Narrowest Significance Pursuit: inference for multiple change-points in linear models

Piotr Fryzlewicz

1Department of Statistics, London School of Economics, Houghton Street, London WC2A 2AE, UK

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

Funding information

We propose Narrowest Significance Pursuit (NSP), a general and flexible methodology for automatically detecting localised regions in data sequences which each must contain a change-point (understood as an abrupt change in the parameters of an underlying linear model), at a prescribed global significance level. NSP works with a wide range of distributional assumptions on the errors, and guarantees important stochastic bounds which directly yield exact desired coverage probabilities, regardless of the form or number of the regressors. In contrast to the widely studied “post-selection inference” approach, NSP paves the way for the concept of “post-inference selection”. Pre-CRAN R code is available at https://github.com/pfryz/nsp.

KEYWORDS
confidence intervals, structural breaks, post-selection inference, wild binary segmentation, narrowest-over-threshold

1 INTRODUCTION

Examining or monitoring data sequences for abrupt changes (change-points) in their behaviour is an important task in a variety of fields. Having to discriminate between significant (real) changes and random fluctuations points to the importance of statistical inference in multiple change-point problems. In this paper, we propose a new generic methodology for determining, for a given data sequence and at a given global significance level, localised regions of the data that each must contain a change-point. We define a change-point in $Y_t$ on an interval $[s, e]$ as a departure, on that interval, from a linear model for $Y_t$ with respect to pre-specified regressors. We now give examples of scenarios covered by the proposed methodology.
**Scenario 1.** Piecewise-constant signal plus noise model.

\[ Y_t = f_t + Z_t, \quad t = 1, \ldots, T, \]  

where \( f_t \) is a piecewise-constant vector with an unknown number \( N \) and locations \( 0 = \eta_0 < \eta_1 < \ldots < \eta_N < \eta_{N+1} = T \) of change-points, and \( Z_t \) is zero-centred noise; we give examples of permitted joint distributions of \( Z_t \) below. The location \( \eta_j \) is a change-point if \( f_{\eta_j-1} = f_{\eta_j} \) but \( f_{\eta_j} \neq f_{\eta_j+1} \).

**Scenario 2.** Piecewise-polynomial (e.g. piecewise-constant or piecewise-linear) signal plus noise model.

In (1), \( f_t \) is a piecewise-polynomial vector, in which the polynomial pieces have a fixed degree \( q \geq 0 \), assumed known to the analyst. The location \( \eta_j \) is a change-point if \( f_t \) can be described as a polynomial vector of degree \( q \) on \([\eta_j - q - 1, \eta_j]\), but not on \([\eta_j - q, \eta_j + 1]\).

**Scenario 3.** Linear regression with piecewise-constant parameters.

For a given design matrix \( X = (X_{t,i}), t = 1, \ldots, T, i = 1, \ldots, p \), the response \( Y_t \) follows the model

\[ Y_t = X_{t,i} \beta^{(i)} + Z_t \quad \text{for} \quad t = \eta_j + 1, \ldots, \eta_{j+1}, \]  

for \( j = 0, \ldots, N \), where the parameter vectors \( \beta^{(i)} = (\beta_1^{(i)}, \ldots, \beta_p^{(i)})' \) are such that \( \beta^{(j)} \neq \beta^{(j+1)} \).

Each of these scenarios is a generalisation of the preceding one: Scenario 3 reduces to Scenario 2 if \( p = q + 1 \) and the \( i \)th column of \( X \) is a polynomial in \( t \) of degree \( i - 1 \). We permit a broad range of distributional assumptions for \( Z_t \); we cover i.i.d. Gaussianity and other light-tailed distributions, and we use self-normalisation to also handle (not necessarily known) distributions within the Gaussian domain of attraction, including under heterogeneity. In addition, in Section 3, we introduce Scenario 4, a generalisation of Scenario 3, which provides a framework for the use of our methodology under regression with autoregression (AR). Finally, Sections 5.1 and 5.2 address the cases in which \( Z_t \)'s are not serially independent but conditionally heteroscedastic and autocorrelated, respectively. We now review the existing literature on uncertainty in multiple change-point problems which seeks to make confidence statements about the existence or locations of change-points in particular regions of the data, or significance statements about their importance.

In the i.i.d. Gaussian piecewise-constant model, SMUCE (Frick et al., 2014) estimates the number \( N \) of change-points as the minimum among those candidate fits \( f_t \) for which the empirical residuals pass a certain test at level \( \alpha \). An issue for SMUCE, discussed e.g. in Chen et al. (2014), is that the smaller the significance level \( \alpha \), the more lenient the test on the empirical residuals, and therefore the higher the risk of underestimating \( N \). This poses problems for the kinds of inferential statements in SMUCE that the authors envisage for it, because for a confidence set of an estimate of \( f_t \) to cover the truth, the authors require that the estimated number of change-points agrees with the truth, which leads to the counter-intuitive behaviour of the coverage properties of SMUCE illustrated in Chen et al. (2014). SMUCE2 (Chen et al., 2014) remedies this issue, but still requires that the number of estimated change-points agrees with the truth for successful coverage. As the accurate estimation of \( N \) is a well-known difficult problem in low signal-to-noise scenarios, SMUCE2 is also at risk of being unable to cover the truth with a high nominal probability requested by the user. In the approach taken in this paper, this issue does not arise as we shift the inferential focus away from \( N \). SMUCE is extended to heterogeneous Gaussian noise in Pein et al. (2017) and to dependent data in Dette et al. (2020).

Some authors approach uncertainty quantification for multiple change-point problems from the point of view of post-selection inference (PSI, a.k.a. selective inference); these include Hyun et al. (2018), Hyun et al. (2021), Jewell et al. (2020) and Duy et al. (2020). To ensure valid inference, PSI conditions on many aspects of the estimation process, which tends to produce \( p \)-values with somewhat complex definitions. PSI also does not permit the selection of the tuning
parameters of the inference procedure (e.g. the bandwidth \( h \) in Jewell et al. (2020)) from the same data. Notwithstanding their usefulness in assessing the significance of previously estimated change-points, these PSI approaches share the following features: (a) they do not consider uncertainties in estimating change-point locations, (b) they do not provide regions of globally significant change in the data, (c) they define significance for each change-point separately, as opposed to globally, (d) they rely on a particular base change-point detection method with its potential strengths or weaknesses, (e) they are known to be non-robust to misspecification of the error distributions (Tibshirani et al., 2018). Our approach explicitly contrasts with these features; in particular, in contrast to post-selection inference, it can be described as enabling “post-inference selection”, as we argue later on.

A number of authors provide simple consistency results for the number and locations of detected change-points, typically stating that on a set of probability tending to 1, for \( T \) large enough and under certain change-point-strength assumptions, \( N \) is estimated correctly and the true change-points must lie within certain distances of the estimated change-points. Examples in the piecewise-constant setting include Yao (1988), Boysen et al. (2009), Hao et al. (2013), Fryzlewicz (2014), Lin et al. (2017), Fryzlewicz (2018), Wang et al. (2020), Cho and Kirch (2021), and Kovács et al. (2020b). There are fewer results of this type beyond Scenario 1: examples include Baranowski et al. (2019) in the piecewise-linear model, and Wang et al. (2021) in the linear regression setting. In inferential terms, such results are usually difficult to use in practice, as the probability statements involve unknown and hard-to-estimate constants, and the significance level used usually tends to 0 with \( T \), rather than being a user-fixable constant.

Some authors go further and provide simultaneous asymptotic distributional results regarding the distance between the estimated change-point locations and the truth. In the linear regression context, this is done in Bai and Perron (1998, 2003), and in the piecewise-constant signal plus noise model – in Eichinger and Kirch (2018). In contrast to these approaches, which are asymptotic, conditional on the estimated change-point locations, and involve unknown quantities, our methodology has a finite-sample nature, makes no assumptions on the signal, and unconditionally and automatically flags up regions of data that must (at a given global significance level) contain change. A further discussion of the differences between our approach and that of Bai and Perron (1998, 2003) can be found in Section 1 of the online supplement.

Inference for multiple change-points is also sometimes posed as control of the False Discovery Rate (FDR), see e.g. Li and Munk (2016), Hao et al. (2013) and Cheng et al. (2020). However, control of FDR is too weak a criterion when one wants to obtain local regions of globally significant change, as we do in this work: FDR is focused on the number of change-points rather than on their locations, and thus permits estimators which over-estimate the number of change-points by a small fraction. This makes it impossible to guarantee that with a large global probability, all regions of significance detected by an FDR-controlled estimator contain at least one change-point each.

Bayesian approaches to uncertainty quantification in multiple change-point problems are considered e.g. in Fearnhead (2006) and Nam et al. (2012) (see also the monograph Ruanaidh and Fitzgerald (1996)), and are particularly useful when clear priors, chosen independently of the data, are available about some features of the signal.

We now summarise our new approach, then situate it in the context of the related literature, and next discuss its novel aspects. The objective of our methodology, called “Narrowest Significance Pursuit” (NSP), is to automatically detect localised regions of the data \( Y_t \), each of which must contain at least one change-point (in a suitable sense determined by the given scenario), at a prescribed global significance level. NSP performs inference without change-point location estimation, and proceeds as follows. A number \( M \) of intervals are drawn from the index domain \([1, \ldots, T]\), with start- and end-points chosen either uniformly at random, or over an equispaced deterministic grid. On each interval drawn, \( Y_t \) is then checked to see whether or not it locally conforms to the prescribed linear model, with any set of parameters. This check is performed through estimating the parameters of the given linear model locally by minimising a particular multiresolution sup-norm loss, and testing the residuals from this fit via the same norm; self-normalisation is
involved if necessary. In the first greedy stage, the shortest interval (if one exists) is chosen on which the test is violated at a certain global significance level $\alpha$. In the second greedy stage, the selected interval is searched for its shortest sub-interval on which a similar test is violated. This sub-interval is then chosen as the first region of global significance, in the sense that it must (at a global level $\alpha$) contain a change-point, or otherwise the local test would not have rejected the linear model. The procedure then recursively draws $M$ intervals to the left and to the right of the chosen region (with or without overlap), and stops when no further local regions of global significance can be found.

The following alternative viewpoint on NSP may be helpful. We start with a norm (or more generally this could be a test statistic) with the property that for an interval $[s, e]$, we have $||y_s e|| = \max_{t \in [s, e]} ||Y_t||$ for any vector $y$. We work in a change-point model generally described as $Y_t = f_t + Z_t$ (where, for example, in Scenario 3 we have $f_t = X_t, \beta^{(j)}$ for $t = \eta_j + 1, \ldots, \eta_{j+1}$). Given sets $\mathcal{F}((s, e)) \subseteq \mathbb{R}^{e-s+1}$ for each interval $[s, e]$, we wish to find a set of intervals $S$ such that with a global probability of at least $1 - \alpha$, for each $[s, e] \in S$, we have $f_s, e \notin S$. For instance, in Scenario 1, we take each $\mathcal{F}((s, e))$ to be constant vectors. To construct $S$, we may in principle take $S = \{ [s, e] : \min_{v \in \mathcal{F}([s, e])} ||Y_{s-e} - v|| > \lambda_\alpha \}$, where $P(||Z|| > \lambda_\alpha) \leq \alpha$: that this $S$ has the desired property follows in a straightforward way from its definition. However, the problems with $S$ defined in this way are that it is computationally infeasible and may contain a large number of overlapping intervals. NSP is an algorithm for extracting a meaningful subset of $S$, whose elements are suitably short, i.e. provide localisation of any change-points in $f_t$. We also explain how $\lambda_\alpha$ may be determined for different classes of noise distribution, and discuss extensions to settings in which autoregression is present.

The theme of searching for globally significant localised regions of the data containing change appears in different versions in the existing literature. This frequently involves multiscale statistics: operators of the same form applied over sub-samples of the data taken at different locations and of differing lengths. Dümbgen and Spokoiny (2001) test locally and at multiple scales for monotonicity or concavity of a curve against a general smooth alternative with an unknown degree of smoothness. Dümbgen and Walther (2008) identify regions of local increases or decreases of a density function. Walther (2010) searches for anomalous spatial clusters in the Bernoulli model using dyadically constructed blocked scan statistics. SIZer (Chaudhuri and Marron, 1999; Kim and Marron, 2006) is an exploratory multiscale data analytic tool for assessing the significance of curve features.

Fang et al. (2020), in the piecewise-constant signal plus i.i.d. Gaussian noise model, approximate the tail probability of the maximum CUSUM statistic over all sub-intervals of the data. They then propose an algorithm, in a few variants, for identifying short, non-overlapping segments of the data on which the local CUSUM exceeds the derived tail bound, and hence the segments identified must contain at least a change-point each, at a given significance level. Fang and Siegmund (2020) present results of similar nature for a Gaussian model with lag-one autocorrelation, linear trend, and features that are linear combinations of continuous, piecewise differentiable shapes. The most important high-level differences between NSP and these two approaches are that (a) 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; (b) NSP searches for any deviations from local model linearity with respect to the regressors provided; (c) the results of Fang et al. (2020) and Fang and Siegmund (2020) 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) 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. We expand on these points in Section 1 of the online supplement.

We also mention below other main distinctive features of NSP in comparison with the existing literature. NSP is specifically constructed to target the shortest possible significant intervals at every stage of the procedure, and to explore as many intervals as possible while remaining computationally efficient. This is achieved by a two-stage greedy mechanism for determining the shortest significant interval, and by using a recursive interval sampling scheme.
Importantly, NSP critically relies on what we believe is a new use of the multiresolution sup-norm. On each interval drawn, NSP locally fits the postulated linear model via multiresolution sup-norm minimisation (as opposed to e.g. the more usual OLS or MLE). It then uses the same norm to test the empirical residuals from this fit, which ensures that, under the local null, their maximum in this norm is bounded by that of the corresponding (unobserved) true residuals on that interval. This ensures the exactness of the coverage statements furnished by NSP, at a prescribed global significance level, regardless of the scenario and for any given regressors \( X \). Also, thanks to the fact that multiresolution sup-norms can be interpreted as Hölder-like norms on certain function spaces, NSP naturally extends to the cases of unknown or heterogeneous distributions of \( Z_t \), using the functional-analytic self-normalisation framework developed in Račkauskas and Suquet (2001), Račkauskas and Suquet (2003) and related papers. Finally, the use of multiresolution sup-norms means that if simulation needs to be used to determine critical values for NSP, then this can be done in a computationally efficient manner.

Section 2 introduces the NSP methodology and provides the relevant coverage theory. Section 3 extends this to NSP under self-normalisation and in the additional presence of autoregression. Section 4 provides finite-sample and traditional large-sample detection consistency and rate optimality results for NSP in Scenarios 1 and 2. Section 5 discusses extensions of NSP to some serially dependent \( Z_t \) models and constructs tighter, \( X \)-dependent bounds for use by the procedure. Section 6 provides extensive numerical examples under a variety of settings. Section 7 describes two real-data case studies. Complete R code implementing NSP is available at https://github.com/pfryz/nspl. There is an online supplement, whose contents are mentioned at appropriate places in the paper. Results whose proofs do not appear in the paper are proved in the supplement.

## 2 | THE NSP INFERENCE FRAMEWORK

Throughout the section, we use the language of Scenario 3, which includes Scenarios 1 and 2 as special cases. In particular, in Scenario 1, the matrix \( X \) in (2) is of dimensions \( T \times 1 \) and has all entries equal to 1. In Scenario 2, the matrix \( X \) is of dimensions \( T \times (q + 1) \) and its \( i \)th column is given by \( (t/T)^{\tau_L}, t = 1, \ldots, T \). Scenario 4 (for NSP in the additional presence of autoregression), which generalises Scenario 3, is dealt with in Section 3.2.

### 2.1 | Generic NSP algorithm

We start with a pseudocode definition of the NSP algorithm, in the form of a recursively defined function NSP. In its arguments, \([s, e]\) is the current interval under consideration and at the start of the procedure, we have \([s, e] = [1, T]\); \( Y \) (of length \( T \)) and \( X \) (of dimensions \( T \times p \)) are as in the model formula (2); \( M \) is the (maximum) number of sub-intervals of \([s, e]\) drawn; \( \lambda_\alpha \) is the threshold corresponding to the global significance level \( \alpha \) (typical values for \( \alpha \) would be 0.05 or 0.1) and \( \tau_L \) (respectively \( \tau_R \)) is a functional parameter used to specify the degree of overlap of the left (respectively right) child interval of \([s, e]\) with respect to the region of significance identified within \([s, e]\), if any. The no-overlap case would correspond to \( \tau_L = \tau_R = 0 \). In each recursive call on a generic interval \([s, e]\), NSP adds to the set \( S \) any globally significant local regions (intervals) of the data identified within \([s, e]\) on which \( Y \) is deemed to depart significantly (at global level \( \alpha \)) from linearity with respect to \( X \). We provide more details underneath the pseudocode below.

```plaintext
1: function NSP(s, e, Y, X, M, \lambda_\alpha, \tau_L, \tau_R)
2:  if e - s < 1 then
3:    RETURN
4:  end if
```
THE HYPOTHESIS OF LINEARITY IS REJECTED LOCALLY AT A GLOBAL LEVEL \( \alpha \). 

CARDINALITY ONE, IN WHICH CASE THE CHOICE IN LINE 18 IS TRIVIAL.)

THE CHOSEN INTERVAL IS DENOTED BY \( \text{the shortest sign interval} \).

LOCAL HYPOTHESIS OF LINEARITY OF \( \text{below and in section 3 of the online supplement.} \)

MESURES OF DEVIATION OBTAINED IN LINE 12 ARE TESTED AGAINST THRESHOLD \( \lambda_\alpha \).

ALGORITHM IS DESCRIBED IN MORE DETAIL IN SECTION 2.2.

THE EMPIRICAL RESIDUALS FROM THIS MODEL ON \( \text{permits at least} \) \( \text{of} \).

\( T, \text{to form a measure of deviation from linearity on this interval.} \)

MINIMIZE \( \text{deviation from linearity on this interval.} \)

END FUNCTION

The NSP algorithm is launched by the pair of calls: \( S := \emptyset; \text{NSP}(1, T, Y, X, M, \lambda_\alpha, \tau_L, \tau_R) \). On completion, the output of NSP is in the variable \( S \). We now comment on the NSP function line by line. In lines 2–4, execution is terminated for intervals that are too short. In lines 5–10, a check is performed to see if \( M \) is at least as large as the number of all sub-intervals of \( [s, e] \). If so, then \( M \) is adjusted accordingly, and all sub-intervals are stored in \( \{[s_m, e_m]\}^M_{m=1} \). Otherwise, a sample of \( M \) sub-intervals \( \{s_m, e_m\} \subseteq [s, e] \) is drawn in which either (a) \( s_m \) and \( e_m \) are obtained uniformly and with replacement from \( [s, e] \), or (b) \( s_m \) and \( e_m \) are all possible pairs from an (approximately) equispaced grid on \( [s, e] \) which permits at least \( M \) such sub-intervals.

In lines 11–13, each sub-interval \( [s_m, e_m] \) is checked to see to what extent the response on this sub-interval (denoted by \( Y_{s_m:e_m} \)) conforms to the linear model (2) with respect to the set of covariates on the same sub-interval (denoted by \( X_{s_m:e_m} \)). For NSP without self-normalisation, described in this section, this check is done by fitting the postulated linear model on \( [s_m, e_m] \) using a certain multiresolution sup-norm loss, and computing the same multiresolution sup-norm of the empirical residuals from this fit, to form a measure of deviation from linearity on this interval. This core step of the NSP algorithm is described in more detail in Section 2.2.

In line 14, the measures of deviation obtained in line 12 are tested against threshold \( \lambda_\alpha \), chosen to guarantee global significance level \( \alpha \). How to choose \( \lambda_\alpha \) depends (only) on the distribution of \( Z_t \); this question is addressed in Section 2.3 below and in Section 3 of the online supplement. The shortest sub-interval(s) \( [s_m, e_m] \) for which the test rejects the local hypothesis of linearity of \( Y \) versus \( X \) at global level \( \alpha \) are collected in set \( M_0 \). In lines 15–17, if \( M_0 \) is empty, then the procedure decides that it has not found regions of significant deviations from linearity on \( [s, e] \), and stops on this interval as a consequence. Otherwise, in line 18, the procedure continues by choosing the sub-interval, from among the shortest significant ones, on which the deviation from linearity has been the largest. (Empirically, \( M_0 \) often has cardinality one, in which case the choice in line 18 is trivial.) The chosen interval is denoted by \( [s_{m_0}, e_{m_0}] \).

In line 19, \( [s_{m_0}, e_{m_0}] \) is searched for its shortest significant sub-interval, i.e. the shortest sub-interval on which the hypothesis of linearity is rejected locally at a global level \( \alpha \). Such a sub-interval certainly exists, as \( [s_{m_0}, e_{m_0}] \) itself
has this property. The structure of this search again follows the workflow of the NSP procedure; more specifically, it proceeds by executing lines 2–18 of NSP, but with \(s_{m_0}, e_{m_0}\) in place of \(s, e\). The chosen interval is denoted by \([\hat{s}, \hat{e}]\). This two-stage search (identification of \([s_{m_0}, e_{m_0}\]) in the first stage and of \([\hat{s}, \hat{e}] \subseteq [s_{m_0}, e_{m_0}\]) in the second stage) is crucial in NSP’s pursuit to force the identified intervals of significance to be as short as possible, without unacceptably increasing the computational cost. The importance of this two-stage solution is illustrated in Section 4 of the online supplement. In line 20, the selected interval \([\hat{s}, \hat{e}]\) is added to the output set \(S\).

In lines 21–22, NSP is executed recursively to the left and to the right of the detected interval \([\hat{s}, \hat{e}]\). However, we optionally allow for some overlap with \([\hat{s}, \hat{e}]\). The overlap, if present, is a function of \([\hat{s}, \hat{e}]\) and, if it involves detection of the location of a change-point within \([\hat{s}, \hat{e}]\), then it is also a function of \(Y\), \(X\). An example of the relevance of this is given in Section 6.1.1.

In NSP, having \(p = p(T)\) growing with \(T\) is possible, but we must have \(p(T) + 1 \leq T\) or otherwise no regions of significance will be found. Section 2 of the online supplement comments on a few other generic aspects of the NSP algorithm. 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. Another reason why the approach of Meijer et al. (2015) does not align with the practicalities of NSP is because of the stage-two sub-search conducted by NSP in line 19, which (in the language of Meijer et al. (2015)) changes the tree structure of the problem, including its leaves, which is not permitted in Meijer et al. (2015).

### 2.2 Measuring deviation from linearity in NSP

This section completes the definition of NSP (in the version without self-normalisation) by describing the **DEVIATION-FROM-LINEARITY** function (NSP algorithm, line 12). Its basic building block is a scaled partial sum statistic, defined for an arbitrary input sequence \(\{y_t\}_{t=1}^T\) by \(U_{x,e}(y) = (e - s + 1)^{-1/2} \sum_{t=s}^{e} y_t\). 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. In this use, scaled partial sum statistics are components of a multiresolution sup-norm used to check this aspect of the empirical residuals (Frick et al., 2014; Davies and Kovac, 2001; Davies et al., 2009; Li, 2016).

We define the scan statistic of an input vector \(y\) (of length \(T\)) with respect to the interval set \(I\) as

\[
\|y\|_I = \max_{[s,e] \in I} |U_{x,e}(y)|. \tag{3}
\]

The set \(I\) used in NSP contains intervals at a range of scales and locations. For computational efficacy, instead of the set \(I^a\) of all subintervals of \([1, T]\), we use the set \(I^d\) of all intervals of dyadic lengths and arbitrary locations, that is \(I^d = \{(s, e) \subseteq [1, T] : e - s = 2^j - 1, \ j = 0, \ldots, \lfloor \log_2 T \rfloor \}\). A simple pyramid algorithm of complexity \(O(T \log T)\) is available for the computation of all \(U_{x,e}(y)\) for \([s, e] \in I^d\). We also define restrictions of \(I^a\) and \(I^d\) to arbitrary intervals \([s, e]\) as \(I^a_{[s,e]} = \{[u,v] \subseteq [s,e] : [u,v] \in I^a\}\), and analogously for \(I^d_{[s,e]}\). We refer to \(\|\cdot\|_{I^d}, \|\cdot\|_{I^a}\) and their restrictions
as multiresolution sup-norms (see Nemirovski (1986) and Li (2016)) or, alternatively, multiscale scan statistics if they are used as operations on data. If the context requires this, the qualifier “dyadic” will be added to these terms when referring to the $I^d$ versions. The facts that, for any interval $[s, e]$ and any input vector $y$ (of length $T$), we have

\[ \|y_{x,e}\|_{I^d} \leq \|y_{x,e}\|_{[s,e]} \leq \|y\|_{I^d} \quad \text{and} \quad \|y_{x,e}\|_{[s,e]} \leq \|y\|_{I^d} \leq \|y\|_{I^s} \]  

(4)

are trivial consequences of the facts that $I^d \subseteq I^s \subseteq I^d$ and $I^d \subseteq I^d \subseteq I^s$. With this notation in place, \textsc{DeviationFromLinearity}($s_m, e_m, Y, X$) is defined as follows.

**Step 1.** Find $\beta_0 = \arg \min_{\beta} \|Y_{s_m:e_m} - X_{s_m:e_m} \beta\|_{I^d}$. This fits the postulated linear model between $X$ and $Y$ restricted to the interval $[s_m, e_m]$. However, we use the multiresolution sup-norm $\|\cdot\|_{I^d}$ as the loss function, rather than the more usual $L_2$ loss. This has important consequences for the exactness of our significance statements, which we explain later below.

**Step 2.** Compute the same multiresolution sup-norm of the empirical residuals from the above fit, $D_{[s_m:e_m]} := \|Y_{s_m:e_m} - X_{s_m:e_m} \beta_0\|_{I^d}$.  

**Step 3.** Return $D_{[s_m:e_m]}$.

Steps 1 and 2 above can obviously also be carried out in a single step as $D_{[s_m:e_m]} = \min_{\beta} \|Y_{s_m:e_m} - X_{s_m:e_m} \beta\|_{I^d}$, however, for comparison with other approaches, it will be convenient for us to use the two-stage process in steps 1 and 2 for the computation of $D_{[s_m:e_m]}$. Computationally, the linear model fit in step 1 can be carried out via simple linear programming; we use the R package \texttt{pSolve}. The following important property lies at the heart of \textsc{NSP}.

**Proposition 2.1** Let the interval $[s, e]$ be such that $\forall j = 1, \ldots, N$ $[\eta_j, \eta_j + 1] \not\subseteq [s, e]$. We have $D_{[s,e]} \leq \|Z_{x,e}\|_{I^d}$.

**Proof.** As $[s, e]$ does not contain a change-point, there is a $\beta^*$ such that $Y_{x,e} = X_{x,e} \beta^* + Z_{x,e}$. Therefore, $D_{[s,e]} = \min_{\beta} \|Y_{x,e} - X_{x,e} \beta\|_{I^d} \leq \|Y_{x,e} - X_{x,e} \beta^*\|_{I^d} = \|Z_{x,e}\|_{I^d}$, which completes the proof. \hfill $\square$

This is a simple but valuable result, which can be read as follows: “under the local null hypothesis of no signal on $[s, e]$, the test statistic $D_{[s,e]}$, defined as the multiresolution sup-norm of the empirical residuals from the same multiresolution sup-norm fit of the postulated linear model on $[s, e]$, is bounded by the multiresolution sup-norm of the true residual process $Z_t$.” This bound is achieved because the same norm is used in the linear model fit and in the residual check, and it is important to note that the corresponding bound would not be available if the postulated linear model were fitted with a different loss function, e.g. via OLS. Having such a bound allows us to transfer our statistical significance calculations to the domain of the unobserved true residuals $Z_t$, which is much easier than working with the corresponding empirical residuals. It is also critical to obtaining global coverage guarantees for \textsc{NSP}, as we now show.

**Theorem 1** Let $S = \{S_1, \ldots, S_R\}$ be a set of intervals returned by the \textsc{NSP} algorithm. The following guarantee holds.

\[ P \left( \exists i = 1, \ldots, R \ \forall j = 1, \ldots, N \ [\eta_j, \eta_j + 1] \not\subseteq S_i \right) \leq P(\|Z\|_{I^d} > \lambda_\alpha) \leq P(\|Z\|_{I^s} > \lambda_\alpha). \]

**Proof.** The second inequality is implied by (4). We now prove the first inequality. On the set $\|Z\|_{I^d} \leq \lambda_\alpha$, each interval $S_i$ must contain a change-point as if it did not, then by Proposition 2.1, we would have to have

\[ D_{S_i} \leq \|Z\|_{I^d} \leq \lambda_\alpha. \]  

(5)
However, the fact that $S_i$ was returned by NSP means, by line 14 of the NSP algorithm, that $D_{S_i} > \lambda_a$, which contradicts (17). This completes the proof.

Theorem 1 should be read as follows. Let $\alpha = P(\|Z\|_{fi} > \lambda_a)$. For a set of intervals returned by NSP, we are guaranteed, with probability of at least $1 - \alpha$, that there is at least one change-point in each of these intervals. Therefore, $S = \{S_1, \ldots, S_R\}$ can be interpreted as an automatically chosen set of regions (intervals) of significance in the data. In the no-change-point case ($N = 0$), the correct reading of Theorem 1 is that the probability of obtaining one of more intervals of significance ($R \geq 1$) is bounded from above by $P(\|Z\|_{fi} > \lambda_a)$.

NSP uses a multiresolution sup-norm fit to be checked via the same multiresolution sup-norm. This leads to exact coverage guarantees for NSP with very simple mathematics. In contrast to the confidence intervals in e.g. Bai and Perron (1998), the NSP regions of significance are not conditional on any particular estimator of $N$, and are in addition of a finite-sample nature. Still, they have a "confidence interval" interpretation in the sense that each must contain at least one change, with a certain prescribed global probability.

For $S_i = [s, e]$, we define $S_i^- = [s, e - 1]$. A simple corollary of Theorem 1 is that for $S = \{S_1, \ldots, S_R\}$, if the corresponding sets $S_i^-$ are mutually disjoint (as is the case e.g. if $\tau_L = \tau_R \equiv 0$), then we must have $N \geq R$ with probability at least $1 - \alpha$. It would be impossible to obtain a similar upper bound on $N$ without order-of-magnitude assumptions on spacings between change-points and magnitudes of parameter changes. Such assumptions are typically difficult to verify, and we do not make them in this work. As a consequence, our result in Theorem 1 does not rely on asymptotics and has a finite-sample character.

$\beta_0$ in Step 1 above does not have to be an accurate estimator of the true local $\beta$ for the bound in Proposition 2.1 to hold; it holds unconditionally and for arbitrary short intervals $[s, e]$. This is in contrast to e.g. an OLS fit, in which we would have to ensure accurate estimation of the local $\beta$ (and therefore: suitably long intervals $[s, e]$) to be able to obtain similar bounds. We return to this important issue in Section 3.2 for comparison with the existing literature.

NSP is not automatically equipped with pointwise estimators of change-point locations. This is an important feature, because thanks to this, it can be so general and work in the same way for any $X$ without a change. If it were to come with meaningful pointwise change-point location estimators, they would have to be designed for each $X$ separately, e.g. using the maximum likelihood principle. (However, NSP can be paired up with such pointwise estimators; see below for details.) We now introduce a few new concepts, to contrast this feature of NSP with the existing concept of post-selection inference.

"Post-inference selection" and "inference without selection". If it can be assumed that an interval $S_i = [s, e] \in S$ only contains a single change-point, its location can be estimated e.g. via MLE performed locally on the data subsample living on $[s, e]$. Naturally, the MLE should be constructed with the specific design matrix $X$ in mind, see Baranowski et al. (2019) for examples in Scenarios 1 and 2. In this construction, "inference", i.e. the execution of NSP, occurs before "selection", i.e. the estimation of the change-point locations, hence the label of "post-inference selection". This avoids the complicated machinery of post-selection inference, as we automatically know that the $p$-value associated with the estimated change-point must be less than $\alpha$. Similarly, "inference without selection" refers to the use of NSP unaccompanied by a change-point location estimator.

"Simultaneous inference and selection" or "in-inference selection". In this construction, change-point location estimation on an interval $[\hat{s}, \hat{e}]$ occurs directly after adding it to $S$. The difference with "post-inference selection" is that this then naturally enables appropriate non-zero overlaps $\tau_L$ and $\tau_R$ in the execution of NSP. More specifically, denoting the estimated location within $[\hat{s}, \hat{e}]$ by $\hat{\eta}$, we can set, for example, $\tau_L(\hat{s}, \hat{e}, Y, X) = \hat{\eta} - \hat{s}$ and $\tau_R(\hat{s}, \hat{e}, Y, X) = \hat{e} - \hat{\eta} - 1$, so that lines 21–22 of the NSP algorithm become, respectively, $\text{NSP}(s, \hat{\eta}, Y, X, M, \lambda_a, \tau_L, \tau_R)$ and $\text{NSP}(\hat{\eta} + 1, e, Y, X, M, \lambda_a, \tau_L, \tau_R)$.
By Theorem 1, the only piece of knowledge required to obtain coverage guarantees in NSP is the distribution of $\|Z\|_{f^*}$ (or $\|Z\|_{f^*}$), regardless of the form of $X$. Much is known about this distribution for various underlying distributions of $Z$; see Section 2.3 below and Section 2 of the online supplement for $Z$ Gaussian and following other light-tailed distributions, respectively. Any future further distributional results of this type would only further enhance the applicability of NSP. However, if the distribution of $\|Z\|_{f^*}$ is unknown, then an approximation can also be obtained by simulation, which is particularly computationally efficient for $\|Z\|_{f^*}$.

## 2.3 | $Z_t \sim \text{i.i.d. } N(0, \sigma^2)$

We now recall distributional results for $\|Z\|_{f^*}$, in the case $Z_t \sim \text{i.i.d. } N(0, \sigma^2)$ with $\sigma^2$ assumed known, which will permit us to choose $\lambda_\alpha = \lambda_\alpha(T)$ so that $P(\|Z\|_{f^*} > \lambda_\alpha(T)) \to \alpha$ as $T \to \infty$. The resulting $\lambda_\alpha(T)$ can then be used in Theorem 1.

The assumption of a known $\sigma^2$ is common in the change-point inference literature, see e.g. Hyun et al. (2018), Fang and Siegmund (2020) and Jewell et al. (2020). Fundamentally, this is because in Scenarios 1 and 2, in which the covariates possess some degree of regularity across $t$, the variance parameter $\sigma^2$ is relatively easy to estimate (see Section 4.1 of Dumbgen and Spokoiny (2001), and Fang and Siegmund (2020), for overviews of the most common approaches). Fryzlewicz (2020) points out potential issues in estimating $\sigma^2$ in the presence of frequent change-points, but they are addressed in Kovács et al. (2020a). Section 3 of the online supplement covers the unknown $\sigma^2$ case; we show there a condition under which Theorem 2 remains valid with an estimated variance $\hat{\sigma}^2$, and give an estimator of $\hat{\sigma}^2$ that satisfies that condition for certain matrices $X$ and parameter vectors $\beta^T$.

Results on the distribution of $\|Z\|_{f^*}$ are given in Siegmund and Venkatraman (1995) and Kabluchko (2007). We recall the formulation from Kabluchko (2007) as it is slightly more explicit.

### Theorem 2 (Theorem 1.3 in Kabluchko (2007))

Let $\{Z_t\}_{t=1}^T$ be i.i.d. $N(0, 1)$. For every $\gamma \in \mathbb{R}$, we have

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

where

$$
a_T = \sqrt{2 \log T} + \frac{1}{2} \log \log T + \log \frac{H}{2\sqrt{\pi}}; \quad b_T = \frac{1}{\sqrt{2 \log T}}; \quad H = \int_0^\infty \exp\left(-4 \sum_{k=1}^\infty \frac{1}{k} \Phi\left(-\frac{k}{2\gamma}\right)\right) dy,
$$

and $\Phi()$ is the standard normal cdf.

We use the approximate value $H \approx 0.82$ in our numerical work. Using the asymptotic independence of the maximum and the minimum (Kabluchko and Wang, 2014), and the symmetry of $Z$, we get the following simple corollary.

$$
P\left( \max_{1 \leq s \leq T} |U_{s, e}(Z)| > a_T + b_T \gamma \right) = 1 - P\left( \max_{1 \leq s \leq T} |U_{s, e}(Z)| \leq a_T + b_T \gamma \right)
$$

$$
= 1 - P\left( \max_{1 \leq s \leq T} U_{s, e}(Z) \leq a_T + b_T \gamma \wedge \min_{1 \leq s \leq T} U_{s, e}(Z) \geq -(a_T + b_T \gamma) \right)
$$

$$
\rightarrow 1 - \exp(-2e^{-T}) \quad (6)
$$

as $T \to \infty$. In light of (6), we obtain $\lambda_\alpha$ for use in Theorem 1 as follows: (a) equate $\alpha = 1 - \exp(-2e^{-T})$ and obtain $\gamma$, (b) form $\lambda_\alpha = \sigma(a_T + b_T \gamma)$. 

3 | NSP WITH SELF-NORMALISATION AND WITH AUTOREGRESSIVE

3.1 | Self-normalised NSP for possibly heavy-tailed, heteroscedastic $Z_t$

Kabluchko and Wang (2014) point out that the square-root normalisation used in $U_{x,e}(y)$ is not natural for distributions with tails heavier than Gaussian. We are interested in obtaining a universal normalisation in $U_{x,e}(y)$ which would work across a wide range of possibly heavy-tailed distributions without requiring their explicit knowledge, including under heterogeneity. One such solution is offered by the self-normalisation framework developed in Račkauskas and Suquet (2003) and related papers. We now recall the basics and discuss the necessary adaptations to our context. We first discuss the relevant distributional results for the true residuals $Z_t$. In this paper, we only cover the case of symmetric distributions of $Z_t$. For the non-symmetric case, which requires a slightly different normalisation, see Račkauskas and Suquet (2003). In Račkauskas and Suquet (2003), the following result is proved. Let $\rho_{\theta,v,c}(\delta) = \delta^\theta \log^v(c/\delta), 0 < \theta < 1, v \in \mathbb{R}$, where $c \geq \exp(v/\theta)$ if $v > 0$ and $c > \exp(-v/(1-\theta))$ if $v < 0$. Further, suppose $\lim_{\epsilon\to0} Z^2_{\rho_{\theta,v,c}^2(2^{-j})}/j = \infty$. This last condition, in particular, is satisfied if $\theta = 1/2$ and $v > 1/2$. The function $\rho_{\theta,v,c}$ will play the role of a modulus of continuity. Let $Z_1, Z_2, \ldots$ be independent and symmetrically distributed with $\mathbb{E}(Z_1) = 0$; note they do not need to be identically distributed. Define $S_0 = Z_1 + \ldots + Z_t$ and $V_1^2 = Z_1^2 + \ldots + Z_t^2$. Assume further $V_t^2 \max_{1 \leq n \leq T} Z_n^2 \to 0$ in probability as $T \to \infty$. Egorov (1997) shows that this last condition is equivalent to $Z_t$ being within the domain of attraction of the normal law. Therefore, the material of this section applies to a much wider class of distributions than the heterogeneous extension of SMUCE in Pein et al. (2017), which only applies to normally distributed $Z_t$.

Let the random polygonal partial sums process $\xi_T$ be defined on $[0, 1]$ as linear interpolation between the knots $(V_i^2/V_j^2, S_i), i = 0, \ldots, T$, where $S_0 = V_0 = 0$, and let $\xi_T^{se} = \xi_T/V_T$. Denote by $H^{\rho_{\theta,v,c}}_{[0, 1]}$ the set of continuous functions $x : [0, 1] \to \mathbb{R}$ such that $\omega_{\rho_{\theta,v,c}}(x, 1) < \infty$, where $\omega_{\rho_{\theta,v,c}}(x, \delta) = \sup_{u,v \in [0,1]} 0 < v < \delta \{ (x(v) - x(u))/\rho_{\theta,v,c}(v - u) \}$. $H^{\rho_{\theta,v,c}}_{[0, 1]}$ is a Banach space in its natural norm $\|x\|_{\rho_{\theta,v,c}} = \|x(0)\| + \omega_{\rho_{\theta,v,c}}(x, 1)$. Define $H^{\rho_{\theta,v,c}}_{0[0, 1]}$, a closed subspace of $H^{\rho_{\theta,v,c}}_{[0, 1]}$, by $H^{\rho_{\theta,v,c}}_{0[0, 1]} = \{ x \in H^{\rho_{\theta,v,c}}_{[0, 1]} : \lim_{\delta \to 0} \omega_{\rho_{\theta,v,c}}(x, \delta) = 0 \}$. $H^{\rho_{\theta,v,c}}_{0[0, 1]}$ is a separable Banach space.

Under the assumptions above, we have the following convergence in distribution as $T \to \infty$:

$$\xi_T^{se} \to W$$

in $H^{\rho_{\theta,v,c}}_{0[0, 1]}$, where $W(u), u \in [0, 1]$ is a standard Wiener process. Define $I_{\rho_{\theta,v,c}}(x, u, v) = |x(v) - x(u)|/\rho_{\theta,v,c}(v - u)$ and, with $e > 0$ and $c = \exp(1 + 2e)$, consider the statistic

$$\sup_{0 \leq i < j \leq T} I_{\rho_{\theta,v,c}}(x, u, v) = \sup_{0 \leq i < j \leq T} |S_j - S_i|/\sqrt{V_j^2 - V_i^2 \log^{1/2\epsilon}(c/(V_j^2/V_i^2 - V_i^2/V_j^2))}$$

In the notation and under the conditions listed above, it is a direct consequence of the distributional convergence (7) in the space $H^{\rho_{\theta,v,c}}_{0[0, 1]}$ that for any level $\gamma$, we have

$$P\left( \sup_{0 \leq i < j \leq T} I_{\rho_{\theta,v,c}}(x, u, v) \geq \gamma \right) \leq P\left( \sup_{u,v \in [0,1]} I_{\rho_{\theta,v,c}}(x, u, v) \geq \gamma \right) \to P\left( \sup_{u,v \in [0,1]} I_{\rho_{\theta,v,c}}(W, u, v) \geq \gamma \right)$$

as $T \to \infty$, and the quantiles of the distribution of $\sup_{u,v \in [0,1]} I_{\rho_{\theta,v,c}}(W, u, v)$, which does not depend on the sample size $T$, can be computed (once) by simulation.
Following the narrative of Sections 2.2 and 2.3, to make these results operational in a new function \textsc{deviation-from-linearity} (SN where ‘SN’ stands for self-normalisation), to replace the function \textsc{deviation-from-linearity} in line 12 of the NSP algorithm, we need the following development. Initially, assume the global residual sum of squares \( \sum_{i,j} \) is known; we discuss later how to estimate it. For a generic interval \([s, e]\) containing no change-points, we need to be able to obtain empirical residuals \( \hat{z}_{i+j}^{(k)} \) for \( k = 1, 2 \) and \( \hat{z}_{i+j}^{(k)} \) for \( k = 3 \) for which we can guarantee that

\[
\sup_{s-1 \leq j \leq e} \frac{|\hat{z}_{i+j}^{(k)}|}{\sqrt{\hat{z}_{i+j}^{(k)} + \ldots + (\hat{z}_{j}^{(k)}) \log^{1/2}\{cV^2_f/((Z_{i+j}^{(1)})^2 + \ldots + (\hat{z}_{j}^{(1)})^2)\}}} \leq \sup_{s-1 \leq j \leq e} \frac{|Z_{i+j} + \ldots + Z_j|}{\sqrt{Z_{i+j}^2 + \ldots + Z_j^2 \log^{1/2}\{cV^2_f/((Z_{i+j}^{(1)})^2 + \ldots + Z_j^2)\}}}.
\]  

(9)

We now outline how this can be achieved.

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

- **\( k = 2 \)**: We use

\[
(\hat{z}_{i+1}^{(2)}, \ldots, \hat{z}_{i}^{(2)}) = (1 + e)(\hat{z}_{i+1}^{(1)}, \ldots, \hat{z}_{i}^{(1)}),
\]  

(10)

which guarantees \( \hat{z}_{i+1}^{(2)} \) for \( s-1 \leq j \leq e \).

In light of the distributional result (8), the relationship between the statistic \( I_{P/(1/2,2/e+c)}(W, u, v) \) and Račkauskas and Suquet (2004)’s statistic \( Ul_{P/(1/2,2/e+c)} \) as well as their Remark 5, we are able to bound \( \sup_{0 \leq s \leq T} I_{P/(1/2,2/e+c)}(\xi_{P/2, V^2_f / V^2_f, V^2_f / V^2_f}) \) by a term of order \( O(\log T) \) on a set of probability \( 1 - O(T^{-1}) \). Making the mild assumption that \( \sup_{0 \leq s \leq T} \log^{1/2}\{cV^2_f/(Z_{i+j}^2 + \ldots + Z_j^2)\} \leq \log T \), we obtain \( Z_{i+j} \geq Z_{i+j}(1 - C(j - i)^{-1}\log T) \) for some constant \( C > 0 \) which can be bounded from below by \( Z_{i+j}(1 + e)^{-1} \), uniformly over all \( j \) for which \( (j - i)^{-1}\log T \to 0 \). This justifies (10) and completes the argument.

**\( k = 3 \)**: Having obtained \( \hat{z}_{i+1}^{(1)} \) and \( \hat{z}_{i}^{(2)} \) as above, the problem of obtaining \( \hat{z}_{i}^{(3)} \) to guarantee

\[
\sup_{s-1 \leq j \leq e} \frac{|\hat{z}_{i+j}^{(3)}|}{\sqrt{\hat{z}_{i+j}^{(2)} + \ldots + (\hat{z}_{j}^{(2)}) \log^{1/2}\{cV^2_f/((Z_{i+j}^{(1)})^2 + \ldots + (\hat{z}_{j}^{(1)})^2)\}}} \leq \sup_{s-1 \leq j \leq e} \frac{|Z_{i+j} + \ldots + Z_j|}{\sqrt{Z_{i+j}^2 + \ldots + Z_j^2 \log^{1/2}\{cV^2_f/((Z_{i+j}^{(1)})^2 + \ldots + Z_j^2)\}}}.
\]  

(12)

which in turn guarantees the bound (9), is practically equivalent to the multiresolution norm minimisation solved in Step 1 of Section 2.2 except it now uses a weighted version of the norm \(|\cdot|_{f=n}^*\), where the weights are given in the
denominator of (12). This weighted problem is solved via linear programming just as easily as Step 1 of Section 2.2, the only difference being that the relevant constraints are multiplied by the corresponding weights. Further practicalities of the self-normalisation are discussed in Section 5 of the online supplement.

3.2 NSP with autoregression

To accommodate autoregression, we introduce the following additional scenario.

Scenario 4. Linear regression with autoregression, with piecewise-constant parameters.

For a given design matrix \( X = (X_t, t = 1, \ldots, T, i = 1, \ldots, p) \), the response \( Y_t \) follows the model

\[
Y_t = X_t \beta^{(j)} + \sum_{k=1}^{r} a_k^{(j)} Y_{t-k} + Z_t \quad \text{for} \quad t = \eta_j + 1, \ldots, \eta_{j+1},
\]

for \( j = 0, \ldots, N \), where the regression parameter vectors \( \beta^{(j)} = (\beta_1^{(j)}, \ldots, \beta_p^{(j)})' \) and the autoregression parameters \( a_k^{(j)} \) are such that either \( \beta^{(j)} \neq \beta^{(j+1)} \) or \( a_k^{(j)} \neq a_k^{(j+1)} \) for some \( k \) (or both types of changes occur).

In this work, we treat the autoregressive order \( r \) as fixed and known to the analyst. Change-point detection in the signal in the presence of serial correlation is a well-known hard problem in change-point analysis and many methods (see e.g. Dette et al. (2020) for an example and a literature review) rely on the accurate estimation of the long-run variance of the noise, a difficult problem. Fang and Siegmund (2020) consider \( r = 1 \) and treat the autoregressive parameter as known, but acknowledge that in practice it is estimated from the data; however, they add that "[it] would also be possible to estimate [the autoregressive parameter] from the currently studied subset of the data, but this estimator appears to be unstable". NSP circumvents this instability issue, as explained below. NSP for Scenario 4 proceeds as follows.

1. Supplement the design matrix \( X \) with the lagged versions of the variable \( Y \), or in other words substitute \( X := [X \, Y_{-1} \, \ldots \, Y_{-r}] \), where \( Y_{-k} \) denotes the respective backshift operation. Omit the first \( r \) rows of the thus-modified \( X \), and the first \( r \) elements of \( Y \).
2. Run the NSP algorithm of Section 2.1 with the new \( X \) and \( Y \) (with a suitable modification to line 12 if using the self-normalised version), with the following single difference. In lines 21 and 22, recursively call the NSP routine on the intervals \([s, \bar{s} + \tau_L(s, \bar{s}, Y, X) - r]\) and \([\bar{s} - \tau_R(s, \bar{s}, Y, X) + r, e]\), respectively. As each local regression is now supplemented with autoregression of order \( r \), we insert the extra "buffer" of size \( r \) between the detected interval \([s, \bar{s}]\) and the next children intervals to ensure that we do not process information about the same change-point in both the parent call and one of the children calls, which prevents double detection.

As the NSP algorithm for Scenario 4 proceeds in exactly the same way as for Scenario 3, the result of Theorem 1 applies to the output of NSP for Scenario 4 too. The NSP algorithm offers a new point of view on change-point analysis in the presence of autocorrelation. This is because unlike the existing approaches, most of which require the accurate estimation of the autoregressive parameters before successful change-point detection can be achieved, NSP circumvents the issue by using the same multiresolution norm in the local regression fits on each \([s, e]\), and in the subsequent tests of the local residuals. In this way, the autoregression parameters do not have to be estimated accurately for the relevant stochastic bound in Proposition 2.1 to hold; it holds unconditionally and for arbitrary short intervals \([s, e]\). Therefore unlike e.g. the method of Fang and Siegmund (2020), NSP is able to deal with autoregression,
stably, on arbitrarily short intervals.

4 DETECTION CONSISTENCY AND LENGTHS OF NSP INTERVALS

This section shows the consistency of NSP in detecting change-points, and the rates at which the lengths of the NSP intervals contract, as the sample size increases. To simplify the exposition, we consider a version of the NSP algorithm that considers all sub-intervals of \([1, T]\), and we provide results in Scenario 1 as well as in Scenario 2 with continuous piecewise-linearity (this parallels the scenarios for which consistency is shown in Baranowski et al. (2019)).

So far in the paper, we avoided introducing any assumptions on the signal: our coverage guarantees in Theorem 1 hold under no conditions on the number of change-points, their spacing, or the sizes of the breaks. This was unsurprising as they amounted to statistical size control. By contrast, the results of this section relate to detection consistency and as such, require assumptions on the minimum strength of the changes.

4.1 Scenario 1 – piecewise constancy

In this section, \(f_t\) falls under Scenario 1. We start with assumptions on the strength of the change-points. For each change-point \(\eta_j, j = 1, \ldots, N\), define

\[
\bar{d}_j = \left\lfloor \frac{16\alpha^2}{|\sigma^{\eta}_{j+1} - \sigma^\eta_j|^2} \right\rfloor + 1. \tag{14}
\]

Recalling that \(\eta_0 = 0\) and \(\eta_{N+1} = T\), we require the following assumption.

**Assumption 4.1** \(\eta_{j+1} - \eta_j \geq 2\bar{d}_{j+1} + 2\bar{d}_j - 2 \ (j = 1, \ldots, N - 1); \quad \eta_1 - \eta_0 \geq 2\bar{d}_1 - 1; \quad \eta_{N+1} - \eta_N \geq 2\bar{d}_N - 1.\)

We have the following theorem.

**Theorem 3** Let Assumption 4.1 hold, with \(\bar{d}_j\) defined in (14). On the set \(\|Z\|_2^2 \leq \lambda_a\), a version of the NSP algorithm that considers all sub-intervals, executed with no overlaps and with threshold \(\lambda_a\), returns exactly \(N\) intervals of significance \([s_1, e_1] < \ldots < [s_N, e_N]\) such that \(\eta_j \in [s_j, e_j - 1]\) and \(e_j - s_j + 1 \leq 2\bar{d}_j\), for \(j = 1, \ldots, N\).

Theorem 3 leads to the following corollary.

**Corollary 4** Let the assumptions of Theorem 3 hold, and in addition let \(Z_t \sim N(0, \sigma^2)\). Let \(\lambda_a = \sigma(1 + \Delta)\sqrt{2\log T}\) for any \(\Delta > 0\). Let \(S\) denote the set of intervals of significance \([s_1, e_1] < \ldots < [s_N, e_N]\) returned by a version of the NSP algorithm that considers all sub-intervals, executed with no overlaps and with threshold \(\lambda_a\). Let \(\mathcal{A} = \{|S| = N \land \forall j = 1, \ldots, N \ \eta_j \in [s_j, e_j - 1] \land e_j - s_j + 1 \leq 2\bar{d}_j\}\). We have \(P(\mathcal{A}) \to 1\) as \(T \to \infty\).

Corollary 4 is a traditional, large-sample consistency result for NSP. Consider first Assumption 4.1, under which it operates. With \(\lambda_a\) as in Corollary 4, Assumption 4.1 permits \(\min_j \{ |\eta_{j+1} - \eta_j|/2 \min(|\sigma^{\eta}_{j+1} - \sigma^\eta_j|, |\sigma^{\eta}_{j+1} - \sigma^{\eta}_{j+1} - \sigma^\eta_j|)\}\), a quantity that characterises the difficulty of the multiple change-point detection problem, to be of order \(O(\log^{1/2} T)\), which is the same as in Baranowski et al. (2019) and minimax-optimal as argued in Chan and Walther (2013). Further, the statement of Corollary 4 implies statistical consistency of NSP in the sense that with probability tending to one with \(T\), NSP estimates the correct number of change-points and each NSP interval contains exactly one true change-point. Moreover, the length of the NSP interval around each \(\eta_j\) is of order \(O(\log T/|\sigma^{\eta}_{j+1} - \sigma^\eta_j|^2)\), which is near-optimal
and the same as in Baranowski et al. (2019). Finally, this also implies that this consistency rate is inherited by any pointwise estimator of \( \eta_j \) that takes its value in the \( j \)th NSP interval of significance; this applies even to naive estimators constructed e.g. as the middle points of their corresponding NSP intervals \([s_j, e_j]\), i.e. \( \hat{\eta}_j = \frac{(s_j + e_j)}{2} \). Of course more refined estimators, e.g. that based on CUSUM maximisation within each NSP interval, can also be used and they will also automatically inherit the consistency and rate.

### 4.2 | Scenario 2 – continuous piecewise linearity

In this section, \( f_t \) falls under Scenario 2 and is piecewise linear and continuous. Naturally, the definition of change-point strength has to be different from Section 4.1. For each change-point \( \eta_j, j = 1, \ldots, N \), put

\[
d_j = \left| C_2 \lambda_a^{2/3} \xi_j^{-2/3} \right|, \tag{15}
\]

where \( \xi_j = |\xi_{j,1} - \xi_{j,2}|/2 \) and \( \xi_{j,1}, \xi_{j,2} \) are, respectively, the slopes of \( f_t \) immediately to the left and to the right of \( \eta_j \), and \( C_2 \) is a certain universal constant (i.e. valid for all \( f_t \)), suitably large. The following theorem holds.

**Theorem 5** Let Assumption 4.1 hold, with \( \delta_j \) defined in (15). On the set \( \|Z\|_{\mathcal{F}_T} \leq \lambda_a \), a version of the NSP algorithm that considers all sub-intervals, executed with no overlaps and with threshold \( \lambda_a \), returns exactly \( N \) intervals of significance \([s_1, e_1] < \ldots < [s_N, e_N] \) such that \( \eta_j \in [s_j, e_j - 1] \) and \( e_j - s_j + 1 \leq 2\delta_j \), for \( j = 1, \ldots, N \).

We note that Assumption 4.1 is model-independent: we require it as much in the piecewise-constant Scenario 1 as in the piecewise-linear Scenario 2 (and in any other scenario), but with \( \delta_j \) defined separately for each scenario. Theorem 5 leads to the following corollary.

**Corollary 6** Let the assumptions of Theorem 5 hold, and in addition let \( Z_t \sim N(0, \sigma^2) \). Let \( \lambda_a = \sigma(1 + \Delta)\sqrt{2\log T} \) for any \( \Delta > 0 \). Let \( S \) denote the set of intervals of significance \([s_1, e_1] < [s_2, e_2] < \ldots \) returned by a version of the NSP algorithm that considers all sub-intervals, executed with no overlaps and with threshold \( \lambda_a \). Let \( \mathcal{A} = \{|S| = N \land \forall j = 1, \ldots, N \, \eta_j \in [s_j, e_j - 1] \land e_j - s_j + 1 \leq 2\delta_j \} \). We have \( P(\mathcal{A}) \to 1 \) as \( T \to \infty \).

Corollary 6 implies that with \( \lambda_a \) as defined therein, and if \( \xi_j \sim T^{-1} \) (a case in which \( f_t \) is bounded; see Baranowski et al. (2019)), we have that the accuracy of change-point localisation via NSP (measured by \( e_j - s_j \)) is \( O(T^{2/3} \log^{1/3} T) \), the same as in Baranowski et al. (2019) and within a logarithmic factor of Raimondo (1998). Our comment (made in Section 4.1) regarding this rate being inherited by any pointwise estimator of \( \eta_j \), as long as it falls within \([s_j, e_j]\), of course equally applies in this case.

We end this section with a brief reflection on the issue of efficient estimation of the model parameters \( \beta^{(j)} \) (in the language of Scenario 3). As long as consistency results of the type established in Theorems 3 or 5 are in place, it is particularly straightforward to estimate the parameters \( \beta^{(j)} \) efficiently, as these theorems guarantee that there are no undetected change-points between each pair of neighbouring NSP intervals returned, or in other words that the true \( \beta^{(j)} \)’s are guaranteed to be constant between each pair of neighbouring NSP intervals of significance (including the regions to the left of the first NSP interval and to the right of the last one). This overcomes the issue of bias in parameter estimation in the presence of change-points, inherently present in methods which do not return intervals that are guaranteed to unconditionally cover each change-point.
5 | FURTHER EXTENSIONS AND GENERALISATIONS OF NSP

5.1 | NSP with conditionally heteroscedastic innovations

Although the results of Section 3.1 permit the use of NSP with innovations $Z_t$ that are heteroscedastic (i.e. have a time-varying variance over $t$), they cannot be applied to $Z_t$’s that are conditionally heteroscedastic, e.g. following an ARCH or GARCH model. In this section, we introduce a modification of NSP that allows its use in models with conditionally heteroscedastic $Z_t$’s. Key to the construction is the following observation.

**Proposition 5.1** Let $Z_t$ be a stochastic process satisfying

$$Z_t = |\sigma_t|\epsilon_t; \quad \sigma_t^2 = g(Z_{t-1}, Z_{t-2}, \ldots),$$

(16)

where $\epsilon_t$ are serially i.i.d. and the measurable function $g() > 0$. Define the sign function $\text{sign}(x) = 0$ if $x = 0$; $x/|x|$ otherwise. Then the stochastic process $\text{sign}(Z_t)$ is serially i.i.d.

**Proof.** Since $g() > 0$, we have $\text{sign}(Z_t) = \text{sign}(\epsilon_t)$. Therefore, $\text{sign}(Z_t)$ is (a) independent of any $\mathcal{F}_{t-1}$-measurable variable, where $\tau > 0$ and $\mathcal{F}_{t}$ is the $\sigma$-field generated by $Z_t, Z_{t-1}, \ldots$, and (b) identically distributed for all $t$, due to (a) and (b) holding for $\epsilon_t$. This completes the proof. □

To be able to use NSP in models in which $Z_t$ satisfies assumption (16) (including in particular (G)ARCH models), we define the sign version of the deviation measure as:

$$D^\text{sign}_{[s,e]} = \min_d \|\text{sign}(Y_{t,e} - X_{t,e}, \beta)\|_{\mathcal{F}_{d}}$$

where the sign function acts component-wise on the elements of its input. By analogy to Proposition 2.1, if there are no change-points in $[s, e]$, then $D^\text{sign}_{[s,e]} \leq \|\text{sign}(Z)\|_{\mathcal{F}_{d}} \leq \|\text{sign}(Z)\|_{\mathcal{F}_{s}}$. This leads to the following analogue of Theorem 1.

**Theorem 7** Let $Z_t$ satisfy (16). Let $S = \{S_1, \ldots, S_R\}$ be a set of intervals returned by the NSP algorithm which uses the deviation measure $D^\text{sign}_{[s,e]}$ in place of $D_{[s,e]}$. The following guarantee holds.

$$P(\exists i = 1, \ldots, N | j, j+1 \in S_i) \leq P(\|\text{sign}(Z)\|_{\mathcal{F}_{d}} > \lambda_\alpha) \leq P(\|\text{sign}(Z)\|_{\mathcal{F}_{s}} > \lambda_\alpha).$$

The proofs directly follows the proof of Theorem 1 and is therefore omitted. By Proposition 5.1, if $\epsilon_t$ satisfy $P(\epsilon_t > 0) = P(\epsilon_t < 0) = 1/2$, then $\text{sign}(Z_t)$ is simply a sequence of independent Rademacher variables, which makes tail bounds for $\|\text{sign}(Z)\|_{\mathcal{F}_{d}}$ particularly simple to obtain; see Kabluchko and Wang (2014) and Fryzlewicz (2021). The computation of $D^\text{sign}_{[s,e]}$ is discussed in Section 8.1 of the online supplement.

5.2 | 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. Section 8.2 of the online supplement explores the use of NSP in cases in which $Z_t$ itself is autocorrelated.

5.3 | Tightening the bounds: adjusting for the number $p$ of covariates

The theoretical thresholds used by NSP, at least in the Gaussian case, appear generous: we frequently obtain the empirical coverage of 100% when e.g. 90% is asked for. The question is whether valid lower threshold could be obtained,
and this calls for the re-examination of Proposition 2.1. Consider the following alternative version.

**Proposition 5.2** Let the interval \([s, e]\) be such that \(\forall j = 1, \ldots, N \ [\eta_j, \eta_j + 1] \not\subseteq [s, e]\). We have \(D_{[s, e]} = \min_{\beta} \|Z_{s:e} - X_{s:e} \beta\|_{J_d} \leq \min_{\beta} \|Z - X \beta\|_{J_d}\).

**Proof.** The inequality is true because for any fixed \(\beta\), the norm \(\|Z - X \beta\|_{J_d}\) is a maximum over a larger set than the maximum in \(\|Z_{s:e} - X_{s:e} \beta\|_{J_d}\). We now prove the inequality. As \([s, e]\) does not contain a change-point, there is a \(\beta^*\) such that \(Y_{s:e} = X_{s:e} \beta^* + Z_{s:e}\). We have

\[
D_{[s, e]} = \min_{\beta} \|Y_{s:e} - X_{s:e} \beta\|_{J_d} = \min_{\beta} \|X_{s:e} (\beta^* + Z_{s:e} - X_{s:e} \beta)\|_{J_d} = \min_{\beta} \|Z_{s:e} - X_{s:e} \beta\|_{J_d}.
\]

This leads to a tighter version of Theorem 1.

**Theorem 8** Let \(S = \{S_1, \ldots, S_R\}\) be a set of intervals returned by the NSP algorithm. The following guarantee holds.

\[
P(\exists i = 1, \ldots, R \quad \forall j = 1, \ldots, N \ [\eta_j, \eta_j + 1] \not\subseteq S_i) \leq P(\min_{\beta} \|Z - X \beta\|_{J_d} > \lambda_\alpha).
\]

**Proof.** On the set \(\min_{\beta} \|Z - X \beta\|_{J_d} \leq \lambda_\alpha\), each interval \(S_i\) must contain a change-point as if it did not, then by Proposition 5.2, we would have to have

\[
D_{S_i} \leq \min_{\beta} \|Z - X \beta\|_{J_d} \leq \lambda_\alpha.
\]

(17)

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 (17). This completes the proof.

The differences with Theorem 1 are as follows. In Theorem 1, the probability \(P(\exists i = 1, \ldots, R \quad \forall j = 1, \ldots, N \ [\eta_j, \eta_j + 1] \not\subseteq S_i)\) is bounded from above by \(P(\|Z\|_{J_d} > \lambda_\alpha)\), which is in turn bounded from above by \(P(\|Z\|_{J_d} > \lambda_\alpha)\). These bounds are independent of the covariates \(X\) (i.e. independent of the scenario). In the Gaussian case, to choose \(\lambda_\alpha\), we approximated the probability \(P(\|Z\|_{J_d} > \lambda_\alpha)\) using Theorem 1.3 in Kabluchko (2007) (Theorem 2 earlier).

By contrast, in Theorem 8, the probability \(P(\exists i = 1, \ldots, R \quad \forall j = 1, \ldots, N \ [\eta_j, \eta_j + 1] \not\subseteq S_i)\) is bounded from above by \(P(\min_{\beta} \|Z - X \beta\|_{J_d} > \lambda_\alpha)\). As \(\min_{\beta} \|Z - X \beta\|_{J_d} \leq \|Z - X\|_{J_d} = \|Z\|_{J_d} \leq \|Z\|_{J_d}\), the threshold \(\lambda_\alpha\) obtained by solving

\[
P(\min_{\beta} \|Z - X \beta\|_{J_d} > \lambda_\alpha) = \alpha
\]

will be lower than that obtained by solving

\[
P(\|Z\|_{J_d} > \lambda_\alpha) = \alpha
\]

(19)

(which was done in Theorem 1). In addition, unlike the solution to (19), the solution to (18) accounts for the number and form of the covariates \(X\). To solve (18), the distribution of \(\min_{\beta} \|Z - X \beta\|_{J_d}\) can be obtained by simulation, separately
6 | NUMERICAL ILLUSTRATIONS

In addition to the settings below, the online supplement contains further examples.

6.1 | Scenario 1 – piecewise constancy

6.1.1 | Low signal-to-noise example

We use the piecewise-constant blocks signal of length $T = 2048$ containing $N = 11$ change-points, defined in Fryzlewicz (2014). We add i.i.d. Gaussian noise with $\sigma = 10$, simulated with random seed set to 1. This represents a difficult setting for multiple change-point detection, with practically all state of the art multiple change-point detection methods failing to estimate all 11 change-points with high probability (Anastasiou and Fryzlewicz, 2021). Therefore, a high degree of uncertainty with regards to the existence and locations of change-points can be expected here.

The NSP procedure with the $\hat{\sigma}_{MAD}$ estimate of $\sigma$, run with the following parameters: $M = 1000$, $\alpha = 0.1$, $\tau_L = \tau_R = 0$, and with a deterministic interval sampling grid, returns 7 intervals of significance, shown in the top left plot of Figure 1. We recall that at a fixed significance level, it is not the aim of the NSP procedure to detect all change-points. The correct interpretation of the result is that we can be at least $100(1 - \alpha)\% = 90\%$ certain that each of the intervals returned by NSP covers at least one true change-point. We note that this coverage holds for this particular sample path, with exactly one true change-point being located within each interval of significance.

NSP enables the following definition of a change-point hierarchy. A hypothesised change-point contained in the detected interval of significance $[\tilde{s}_1, \tilde{e}_1]$ is considered more prominent than one contained in $[\tilde{s}_2, \tilde{e}_2]$ if $[\tilde{s}_1, \tilde{e}_1]$ is shorter than $[\tilde{s}_2, \tilde{e}_2]$. The bottom left plot of Figure 1 shows a “prominence plot” for this output of the NSP procedure, in which the lengths of the detected intervals of significance are arranged in the order from the shortest to the longest.

It is unsurprising that the intervals returned by NSP do not cover the remaining 4 change-points, as from a visual inspection, it appears that all of them are located towards the edges of data sections situated between the intervals of significance. Executing NSP without an overlap, i.e. with $\tau_L = \tau_R = 0$, means that the procedure runs, in each recursive step, wholly on data sections between (and only including the end-points of) the previously detected intervals of significance. Therefore, in light of the close-to-the-edge locations of the remaining 4 change-points within such data sections, and the low signal-to-noise ratio, any procedure would struggle to detect them there.

This shows the importance of allowing non-zero overlaps $\tau_L$ and $\tau_R$ in NSP. We try the following.

$$\tau_L(\tilde{s}, \tilde{e}) = \lfloor(\tilde{s} + \tilde{e})/2 \rfloor - \tilde{s}; \quad \tau_R(\tilde{s}, \tilde{e}) = \lfloor(\tilde{s} + \tilde{e})/2 \rfloor + 1 - \tilde{e}. \quad (20)$$

This setting means that upon detecting a generic interval of significance $[\tilde{s}, \tilde{e}]$ within $[s, e]$, the NSP algorithm continues on the left interval $[s, \lfloor(\tilde{s} + \tilde{e})/2 \rfloor]$ and the right interval $[\lfloor(\tilde{s} + \tilde{e})/2 \rfloor + 1, e]$ (recall that the no-overlap case results uses the left interval $[s, \tilde{s}]$ and the right interval $[\tilde{e}, e]$). The outcome of the NSP procedure with the overlap functions in (20) but otherwise the same parameters as earlier is shown in the top right plot of Figure 1. This version of the procedure returns 10 intervals of significance, such that (a) each interval covers at least one true change-point, and (b) they collectively cover 10 of the signal’s $N = 11$ change-points, the only exception being $\eta_3 = 307$.

We briefly remark that one of the returned intervals of significance, $[\tilde{s}, \tilde{e}] = [837, 1303]$, is much longer than the
others, but this should not surprise given that the (only) change-point it covers, \( \eta_7 = 901 \), is barely, if at all, suggested by the visual inspection of the data. The data section \( y_{837:1303} \) is shown in the bottom right plot of Figure 1.

Finally, we mention computation times for this particular example, on a standard 2015 iMac: 14 seconds (\( M = 1000 \), no overlap), 24 seconds (\( M = 1000 \), overlap as above), 1.6 seconds (\( M = 100 \), no overlap), and 2.6 seconds (\( M = 100 \), overlap as above).

6.1.2 NSP vs SMUCE: coverage comparison

For the NSP procedure, Theorem 1 promises that the probability of detecting an interval of significance which does not cover a true change-point is bounded from above by \( P(\|Z\|_{I_o} > \lambda_2) \), regardless of the value of \( M \) and of the overlap parameters \( \tau_L, \tau_R \). In this section, we set \( P(\|Z\|_{I_o} > \lambda_2) = \alpha = 0.1 \).

We now show that a similar coverage guarantee is not available in SMUCE, even if we move away from its focus on \( N \) as an inferential quality, thereby obtaining a more lenient performance test for SMUCE. In R, SMUCE is implemented in the package \texttt{stepR}, available from CRAN. For a generic data vector \( y \), the start- and end-points of the confidence intervals for the SMUCE-estimated change-point locations (at significance level \( \alpha = 0.1 \)) are available in columns 3 and 4 of the table returned by the call \texttt{jumpint(stepFit(y, alpha=0.1, confband=T))} with the exception of its final row.

In this numerical example, we consider again the \texttt{blocks} signal with \( \sigma = 10 \). For each of 100 simulated sample paths, we record a “1” for SMUCE if each interval defined above contains at least one true change-point, and a “0” otherwise. Similarly, we record a “1” for NSP if each interval \( \hat{s}_i^- \) contains at least one true change-point, where \( S = \{S_1, \ldots, S_R\} \).
is the set of intervals returned by NSP, and a "0" otherwise. As before, in NSP, we use $M = 1000, \tau_L = \tau_R = 0$, and a deterministic interval sampling grid.

With the random seed set to 1 prior to the simulation of the sample paths, the percentages of "1"'s obtained for SMUCE and NSP are: 52 and 100, respectively. While NSP (generously) keeps its promise of delivering a "1" with the probability of at least 0.9, the same cannot be said for SMUCE, for which the result of 52% makes the interpretation of its significance parameter $\alpha = 0.1$ difficult.

### 6.2 | Scenario 2 – piecewise linearity

We consider the continuous, piecewise-linear shortwave2 signal, defined as the first 450 elements of the wave2 signal from Baranowski et al. (2019), contaminated with i.i.d. Gaussian noise with $\sigma = 0.5$. The signal and a sample path are shown in Figure 2. In this model, we run the NSP procedure, with no overlaps and with the other parameters set as in Section 6.1.1, (wrongly or correctly) assuming the following, where $q$ denotes the postulated degree of the underlying piecewise polynomial: (a) $q = 0$, which wrongly assumes that the true signal is piecewise constant; (b) $q = 1$, which assumes the correct degree of the polynomial pieces making up the signal; $q = 2$, which over-specifies the degree. We denote the resulting versions of the NSP procedure by NSP$_q$ for $q = 0, 1, 2$. The intervals of significance returned by all three NSP$_q$ methods are shown in Figure 2. Theorem 1 guarantees that the NSP$_1$ intervals each cover a true change-point with probability of at least $1 - \alpha = 0.9$ and this behaviour takes place in this particular realisation. The same guarantee holds for the over-specified situation in NSP$_2$, but there is no performance guarantee for NSP$_0$.

The total length of the intervals of significance returned by NSP$_q$ for a range of $q$ can potentially be used to aid the selection of the ‘best’ $q$. To illustrate this potential use, note that the total length of the NSP$_1$ intervals of significance is much larger than that of NSP$_1$ or NSP$_2$, and therefore the piecewise-constant model would not be preferred here on the grounds that the data deviates from it over a large proportion of its domain. The total lengths of the intervals of significance for NSP$_1$ and NSP$_2$ are very similar, and hence the piecewise-linear model might (correctly) be preferred here as offering a good description of a similar portion of the data, with fewer parameters than the piecewise-quadratic model.
FIGURE 3 Left: squarewave signal with heterogeneous $t_4$ noise (black), self-normalised NSP intervals of significance (shaded red), true change-points (blue); see Section 6.3 for details. Right: time series $Q_t$ for $t = 1, \ldots, 131$. Red: the centre of the (single) NSP interval of significance. See Section 7.2 for details.

6.3 | Self-normalised NSP

We briefly illustrate the performance of the self-normalised NSP. We define the piecewise-constant squarewave signal as taking the values of 0, 10, 0, 10, each over a stretch of 200 time points. With the random seed set to 1, we contaminate it with a sequence of independent $t$-distributed random variables with 4 degrees of freedom, with the standard deviation changing linearly from $\sigma_1 = 2\sqrt{2}$ to $\sigma_{1300} = 8\sqrt{2}$. The simulated dataset, showing the “spiky” nature of the noise, is in the left plot of Figure 3.

We run the self-normalised version of NSP with the following parameters: a deterministic equispaced interval sampling grid, $M = 1000$, $\alpha = 0.1$, $\epsilon = 0.03$, no overlap; the outcome is in the left plot of Figure 3. Each true change-point is correctly contained within a (separate) NSP interval of significance, and we note that no spurious intervals get detected despite the heavy-tailed and heterogeneous character of the noise.

A typical feature of the self-normalised NSP intervals of significance, exhibited also in this example, is their relatively large width in comparison to the standard (non-self-normalised) NSP. In practice, we rarely came across a self-normalised NSP interval of significance of length below 60. This should not surprise given the fact that the self-normalised NSP is distribution-agnostic in the sense that the data transformation it uses is valid for a wide range of distributions of $Z_t$, and leads to the same limiting distribution under the null. Therefore, the relative large width of self-normalised intervals of significance arises naturally as a protection against mistaking potential heavy-tailed noise for signal. We emphasise that the user does not need to know the distribution of $Z_t$ to perform the self-normalised NSP.

7 | DATA EXAMPLES

7.1 | The US ex-post real interest rate

We re-analyse the time series of US ex-post real interest rate (the three-month treasury bill rate deflated by the CPI inflation rate) considered in García and Perron (1996) and Bai and Perron (2003). The dataset is available at http://qed.econ.queensu.ca/jae/datasets/bai001/. The dataset $Y_t$, shown in the left plot of Figure 4, is quarterly and the range is 1961:1–1986:3, so $t = 1, \ldots, T = 103$. The arguments outlined in Section 9 of the online supplement justify the applicability of NSP in this context.
We first perform a naive analysis in which we assume our Scenario 1 (piecewise-constant mean) plus i.i.d. $N(0, \sigma^2)$ innovations. This is only so we can obtain a rough segmentation which we can then use to adjust for possible heteroscedasticity of the innovations in the next stage. We estimate $\sigma^2$ via $\hat{\sigma}^2_{MAD}$ and run the NSP algorithm (with random interval sampling but having set the random seed to 1, for reproducibility) with the following parameters: $M = 1000$, $\alpha = 0.1, \tau_L = \tau_R = 0$. This returns the set $\mathcal{S}_0$ of two significant intervals: $\mathcal{S}_0 = \{[31, 62], [78, 84]\}$. We estimate the locations of the change-points within these two intervals via CUSUM fits on $Y_{31:62}$ and $Y_{78:84}$; this returns $\hat{\eta}_1 = 47$ and $\hat{\eta}_2 = 82$. The corresponding fit is in the left plot of Figure 4. We then produce an adjusted dataset, in which we divide $Y_{1:47}, Y_{48:82}, Y_{83:103}$ by the respective estimated standard deviations of these sections of the data. The adjusted dataset $\bar{Y}_t$ is shown in the right plot of Figure 4 and has a visually homoscedastic appearance. NSP run on the adjusted dataset with the same parameters (random seed 1, $M = 1000$, $\alpha = 0.1, \tau_L = \tau_R = 0$) produces the significant interval set $\mathcal{S}_0 = \{[23, 54], [76, 84]\}$. CUSUM fits on the corresponding data sections $\bar{Y}_{23:54}, \bar{Y}_{76:84}$ produce identical estimated change-point locations $\hat{\eta}_1 = 47, \hat{\eta}_2 = 82$. The fit is in the right plot of Figure 4.

We could stop here and agree with Garcia and Perron (1996), who also conclude that there are two change-points in this dataset, with locations within our detected intervals of significance. However, we note that the first interval, $[23, 54]$, is relatively long, so one question is whether it could be covering another change-point to the left of $\hat{\eta}_1 = 47$. To investigate this, we re-run NSP with the same parameters on $\bar{Y}_{1:47}$ but find no intervals of significance (not even with the lower thresholds induced by the shorter sample size $T_1 = 47$ rather than the original $T = 103$). Our lack of evidence for a third change-point contrasts with Bai and Perron (2003)’s preference for a model with three change-points.

However, the fact that the first interval of significance $[23, 54]$ is relatively long could also be pointing to model misspecification. If the change of level over the first portion of the data were gradual rather than abrupt, we could naturally expect longer intervals of significance under the misspecified piecewise-constant model. To investigate this further, we now run NSP on $\bar{Y}_1$ but in Scenario 2, initially in the piecewise-linear model ($q = 1$), which leads to one interval of significance: $\mathcal{S}_1 = \{[73, 99]\}$.

This raises the prospect of modelling the mean of $\bar{Y}_{1:73}$ as linear. We produce such a fit, in which in addition the mean of $\bar{Y}_{74:103}$ is modelled as piecewise-constant, with the change-point location $\hat{\eta}_2 = 79$ found via a CUSUM fit on $\bar{Y}_{74:103}$. As the middle section of the estimated signal between the two change-points ($\eta_1 = 73, \eta_2 = 79$) is relatively short, we also produce an alternative fit in which the mean of $\bar{Y}_{1:76}$ is modelled as linear, and the mean of $\bar{Y}_{77:103}$ as constant (the starting point for the constant part was chosen to accommodate the spike at $t = 77$). This is in the right plot of Figure 5.
and has a lower BIC value (9.28) than the piecewise-constant fit from the right plot of Figure 4 (10.57). This is because the linear+constant fit uses four parameters, whereas the piecewise-constant fit uses five.

The viability of the linear+constant model for the scaled data $\tilde{Y}_t$ is encouraging because it raises the possibility of a model for the original data $Y_t$ in which the mean of $Y_t$ evolves smoothly in the initial part of the data. We construct a simple example of such a model by fitting the best quadratic on $Y_{1:76}$ (resulting in a strictly decreasing, concave fit), followed by a constant on $Y_{77:104}$. The change-point location, 77, is the same as in the linear+constant fit for $\tilde{Y}_t$. The fit is in the left plot of Figure 5. It is interesting to see that the quadratic+constant model for $Y_t$ leads to a lower residual variance than the piecewise-constant model (4.83 to 4.94). Both models use five parameters. We conclude that more general piecewise-polynomial modelling of this dataset can be a viable alternative to the piecewise-constant modelling used in Garcia and Perron (1996) and Bai and Perron (2003). This example shows how NSP, beyond its usual role as an automatic detector of regions of significance, can also serve as a useful tool in achieving improved model selection.

### 7.2 House prices in London Borough of Newham

We consider the average monthly property price $P_t$ in the London Borough of Newham, for all property types, recorded from January 2010 to November 2020 ($T = 131$) and accessed on 1st February 2021. The data is available on [https://landregistry.data.gov.uk/](https://landregistry.data.gov.uk/). We use the logarithmic scale $Q_t = \log P_t$ and are interested in the stability of the autoregressive model $Q_t = b + a Q_{t-1} + Z_t$. Again, the arguments of Section 9 of the online supplement justify the applicability of NSP here.

NSP, run on a deterministic equispaced interval sampling grid, with $M = 1000$ and $\alpha = 0.1$, with the $\hat{\sigma}^2_{MOLS}$ estimator of the residual variance (see Section 3 of the online supplement) and both with no overlap and with an overlap as defined in formula (20), returns a single interval of significance [24, 96], which corresponds to a likely change-point location between December 2011 and December 2017. Assuming a possible change-point in the middle of this interval, i.e. in December 2014, we run two autoregressions (up to December 2014 and from January 2015 onwards) and compare the coefficients. Table 1 shows the estimated regression coefficients (with their standard errors) over the two sections. A goodness-of-fit analysis (not shown) reveals a satisfactory fit of this single-change-point model with $Z_t$ modelled as i.i.d. Gaussian.

It appears that both the intercept and the autoregressive parameter change significantly at the change-point. In
### Table 1
Parameter estimates (standard error in brackets) in the autoregressive model of Section 7.2.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(b)</td>
<td>-0.35 (0.2)</td>
<td>0.66 (0.23)</td>
</tr>
<tr>
<td>(a)</td>
<td>1.03 (0.02)</td>
<td>0.95 (0.02)</td>
</tr>
</tbody>
</table>

Particular, the change in the autoregressive parameter from 1.03 (standard error 0.02) to 0.95 (0.02) suggests a shift from a unit-root process to a stationary one. This agrees with a visual assessment of the character of the process in the right plot of Figure 3, where it appears that the process is more ‘trending’ before the change-point than it is after, where it exhibits a conceivably stationary behaviour, particularly from the middle of 2016 or so. Indeed, the average monthly change in \(Q_t\) over the time period Jan 2010 – Dec 2014 is 0.0061, larger than the corresponding average change of 0.0052 over Jan 2015 – Nov 2020.

### References


