O_{L,l}+1}^{\hat{\tau}^O_{L,l+1}-1} \left|\mu^E_i - \mu^O_i\right| \le \Delta_{k(l)} + \Delta_{k(l)+1},$$
$$\max_{i=\hat{\tau}^O_{L,l}+1,\ldots,\hat{\tau}^O_{L,l+1}-1} \left|\mu^E_i - \overline{\mu}^E_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}-1}\right| \le \Delta_{k(l)} + \Delta_{k(l)+1},$$
$$\max_{i=\hat{\tau}^O_{L,l}+1,\ldots,\hat{\tau}^O_{L,l+1}} \left|\mu^O_i - \overline{\mu}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}}\right| \le \Delta_{k(l)} + \Delta_{k(l)+1}.$$

Also, because of Assumption 3(iv), if $\overline{\sigma}^2 \log\log\overline{\lambda}/(K\Delta^2_{k(l)}) \le C_n$, we have that $\tau^O_{k(l)} = \hat{\tau}^O_{L,l}$. Hence, we can replace above all $\Delta_{k(l)}$'s by $\min\left\{\Delta_{k(l)},\ \left(C_n^{-1}\overline{\sigma}^2\log\log\overline{\lambda}/K\right)^{1/2}\right\}$.

Thus we obtain for instance that on the event Ω_n,

$$\max_{L=K,\ldots,K_{\max}} \sum_{l=0}^{L} \left(\hat{n}^O_l \left|\overline{\mu}^E_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}-1} - \overline{\mu}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}}\right|\right)^2 = o_{\mathbb{P}}\left(\overline{\sigma}^2\log\log\overline{\lambda}\right). \qquad (89)$$

Moreover, we obtain from Assumption 4 that

$$\max_{L=K+1,\ldots,K_{\max}} \frac{\sum_{l=0}^{L}\left(\hat{n}_l^O\left|\overline{\mu}^E_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}-1}-\overline{\mu}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}}\right|\right)^2}{S_{\varepsilon^O}\left(\mathcal{T}_K^O\right)-S_{\varepsilon^O}\left(\hat{\mathcal{T}}_L^O\cup\mathcal{T}_K^O\right)}=o_{\mathbb{P}}(1).$$

Note that the same inequality applies when we replace the maximum by $L=K$. We will use such inequalities in the arguments to follow.

Let us fix $L<K$ now. Due to Assumption 3(v), there exists a constant $A>0$ such that for each $k=1,\ldots,K$ either $\Delta_k^2\le\overline{\sigma}^2\log\log\overline{\lambda}/(KC_n)$ or there exists an $l=1,\ldots,L$ such that $\tau_k^O=\hat{\tau}_{L,l}^O$ or

$$\sum_{i=\tau_k^O-\frac{\lambda}{4}+1}^{\tau_k^O+\frac{\lambda}{4}}(\mu_i^O-\overline{\mu}_{L,i}^O)^2\ge A\underline{\lambda}\Delta_k^2.$$

Furthermore, let $0=\hat{k}_0\le\cdots\le\hat{k}_{L+1}=K+1$ be such that $\tau^O_{\hat{k}_l}\le\hat{\tau}^O_{L,l}<\tau^O_{\hat{k}_l+1}$. Hence for each $l=0,\ldots,L$ we have that $\tau^O_{\hat{k}_l}\le\hat{\tau}^O_{L,l}<\hat{\tau}^O_{L,l+1}<\tau^O_{\hat{k}_{l+1}+1}$. Thus, there exists sequences $\mathcal{I}_L\subseteq\{1,\ldots,K\}$ such that

$$\sum_{i=\tau_k^O-\frac{\lambda}{4}+1}^{\tau_k^O+\frac{\lambda}{4}}(\mu_i^O-\overline{\mu}_{L,i}^O)^2\ge A\underline{\lambda}\Delta_k^2\ \forall\ k\in\mathcal{I}_L$$

and

$$\hat{n}_l^O\left|\overline{\mu}^E_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}-1}-\overline{\mu}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}}\right|\le\sum_{\substack{k=\hat{k}_l+1,\ldots,\hat{k}_{l+1}\\k\in\mathcal{I}_L}}\Delta_k+\left(\hat{k}_{l+1}-\hat{k}_l\right)\sqrt{\overline{\sigma}^2\log\log\overline{\lambda}/(KC_n)},$$

$$\sum_{i=\hat{\tau}^O_{L,l}+1}^{\hat{\tau}^O_{L,l+1}-1}\left|\mu_i^E-\mu_i^O\right|\le\sum_{\substack{k=\hat{k}_l+1,\ldots,\hat{k}_{l+1}\\k\in\mathcal{I}_L}}\Delta_k+\left(\hat{k}_{l+1}-\hat{k}_l\right)\sqrt{\overline{\sigma}^2\log\log\overline{\lambda}/(KC_n)},$$

$$\max_{i=\hat{\tau}^O_{L,l}+1,\ldots,\hat{\tau}^O_{L,l+1}-1}\left|\mu_i^E-\overline{\mu}^E_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}-1}\right|\le\sum_{\substack{k=\hat{k}_l+1,\ldots,\hat{k}_{l+1}\\k\in\mathcal{I}_L}}\Delta_k+\left(\hat{k}_{l+1}-\hat{k}_l\right)\sqrt{\overline{\sigma}^2\log\log\overline{\lambda}/(KC_n)},$$

$$\max_{i=\hat{\tau}^O_{L,l}+1,\ldots,\hat{\tau}^O_{L,l+1}}\left|\mu_i^O-\overline{\mu}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}}\right|\le\sum_{\substack{k=\hat{k}_l+1,\ldots,\hat{k}_{l+1}\\k\in\mathcal{I}_L}}\Delta_k+\left(\hat{k}_{l+1}-\hat{k}_l\right)\sqrt{\overline{\sigma}^2\log\log\overline{\lambda}/(KC_n)}.$$

Consequently, by using $(x+y)^2 \leq 2x^2 + 2y^2$ we obtain for instance

$$\begin{aligned}
&\sum_{l=0}^{L}\left(\hat{n}_l^O\left|\overline{\mu}^E_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}-1} - \overline{\mu}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}}\right|\right)^2 \\
\leq &\sum_{l=0}^{L}\left(\sum_{\substack{k=\hat{k}_l+1,\ldots,\hat{k}_{l+1}\\ k\in\mathcal{I}_L}}\Delta_k + (\hat{k}_{l+1}-\hat{k}_l)\sqrt{\overline{\sigma}^2\log\log\overline{\lambda}/(KC_n)}\right)^2 \\
= &2K\sum_{k\in\mathcal{I}_L}\Delta_k^2 + 2\overline{\sigma}^2\log\log\overline{\lambda}.
\end{aligned}$$

Note that each $\mathcal{I}_L$ is non-empty, since $L < K$. Hence, $K\sum_{k\in\mathcal{I}_L}\Delta_k^2 \geq K\Delta_{(1)}^2$. Moreover, Assumptions 5 yields

$$\overline{\sigma}^2\log\log\overline{\lambda} = o_{\mathbb{P}}\left(\min_{L=0,\ldots,K-1}\underline{\lambda}\sum_{k\in\mathcal{I}_L}\Delta_k^2\right)$$

and hence for example, because of Assumption 1(ii),

$$\max_{L=0,\ldots,K-1}\left(\sum_{k\in\mathcal{I}_L}\Delta_k^2\right)^{-1}\sum_{l=0}^{L}\left(\hat{n}_l^O\left|\overline{\mu}^E_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}-1} - \overline{\mu}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}}\right|\right)^2 = o_{\mathbb{P}}\left(\underline{\lambda}\right).$$

We will use such inequalities in the arguments to follow. From (89) and similar arguments we also obtain that for example

$$\sum_{k=0}^{K}\left(\hat{n}_k^O\left|\overline{\mu}^E_{\hat{\tau}^O_{K,k}:\hat{\tau}^O_{K,k+1}-1} - \overline{\mu}^O_{\hat{\tau}^O_{K,k}:\hat{\tau}^O_{K,k+1}}\right|\right)^2 = o_{\mathbb{P}}\left(\underline{\lambda}\min_{L=0,\ldots,K-1}\sum_{k\in\mathcal{I}_L}\Delta_k^2\right).$$

Furthermore, from the same arguments it follows that

$$K\overline{\sigma}^2(\log\overline{\lambda})^2 = o_{\mathbb{P}}\left(\min_L\underline{\lambda}\sum_{k\in\mathcal{I}_L}\Delta_k^2\right). \tag{90}$$

This completes bounding $A^{(n)}_{1,L}$ as noted before.

Bounding $A^{(n)}_{2,L}$

$$\begin{aligned}
A^{(n)}_{2,L} \leq &\left|\sum_{l=0}^{L}\sum_{i=\hat{\tau}^O_{L,l}+1}^{\hat{\tau}^O_{L,l+1}-1}\frac{\hat{n}_l^O}{\hat{n}_l^O-1}\left(\mu_i^E - \overline{\mu}^E_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}-1}\right)^2 - \sum_{l=0}^{L}\sum_{i=\hat{\tau}^O_{L,l}+1}^{\hat{\tau}^O_{L,l+1}-1}\frac{\hat{n}_l^O}{\hat{n}_l^O-1}\left(\mu_i^E - \overline{\mu}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}}\right)^2\right| \\
&+\left|\sum_{l=0}^{L}\sum_{i=\hat{\tau}^O_{L,l}+1}^{\hat{\tau}^O_{L,l+1}-1}\frac{\hat{n}_l^O}{\hat{n}_l^O-1}\left(\mu_i^E - \overline{\mu}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}}\right)^2 - \sum_{l=0}^{L}\sum_{i=\hat{\tau}^O_{L,l}+1}^{\hat{\tau}^O_{L,l+1}-1}\frac{\hat{n}_l^O}{\hat{n}_l^O-1}\left(\mu_i^O - \overline{\mu}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}}\right)^2\right| \\
&+\left|\sum_{l=0}^{L}\sum_{i=\hat{\tau}^O_{L,l}+1}^{\hat{\tau}^O_{L,l+1}-1}\frac{\hat{n}_l^O}{\hat{n}_l^O-1}\left(\mu_i^O - \overline{\mu}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}}\right)^2 - \sum_{l=0}^{L}\sum_{i=\hat{\tau}^O_{L,l}+1}^{\hat{\tau}^O_{L,l+1}}\left(\mu_i^O - \overline{\mu}^O_{\hat{\tau}^O_{L,l}:\hat{\tau}^O_{L,l+1}}\right)^2\right|.
\end{aligned}$$

The r.h.s. is bounded from above by

$$
\begin{aligned}
&2\left|\sum_{l=0}^{L}\sum_{i=\hat\tau^O_{L,l}+1}^{\hat\tau^O_{L,l+1}-1}\frac{\hat n^O_l}{\hat n^O_l-1}\left(\mu^E_i-\overline\mu^E_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}-1}\right)\left(\overline\mu^E_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}-1}-\overline\mu^O_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}}\right)\right|\\
&+\left|\sum_{l=0}^{L}\sum_{i=\hat\tau^O_{L,l}+1}^{\hat\tau^O_{L,l+1}-1}\frac{\hat n^O_l}{\hat n^O_l-1}\left(\overline\mu^E_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}-1}-\overline\mu^O_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}}\right)^2\right|\\
&+\left|\sum_{l=0}^{L}\sum_{i=\hat\tau^O_{L,l}+1}^{\hat\tau^O_{L,l+1}-1}\frac{\hat n^O_l}{\hat n^O_l-1}\left(\mu^E_i-\mu^O_i\right)^2\right|\\
&+2\left|\sum_{l=0}^{L}\sum_{i=\hat\tau^O_{L,l}+1}^{\hat\tau^O_{L,l+1}-1}\frac{\hat n^O_l}{\hat n^O_l-1}\left(\mu^E_i-\mu^O_i\right)\left(\mu^O_i-\overline\mu^O_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}}\right)\right|\\
&+\left|\sum_{l=0}^{L}\sum_{i=\hat\tau^O_{L,l}+1}^{\hat\tau^O_{L,l+1}-1}\frac{1}{\hat n^O_l-1}\left(\mu^O_i-\overline\mu^O_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}}\right)^2-\sum_{l=0}^{L}\left(\mu^O_{\hat\tau^O_{L,l+1}}-\overline\mu^O_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}}\right)^2\right|,
\end{aligned}
$$

which is in turn bounded from above by

$$
\begin{aligned}
&2\sum_{l=0}^{L}\left(\max_{i=\hat\tau^O_{L,l}+1,\dots,\hat\tau^O_{L,l+1}-1}\left|\mu^E_i-\overline\mu^E_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}-1}\right|\right)\left(\hat n^O_l\left|\overline\mu^E_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}-1}-\overline\mu^O_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}}\right|\right)\\
&+\sum_{l=0}^{L}\left(\hat n^O_l\left|\overline\mu^E_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}-1}-\overline\mu^O_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}}\right|\right)^2\\
&+2\sum_{l=0}^{L}\left(\sum_{i=\hat\tau^O_{L,l}+1}^{\hat\tau^O_{L,l+1}-1}\left|\mu^E_i-\mu^O_i\right|\right)^2\\
&+4\sum_{l=0}^{L}\left(\max_{i=\hat\tau^O_{L,l}+1,\dots,\hat\tau^O_{L,l+1}-1}\left|\mu^O_i-\overline\mu^O_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}}\right|\right)\left(\sum_{i=\hat\tau^O_{L,l}+1}^{\hat\tau^O_{L,l+1}-1}\left|\mu^E_i-\mu^O_i\right|\right)\\
&+\sum_{l=0}^{L}\left(\max_{i=\hat\tau^O_{L,l}+1,\dots,\hat\tau^O_{L,l+1}-1}\left|\mu^O_i-\overline\mu^O_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}}\right|\right)^2.
\end{aligned}
$$

Thus, from the preliminary calculations it follows that $A^{(n)}_{2,L}$ satisfies the same bounds as $A^{(n)}_{L}$ in the statement of the lemma.

Bounding $A^{(n)}_{3,L}$ Since

$$
\overline\varepsilon^E_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}}\left(\mu^O_i-\overline\mu^O_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}}\right)=0=\overline\varepsilon^E_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}-1}\left(\mu^O_i-\overline\mu^O_{\hat\tau^O_{L,l}:\hat\tau^O_{L,l+1}}\right),
$$

we have

$$
\begin{aligned}
A_{3,L}^{(n)} \leq & \left| \sum_{l=0}^{L} \sum_{i=\hat{\tau}_{L,l}^{O}+1}^{\hat{\tau}_{L,l+1}^{O}-1} \frac{\hat{n}_l^O}{\hat{n}_l^O - 1} \left(\varepsilon_i^E - \overline{\varepsilon}^E_{\hat{\tau}_{L,l}^{O}:\hat{\tau}_{L,l+1}^{O}-1} \right) \left(\mu_i^E - \overline{\mu}^E_{\hat{\tau}_{L,l}^{O}:\hat{\tau}_{L,l+1}^{O}-1} \right) \right. \\
& - \left. \sum_{l=0}^{L} \sum_{i=\hat{\tau}_{L,l}^{O}+1}^{\hat{\tau}_{L,l+1}^{O}-1} \frac{\hat{n}_l^O}{\hat{n}_l^O - 1} \left(\varepsilon_i^E - \overline{\varepsilon}^E_{\hat{\tau}_{L,l}^{O}:\hat{\tau}_{L,l+1}^{O}-1} \right) \left(\mu_i^O - \overline{\mu}^O_{\hat{\tau}_{L,l}^{O}:\hat{\tau}_{L,l+1}^{O}} \right) \right| \\
& + \left| \sum_{l=0}^{L} \sum_{i=\hat{\tau}_{L,l}^{O}+1}^{\hat{\tau}_{L,l+1}^{O}-1} \frac{\hat{n}_l^O}{\hat{n}_l^O - 1} \left(\varepsilon_i^E - \overline{\varepsilon}^E_{\hat{\tau}_{L,l}^{O}:\hat{\tau}_{L,l+1}^{O}-1} \right) \left(\mu_i^O - \overline{\mu}^O_{\hat{\tau}_{L,l}^{O}:\hat{\tau}_{L,l+1}^{O}} \right) \right. \\
& - \left. \sum_{l=0}^{L} \sum_{i=\hat{\tau}_{L,l}^{O}+1}^{\hat{\tau}_{L,l+1}^{O}} \left(\varepsilon_i^E - \overline{\varepsilon}^E_{\hat{\tau}_{L,l}^{O}:\hat{\tau}_{L,l+1}^{O}-1} \right) \left(\mu_i^O - \overline{\mu}^O_{\hat{\tau}_{L,l}^{O}:\hat{\tau}_{L,l+1}^{O}} \right) \right|.
\end{aligned}
$$

Using the same arguments as for the first term gives us

$$
\begin{aligned}
& A_{3,L}^{(n)} \left(\sum_{l=0}^{L} \left(\sum_{i=\hat{\tau}_{L,l}^{O}+1}^{\hat{\tau}_{L,l+1}^{O}-1} \left| \mu_i^E - \mu_i^O \right| + \hat{n}_l^O \left| \overline{\mu}^O_{\hat{\tau}_{L,l}^{O}:\hat{\tau}_{L,l+1}^{O}} - \overline{\mu}^E_{\hat{\tau}_{L,l}^{O}:\hat{\tau}_{L,l+1}^{O}-1} \right| \right. \right. \\
& \qquad \left. \left. + \max_{i=\hat{\tau}_{L,l}^{O}+1,\ldots,\hat{\tau}_{L,l+1}^{O}} \left| \mu_i^O - \overline{\mu}^O_{\hat{\tau}_{L,l}^{O}:\hat{\tau}_{L,l+1}^{O}} \right| \right) \right)^{-1} \\
= & O_{\mathbb{P}} \left(\log(K_{\max}) \overline{\sigma} \right).
\end{aligned}
$$

Hence, from the preliminary calculations and the same simplifications as seen before it follows that $A_{3,L}^{(n)}$ satisfies the same bounds as $A_L^{(n)}$ in the statement of the lemma.

Bounding $A_{4,L}^{(n)}$ Applying the triangle and Cauchy–Schwarz inequalities gives

$$
\begin{aligned}
A_{4,L}^{(n)} \leq & \left| \sum_{l=0}^{L} \sum_{i=\hat{\tau}_{L,l}^{O}+1}^{\hat{\tau}_{L,l+1}^{O}-1} \frac{\hat{n}_l^O}{\hat{n}_l^O - 1} \varepsilon_i^E \left(\overline{\mu}^E_{\hat{\tau}_{L,l}^{O}:\hat{\tau}_{L,l+1}^{O}-1} - \overline{\mu}^O_{\hat{\tau}_{L,l}^{O}:\hat{\tau}_{L,l+1}^{O}} \right) \right| \\
& + \left| \sum_{l=0}^{L} \hat{n}_l^O \overline{\varepsilon}^O_{\hat{\tau}_{L,l}^{O}:\hat{\tau}_{L,l+1}^{O}} \left(\overline{\mu}^E_{\hat{\tau}_{L,l}^{O}:\hat{\tau}_{L,l+1}^{O}-1} - \overline{\mu}^O_{\hat{\tau}_{L,l}^{O}:\hat{\tau}_{L,l+1}^{O}} \right) \right| \\
\leq & \left| \sum_{l=0}^{L} \sum_{i=\hat{\tau}_{L,l}^{O}+1}^{\hat{\tau}_{L,l+1}^{O}-1} \frac{\hat{n}_l^O}{\hat{n}_l^O - 1} \varepsilon_i^E \left(\overline{\mu}^E_{\hat{\tau}_{L,l}^{O}:\hat{\tau}_{L,l+1}^{O}-1} - \overline{\mu}^O_{\hat{\tau}_{L,l}^{O}:\hat{\tau}_{L,l+1}^{O}} \right) \right| \\
& + \left(\sum_{l=0}^{L} \left(\overline{\varepsilon}^O_{\hat{\tau}_{L,l}^{O}:\hat{\tau}_{L,l+1}^{O}} \right)^2 \right)^{1/2} \left(\sum_{l=0}^{L} \left(\hat{n}_l^O \right)^2 \left(\overline{\mu}^E_{\hat{\tau}_{L,l}^{O}:\hat{\tau}_{L,l+1}^{O}-1} - \overline{\mu}^O_{\hat{\tau}_{L,l}^{O}:\hat{\tau}_{L,l+1}^{O}} \right)^2 \right)^{1/2}.
\end{aligned}
$$

Then it follows from the same arguments as before that $A_{4,L}^{(n)}$ satisfies the same bounds as $A_L^{(n)}$ in the statement of the lemma.

Bounding $A_{5,L}^{(n)}$ From the preliminary calculations it follows that $A_{5,L}^{(n)}$ satisfies the same bounds as $A_L^{(n)}$ in the statement of the lemma.

Proving the statement Combining the bounds for $A_L^{(n)}$, $A_{1,L}^{(n)}$, $A_{2,L}^{(n)}$, $A_{3,L}^{(n)}$, $A_{4,L}^{(n)}$ and $A_{5,L}^{(n)}$ gives that $A_L^{(n)}$ satisfies the bounds given in the statement of the lemma. □

Proof of Theorem 4. Recall $\mathrm{CV}_{\mathrm{mod}}^{O}(L)$, $\mathrm{CV}_{\mathrm{mod}}^{E}(L)$, $\widetilde{\mathrm{CV}}_{\mathrm{mod}}^{O}(L)$ and $\widetilde{\mathrm{CV}}_{\mathrm{mod}}^{E}(L)$ from (61) and (62), respectively.

Then, $\mathrm{CV}_{\mathrm{mod}}(L) = \mathrm{CV}_{\mathrm{mod}}^{O}(L) + \mathrm{CV}_{\mathrm{mod}}^{E}(L)$. Thus,

$$
\begin{aligned}
&\mathbb{P}\left(\hat{K} = K\right)\\
=&\mathbb{P}\left(\min_{\substack{L=0,\dots,K_{\max}\\ L\neq K}} \mathrm{CV}_{\mathrm{mod}}(L) - \mathrm{CV}_{\mathrm{mod}}(K) > 0\right)\\
\geq&\mathbb{P}\left(\min_{\substack{L=0,\dots,K_{\max}\\ L\neq K}} \mathrm{CV}_{\mathrm{mod}}^{O}(L) - \mathrm{CV}_{\mathrm{mod}}^{O}(K) > 0, \min_{\substack{L=0,\dots,K_{\max}\\ L\neq K}} \mathrm{CV}_{\mathrm{mod}}^{E}(L) - \mathrm{CV}_{\mathrm{mod}}^{E}(K) > 0\right)\\
=&\mathbb{P}\left(\min_{\substack{L=0,\dots,K_{\max}\\ L\neq K}} \mathrm{CV}_{\mathrm{mod}}^{O}(L) - \mathrm{CV}_{\mathrm{mod}}^{O}(K) > 0\right) + \mathbb{P}\left(\min_{\substack{L=0,\dots,K_{\max}\\ L\neq K}} \mathrm{CV}_{\mathrm{mod}}^{E}(L) - \mathrm{CV}_{\mathrm{mod}}^{E}(K) > 0\right)\\
&- \mathbb{P}\left(\min_{\substack{L=0,\dots,K_{\max}\\ L\neq K}} \mathrm{CV}_{\mathrm{mod}}^{O}(L) - \mathrm{CV}_{\mathrm{mod}}^{O}(K) > 0 \text{ or } \min_{\substack{L=0,\dots,K_{\max}\\ L\neq K}} \mathrm{CV}_{\mathrm{mod}}^{E}(L) - \mathrm{CV}_{\mathrm{mod}}^{E}(K) > 0\right).
\end{aligned}
$$

In the following we will show that

$$
\mathbb{P}\left(\min_{\substack{L=0,\dots,K_{\max}\\ L\neq K}} \mathrm{CV}_{\mathrm{mod}}^{O}(L) - \mathrm{CV}_{\mathrm{mod}}^{O}(K) > 0\right) \to 1, \text{ as } n \to \infty. \tag{91}
$$

Note that this completes the proof as it implies

$$
\mathbb{P}\left(\min_{\substack{L=0,\dots,K_{\max}\\ L\neq K}} \mathrm{CV}_{\mathrm{mod}}^{E}(L) - \mathrm{CV}_{\mathrm{mod}}^{E}(K) > 0\right) \to 1, \text{ as } n \to \infty,
$$

since $\mathrm{CV}_{\mathrm{mod}}^{O}(L)$ and $\mathrm{CV}_{\mathrm{mod}}^{E}(L)$ are symmetric and for both terms the same properties follow from Assumptions 1–5. In fact, one way to see this is to consider the observations in reverse order, which interchanges Y_i^O and Y_i^E, but Assumptions 1–5 are not altered. Hence, (91) implies $\mathbb{P}\left(\hat{K} = K\right) \to 1$, as $n \to \infty$.

It remains to show (91). We will consider $L > K$ and $L < K$ separately. Let $L > K$. Recall the notation $A_L^{(n)} = \left|\mathrm{CV}_{\mathrm{mod}}^{O}(L) - \widetilde{\mathrm{CV}}_{\mathrm{mod}}^{O}(L)\right|$ from Lemma 15. It

follows from Lemmas 14 and 15 that

$$\begin{aligned}
&\min_{L=K+1,\ldots,K_{\max}} \mathrm{CV}^O_{\mathrm{mod}}(L) - \mathrm{CV}^O_{\mathrm{mod}}(K) \\
&= \min_{L=K+1,\ldots,K_{\max}} \left\{\widetilde{\mathrm{CV}}^O_{\mathrm{mod}}(L) - A_L^{(n)}\right\} - \widetilde{\mathrm{CV}}^O_{\mathrm{mod}}(K) + A_K^{(n)} \\
&\geq \min_{L=K+1,\ldots,K_{\max}} \left\{S_{\varepsilon^O}\left(\mathcal{T}_K^O\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}_L^O \cup \mathcal{T}_K^O\right)\right\}(1 + o_{\mathbb{P}}(1))\,.
\end{aligned}$$

It follows from Assumption 4 that the r.h.s. is positive with probability converging to 1 as $n \to \infty$. Hence, (91) holds when $L > K$. Note that thus far, Assumptions 3 and 5 have not been used directly, and moreover they are not required for the referenced Lemmas when $K = 0$ for all n.

Let $L < K$ now, which in particular requires $K \geq 1$. Lemma 15 yields that there exist sequences of stochastic non-empty sets $\mathcal{I}_L \subseteq \{1, \ldots, K\}$ and a constant $A > 0$ such that

$$\mathbb{P}\left(\forall\, L < K,\ \sum_{i=\tau_k^O - \frac{\lambda}{4}+1}^{\tau_k^O + \frac{\lambda}{4}} (\mu_i^O - \overline{\mu}_{L,i}^O)^2 \geq A\underline{\lambda}\Delta_k^2\ \forall\ k \in \mathcal{I}_L\right) \to 1,$$

where $\overline{\mu}_{L,i}^O := \sum_{l=0}^{L} \mathbb{1}_{\{\hat{\tau}_{L,l}^O + 1 \leq i \leq \hat{\tau}_{L,l+1}^O\}} \overline{\mu}^O_{\hat{\tau}_{L,l}^O : \hat{\tau}_{L,l+1}^O}$. Thus, it follows from Lemmas 14 and 15 that

$$\begin{aligned}
&\min_{L=0\ldots,K-1} \mathrm{CV}^O_{\mathrm{mod}}(L) - \mathrm{CV}^O_{\mathrm{mod}}(K) \\
&= \min_{L=0\ldots,K-1} \left\{\widetilde{\mathrm{CV}}^O_{\mathrm{mod}}(L) - A_L^{(n)}\right\} - \widetilde{\mathrm{CV}}^O_{\mathrm{mod}}(K) + A_K^{(n)} \\
&\geq \min_{L=0\ldots,K-1} \left\{\underline{\lambda} \sum_{k \in \mathcal{I}_L} \Delta_k^2\right\}(A + o_{\mathbb{P}}(1))\,.
\end{aligned}$$

The r.h.s. is positive, since $|\mathcal{I}_L| \geq 1,\ \forall L = L = 0\ldots, K-1$ and $A > 0$. Hence, (91) holds also when $L < K$. This completes the proof as noted before. □

S6. Proof of Theorem 3

The first statement in Theorem 3 follows directly from Theorem 4 if its assumptions are satisfied by least squares estimation under the given setting. To show this, we will use Theorem 5 and Lemma 16. The latter is a shorter version of Theorem 2 in Zou, Wang and Li (2020). It shows that Assumption 4 is indeed satisfied in the given set-up.

Lemma 16. *Suppose the noise assumption in Section 3.2, i.e. sub-Gaussian noise, constant variance on each segment and ration between the smallest variance and largest variance proxy is bounded from below. Then, for all $\epsilon > 0$,*

$$\mathbb{P}\left(\frac{\min_{L=K+1,\ldots,K_{\max}} \left\{S_{\varepsilon^O}\left(\mathcal{T}_K^O\right) - S_{\varepsilon^O}\left(\hat{\mathcal{T}}_L^O \cup \mathcal{T}_K^O\right)\right\}}{\overline{\sigma}^2 \log\log \overline{\lambda}} < \epsilon\right) \to 0$$

and as above but with all instances of O replaced by E.

Proof. Let k^* be the index of the longest segment, so $\overline{\lambda}/2-1 \leq \tau^O_{k^*+1}-\tau^O_{k^*} \leq \overline{\lambda}/2+1$. Then, it follows from the fact that adding change-points only decreases the costs and the definition of least squares estimation that for every $L = K+1, \ldots, K_{\max}$,

$$\begin{aligned} S_{\varepsilon O}\left(\hat{\mathcal{T}}^O_L \cup \mathcal{T}^O_K\right) &= S_{YO}\left(\hat{\mathcal{T}}^O_L \cup \mathcal{T}^O_K\right) \leq S_{YO}\left(\hat{\mathcal{T}}^O_{K+1}\right) \\ &\leq \min_{\tau^O_{k^*}<t<\tau^O_{k^*+1}} S_{YO}\left(\mathcal{T}^O_K \cup \{t\}\right) \\ &= \min_{\tau^O_{k^*}<t<\tau^O_{k^*+1}} S_{\varepsilon O}\left(\mathcal{T}^O_K \cup \{t\}\right). \end{aligned}$$

Thus,

$$\begin{aligned} & S_{\varepsilon O}\left(\mathcal{T}^O_K\right) - S_{\varepsilon O}\left(\hat{\mathcal{T}}^O_L \cup \mathcal{T}^O_K\right) \\ \geq & S_{\varepsilon O}\left(\mathcal{T}^O_K\right) - \min_{\tau^O_{k^*}<t<\tau^O_{k^*+1}} S_{\varepsilon O}\left(\mathcal{T}^O_K \cup \{t\}\right) \\ \geq & \max_{\tau^O_{k^*}<t<\tau^O_{k^*+1}} \left\{ (t-\tau^O_{k^*})\left(\overline{\varepsilon}^O_{\tau^O_{k^*}:t}\right)^2 + (\tau^O_{k^*+1}-t)\left(\overline{\varepsilon}^O_{t:\tau^O_{k^*+1}}\right)^2 \right\} - (\tau^O_{k^*+1}-\tau^O_{k^*})\left(\overline{\varepsilon}^O_{\tau^O_{k^*}:\tau^O_{k^*+1}}\right)^2. \end{aligned}$$

It follows from Zou, Wang and Li (2020, Lemma 2) (see also Horváth (1993, Lemma 2.1)) that there exists a constant $c > 0$ such that

$$\mathbb{P}\left(\max_{\tau^O_{k^*}<t<\tau^O_{k^*+1}} \left\{ (t-\tau^O_{k^*})\left(\overline{\varepsilon}^O_{\tau^O_{k^*}:t}\right)^2 + (\tau^O_{k^*+1}-t)\left(\overline{\varepsilon}^O_{t:\tau^O_{k^*+1}}\right)^2 \right\} > c\,\underline{\sigma}^2 \log\log\overline{\lambda} \right) \to 1.$$

Moreover, $(\tau^O_{k^*+1}-\tau^O_{k^*})\left(\overline{\varepsilon}^O_{\tau^O_{k^*}:\tau^O_{k^*+1}}\right)^2 = O_{\mathbb{P}}(\sigma)$. Hence, the stated formula follows by using that $\limsup_{n\to\infty} \sigma/\underline{\sigma} < \infty$.

The same argument holds with all instances of O replaced by E. □

Proof of Theorem 3. For the first part we will use Theorem 4 and hence, we first verify its assumptions in the following. Note that Assumptions 1 and 5 are assumed in Theorem 3 as well. Since $\varepsilon_1, \ldots, \varepsilon_n$ are sub-Gaussian with uniformly bounded variance proxy σ^2, Assumption 2 follows. Moreover, Lemma 16 shows that Assumption 4 holds as well. It remains to show Assumption 3.

As noted in the discussion following Theorem 4, Assumptions 3 is not required for the conclusion when $K = 0$. Hence, we can assume that $K > 0$. Since $K = o(\underline{\lambda})$ and due to the first part of (17), it follows from Theorem 5 that Assumption 3(i) is satisfied for any sequence $\delta_{q,k}$ that satisfies

$$\max_{k=1,\ldots,K} \left(\delta_{0,k}\right)^{-1} (\log(K) \vee 1) \frac{\sigma^2}{\Delta_k^2} = o(1), \tag{92}$$

$$\max_{q=1,\ldots,K_{\max}-K} \max_{k=1,\ldots,K} \left(\delta_{q,k}\right)^{-1} \left(\log\left(K \frac{\sigma^2}{\Delta_k^2} \right) \vee 1 \right) \frac{\sigma^2}{\Delta_k^2} = o(1). \tag{93}$$

Now let a_n be a sequence such that

$$a_n K \log\log\left((\log(K)\vee 1)\frac{\sigma^2}{\Delta_{(1)}^2}\vee e\right) = o(\log\log\overline{\lambda}), \tag{94}$$

$$a_n K(\log K\vee 1) = o(\log\log\overline{\lambda}), \tag{95}$$

but $a_n\to\infty$; the existence of such a sequence is guaranteed by (17) and Assumption 1(ii). Then, there exists $\delta_{q,k}$ satisfying Assumption 3(i) for which

$$\max_{0\le q\le Q, 1\le k\le K} K\log\log(\delta_{q,k}\vee e) \le \max_{1\le k\le K} K\log\log\left(a_n\left(\log\left(K\frac{\sigma^2}{\Delta_k^2}\right)\vee(\log K)\vee 1\right)\frac{\sigma^2}{\Delta_k^2}\vee e\right) \tag{96}$$

$$\max_{k=1,\dots,K}\delta_{0,k}\Delta_k^2 \le a_n\sigma^2(\log(K)\vee 1). \tag{97}$$

Now from (94), the r.h.s. of (96) is

$$\begin{aligned}
&K\log\log\left(a_n\left(\log\left(K\frac{\sigma^2}{\Delta_{(1)}^2}\right)\vee(\log K)\vee 1\right)\frac{\sigma^2}{\Delta_{(1)}^2}\vee e\right)\\
\le& O\left(Ka_n\log\log\left((\log(K)\vee 1)\frac{\sigma^2}{\Delta_{(1)}^2}\vee e\right)\right)\\
=& o(\log\log\overline{\lambda}),
\end{aligned}$$

showing Assumption 3(ii). Also, from (97),

$$\sum_{k=1}^K\delta_{0,k}\Delta_k^2 \le a_n K\sigma^2(\log(K)\vee 1) = o(\log\log\overline{\lambda}),$$

using (95), which gives Assumption 3(iii). Finally, if $\overline{\sigma}^2\log\log\overline{\lambda}/(K\Delta_k^2)\le C_n$ and if n is large enough such that $C_n<1$, then $\overline{\sigma}^2/\Delta_k^2<1$ and $(\log(K)\vee 1)\frac{\sigma^2}{\Delta_k^2}\le C_n<1$ because of $K\log(K)=o(\log\log\overline{\lambda})$ due to Assumption 1(ii). Thus, (92) and (93) imply $\delta_{q,k}=0$ for all $q=0,\dots,Q$, which is needed for Assumption 3(iv). Using similar arguments and the second part from Theorem 5, it follows that that Assumption 3(v) is satisfied as well. Hence, Assumption 3 holds.

Consequently, it follows from Theorem 4 that $\mathbb{P}\left(\hat{K}=K\right)\to 1$, as $n\to\infty$.

It remains to show (18). Let us write $\hat{\tau}_k := \hat{\tau}_{K,k}$ for $k=0,\dots,K+1$ and $\hat{\delta}_k := |\hat{\tau}_k-\tau_k|$. Recall that $\hat{f}_K:\ [0,1]\to\mathbb{R},\ t\mapsto\sum_{k=0}^K\overline{Y}_{\hat{\tau}_k:\hat{\tau}_{k+1}}\mathbb{1}_{(\hat{\tau}_k/n,\hat{\tau}_{k+1}/n]}(t)$. We have that

$$\begin{aligned}
&n\int_0^1\left(\hat{f}_K(t)-f(t)\right)^2dt\\
\le&\sum_{k=0}^K(\hat{\tau}_{k+1}-\hat{\tau}_k)|\overline{Y}_{\hat{\tau}_k:\hat{\tau}_{k+1}}-\beta_k|^2+\sum_{k=1}^K\hat{\delta}_k\left(\Delta_k+\max(|\overline{Y}_{\hat{\tau}_{k-1}:\hat{\tau}_k}-\beta_{k-1}|,\ |\overline{Y}_{\hat{\tau}_k:\hat{\tau}_{k+1}}-\beta_k|)\right)^2.
\end{aligned}$$

In the following we will bound these terms. It follows from Theorem 5 and the fact that $\mathbb{P}\left(\hat{K} = K\right) \to 1$, that for $\gamma_k := c_n(\log(K) \vee 1)\sigma^2/\Delta_k^2$, where c_n can be any sequence such that $c_n \to \infty$, as $n \to \infty$,

$$\mathbb{P}\left(\hat{\delta}_k \leq \gamma_k \ \forall\ k = 1, \ldots, K \text{ and } \hat{K} = K\right) \to 1. \tag{98}$$

Furthermore, (32) implies $\gamma_k < \underline{\lambda}/2$ for a suitable chosen c_n. In the following, we work on the sequence of events in (98). Consequently, $\hat{\tau}_{k+1} - \hat{\tau}_k \geq \underline{\lambda}/2$.

We also have that

$$|\overline{Y}_{\hat{\tau}_k:\hat{\tau}_{k+1}} - \beta_k| \leq \frac{\hat{\delta}_k\Delta_k + \hat{\delta}_{k+1}\Delta_{k+1}}{\hat{\tau}_{k+1} - \hat{\tau}_k} + |\overline{\varepsilon}_{\hat{\tau}_k:\hat{\tau}_{k+1}}|.$$

For the following calculation we assume w.l.o.g. that $\hat{\tau}_k < \tau_k < \tau_{k+1} < \hat{\tau}_{k+1}$, since other cases lead to the same bound. Then,

$$\begin{aligned}
&(\hat{\tau}_{k+1} - \hat{\tau}_k)|\overline{\varepsilon}_{\hat{\tau}_k:\hat{\tau}_{k+1}}|^2 \\
\leq& 3\left((\tau_k - \hat{\tau}_k)|\overline{\varepsilon}_{\hat{\tau}_k:\tau_k}|^2 + (\tau_{k+1} - \tau_k)|\overline{\varepsilon}_{\tau_k:\tau_{k+1}}|^2 + (\hat{\tau}_{k+1} - \tau_{k+1})|\overline{\varepsilon}_{\tau_{k+1}:\hat{\tau}_{k+1}}|^2\right) \\
\leq& \max_{t=\tau_k-\gamma_k,\ldots,\tau_k-1}\{(\tau_k - t)|\overline{\varepsilon}_{t:\tau_k}|^2\} + (\tau_{k+1} - \tau_k)|\overline{\varepsilon}_{\tau_k:\tau_{k+1}}|^2 \\
&\quad + \max_{t=\tau_{k+1}+1,\ldots,\tau_{k+1}+\gamma_{k+1}}\{(t - \tau_{k+1})|\overline{\varepsilon}_{\tau_{k+1}:t}|^2\}.
\end{aligned}$$

Hence, due to sub-Gaussianity and independence,

$$\max_{k=1,\ldots,K}\left(\log\log\gamma_k + 1 + \log\log\gamma_{k+1}\right)^{-1}(\hat{\tau}_{k+1} - \hat{\tau}_k)|\overline{\varepsilon}_{\hat{\tau}_k:\hat{\tau}_{k+1}}|^2 = O_{\mathbb{P}}(\log(K)\sigma^2).$$

Note that here and in the following we have written $\log\log\gamma_k$ instead of $\log\log(\gamma_k \vee e)$ to improve readability.

Thus, using $(x + y)^2 \leq 2x^2 + 2y^2$,

$$\sum_{k=0}^{K}(\hat{\tau}_{k+1} - \hat{\tau}_k)|\overline{Y}_{\hat{\tau}_k:\hat{\tau}_{k+1}} - \beta_k|^2 = O_{\mathbb{P}}\left(\sum_{k=1}^{K}\frac{\gamma_k\Delta_k^2}{\underline{\lambda}}\right) + O_{\mathbb{P}}\left(\sum_{k=1}^{K}(1 + \log\log\gamma_k)\log(K)\sigma^2\right),$$

and

$$\begin{aligned}
&\sum_{k=1}^{K}\hat{\delta}_k\left(\Delta_k + \max(|\overline{Y}_{\hat{\tau}_{k-1}:\hat{\tau}_k} - \beta_{k-1}|,\ |\overline{Y}_{\hat{\tau}_k:\hat{\tau}_{k+1}} - \beta_k|)\right)^2 \\
=& O_{\mathbb{P}}\left(\sum_{k=1}^{K}\gamma_k\Delta_k^2\right) + O_{\mathbb{P}}\left(\sum_{k=1}^{K}\gamma_k\frac{\left((\gamma_{k-1})^2\Delta_{k-1}^2 + (\gamma_k)^2\Delta_k^2 + (\gamma_{k+1})^2\Delta_{k+1}^2\right)}{\underline{\lambda}^2}\right) \\
&+ O_{\mathbb{P}}\left(\sum_{k=1}^{K}\gamma_k(1 + \log\log\gamma_{k-1} + \log\log\gamma_k + \log\log\gamma_{k+1})\log(K)\sigma^2/\underline{\lambda}\right),
\end{aligned}$$

where we have used the notation $\Delta_0 = \Delta_{K+1} = 0$ and $\gamma_0 = \gamma_{K+1} = e$.

Since $\gamma_k \leq \underline{\lambda}$, it follows that

$$\begin{aligned}
&n \int_0^1 \left(\hat{f}_K(t) - f(t)\right)^2 dt \\
=&O_{\mathbb{P}}\left(\sum_{k=1}^{K} \gamma_k \Delta_k^2\right) + O_{\mathbb{P}}\left(\sum_{k=1}^{K} (1 + \log\log\gamma_k)\log(K)\sigma^2\right) \\
=&O_{\mathbb{P}}\left(\sum_{k=1}^{K} (c_n + \log\log\gamma_k)(\log(K) \vee 1)\sigma^2\right) \\
=&O_{\mathbb{P}}\left(\left(c_n + \log\log\left(c_n(\log(K) \vee 1)\sigma^2/\Delta_{(1)}^2\right)\right) K(\log(K) \vee 1)\sigma^2\right).
\end{aligned}$$

By using the second part of (17) and that $\log\log\overline{\lambda} \to \infty$, as $n \to \infty$, as well as by choosing a c_n that increases slowly enough, this simplifies to the claimed bound. □

References

Balsubramani, A. (2014). Sharp finite-time iterated-logarithm martingale concentration. *arXiv preprint arXiv:1405.2639*.

Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Horváth, L. (1993). The maximum likelihood method for testing changes in the parameters of normal observations. *Ann. Stat.* 671–680.

Verzelen, N., Fromont, M., Lerasle, M. and Reynaud-Bouret, P. (2020). Optimal Change-Point Detection and Localization. *arXiv preprint arXiv:2010.11470*.

Wang, G., Zou, C. and Qiu, P. (2021). Data-Driven Determination of the Number of Jumps in Regression Curves. *Technometrics* 1–11.

Zhang, H. and Chen, S. X. (2020). Concentration inequalities for statistical inference. *arXiv preprint arXiv:2011.02258*.

Zhang, H. and Wei, H. (2021). Sharper Sub-Weibull Concentrations: Non-asymptotic Bai-Yin Theorem. *arXiv preprint arXiv:2102.02450*.

Zou, C., Wang, G. and Li, R. (2020). Consistent selection of the number of change-points via sample-splitting. *Ann. Stat.* **48** 413–439.