Supplement to “Narrowest Significance Pursuit: inference for multiple change-points in linear models”

Piotr Fryzlewicz*

May 4, 2023

Abstract

This supplement discusses a number of aspects of the NSP method.

Keywords: confidence intervals, structural breaks, post-selection inference, wild binary segmentation, narrowest-over-threshold.

1 Additional literature review

We first comment in more detail on the UD max and WD max tests of Bai and Perron (1998) and Bai and Perron (2003) and their relationship to NSP. Bai and Perron (2003) write:

> A useful strategy is to first look at the UD max or WD max tests to see if at least one break is present. If these indicate the presence of at least one break, then the number of breaks can be decided based upon a sequential examination of the sup $F(l+1|l)$ statistics constructed using global minimizers for the break

*Department of Statistics, London School of Economics, Houghton Street, London WC2A 2AE, UK. Email: p.fryzlewicz@lse.ac.uk.

dates (i.e. ignore the test $F(1|0)$ and select m such that the tests sup $F(l + 1|l)$ are insignificant for $l \geq m$. This method leads to the best results and is recommended for empirical applications.

For the purpose of this discussion, we label the process above the 'Improved Sequential Procedure' (ISP). Bai and Perron (2003) do not formulate or prove the inferential properties of the m selected by ISP. For a procedure that selects the number of change-points, the control of global significance would have to mean, in particular, a guarantee that the true number of change-points is at least as high as the estimated number, with at least $1 - \alpha$ probability. NSP provides such a statement as a simple corollary of Theorem 2.1 in the main paper, but ISP is a complex sequential process put together from separate, non-independent, conditionally applied tests, and the exact guarantees for the resulting output (m) have not been shown.

The next difference is that the UD max and WD max tests require the provision of the maximum number of change-points, but NSP does not require this, thereby eliminating the risk of providing too low a maximum by the user.

Furthermore, the ISP test only concerns the number of change-points, but not their locations: inference for locations in Bai and Perron (1998) and Bai and Perron (2003) is carried out later, *conditionally* on the number of change-points and on their estimated locations. Not only that, but also the obtained conditional confidence intervals are asymptotic in nature and are only valid for large sample sizes (unknown to the user). By contrast, NSP provides a single, clear, joint, finite-sample guarantee for the number of change-points and for their locations: it flags up disjoint regions in the data, each of which must contain at least one change-point with a global probability specified by the user. The NSP intervals of significance serve as "unconditional" confidence intervals (in contrast to the conditional CIs of Bai and Perron (1998) and Bai and Perron (2003), whose conditionality on the number of estimated change-points and the estimated locations means that the user cannot be sure whether they contain change-points with a certain probability). The NSP guarantees are valid for any, even small, sample sizes.

Next, we discuss in more detail the most important high-level differences between NSP and the approaches of Fang et al. (2020) and Fang and Siegmund (2020).

(a) While Fang et al. (2020) and Fang and Siegmund (2020) perform change-point location estimation as well as inference, NSP works on the principle of "inference without location estimation". This is a key property of NSP, which enables it to use an all-purpose multiscale test, whose distribution under the null is stochastically bounded by the scan statistic of the corresponding true residuals Z_t, and is therefore independent of the scenario and of the design matrix X used. This means that NSP is ready for use with any user-provided design matrix X, and this will require no new calculations or coding, and will yield correct coverage probabilities. This is in contrast to the approach taken in Fang et al. (2020) and Fang and Siegmund (2020), in which, because of their focus on location estimation, each new scenario not already covered would involve new and fairly complicated approximations of the null distribution. (We note that outside the change-point context, the method for constructing confidence intervals for groups of variables in sparse high dimensional regression by Meinshausen (2015) shares with NSP the attractive property of providing valid error control without assumptions on the design matrix.)

(b) While in Fang et al. (2020) and Fang and Siegmund (2020), the user needs to be able to specify the significant signal shapes to look for, NSP searches for any deviations from local model linearity with respect to specific regressors.

(c) Out of our scenarios, Fang et al. (2020) and Fang and Siegmund (2020) provide results under our Scenario 1 and Scenario 2 with linearity and continuity. Their results do not cover our Scenario 3 (linear regression with arbitrary X) or Scenario 2 with linearity but not necessarily continuity, or Scenario 2 with higher-than-linear polynomials.

(d) Thanks to its double use of the multiresolution sup-norm (in the local linear fit, and then in the test of this fit), NSP is able to handle regression with autoregression practically in the same way as without, in a stable manner and on arbitrarily short intervals, and does not suffer from having to estimate the unknown (nuisance) AR coefficients accurately. This is of importance, as change-point analysis under serial dependence in the data is a problem known to be difficult, and NSP offers a new approach to it, thanks to this feature.

Finally, we provide additional references on the use of scan statistics. In the literature, scaled partial sum statistics acting directly on the data are often combined into variants of scan statistics (Siegmund and Venkatraman, 1995; Arias-Castro et al., 2005; Jeng et al., 2010; Walther, 2010; Chan and Walther, 2013; Sharpnack and Arias-Castro, 2016; König et al., 2020; Munk et al., 2020). They are also used in estimators represented as the simplest (from the point of view of a certain regularity or smoothness functional) fit to the data for which the empirical residuals are deemed to behave like the true residuals (Frick et al., 2014; Davies and Kovac, 2001; Davies et al., 2009; Li, 2016).

2 Discussion of the NSP algorithm

We now comment on a few generic aspects of the NSP algorithm as defined in the main paper.

Length check for $[s, e]$ in line 2 Consider an interval $[s, e]$ with $e - s < p$. If it is known that the matrix $X_{s:e,\cdot}$ is of rank $e - s + 1$ (as is the case, for example, in Scenario 2, for all such s, e) then it is safe to disregard $[s, e]$, as the response $Y_{s:e}$ can then be explained exactly as a linear combination of the columns of $X_{s:e,\cdot}$, so it is impossible to assess any deviations from linearity of $Y_{s:e}$ with respect to $X_{s:e,\cdot}$. Therefore, if this rank condition holds, the check in line 2 of NSP can be replaced with $e - s < p$, which (together with the corresponding modifications in lines 5–10) will reduce the computational effort if $p > 1$. Having $p = p(T)$ growing with T is possible in NSP, but by the above discussion, we must have $p(T) + 1 \leq T$ or otherwise no regions of significance will be found.

Sub-interval sampling Sub-interval sampling in lines 5–10 of the NSP algorithm is done to reduce the computational effort. In the change-point detection literature (without inference considerations), Wild Binary Segmentation (WBS, Fryzlewicz, 2014) uses a random interval sampling mechanism in which all or almost all intervals are sampled at the start of the procedure, i.e. with all or most intervals not being sampled recursively. The same style of interval sampling is used in the Narrowest-Over-Threshold change-point detection (note: not change-point inference) algorithm (Baranowski et al., 2019) and is mentioned in passing in Fang et al. (2020). Instead, NSP uses a different, recursive interval sampling mechanism, introduced in the change-point detection (not inference) context in Wild Binary Segmentation 2 (WBS2, Fryzlewicz, 2020). In NSP (lines 5–10), intervals are sampled separately in each recursive call of the NSP routine. As argued in Fryzlewicz (2020), this enables more thorough exploration of the domain $\{1, \ldots, T\}$ and hence better feature discovery than the non-recursive sampling style. We note that NSP can equally use random or deterministic interval selection mechanisms; a specific example of a deterministic interval sampling scheme in a change-point detection context can be found in Kovács et al. (2023). Our general preference is for NSP to be used with deterministic sampling as it leads to reproducible results without the user having to fix the random seed.

Relationship to NOT The Narrowest-Over-Threshold (NOT) algorithm of Baranowski et al. (2019) is a change-point detection procedure (valid in Scenarios 1 and 2) and comes with no inference considerations. The common feature shared by NOT and NSP is that in their respective aims (change-point detection for NOT; locating regions of global significance for NSP) they iteratively focus on the narrowest intervals on which a certain test (a change-point locator for NOT; a multiscale scan statistic on multiresolution sup-norm fit residuals for NSP) exceeds a threshold, but this is where similarities end: apart from this common feature, the objectives, scopes and modi operandi of both methods are different.

Focus on the smallest significant regions Some authors in the inference literature also identify the shortest intervals (or smallest regions) of significance in data. For example, Dümbgen and Walther (2008) plot minimal intervals on which a density function significantly decreases or increases. Walther (2010) plots minimal significant rectangles on which the probability of success is higher than a baseline, in a two-dimensional spatial model. Fang et al. (2020) mention the possibility of using the interval sampling scheme from

Fryzlewicz (2014) to focus on the shortest intervals in their CUSUM-based determination of regions of significance in Scenario 1. In addition to NSP's new definition of significance involving the multiresolution sup-norm fit (whose benefits are explained in Section 2.2 of the main paper), NSP is also different from these approaches in that its pursuit of the shortest significant intervals is at its algorithmic core and is its main objective. To achieve it, NSP uses a number of solutions which, to the best of our knowledge, either are new or have not been considered in this context before. These include the two-stage search for the shortest significant subinterval (NSP routine, line 19) and the recursive sampling (lines 5–10, proposed previously but in a non-inferential context by Fryzlewicz (2020)).

Lack of penalisation for fine scales. Instead of using multiresolution sup-norms (multiscale scan statistics) as defined in the main paper, some authors, including Walther (2010) and Frick et al. (2014), use alternative definitions which penalise fine scales (i.e. short intervals) in order to enhance detection power at coarser scales. We do not pursue this route, as NSP aims to discover significant intervals that are as short as possible, and hence we are interested in retaining good detection power at fine scales. However, some natural penalisation of fine scales necessarily occurs in the self-normalised case; see Section 3.1 of the main paper.

Upper bounds for p-values on non-detection intervals. By calculating the quantity $D_{[s,e]}$ on each data section $[s, e]$ delimited by the detected intervals of significance, an upper bound on the p-value for the existence of a change-point in $[s, e]$ can be obtained as $P(\|Z\|_{\mathcal{I}^a} > D_{[s,e]})$. If the interval $[s, e]$ were considered by NSP before (as would be the case e.g. if $\tau_L = \tau_R = 0$ and the deterministic sampling grid were used), from the non-detection on $[s, e]$, we would necessarily have $P(\|Z\|_{\mathcal{I}^a} > D_{[s,e]}) \geq \alpha$.

Bottom-up implementation of NSP Our implementation of NSP is "bottom-up", in the sense that at each recursive stage, we consider the intervals $[s_m, e_m]$ in non-decreasing order of their lengths, and exit the current recursive stage (if and) as soon as significance is declared, rather than moving on to longer intervals. This aligns with the objective of looking for the shortest intervals (so the examination of longer intervals is unnecessary if shorter significant intervals have been found). Any non-bottom-up implementation of NSP would therefore unnecessarily be wasting computational resources. This is in contrast to, for example, the region-based multiple testing method of Meijer et al. (2015), in which the successive p-value adjustments (which lead to power improvements) are only possible because of the top-down character of that approach.

3 Proofs of results of Section 2

Proof of Proposition 2.1. As $[s, e]$ does not contain a change-point, there is a β^* such that $Y_{s:e} = X_{s:e,\cdot}\beta^* + Z_{s:e}$. Therefore, $D_{[s,e]} = \min_\beta \|Y_{s:e} - X_{s:e,\cdot}\beta\|_{\mathcal{I}^d_{[s,e]}} \leq \|Y_{s:e} - X_{s:e,\cdot}\beta^*\|_{\mathcal{I}^d_{[s,e]}} = \|Z_{s:e}\|_{\mathcal{I}^d_{[s,e]}}$, which completes the proof. □

Proof of Theorem 2.1. The second inequality is implied by (5) in the main paper. We now prove the first inequality. On the set $\|Z\|_{\mathcal{I}^d} \leq \lambda_\alpha$, each interval S_i must contain a change-point as if it did not, then by Proposition 2.1, we would have to have

$$D_{S_i} \leq \|Z\|_{\mathcal{I}^d} \leq \lambda_\alpha. \tag{1}$$

However, the fact that S_i was returned by NSP means, by line 14 of the NSP algorithm, that $D_{S_i} > \lambda_\alpha$, which contradicts (1). This completes the proof. □

Proof of Proposition 2.2. The inequality is true because for any fixed β, the norm $\|Z - X\beta\|_{\mathcal{I}^d}$ is a maximum over a larger set than the maximum in $\|Z_{s:e} - X_{s:e,\cdot}\beta\|_{\mathcal{I}^d_{[s,e]}}$. We now prove the equality. As $[s, e]$ does not contain a change-point, there is a β^* such that $Y_{s:e} = X_{s:e,\cdot}\beta^* + Z_{s:e}$. We have

$$\begin{aligned} D_{[s,e]} &= \min_\beta \|Y_{s:e} - X_{s:e,\cdot}\beta\|_{\mathcal{I}^d_{[s,e]}} = \min_\beta \|X_{s:e,\cdot}\beta^* + Z_{s:e} - X_{s:e,\cdot}\beta\|_{\mathcal{I}^d_{[s,e]}} \\ &= \min_\beta \|Z_{s:e} - X_{s:e,\cdot}(\beta - \beta^*)\|_{\mathcal{I}^d_{[s,e]}} = \min_{\beta-\beta^*} \|Z_{s:e} - X_{s:e,\cdot}(\beta - \beta^*)\|_{\mathcal{I}^d_{[s,e]}} = \min_\beta \|Z_{s:e} - X_{s:e,\cdot}\beta\|_{\mathcal{I}^d_{[s,e]}}. \end{aligned}$$

□

Proof of Theorem 2.3. On the set $\min_\beta \|Z - X\beta\|_{\mathcal{I}^d} \leq \lambda_\alpha$, each interval S_i must contain a change-point as if it did not, then by Proposition 2.2, we would have to have

$$D_{S_i} \leq \min_\beta \|Z - X\beta\|_{\mathcal{I}^d} \leq \lambda_\alpha. \tag{2}$$

However, the fact that S_i was returned by NSP means, by line 14 of the NSP algorithm, that $D_{S_i} > \lambda_\alpha$, which contradicts (2). This completes the proof. □

4 Estimated σ^2, and other light-tailed distributions

We first show under what condition Theorem 2.2 in the main paper remains valid with an estimated variance σ^2, and give an estimator of σ^2 that satisfies this condition for certain matrices X and parameter vectors $\beta^{(j)}$. Similar considerations are possible for the light-tailed distributions from the latter part of this section, but we omit them here. With $\{Z_t\}_{t=1}^T \sim N(0, \sigma^2)$ rather than $N(0, 1)$, the statement of Theorem 2.2 of the main paper trivially modifies to $\lim_{T\to\infty} P\left(\max_{1\leq s\leq e\leq T}\; U_{s,e}(Z) \leq \sigma(a_T + b_T\,\gamma)\right) = \exp(-e^{-\gamma})$. From the form of the limiting distribution, it is clear that the theorem remains valid if $\gamma_T \underset{T\to\infty}{\longrightarrow} \gamma$ is used in place of γ, yielding

$$\lim_{T\to\infty} P\left(\max_{1\leq s\leq e\leq T}\; U_{s,e}(Z) \leq \sigma(a_T + b_T\,\gamma_T)\right) = \exp(-e^{-\gamma}). \tag{3}$$

With σ estimated via a generic estimator $\hat{\sigma}$, we ask under what circumstances

$$\lim_{T\to\infty} P\left(\max_{1\leq s\leq e\leq T}\; U_{s,e}(Z) \leq \hat{\sigma}(a_T + b_T\,\gamma)\right) = \exp(-e^{-\gamma}). \tag{4}$$

In light of (3), it is enough to solve for γ_T in $\sigma(a_T + b_T\,\gamma_T) = \hat{\sigma}(a_T + b_T\,\gamma)$, yielding $\gamma_T = \frac{a_T}{b_T}\left(\frac{\hat{\sigma}}{\sigma} - 1\right) + \frac{\hat{\sigma}}{\sigma}\gamma$. In view of the form of a_T and b_T defined in Theorem 2.2 of the main paper, we have $\gamma_T \underset{T\to\infty}{\longrightarrow} \gamma$ on a set large enough for (4) to hold if

$$\left|\frac{\hat{\sigma}}{\sigma} - 1\right| = o_P(\log^{-1} T), \quad \text{or equivalently} \quad \left|\frac{\hat{\sigma}^2}{\sigma^2} - 1\right| = o_P(\log^{-1} T). \tag{5}$$

After Rice (1984) and Dümbgen and Spokoiny (2001), define $\hat{\sigma}_R^2 = \frac{1}{2(T-1)} \sum_{t=1}^{T-1}(Y_{t+1} - Y_t)^2$. Define the signal in model (2) of the main paper by $f_t = X_{t,\cdot}\beta^{(j)}$ for $t = \eta_j + 1, \ldots, \eta_{j+1}$, for $j = 0, \ldots, N$. The total variation of a vector $\{f_t\}_{t=1}^T$ is defined by $TV(f) = \sum_{t=1}^{T-1} |f_{t+1} - f_t|$. As in Dümbgen and Spokoiny (2001), we have $\mathbb{E}\{(\hat{\sigma}_R^2/\sigma^2 - 1)^2\} = O(T^{-1}\{1 + TV^2(f)\})$, from which (5) follows, by Markov inequality, if

$$TV(f) = o(T^{1/2} \log^{-1} T). \tag{6}$$

By way of a simple example, in Scenario 1, $TV(f) = \sum_{j=1}^N |f_{\eta_j} - f_{\eta_j+1}|$, and therefore (6) is satisfied if the sum of jump magnitudes in f is $o(T^{1/2}\log^{-1} T)$. Note that if f is bounded with a number of change-points that is finite in T, then $TV(f) = \mathrm{const}(T)$. Similar arguments apply in Scenario 2, and in Scenario 3 for some matrices X.

Without formal theoretical justifications, we also mention two further estimators of σ^2 (or σ) which we use in our numerical work. In Scenarios 1 and 2, we use $\hat{\sigma}_{MAD}$, the Median Absolute Deviation (MAD) estimator as implemented in the R routine `mad`, computed on the sequence $\{2^{-1/2}(Y_{t+1} - Y_t)\}_{t=1}^{T-1}$. Empirically, $\hat{\sigma}_{MAD}$ is more robust than $\hat{\sigma}_R$ to the presence of change-points in f_t, but is also more sensitive to departures from the Gaussianity of Z_t. In Scenario 3, in settings outside Scenarios 1 and 2, we use the following estimator. In model (2) of the main paper, we estimate σ via least squares, on a rolling window basis, using the window of size $w = \min\{T, \max([T^{1/2}], 20)\}$, to obtain the sequence of estimators $\hat{\sigma}_1, \ldots, \hat{\sigma}_{T-w+1}$. We take $\hat{\sigma}_{MOLS} = \mathrm{median}(\hat{\sigma}_1, \ldots, \hat{\sigma}_{T-w+1})$, where MOLS stands for 'Median of OLS estimators'. The hope is that most of the local estimators $\hat{\sigma}_1, \ldots, \hat{\sigma}_{T-w+1}$ are computed on change-point-free sections of the data, and therefore the median of these local estimators should serve as an accurate estimator of the true σ. Empirically, $\hat{\sigma}_{MOLS}$ is a useful alternative to $\hat{\sigma}_R$ in settings in which condition (6) is not satisfied.

Kabluchko and Wang (2014) provide a result similar to Theorem 2.2 of the main paper for distributions of Z dominated by the Gaussian in a sense specified below. These include, after scaling so that $\mathbb{E}(Z) = 0$ and $\mathrm{Var}(Z) = 1$, the symmetric Bernoulli, symmetric binomial and uniform distributions, amongst others. We now briefly summarise it. Consider the cumulant-generating function of Z defined by $\varphi(u) = \log \mathbb{E}(e^{uZ})$ and assume that for some $\sigma_0 > 0$, we have $\varphi(u) < \infty$ for all $u \geq -\sigma_0$. Assume further that for all $\varepsilon > 0$, $\sup_{u \geq \varepsilon} \varphi(u)/(u^2/2) < 1$. Finally, assume

$$\varphi(u) = \frac{u^2}{2} - \kappa u^d + o(u^d), \quad u \downarrow 0,$$

for some $d \in \{3, 4, \ldots\}$ and $\kappa > 0$. Typical values of d for non-symmetric and symmetric distributions, respectively, are 3 and 4. Under these assumptions, we have

$$\lim_{T \to \infty} P \left(\frac{1}{2} \left\{ \max_{1 \leq s \leq e \leq T} U_{s,e}(Z) \right\}^2 \leq \log \left\{ T \log^{\frac{d-6}{2(d-2)}} T \right\} + \gamma \right) = \exp(-\Lambda_{d,\kappa} e^{-\gamma}),$$

for all $\gamma \in \mathbb{R}$, where $\Lambda_{d,\kappa} = \pi^{-1/2} \Gamma(d/(d-2))(2\kappa)^{2/(d-2)}$. After simple algebraic manipulations, this result permits a selection of λ_α for use in Theorem 2.1 of the main paper, similarly to Section 2.3 of the main paper.

5 Importance of two-stage search for shortest interval of significance

We next illustrate the importance of the two-stage search for the shortest interval of significance, whose stage two is performed in line 19 of the NSP algorithm via the call

$$[\tilde{s}, \tilde{e}] := \textsc{ShortestSignificantSubinterval}(s_{m_0}, e_{m_0}, Y, X, M, \lambda_\alpha).$$

Consider the Blocks signal referred to in the main paper but with the much smaller noise standard deviation $\sigma = 1$. A realisation Y_t is shown in the left plot of Figure 1. All $N = 11$ change-points are visually obvious and hence we would expect NSP to return 11 intervals $[\tilde{s}_i, \tilde{e}_i]$, exactly covering the true change-points, for which we would have $\tilde{e}_i - \tilde{s}_i = 1$ for most if not all i. As shown in the middle plot of Figure 1, the NSP procedure with no overlap and with the same parameters as in Section 5.1 of the main paper returns 11 intervals of significance with $\tilde{e}_i - \tilde{s}_i = 1$ for $i = 1, \ldots, 10$ and $\tilde{e}_{11} - \tilde{s}_{11} = 2$. The 11 intervals of significance cover the true change-points.

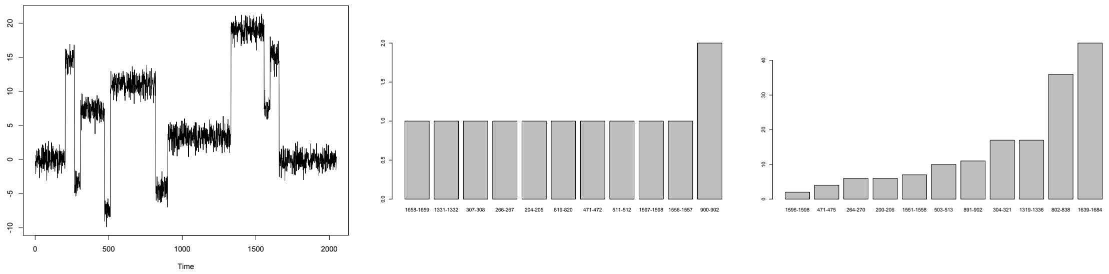

Figure 1: Left: realisation Y_t of noisy Blocks with $\sigma = 1$. Middle: prominence plot of NSP-detected intervals. Right: the same for NSP(1). See Section 5 for more details.

However, consider now an alternative version of NSP, labelled NSP(1), which only performs a one-stage search for the shortest interval of significance. NSP(1) proceeds by replacing line 19 of the NSP algorithm by

$$[\tilde{s}, \tilde{e}] := [s_{m_0}, e_{m_0}].$$

In other words, $[s_{m_0}, e_{m_0}]$ is not searched for its shortest sub-interval of significance, but is added to $\mathcal{S}$ as it is. The output of NSP(1) on Y_t is shown in the right plot of Figure 1. The intervals of significance returned by NSP(1) are unreasonably long from the statistical point of view, with $\tilde{e}_i - \tilde{s}_i$ varying from 2 to 45. However, this has a clear explanation from the point of view of the algorithmic construction of NSP(1). For example, in the first recursive stage, in which $[s, e] = [1, T]$, the spacing of the (approximately) equispaced grid from which the candidate intervals $[s_m, e_m]$ are drawn varies between 45 and 46. Therefore, it is unsurprising that the first detection performed by NSP(1) is such that $\tilde{e}_i - \tilde{s}_i = 45$.

This issue would not arise in NSP, as NSP would then search this detection interval for its shortest significant sub-interval. From the output of the NSP procedure, we can see that this second-stage search drastically reduced the length of this detection interval, which is unsurprising given how obvious the change-points are in this example. This illustrates the importance of the two-stage search in NSP.

For very long signals, it is conceivable that an analogous three-stage search may be a better option, possibly combined with a reduction in M to enhance the speed of the procedure.

6 Self-normalised NSP – further discussion

We now outline the construction of $\hat{Z}^{(k)}$ for $k = 1, 2, 3$ so that (11) in the main paper is guaranteed, and propose a suitable estimator of V_T^2 for use in (11) in the main paper.

$k = 1$. Let $(\hat{Z}_{i+1}^{(1)}, \ldots, \hat{Z}_j^{(1)})$ be the ordinary least-squares residuals from regressing $Y_{(i+1):j}$ on $X_{(i+1):j,\cdot}$, where $j - i > p$. As $[s, e]$ contains no change-point, we have $(\hat{Z}_{i+1}^{(1)})^2 + \ldots + (\hat{Z}_j^{(1)})^2 \leq Z_{i+1}^2 + \ldots + Z_j^2$ and hence $\log^{1/2+\epsilon}\{cV_T^2/((\hat{Z}_{i+1}^{(1)})^2 + \ldots + (\hat{Z}_j^{(1)})^2)\} \geq \log^{1/2+\epsilon}\{cV_T^2/(Z_{i+1}^2 + \ldots + Z_j^2)\}$.

$k = 2$. We use

$$(\hat{Z}_{i+1}^{(2)}, \ldots, \hat{Z}_j^{(2)}) = (1 + \epsilon)(\hat{Z}_{i+1}^{(1)}, \ldots, \hat{Z}_j^{(1)}), \tag{7}$$

which guarantees $(\hat{Z}_{i+1}^{(2)})^2 + \ldots + (\hat{Z}_j^{(2)})^2 \geq Z_{i+1}^2 + \ldots + Z_j^2$ for ϵ and $j - i$ suitably large, for a range of distributions of Z_t and design matrices X. We now briefly sketch the argument justifying this for Scenario 1; similar considerations are possible in Scenario 2 but are notationally much more involved and we omit them here. The argument relies again on self-normalisation. From standard least-squares theory (in any Scenario), we have $(\hat{Z}_{(i+1):j}^{(1)})^\top \hat{Z}_{(i+1):j}^{(1)} = Z_{(i+1):j}^\top Z_{(i+1):j} - Z_{(i+1):j}^\top X_{(i+1):j,\cdot}(X_{(i+1):j,\cdot}^\top X_{(i+1):j,\cdot})^{-1} X_{(i+1):j,\cdot}^\top Z_{(i+1):j}$. In Scenario 1, $(X_{(i+1):j,\cdot}^\top X_{(i+1):j,\cdot})^{-1} = (j - i)^{-1}$, and hence $Z_{(i+1):j}^\top X_{(i+1):j,\cdot}(X_{(i+1):j,\cdot}^\top X_{(i+1):j,\cdot})^{-1} X_{(i+1):j,\cdot}^\top Z_{(i+1):j} = U_{i+1,j}(Z)^2$. From the above, we obtain

$$\begin{aligned}
(\hat{Z}_{(i+1):j}^{(1)})^\top \hat{Z}_{(i+1):j}^{(1)} &= Z_{(i+1):j}^\top Z_{(i+1):j} \left(1 - \frac{U_{i+1,j}(Z)^2}{Z_{(i+1):j}^\top Z_{(i+1):j}}\right) \\
&= Z_{(i+1):j}^\top Z_{(i+1):j} \left(1 - \frac{1}{j - i} \log^{1+2\epsilon}\{cV_T^2/(Z_{i+1}^2 + \ldots + Z_j^2)\}\right. \\
&\times \left. I_{\rho_{1/2,1/2+\epsilon,c}}^2(\zeta_T^{\text{se}}, V_i^2/V_T^2, V_j^2/V_T^2)\right).
\end{aligned} \tag{8}$$

In light of the distributional result (10) of the main paper, the relationship between the statistic $I_{\rho_{1/2,1/2+\epsilon,c}}(W, u, v)$ and Račkauskas and Suquet (2004)'s statistic $\text{UI}(\rho_{1/2,1/2+\epsilon,c})$, as well as their Remark 5, we are able to bound $\sup_{0 \leq i < j \leq T} I_{\rho_{1/2,1/2+\epsilon,c}}^2(\zeta_T^{\text{se}}, V_i^2/V_T^2, V_j^2/V_T^2)$ by a term of order $O(\log T)$ on a set of probability $1 - O(T^{-1})$. Making the mild assumption that $\sup_{0\le i<j\le T}\log^{1+2\epsilon}\{cV_T^2/(Z_{i+1}^2+\ldots+Z_j^2)\}\asymp l_T=o_P(T\log^{-1}T)$ and continuing from (8), we obtain $(\hat{Z}^{(1)}_{(i+1):j})^\top\hat{Z}^{(1)}_{(i+1):j}\ \ge\ Z^\top_{(i+1):j}Z_{(i+1):j}\left(1-C(j-i)^{-1}l_T\log T\right)$ for a certain constant $C>0$, which can be bounded from below by $Z^\top_{(i+1):j}Z_{(i+1):j}(1+\epsilon)^{-2}$, uniformly over those i,j for which $(j-i)^{-1}l_T\log T\to 0$. This justifies (7) and completes the argument.

$k=3$. Having obtained $\hat{Z}^{(1)}_{(i+1):j}$ and $\hat{Z}^{(2)}_{(i+1):j}$ as above, the problem of obtaining $\hat{Z}^{(3)}_{s:e}$ to guarantee

$$\sup_{s-1\le i<j\le e}\frac{|\hat{Z}^{(3)}_{i+1}+\ldots+\hat{Z}^{(3)}_j|}{\sqrt{(\hat{Z}^{(2)}_{i+1})^2+\ldots+(\hat{Z}^{(2)}_j)^2}\log^{1/2+\epsilon}\{cV_T^2/((\hat{Z}^{(1)}_{i+1})^2+\ldots+(\hat{Z}^{(1)}_j)^2)\}}$$
$$\le\sup_{s-1\le i<j\le e}\frac{|Z_{i+1}+\ldots+Z_j|}{\sqrt{(\hat{Z}^{(2)}_{i+1})^2+\ldots+(\hat{Z}^{(2)}_j)^2}\log^{1/2+\epsilon}\{cV_T^2/((\hat{Z}^{(1)}_{i+1})^2+\ldots+(\hat{Z}^{(1)}_j)^2)\}},\qquad(9)$$

which in turn guarantees the bound (11) in the main paper, is practically equivalent to the multiresolution norm minimisation solved in Step 1 of Section 2.2 of the main paper except it now uses a weighted version of the norm $\|\cdot\|_{\mathcal{I}^a_{[s,e]}}$, where the weights are given in the denominator of (9). This weighted problem is solved via linear programming just as easily as Step 1 of Section 2.2 of the main paper, the only difference being that the relevant constraints are multiplied by the corresponding weights.

We now discuss further practicalities of the self-normalisation. In the exposition of the main paper, we use all intervals $[i+1,j]\subseteq[s,e]$, i.e. the set $\mathcal{I}^a_{[s,e]}$. In practice, for computational reasons, we compute the supremum on the LHS of (11) in the main paper over the dyadic set $\mathcal{I}^d_{[s,e]}$, which does not alter the validity of the bound. Our empirical experience is that the statistic on the LHS of (11) of the main paper is fairly robust to the choice of V_T^2, as the latter only enters through the (close to) square-root logarithmic term in the denominator. In addition, over-estimation of V_T^2 for use on the LHS of (11) of the main paper is permitted as it only strengthens the bound in (11) of the main paper. For these reasons, we do not dwell on the accurate estimation of V_T^2 here, but use the rough estimate $\hat{V}_T^2=\frac{T}{T-w+1}\sum_{t=1}^{T-w+1}\hat{\sigma}_t^2$, where the $\hat{\sigma}_t$'s are the constituents of the $\hat{\sigma}_{MOLS}$ estimator from Section 4. As clarified earlier, the use of (7) requires that small values of $j-i$ do not enter in the computation of the supremum on the LHS of (11) of the main paper. In practice, however, we use

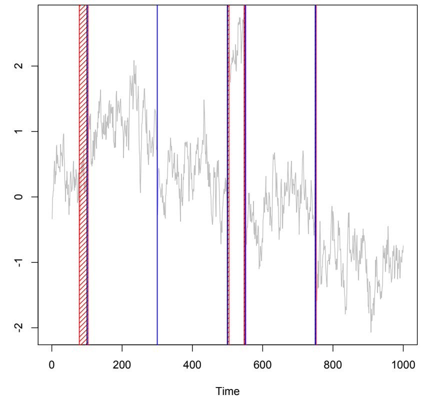

Figure 2: Piecewise-constant signal from Dette et al. (2020) with Gaussian AR(1) noise with coefficient 0.9 and standard deviation $(1 - 0.9^2)^{-1/2}/5$ (light grey), NSP intervals of significance (shaded red), true change-points (blue); see Section 7 for details.

all $[i+1, j] \in \mathcal{I}^d_{[s,e]}$. This is because the function $I_{\rho_{1/2,1/2+\epsilon,c}}(\zeta_T^{\mathrm{se}}, V_i^2/V_T^2, V_j^2/V_T^2)$ naturally penalises small scales (i.e. short intervals $[i+1, j]$) through the use of the logarithmic term in the denominator. Therefore, in practice, short intervals $[i+1, j]$ do not tend to achieve the supremum on the LHS of (11) of the main paper and as a result, we have found further exclusion of such short intervals unnecessary. Finally, we have experimented with ϵ in the range $[0.03, 0.1]$ and found little difference in practical performance. Our code uses $\epsilon = 0.03$ as a default.

7 NSP with autoregression

We use the piecewise-constant signal of length $T = 1000$ from the first simulation setting in Dette et al. (2020), contaminated with Gaussian AR(1) noise with coefficient 0.9 and standard deviation $(1 - 0.9^2)^{-1/2}/5$. A sample path, together with the true change-point locations, is shown in Figure 2.

We run the AR version of the NSP algorithm (as outlined in Section 3.2 of the main paper), with the following parameters: a deterministic equispaced interval sampling grid, $M = 100$, $\alpha = 0.1$, no overlap, $\hat{\sigma}^2_{MOLS}$ estimator of the residual variance. The resulting intervals are shown in Figure 2; NSP intervals cover four out of the five true change-points, and there

Table 1: Percentage of sample paths with the given numbers of NSP-detected intervals in the autoregressive example of Section 7.

no. of intervals of significance	2	3	4	5
percentage of sample paths	11	32	42	15

are no spurious intervals.

We simulate from this model 100 times and obtain the following results. In 100% of the sample paths, each NSP interval of significance covers one true change-point (which fulfils the promise of Theorem 2.1 of the main paper). The distribution of the detected numbers of intervals is as in Table 1; we recall that NSP, with a fixed significance level, does not promise to detect the number of intervals equal to the number of true change-points in the underlying process.

8 Computation of the NSP threshold by simulation

In a number of locations in the main paper, we mention the possibility of obtaining the NSP thresholds by simulation. We now clarify how this is done. For example, to solve

$$P(\|Z\|_{\mathcal{I}^d} > \lambda_\alpha) = \alpha$$

for λ_α (see e.g. Theorem 2.1 of the main paper) by simulation, we would simulate multiple realisations of $\|Z\|_{\mathcal{I}^d}$ and choose λ_α as the $100(1-\alpha)\%$ empirical quantile of the sample. We proceed similarly in Section 2.4, in which the task is to approximate the distribution of $\min_\beta \|Z - X\beta\|_{\mathcal{I}^d}$. It is important to note that this can easily be done for any distribution of Z (assumed known), not just Gaussian. (If there is uncertainty regarding the distribution of Z and there are a few plausible candidates, the corresponding threshold can be computed for each of them and the largest one among them chosen for use in the NSP algorithm.)

This threshold selected as the empirical quantile of $\min_\beta \|Z - X\beta\|_{\mathcal{I}^d}$, for the Gaussian case in Scenarios 1 and 2, is implemented in the R package `nsp` and can be used upon setting `thresh.type = "sim"` in the `nsp_poly` routine.

One remaining question is whether it is possible to use the standard (non-self-normalised)

NSP without knowledge of the distribution of the innovations Z. Here, the following simple practical procedure for determining the threshold via simulation may help.

1. Pre-estimate the time-varying signal $X\beta$ via a localised moving-window fit; then pre-estimate the innovations $\hat{Z}$.

2. Re-sample the innovations to estimate the distribution of the multiscale deviation measure $\|\hat{Z}\|_{\mathcal{I}^d}$.

3. Use a suitable empirical quantile of this distribution as the NSP threshold.

9 Detection consistency and lengths of NSP intervals – proofs and discussion

Proof of Theorem 4.1 (main paper). Assume initially that f_t has a single change-point η_1. As NSP considers all intervals by the assumption of the theorem, it will certainly consider intervals symmetric about the true change-point, i.e. $[\eta_1 - d + 1, \eta_1 + d]$, for all appropriate d. In Scenario 1, there is an explicit formula for the deviation measure $D_{[s,e]}$ on any interval $[s, e]$, given by

$$D_{[s,e]} = \max_{\tau \in \{1,\ldots,e-s+1\}} \frac{1}{2\sqrt{\tau}} \left(\max_{s_1 \in \{s,\ldots,e+1-\tau\}} \sum_{t=s_1}^{s_1+\tau-1} Y_t - \min_{s_1 \in \{s,\ldots,e+1-\tau\}} \sum_{t=s_1}^{s_1+\tau-1} Y_t \right). \quad (10)$$

Without loss of generality, assume $f_{\eta_1} > f_{\eta_1+1}$. Representation (10) implies

$$\begin{aligned} D_{[\eta_1-d+1,\eta_1+d]} &\geq \frac{1}{2\sqrt{d}} \left(\max_{s_1 \in \{\eta_1-d+1,\ldots,\eta_1+1\}} \sum_{t=s_1}^{s_1+d-1} Y_t - \min_{s_1 \in \{\eta_1-d+1,\ldots,\eta_1+1\}} \sum_{t=s_1}^{s_1+d-1} Y_t \right) \\ &\geq \frac{1}{2\sqrt{d}} \left(\sum_{t=\eta_1-d+1}^{\eta_1} Y_t - \sum_{t=\eta_1+1}^{\eta_1+d} Y_t \right) \\ &\geq \frac{1}{2} |f_{\eta_1+1} - f_{\eta_1}| \sqrt{d} - \|Z\|_{\mathcal{I}^a}. \end{aligned} \quad (11)$$

On the set $\|Z\|_{\mathcal{I}^a} \leq \lambda_\alpha$, (11) is further bounded from below by $\frac{1}{2}|f_{\eta_1+1} - f_{\eta_1}|\sqrt{d} - \lambda_\alpha$. From the definition of the NSP algorithm, detection on $[s, e]$ is triggered by the event $D_{[s,e]} > \lambda_\alpha$, so detection on $[\eta_1 - d + 1, \eta_1 + d]$ is triggered if (note: not "only if" as we are using lower bounds here) $\frac{1}{2}|f_{\eta_1+1} - f_{\eta_1}|\sqrt{d} - \lambda_\alpha > \lambda_\alpha$, or

$$|f_{\eta_1+1} - f_{\eta_1}|\sqrt{d} > 4\lambda_\alpha. \qquad (12)$$

As NSP looks for shortest intervals of detection, the NSP interval of significance around η_1 will definitely be no longer than $2d = |[\eta_1 - d + 1, \eta_1 + d]|$. However, from (12), it is sufficient for detection to be triggered if $d > \frac{16\lambda_\alpha^2}{|f_{\eta_1+1} - f_{\eta_1}|^2}$. This shows that the maximum length of an NSP interval of significance will not exceed $2\bar{d}$, where $\bar{d} = \left\lceil \frac{16\lambda_\alpha^2}{|f_{\eta_1+1} - f_{\eta_1}|^2} \right\rceil + 1$. We now turn our attention to the multiple change-point case. For each change-point η_j, define its corresponding $\bar{d}_j$ as in formula (13) of the main paper. Recall we are on the set $\|Z\|_{\mathcal{I}^a} \le \lambda_\alpha$. Note first that even though the NSP interval of significance around η_j is guaranteed to be of length at most $2\bar{d}_j$, it will not necessarily be a subinterval of $[\eta_j - \bar{d}_j + 1, \eta_j + \bar{d}_j]$ (as NSP simply looks for the shortest intervals of significance and interval symmetry around the true change-point is not explicitly promoted). Therefore, in order that an interval detection around η_j does not interfere with detections around η_{j-1} and η_{j+1}, the distances $\eta_j - \eta_{j-1}$ and $\eta_{j+1} - \eta_{j-1}$ must be suitably long, but this is guaranteed by Assumption 4.1 from the main paper. This completes the proof. □

As an aside, note in addition that in the Gaussian case $Z_t \sim N(0, 1)$, Theorem 2.2 of the main paper implies $\lambda_\alpha = O(\log^{1/2} T)$; in fact for $\alpha = 0.05$, we have $\lambda_\alpha \le 1.33\sqrt{2 \log T}$ for $T \ge 100$, for $\alpha = 0.1$, we have $\lambda_\alpha \le 1.25\sqrt{2 \log T}$ over the same range of T.

Proof of Corollary 4.1 (main paper). From Lemma 1 in Yao (1988), we have

$$P(\|Z\|_{\mathcal{I}^a} \le \sigma(1 + \Delta)\sqrt{2 \log T}) \to 1$$

as $T \to \infty$. This combined with the statement of Theorem 4.1 in the main paper proves the result. □

Proof of Theorem 4.2 (main paper). Assume initially that f_t has a single change-point

η_1. In the same way in which the NSP procedure is "blind" to constant shifts in the data in Scenario 1, it is also invariant to the addition of linear trends in the piecewise-linear Scenario 2. Assume, therefore, that we have added a linear trend to Y_t in such a way that the true signal is symmetric around the true change-point η_1. The case that will lead to the longest interval is one in which the change-point leads to a trapezoid shape of the true signal (as in, for example, $1, 2, 3, 3, 2, 1$) rather than one with a single peak or trough (e.g. $1, 2, 3, 2, 1$). Therefore we assume the former case as the "worst case" (whether this is or is not assumed will only lead to $O(1)$ differences in the length of the NSP intervals, so is irrelevant from the point of view of rates). Note that for such a trapezoid signal, the location of η_1 is unambiguous (in the cartoon example above, it must be at the first 3). For such a transformed signal (a transformation which does not change the output of the NSP algorithm), consider intervals symmetric around the true change-point, i.e. $[\eta_1-d+1, \eta_1+d]$, which will be considered by this version of NSP as it considers all intervals. We have

$$D_{[\eta_1-d+1,\eta_1+d]} = \min_{\tilde{f}_{(\eta_1-d+1):(\eta_1+d)}} \|Y_{(\eta_1-d+1):(\eta_1+d)} - \tilde{f}_{(\eta_1-d+1):(\eta_1+d)}\|_{\mathcal{I}^a_{[\eta_1-d+1,\eta_1+d]}}, \tag{13}$$

where the minimum is taken with respect to all linear fits on $[\eta_1 - d + 1, \eta + d]$. Consider a single scale τ. Observing that taking moving partial sums does not change the linearity of $\tilde{f}$, and continuing from (13), we have

$$\begin{aligned} D_{[\eta_1-d+1,\eta_1+d]} &\geq \min_{\tilde{f}_{(\eta_1-d+1):(\eta_1+d)}} \max_{s_1\in\{\eta_1-d+1,\ldots,\eta_1+d+1-\tau\}} \left|\tau^{-1/2}\sum_{t=s_1}^{s_1+\tau-1} Y_t - \tilde{f}_{(\eta_1-d+1):(\eta_1+d)}\right| \\ &\geq \min_{\tilde{f}_{(\eta_1-d+1):(\eta_1+d)}} \max_{s_1\in\{\eta_1-d+1,\ldots,\eta_1+d+1-\tau\}} \left|\tau^{-1/2}\sum_{t=s_1}^{s_1+\tau-1} f_t - \tilde{f}_{(\eta_1-d+1):(\eta_1+d)}\right| \\ &- \|Z\|_{\mathcal{I}^a}. \end{aligned} \tag{14}$$

Observe now that since f_t is symmetric around η_1, the minimising $\tilde{f}$ must be constant. So restrict the class of candidate fits $\tilde{f}$ to constant. Denote the slope of f_t before the change-point by ξ. We have

$$\min_{\tilde{f}_{(\eta_1-d+1):(\eta_1+d)}} \max_{s_1 \in \{\eta_1-d+1,\ldots,\eta_1+d+1-\tau\}} \left| \tau^{-1/2} \sum_{t=s_1}^{s_1+\tau-1} f_t - \tilde{f}_{(\eta_1-d+1):(\eta_1+d)} \right| =$$
$$\frac{\tau^{1/2}}{2} \left(\frac{1}{\tau} \max_{s_1 \in \{\eta_1-d+1,\ldots,\eta_1+d+1-\tau\}} \sum_{t=s_1}^{s_1+\tau-1} f_t - \frac{1}{\tau} \min_{s_1 \in \{\eta_1-d+1,\ldots,\eta_1+d+1-\tau\}} \sum_{t=s_1}^{s_1+\tau-1} f_t \right) =$$
$$\frac{\tau^{1/2}}{2} \xi(d-\tau). \tag{15}$$

Take $\tau = Cd$ for $C \in (0,1)$. (14) and (15) together imply $D_{[\eta_1-d+1,\eta_1+d]} \geq C_1 \xi d^{3/2} - \|Z\|_{\mathcal{I}^a}$ for a certain universal constant C_1. Therefore, on $\|Z\|_{\mathcal{I}^a} \leq \lambda_\alpha$, detection on $[\eta_1-d+1, \eta_1+d]$ will be triggered if $C_1 \xi d^{3/2} > 2\lambda_\alpha$, or in other words if $d \geq C_2 \lambda_\alpha^{2/3} \xi^{-2/3}$, for a large enough constant C_2. This shows that the NSP interval of significance will be of length $O(\lambda_\alpha^{2/3} \xi^{-2/3})$.

We now discuss the slope ξ. Suppose before the symmetrisation the slopes around η_1 were ξ_1 and ξ_2. After the symmetrisation, they are now $\xi_1 + \xi_3$ and $\xi_2 + \xi_3$ where $\xi_1 + \xi_3 = -(\xi_2 + \xi_3)$, which means $\xi = |\xi_1 - \xi_2|/2$ (w.l.o.g., $\xi > 0$). Typically, if $f_t = f(t/T)$ for a certain piecewise-linear function $f(u) : (0,1] \to \mathbb{R}$, then $\xi = O(T^{-1})$. In the Gaussian case, we have $\lambda_\alpha = O(\sqrt{\log T})$. Therefore, if $\xi = O(T^{-1})$, then the NSP interval of significance will have the length $O(T^{2/3} \log^{1/3} T)$.

In the multiple change-point case, the argument about the relevance of Assumption 4.1 from the proof of Theorem 4.1 (main paper) still applies here, and this completes the proof of the theorem. □

Proof of Corollary 4.2 (main paper). The argument is identical to the proof of Corollary 4.1 from the main paper. □

10 NSP with autocorrelated innovations

Scenario 4 permits the use of NSP in settings in which autocorrelation is present, but this is done through the use of the lagged response as an additional covariate, rather than through allowing the innovations Z_t to be autocorrelated. We now briefly explore the case in which the Z_t's themselves are serially correlated. This presents an alternative to the discussion of Section 2.3 of the main paper.

Suppose that Z_t can be modelled as an autoregressive process as follows.

$$U_t = Z_t - a_1 Z_{t-1} - \ldots - a_r Z_{t-r} =: a(L) Z_t,$$

where U_t is independent (not necessarily identically distributed) noise distribution acceptable to NSP in Scenarios 1, 2 or 3, and L is the lag operator. We propose the following iterative scheme which builds on the NSP procedure for independent innovations. We use the (most general) language of Scenario 3.

Clearly, if the user knew r and $(a_1, \ldots, a_r)$, they would be able to transform the regression problem (2) from the main paper into

$$\begin{aligned} a(L)Y_t &= a(L)X_{t,\cdot}\beta^{(j)} + U_t \quad \text{for} \quad t = \eta_j + 1 + r, \ldots, \eta_{j+1}, \\ a(L)Y_t &= a(L)X_{t,\cdot}\beta^{(j,t)} + U_t \quad \text{for} \quad t = \eta_j + 1, \ldots, \eta_j + r. \end{aligned} \tag{16}$$

Due to the smoothing action of the filter $a(L)$, this now only approximates a piecewise-constant parameter regression setting, as it features the short "smooth transition" sections indexed $t = \eta_j + 1, \ldots, \eta_j + r$. However, the presence of these smooth transitions does not spoil the applicability of NSP, with the intervals of significance obtained on the regression problem (16) having a similar interpretation as in the case of exactly abrupt transitions.

In practice, r or $(a_1, \ldots, a_r)$ will be unknown to the analyst. We suggest the following scheme, in which these are treated as nuisance parameters and estimated from the data, as in Fang and Siegmund (2020).

1. Similarly to Fang and Siegmund (2020), estimate r and $(a_1, \ldots, a_r)$ (to obtain, respectively, $\hat{r}$ and $\hat{a} = (\hat{a}_1, \ldots, \hat{a}_{\hat{r}})$) on a stretch of the data believed to contain no change-points.

2. Transform the regression problem using the estimated operator $\hat{a}(L)$ to obtain a problem of the form (16).

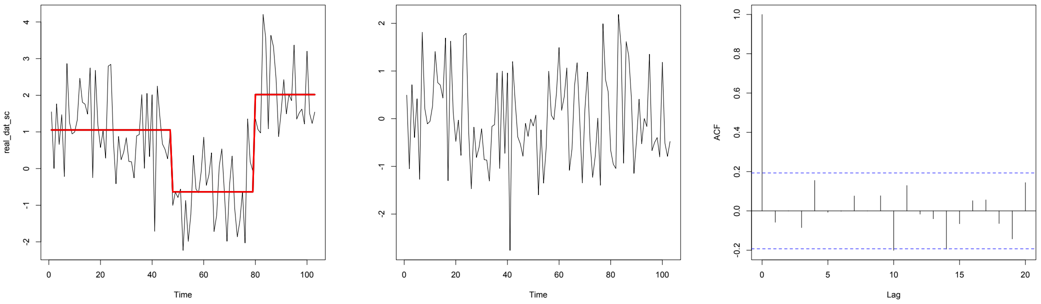

Figure 3: Left: scaled interest rate data (black) with a change-point fit obtained in R package `breakfast` (red); middle: residuals from the fit; right: their sample acf.

3. Run NSP suitable for independent innovations on the transformed problem, to obtain a set $\mathcal{S}$ of the NSP intervals of significance.
4. Re-estimate r and $(a_1, \ldots, a_r)$ on the longest stretch of data outside the NSP intervals of significance.
5. Go back to step 2. and iterate until no changes are seen in the NSP intervals of significance.

11 Additional arguments regarding the real-data analysis

In this section, we show that the application of NSP to the real-data examples of Section 6 of the main paper is justified as the errors do not exhibit significant serial correlation in the interest rate case or conditional heteroskedasticity in the price series case. Figure 3 demonstrates this for the interest rate data (note NSP was used on the scaled data shown in Figure 3, where the scaling had been performed to remove heteroscedasticity). Figure 4 shows this for the Newham house price data example (the presence of significant autocorrelation in the squared empirical residuals could have been indicative of heteroscedasticity).

12 Discussion

We conclude with a brief discussion of a few speculative aspects of NSP.

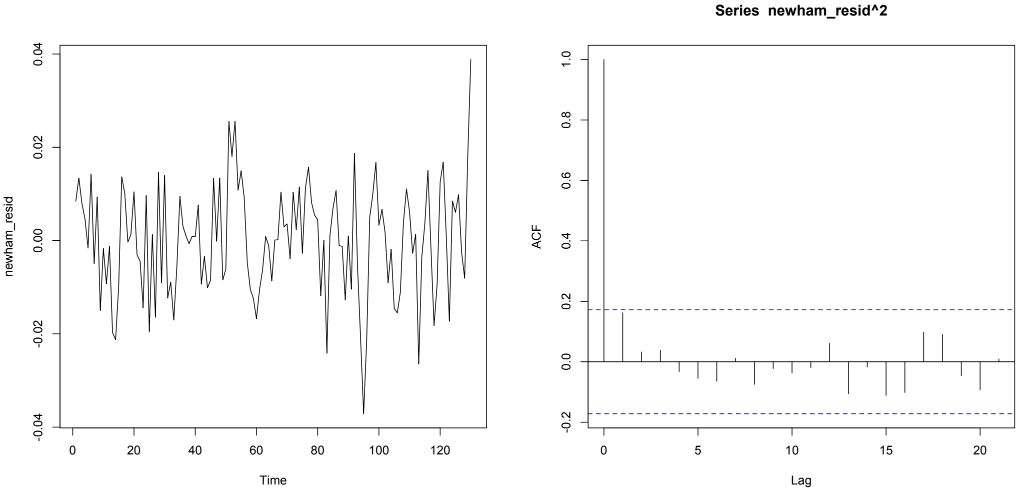

Figure 4: Left: concatenated residuals from two linear regression fits, before and after the change-point (time $t = 60$, as in the paper), in the Newham house price data; right: the sample acf of their squares.

Possible use of NSP in online monitoring for changes NSP can in principle be used in the online setting, in which ‘alarm’ should be raised as soon as Y starts deviating from linearity with respect to X. In particular, consider the following simple construction: having observed (Y_t, X_t), $t = 1, \ldots, T$, successively run NSP on the intervals $[T-1, T]$, $[T-2, T]$, $\ldots$, until either the first interval of significance is discovered, or $[1, T]$ is reached. This will provide an answer to the question of whether the most recently observed data deviates from linearity and if so, over what time interval.

Using and interpreting NSP in the presence of gradual change If NSP is used in the absence of change-points but in the presence of gradual change, obtaining a significant interval means that it must (at global significance level α) contain some of the period of gradual change. However, this does not necessarily mean that the entire period of gradual change is contained within the given interval of significance. Note that this is the situation portrayed in Section 5.2 of the main paper, in which the simulation model used is a ‘gradual change’ model from the point of view of the NSP$_0$ method, but an ‘abrupt change’ model from the point of view of NSP$_1$ and NSP$_2$.

Possible use of NSP in testing for time series stationarity It is tempting to ask whether NSP can serve as a tool in the problem of testing for second-order stationarity of a time series. In this problem, the response Y_t would be the time series in question, while the covariates X_t would be the Fourier basis. The performance of NSP in this setting will be reported in future work.

Does the principle of NSP extend to other settings? NSP is an instance of a statistical procedure which produces intervals of significance (rather than point estimators) as an output. It is an interesting open question to what extent this emphasis on "intervals of significance before point estimators" may extend to other settings, e.g. the problem of parameter inference in high-dimensional regression.

References

E. Arias-Castro, D. Donoho, and X. Huo. Near-optimal detection of geometric objects by fast multiscale methods. *IEEE Trans. Inf. Th.*, 51:2402–2425, 2005.

J. Bai and P. Perron. Estimating and testing linear models with multiple structural changes. *Econometrica*, 66:47–78, 1998.

J. Bai and P. Perron. Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18:1–22, 2003.

R. Baranowski, Y. Chen, and P. Fryzlewicz. Narrowest-Over-Threshold detection of multiple change-points and change-point-like features. *J. Roy. Stat. Soc. Ser. B*, 81:649–672, 2019.

H. P. Chan and G. Walther. Detection with the scan and the average likelihood ratio. *Statistica Sinica*, 23: 409–428, 2013.

P. L. Davies and A. Kovac. Local extremes, runs, strings and multiresolution. *Ann. Statist.*, 29:1–48, 2001.

P. L. Davies, A. Kovac, and M. Meise. Nonparametric regression, confidence regions and regularization. *Ann. Stat.*, 37:2597–2625, 2009.

H. Dette, T. Eckle, and M. Vetter. Multiscale change point detection for dependent data. *Scand. J. Statist.*, 47:1243–1274, 2020.

L. Dümbgen and V. Spokoiny. Multiscale testing of qualitative hypotheses. *Ann. Statist.*, 29:124–152, 2001.

L. Dümbgen and G. Walther. Multiscale inference about a density. *Ann. Stat.*, 36:1758–1785, 2008.

X. Fang and D. Siegmund. Detection and estimation of local signals. *Preprint*, 2020.

X. Fang, J. Li, and D. Siegmund. Segmentation and estimation of change-point models: false positive control and confidence regions. *Ann. Stat.*, 48:1615–1647, 2020.

K. Frick, A. Munk, and H. Sieling. Multiscale change-point inference (with discussion). *Journal of the Royal Statistical Society Series B*, 76:495–580, 2014.

P. Fryzlewicz. Wild Binary Segmentation for multiple change-point detection. *Ann. Stat.*, 42:2243–2281, 2014.

P. Fryzlewicz. Detecting possibly frequent change-points: Wild Binary Segmentation 2 and steepest-drop model selection. *Journal of the Korean Statistical Society*, 49:1027–1070, 2020.

X. Jeng, T. Cai, and H. Li. Optimal sparse segment identification with application in copy number variation analysis. *J. Am. Stat. Assoc.*, 105:1156–1166, 2010.

Z. Kabluchko and Y. Wang. Limiting distribution for the maximal standardized increment of a random walk. *Stoch. Proc. Appl.*, 124:2824–2867, 2014.

C. König, A. Munk, and F. Werner. Multidimensional multiscale scanning in exponential families: limit theory and statistical consequences. *Ann. Stat.*, 48:655–678, 2020.

S. Kovács, H. Li, P. Bühlmann, and A. Munk. Seeded binary segmentation: A general methodology for fast and optimal change point detection. *Biometrika*, 110:249–256, 2023.

H. Li. *Variational Estimators in Statistical Multiscale Analysis*. PhD thesis, Georg August University of Göttingen, 2016.

R. Meijer, T. Krebs, and J. Goeman. A region-based multiple testing method for hypotheses ordered in space or time. *Stat. Appl. Genet. Mol. Biol.*, 14:1–19, 2015.

N. Meinshausen. Group bound: confidence intervals for groups of variables in sparse high dimensional regression without assumptions on the design. *Journal of the Royal Statistical Society Series B*, 77: 923–945, 2015.

A. Munk, K. Proksch, H. Li, and F. Werner. Photonic imaging with statistical guarantees: From multiscale testing to multiscale estimation. In T. Salditt, A. Egner, and D. Luke, editors, *Nanoscale Photonic Imaging*, volume 134 of *Topics in Applied Physics*. Springer, 2020.

A. Račkauskas and C. Suquet. Hölder norm statistics for epidemic change. *Stat. Plan. Inf*, 126:495–520, 2004.

J. Rice. Bandwidth choice for nonparametric regression. *Ann. Statist.*, 12:1215–1230, 1984.

J. Sharpnack and E. Arias-Castro. Exact asymptotics for the scan statistic and fast alternatives. *Electronic Journal of Statistics*, 10:2641–2684, 2016.

D. Siegmund and E. S. Venkatraman. Using the generalized likelihood ratio statistic for sequential detection of a change-point. *Ann. Stat.*, 23:255–271, 1995.

G. Walther. Optimal and fast detection of spatial clusters with scan statistics. *Ann. Stat.*, 38:1010–1033, 2010.

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