Electronic Journal of Statistics
Vol. 13 (2019) 3254–3296
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1608

Multiscale change-point segmentation: beyond step functions

Housen Li [†,*] **Qinghai Guo**[†] **and Axel Munk**[†,‡]

University of Göttingen[†] *and Max Planck Institute for Biophysical Chemistry*[‡]
Göttingen, Germany
e-mail: hli1@uni-goettingen.de; qguo@gwdg.de; munk@math.uni-goettingen.de

Abstract: Modern multiscale type segmentation methods are known to detect multiple change-points with high statistical accuracy, while allowing for fast computation. Underpinning (minimax) estimation theory has been developed mainly for models that assume the signal as a piecewise constant function. In this paper, for a large collection of multiscale segmentation methods (including various existing procedures), such theory will be extended to certain function classes beyond step functions in a nonparametric regression setting. This extends the interpretation of such methods on the one hand and on the other hand reveals these methods as robust to deviation from piecewise constant functions. Our main finding is the adaptation over nonlinear approximation classes for a universal thresholding, which includes bounded variation functions, and (piecewise) Hölder functions of smoothness order $0 < \alpha \leq 1$ as special cases. From this we derive statistical guarantees on feature detection in terms of jumps and modes. Another key finding is that these multiscale segmentation methods perform nearly (up to a log-factor) as well as the oracle piecewise constant segmentation estimator (with known jump locations), and the best piecewise constant approximants of the (unknown) true signal. Theoretical findings are examined by various numerical simulations.

MSC 2010 subject classifications: 62G08, 62G20, 62G35.
Keywords and phrases: Change-point regression, adaptive estimation, oracle inequality, jump detection, model misspecification, multiscale inference, approximation spaces, robustness.

Received January 2019.

1. Introduction

Throughout we assume that observations are given through the regression model

$$y_i^n = \bar{f}_i^n + \xi_i^n, \qquad i = 0, \ldots, n-1, \tag{1}$$

where

$$\bar{f}_i^n = n \int_{[i/n,(i+1)/n)} f_0(x)dx\,, \tag{2}$$

and $\xi^n = (\xi_0^n, \ldots, \xi_{n-1}^n)$ are independent (not necessarily i.i.d.) centered sub-Gaussian random variables with scale parameter σ, that is,

$$\mathbb{E}\left[e^{u\xi_i^n}\right] \leq e^{u^2\sigma^2/2}, \qquad \text{for every } u \in \mathbb{R}.$$

*Corresponding author.

For simplicity, the scale parameter σ in model (1) is assumed to be known. In fact, if the noise distribution is known, say Gaussian, σ can be easily pre-estimated $\sqrt{n}$-consistently from the data via local differences (see e.g. Müller and Stadtmüller, 1987; Hall and Marron, 1990; Munk et al., 2005; Tecuapetla-Gómez and Munk, 2017). For the general sub-Gaussian case, estimation of σ is not obvious, but the missing knowledge of σ will actually not affect our asymptotic results (cf. Remark 3), as only an upper bound of σ is needed.

In this paper we are concerned with potentially discontinuous signals $f_0 : [0,1) \to \mathbb{R}$ in (2). As a minimal condition, we always assume that the underlying (unknown) signal f_0 lies in $\mathcal{D} \equiv \mathcal{D}([0,1))$, the space of càdlàg functions on $[0,1)$, which are right-continuous and have left-sided limits (cf. Billingsley, 1999, Chapter 3). In (2), we embed for simplicity the sampling points $x_{i,n} = i/n$ equidistantly in the unit interval. However, we stress that all our results can be transferred to more general domains ($\subseteq \mathbb{R}$) and sampling schemes, also for random $x_{i,n}$. For technical simplicity, we consider local averages $\bar{f}_i^n$ in model (1). The difference to point evaluation is asymptotically ignorable, since $\lim_{n\to\infty} n \int_{[x,x+1/n)} f(t)dt = f(x)$ for $x \in [0,1)$ and $f \in \mathcal{D}$. In many applications, e.g., nuclear magnetic resonance spectroscopy (Spraul et al., 1994), the local averages (a.k.a. data binning) are the typical measurements.

For the particular case that f is piecewise constant with a finite but unknown number of jumps, model (1) has been of particular interest throughout the past and is often referred to as *change-point regression model*. The related problem of estimating the number, locations and sizes of change-points (i.e. its locations of discontinuity) has a long and rich history in the statistical literature. Tukey (1961) already phrased the problem of segmenting a data sequence into constant pieces as the "regressogram problem" and it occurs in a plenitude of applications. From a risk minimization point of view it is well known that certain Bayesian estimators are (asymptotically) optimal (see e.g. Ibragimov and Has'minskiĭ (1981, Chapter VII) and Korostelev and Korosteleva (2011, Chapter 5)); however, they are not easily accessible from a computational point of view, particularly when it comes to multiple change-point recovery (Antoch and Hušková, 2000). Therefore, recent years have witnessed a renaissance in change-point inference motivated by several applications which require *computationally fast* and *statistically efficient* finding of potentially *many* change-points in large data sets, see e.g. Olshen et al. (2004), Siegmund (2013) and Behr, Holmes and Munk (2018) for its relevance to cancer genetics, Chen and Zhang (2015) for network analysis, Aue et al. (2014) for econometrics, and Hotz et al. (2013) for electrophysiology, to name a few. A major challenge for statistical methodology is the multiscale nature of these problems (i.e. change-points occur at different e.g. temporal scales and their number can be potentially large) and the large number of data points (a few millions or more), requiring computationally efficient methods.

Such computationally efficient segmentation methods which provide at the same hand certain statistical guarantees for their findings are either based on dynamic programming (Boysen et al., 2009; Killick, Fearnhead and Eckley, 2012;

Frick, Munk and Sieling, 2014; Du, Kao and Kou, 2016; Li, Munk and Sieling, 2016; Maidstone et al., 2016; Haynes, Eckley and Fearnhead, 2017), local search (Scott and Knott, 1974; Olshen et al., 2004; Fryzlewicz, 2014; Fang, Li and Siegmund, 2019) or convex optimization (Harchaoui and Lévy-Leduc, 2008; Tibshirani and Wang, 2008; Harchaoui and Lévy-Leduc, 2010) resulting from a convex ℓ_1 relaxation of the combinatorial ℓ_0 search problem of the best fitting change-points.

Typically, the statistical justification for all the aforementioned methods is given for models which assume that the underlying truth is a piecewise constant function with a fixed but unknown number of changes. For extensions to increasing number of changes of the truth (as the number of observations increases), see e.g. Zhang and Siegmund (2012), Frick, Munk and Sieling (2014) and Fryzlewicz (2014), or Cai, Jeng and Li (2012) under an additional sparsity assumption. However, in general, nothing is known for such segmentation methods in the general nonparametric regression setting as in (1) when f is not a piecewise constant function, e.g. a smooth function. Notable exceptions are the jump-penalized least square estimator in Boysen et al. (2009), and the unbalanced Haar wavelets based estimator in Fryzlewicz (2007), for which the L^2-risk has been analyzed for functions which can be sufficiently fast approximated by piecewise constant functions (in our notation this corresponds to functions in the space $\mathcal{A}_2^\gamma$, see section 3.2 for the definition).

Intending to fill such a gap, we provide a comprehensive risk analysis for a range of multiscale change-point methods when f in (1) is not a change-point function a priori. To this end, we introduce in a first step a general class of *multiscale change-point segmentation (MCPS) methods*, with scales specified by general c-normal systems (adopted from Nemirovski (1985), see Definition 1), unifying several previous methods. This includes particularly the simultaneous multiscale change-point estimator (SMUCE) by Frick, Munk and Sieling (2014) which minimizes the number of change-points under a side constraint that is based on a simultaneous multiple testing procedure on all scales (length of subsequent observations). Further examples are estimators which are built on different multiscale systems (Walther, 2010), or multiscale type penalties (Li, Munk and Sieling, 2016). These methods can be viewed also as a natural multiscale extension of certain jump penalized estimators via convex duality (see Boysen et al., 2009; Killick, Fearnhead and Eckley, 2012). Implemented by accelerated dynamic programming algorithms, these methods often have a runtime $O(n \log n)$, and are found empirically promising in various applications (see e.g. Hotz et al., 2013; Futschik et al., 2014; Behr and Munk, 2017; Killick, Fearnhead and Eckley, 2012). In case that f in model (1) is a step function, the statistical theory for these methods is well-understood meanwhile. For example, minimax optimality of estimating the change-point locations and sizes has been shown, which is based on exponential deviation bounds on the number, and the locations of change-points. Furthermore, these methods also obey optimal minimax detection properties (in the sense of testing) of vanishing signals, and provide simultaneous confidence statements for all unknown quantities (see Frick, Munk and Sieling, 2014; Li, Munk and Sieling, 2016; Pein, Sieling and Munk, 2017).

To complement the understanding of such methods, this work aims to analyze their behavior when the true regression function f_0 is beyond a piecewise constant function. To this end, we derive convergence rates for sequences of piecewise constant functions with increasing number of changes (Theorem 1), and for functions in certain approximation spaces (Theorem 2), well-known in approximation theory, cf. DeVore and Lorentz (1993, Chapter 12), (see Section 3). Further, we generalize the above mentioned results for quadratic risk to general L^p-risk ($0 < p < \infty$). As a consequence, we obtain the minimax optimal rates $n^{-2/3\cdot\min\{1/2,1/p\}}$ and $n^{-2\alpha/(2\alpha+1)}$ (up to a log-factor) in terms of L^p-loss both almost surely and in expectation for the cases that f has bounded variation ($0 < p < \infty$) (see Mammen and van de Geer, 1997), and that f is (piecewise) Hölder continuous of order $0 < \alpha \leq 1$ ($0 < p \leq 2$), respectively. Most importantly, the discussed MCPS methods are universal (i.e. independent of the smoothness assumption of the unknown truth signal), as the only tuning parameter η (which serves as a threshold, see Section 2 for further details) can be chosen as $\eta \asymp \sqrt{\log n}$. We will show that for this choice, these methods automatically *adapt* to the unknown "smoothness" of the underlying function in an asymptotically optimal way, no matter whether it is piecewise constant or it lies in the aforementioned function spaces. As an illustration, we present the performance of SMUCE (Frick, Munk and Sieling, 2014) with universal parameter choice $\eta = 0.42\sqrt{\log n}$, on different signals in Figure 1. It clearly shows that SMUCE, although designed to provide a piecewise constant solution, successfully recovers the shape of all underlying signals no matter whether they are locally constant or not, as suggested by our theoretical findings.

Further, the developed theory allows us to derive statistical guarantees for feature detection, see Section 4. More precisely, we show for general (incl. piecewise constant) signals in approximation spaces that the discussed methods recover at least as many jumps and modes (or troughs) as the truth, as the sample size tends to infinity (Theorem 3); This statement should be interpreted with the built-in parsimony (i.e., minimization of number of jumps) of these methods, which suggests that the number of artificial jumps and modes (or troughs) is "minimal". At the same hand, large increases (or decreases) of the discussed estimators imply increases (or decreases) of the true signal with high confidence (Theorem 4). In Figure 2, based on our theoretical finding, one can claim, for example, that the two *large* jumps (marked by solid vertical lines) are significant with confidence at least 90% (see Remark 8). In the particular case of step signals, we further show the consistency in estimating the number of jumps, and an error bound of the best known order (in terms of sample sizes) on the estimation accuracy of change-point locations (Proposition 1).

Finally, we address the issue how to benchmark properly the investigated methods. We show that the MCPS methods with a universal threshold perform nearly no worse than piecewise constant segmentation estimators whose change-point locations are provided by an oracle. By considering such oracles, we discover a *saturation* phenomenon (Theorem 5 and Example 2) for the class of all piecewise constant segmentation estimators: only the suboptimal rate $n^{-2/3}$

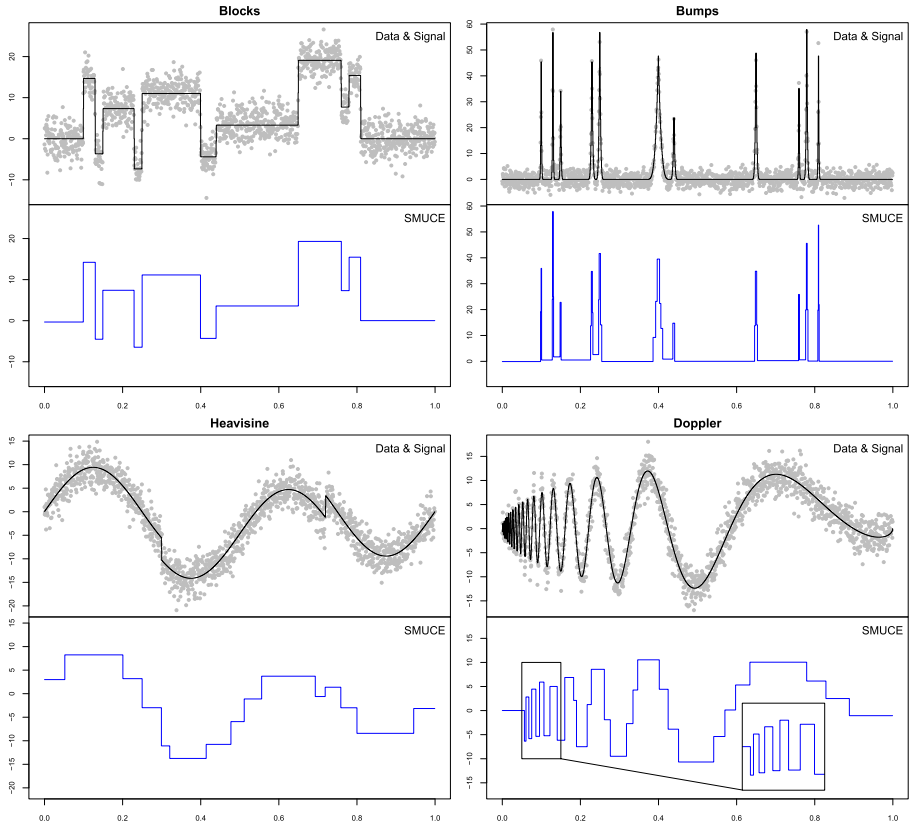


FIG 1. *Estimation by the multiscale change-point segmentation method SMUCE (Frick, Munk and Sieling, 2014) for Blocks, Bumps, Heavisine, and Doppler signals (Donoho and Johnstone, 1994) with sample size* $n = 1{,}500$, *and signal-to-noise ratio* $\|f\|_{L^2}/\sigma = 3.5$.

is attainable for smoother functions in Hölder classes with smoothness order $\alpha > 1$. From a slightly different perspective, we show that the MCPS methods perform nearly as well as the best (deterministic) piecewise constant approximant of the true signal with the same number of jumps or less (Proposition 2).

Besides such theoretical interest (cf. also Linton and Seo, 2014; Farcomeni, 2014), the study of these estimators in models beyond piecewise constant functions is also of particular practical importance, since a piecewise constant function is actually known to be only an approximation of the underlying signal in many applications. For example, in DNA copy number analysis, for which the change-point regression model with locally constant signal is commonly assumed (see e.g. Olshen et al., 2004; Lai et al., 2005), a periodic trend distortion with small amplitude (known as genomic waves) is well known to be present (Diskin et al., 2008). Thus our work can be also regarded as examination of the robustness of such segmentation methods against model misspecification.

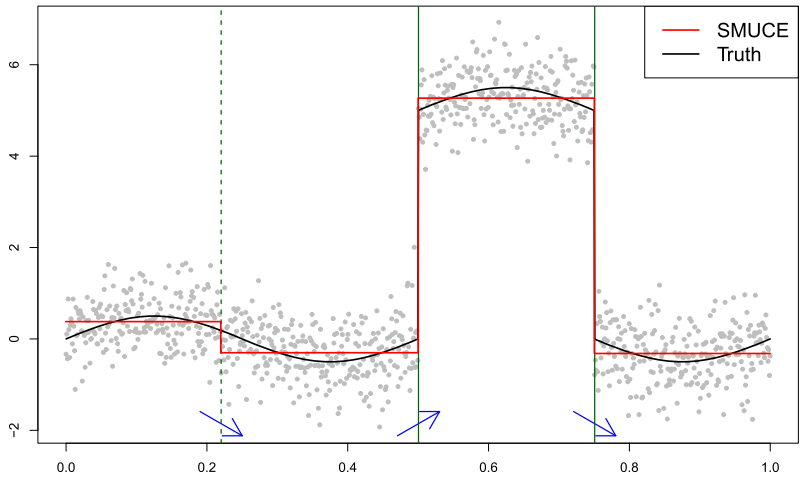


FIG 2. *Feature detection by SMUCE with threshold* $\eta(0.1)$ *by* (7) *(sample size* $n = 1,000$*, SNR* $= 5$*). The solid vertical lines mark significant jumps, while the dashed one marks an insignificant jump; and the arrows at the bottom indicate significant increases and decreases; with simultaneous confidence at least* 90%*. See Remark 8 in Section 4 for details.*

We consider a piecewise constant estimator as robust, if it recovers the majority of interesting features of the underlying true regression function with as small number of jumps as possible. For instance, Figure 3 shows the performance of SMUCE on a typical signal from DNA copy number analysis, where a locally constant function is slightly perturbed, in cases of different noise levels. Visually, SMUCE seems to recover the major features, and the recovered signal provides a simple yet informative summary of the data, meanwhile staying close to the true signal, which confirms our theoretical findings. We note that our viewpoint here complements a recent work by Song, Banerjee and Kosorok (2016) who considered a *reverse* scenario: a sequence of smooth functions approaches a step function in the limit.

In summary, we show that a large class of multiscale change-point segmentation methods with a universal parameter choice are *adaptively minimax optimal* (up a log-factor) for step signals (possibly with unbounded number of change-points) and for (piecewise) smooth signals in certain approximation spaces (Theorems 1 and 2) with respect to general L^p-risk. Building on this, we obtain statistical guarantees on feature detection, such as recovery of the number of discontinuities, or modes (Proposition 1 and Theorems 3 and 4), which explain well-known empirical findings. Moreover, in the particular case of L^2 distance, we show oracle inequalities for such multiscale change-point segmentation methods in terms of both segmentation and approximation of the true signal (Theorem 5 and Proposition 2).

The paper is organized as follows. In Section 2, we introduce a general class of multiscale change-point segmentation methods, discuss examples and pro-

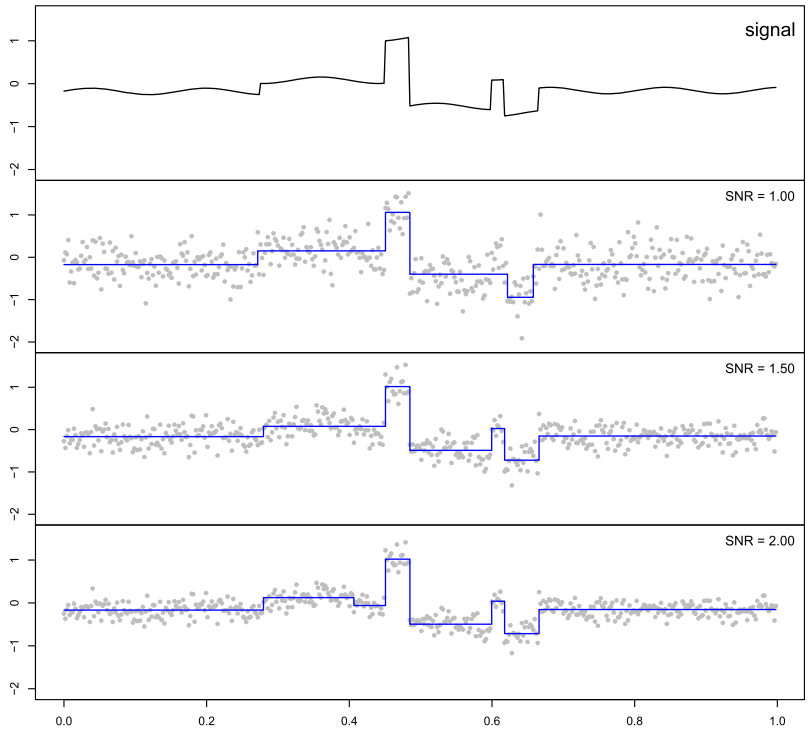


FIG 3. *Estimation by SMUCE with threshold* $\eta(0.1)$ *for the signal in Olshen et al. (2004) and Zhang and Siegmund (2007) with various signal-to-noise ratios* $\|f\|_{L^2}/\sigma$, *cf. Section 6.*

vide technical assumptions. We derive uniform bounds on the L^p-loss over step functions with possibly increasing number of change-points and over certain approximation spaces in Section 3, and present their implication on feature detection in Section 4. Section 5 focuses on the oracle properties of multiscale change-point segmentation methods from a segmentation and an approximation perspective, respectively. Our theoretical findings are investigated for finite samples by a simulation study in Section 6. The paper ends with a conclusion in Section 7. Technical proofs are collected in the appendix.

2. Multiscale change-point segmentation

To ease presentation, we introduce some notation. For $x, y \in \mathbb{R}$, let

$$x \wedge y := \min\{x, y\}, \quad x \vee y := \max\{x, y\} \quad \text{and } (x)_+ := x \vee 0.$$

Recall model (1) and let f now in $\mathcal{S} \equiv \mathcal{S}([0,1))$, the space of right-continuous step functions f on $[0,1)$ with a finite (but possibly unbounded) number of

jumps, that is, for some $k \in \mathbb{N}$

$$f = \sum_{i=0}^{k} c_i \mathbf{1}_{[\tau_i, \tau_{i+1})} \quad \text{with } 0 = \tau_0 < \ldots < \tau_{k+1} = 1, \text{ and } c_i \neq c_{i+1} \text{ for all } i. \tag{3}$$

Here $J(f) \coloneqq \{\tau_1, \ldots, \tau_k\}$ denotes the set of change-points of f. By *intervals* we always refer to those of the form $[a, b)$, $0 \leq a < b \leq 1$. In the following we introduce a general class of multiscale change-point estimators comprising various methods recently developed. To this end, we fix a system $\mathcal{I}$ of subintervals of $[0, 1)$ in the first step (cf. Definition 1). Given $\mathcal{I}$, we introduce a general class of *multiscale change-point segmentation* (MCPS) estimators $\hat{f}_n$ (see Frick, Munk and Sieling, 2014; Li, Munk and Sieling, 2016; Pein, Sieling and Munk, 2017) as a solution to the (nonconvex) optimization problem

$$\min_{f \in \mathcal{S}} \#J(f) \qquad \text{subject to } T_{\mathcal{I}}(y^n; f) \leq \eta. \tag{4}$$

Here $y^n \coloneqq \{y_i^n\}_{i=0}^{n-1}$ is the observational vector, and $\eta \in \mathbb{R}$ is a threshold to be defined later. As a convention, we consider in (4) only those candidate functions f whose change-points lie on the grid $\{i/n\}_{i=0}^{n}$. The side constraint in (4) is defined by a multiscale test statistic

$$T_{\mathcal{I}}(y^n; f) \coloneqq \max_{\substack{I \in \mathcal{I} \\ f \equiv c_I \text{ on } I}} \left\{ \frac{1}{\sqrt{n|I|}} \Big| \sum_{i/n \in I} (y_i^n - c_I) \Big| - s_I \right\}, \tag{5}$$

with $s_I \in \mathbb{R}$ a scale penalty, which can be deterministic or random, and might even depend on the candidate f and the data y^n. The maximum in (4) is taken over all intervals $I \in \mathcal{I}$, on which the candidate function f is constant. We will assume that $\eta + s_I \geq 0$ for all $I \in \mathcal{I}$, see Definition 2. In this case, the constraint in (4) is not empty, as it contains $f = \sum_{i=0}^{n} y_i^n \mathbf{1}_{[i/n,(i+1)/n)}$. Thus, the solution to the optimization problem (4) always exists but might be *non-unique*, in which case one could pick an arbitrary solution.

The side constraint in (4) originates from testing simultaneously the residuals of a candidate f with values c_I on the multiscale system $\mathcal{I}$. In model (1) under a Gaussian error, this combines all the local likelihood ratio tests whether the local mean of f_0 on I equals to a given c_I for every $I \in \mathcal{I}$. Hence, this provides a criterion for testing the constancy of f_0 on each of its segments in $\mathcal{I}$ (for a detailed account see Frick, Munk and Sieling, 2014). The choice of the scale penalties s_I determines the estimator. It balances the detection power over different scales, see Dümbgen and Spokoiny (2001), Walther (2010) and Frick, Munk and Sieling (2014) for several choices, and Davies, Höhenrieder and Krämer (2012) for the unpenalized estimators (i.e. $s_I \equiv 0$) in a slightly different model. Thus, any MCPS method amounts to search for the most parsimonious candidate over the acceptance region of the multiple tests on the right hand side in (4) performed over the system $\mathcal{I}$. The threshold η in (4) provides a trade-off between data-fit and parsimony, and can be chosen such that the truth f_0 satisfies the side constraint with a pre-specified probability $1 - \beta$. To this end, $\eta \equiv \eta(\beta)$ is chosen as

the $(1-\beta)$-quantile of the distribution of $T_{\mathcal{I}}(\xi^n;0)$, which can be determined by Monte-Carlo simulations or asymptotic considerations (Frick, Munk and Sieling, 2014; Pein, Sieling and Munk, 2017). Then the choice of significance level β provides an upper bound on the family-wise error rate of the aforementioned multiple test. It immediately provides for $\hat{f}_n$ a control of overestimating the number of jumps $\#J(f_0)$ of f_0, i.e.

$$\mathbb{P}\left\{\#J(\hat{f}_n) \leq \#J(f_0)\right\} \geq 1-\beta \qquad \text{uniformly over all } f_0 \in \mathcal{S}.$$

Also, with a different penalty, it is possible to control instead the false discovery rate by means of *local* quantiles, see Li, Munk and Sieling (2016) for details. We will see that for the asymptotic analysis of all these estimators it is sufficient to work with a universal threshold $\eta \asymp \sqrt{\log n}$ in (4) (see Definition 2 and Section 3).

The system $\mathcal{I}$ will be required to be truly *multiscale*, i.e. the MCPS methods in (4) require the associated interval system $\mathcal{I}$ to contain different scales, the richness of which can be characterized by the concept of *normality*.

Definition 1 (Nemirovski (1985))**.** A system $\mathcal{I} \equiv \mathcal{I}_n$ of intervals is called *normal* (or *c-normal*) for some constant $c > 1$, provided that it satisfies the following requirements.

(i) For every interval $I \subseteq [0,1)$ with length $|I| > c/n$, there is an interval $\tilde{I}$ in $\mathcal{I}$ such that $\tilde{I} \subseteq I$ and $|\tilde{I}| \geq c^{-1}|I|$.
(ii) The end-points of each interval in $\mathcal{I}$ lie on the grid $\{i/n : i = 0,\dots,n-1\}$.
(iii) The system $\mathcal{I}$ contains all intervals $[i/n,(i+1)/n)$, $i = 0,\dots,n-1$.

Remark 1 (Normal systems)**.** The requirement (i) in the above definition is crucial, while (ii) and (iii) are of technical nature due to the discrete sampling locations $\{i/n\}_{i=0}^{n-1}$ and can be generalized. Examples of normal systems include the highly redundant system $\mathcal{I}^0$ of all intervals whose end-points lie on the grid $\{i/n\}_{i=0}^{n-1}$ (suggested by e.g. Siegmund and Yakir, 2000; Dümbgen and Spokoiny, 2001; Frick, Munk and Sieling, 2014) of order $O(n^2)$, and less redundant but still asymptotically efficient systems (Davies and Kovac, 2001; Walther, 2010; Rivera and Walther, 2013), typically of order $O(n\log n)$. Remarkably, there are even normal systems with cardinality of order $O(n)$, such as the *dyadic partition system*

$$\left\{\left[\frac{i}{n}\lceil 2^{-j}n\rceil, \frac{i+1}{n}\lceil 2^{-j}n\rceil\right) : i = 0,\dots,2^j-1,\ j = 0,\dots,\lfloor \log_2 n\rfloor\right\},$$

which can be shown to be 2-normal, see Grasmair, Li and Munk (2018). We further stress that the choice of $\mathcal{I}$ in general poses no restriction on the change-point locations of solutions to (4), which is in sharp contrast to the wavelet thresholding approaches (e.g. Abramovich, Antoniadis and Pensky, 2007) and the local/reverse segmentation approach by Chan and Chen (2017).

Definition 2 (Multiscale change-point segmentation estimator)**.** Any estimator satisfying (4) is denoted as a *multiscale change-point segmentation (MCPS) estimator*, if

(i) the interval system $\mathcal{I}$ is c-normal for some constant $c > 1$;
(ii) the scale penalties s_I satisfy almost surely that

$$\max_{I\in\mathcal{I}}|s_I| \le \delta\sqrt{\log n} \qquad \text{for some constant } \delta > 0.$$

(iii) the threshold η is chosen as

$$\eta = a\sqrt{\log n} \qquad \text{with } a > \delta + \sigma\sqrt{2r_0+4}, \tag{6}$$

for some fixed $r_0 \in (0,\infty)$.

Remark 2 (Scale penalization)**.** For sub-Gaussian error ξ^n

$$\max_{I\in\mathcal{I}} \frac{1}{\sqrt{n|I|}}\Big|\sum_{i/n\in I}\xi_i^n\Big|$$

is at most of order $\sqrt{\log n}$ (see e.g. Shao, 1995, Theorem 1), so Definition 2 (ii) is quite natural. In particular, Definition 2 (ii) includes many common scale penalties. For instance, SMUCE (Frick, Munk and Sieling, 2014) and FDRSeg (Li, Munk and Sieling, 2016) are special cases. More precisely, for SMUCE, it amounts to select $\mathcal{I} = \mathcal{I}^0$, the system of all possible intervals, and $s_I = \sqrt{2\log(e/|I|)}$, and for FDRSeg, the same system $\mathcal{I} = \mathcal{I}^0$ but a different scale penalty $s_I = \sqrt{2\log(e|I_0|/|I|)}$ with I_0 being the constant segment, which contains I, of the candidate solution. The case $s_I \equiv 0$ is also included and has been suggested by Davies, Höhenrieder and Krämer (2012).

Remark 3 (Choice of threshold η)**.** Note that the choice of the only parameter η in Definition 2 (iii) is completely independent of the unknown truth f_0, while it depends on the distribution of the noise ξ^n via the scale parameter σ. In fact, it is also possible to choose η independent of ξ^n by $\eta = b_n\sqrt{\log n}$ with $b_n \to \infty$ arbitrarily slow (e.g. $b_n = \log\log n$); This will lead to a factor b_n in front of the convergence rates in later sections.

Alternatively, as mentioned earlier, we recommend to choose

$$\eta = \eta(\beta) \quad \text{the } (1-\beta)\text{-quantile of } T_{\mathcal{I}}(\xi^n;0), \tag{7}$$

with $\beta = \beta_n \in (0,1)$, such that

$$\mathbb{P}\{T_{\mathcal{I}}(\xi^n;0) > \eta(\beta) - O(1)\} \le \mathcal{O}(n^{-r_0}). \tag{8}$$

By Shao's theorem (Shao, 1995, Theorem 1), it then holds that

$$\eta(\beta) \le (\delta+\sigma\sqrt{2})\sqrt{\log n} + O(1) \le (\delta+2\sigma)\sqrt{\log n} \tag{9}$$

for large enough n. To ease presentation, we will state and prove our theoretical results for η given in (6), and emphasize that all the results also hold for η given in (7), the proofs of which are essentially the same (thus omitted) but relying on (8) and (9) instead. In addition, we note that a more refined analysis of $\eta(\beta)$

is even possible, although not necessary for our purposes. For instance, in case of no scale penalization, standard Gaussian noise and $\mathcal{I} = \mathcal{I}^0$ consisting of all intervals, it follows from Kabluchko (2007, Theorem 1.3) that

$$\eta(\beta) \sim \sqrt{2\log n} + \frac{\log\log n + \log\frac{\lambda}{4\pi} - 2\log\log(1/\beta)}{2\sqrt{2\log n}} \qquad \text{as } n \to \infty,$$

with constant $\lambda \in (0, \infty)$, see also Siegmund and Venkatraman (1995, Proposition 1) for approximation of $\eta(\beta)$ for finite sample sizes.

3. Asymptotic error analysis

This section mainly provides convergence rates of the MCPS methods for the model (1).

3.1. Convergence rates for step functions

We consider first locally constant change-point regression, i.e. the underlying signal $f_0 \in \mathcal{S}$ in model (1). We introduce the class of uniformly bounded piecewise constant functions, cf. (3), with up to k jumps

$$\mathcal{S}_L(k) := \Big\{ f \in \mathcal{S} : \#J(f) \le k, \text{ and } \|f\|_{L^\infty} \le L \Big\},$$

for $k \in \mathbb{N}_0$ and $L \in (0, \infty)$. If the number of change-points is bounded, i.e. k is known beforehand, the estimation problem is, roughly speaking, parametric, by interpreting change-point locations and function values as parameters. A rather complete analysis of this situation is provided either from a Bayesian viewpoint (see e.g. Ibragimov and Has'minskiĭ, 1981; Hušková and Antoch, 2003, Chapter VII) or from a likelihood viewpoint (see e.g. Yao and Au (1989); Braun, Braun and Mueller (2000); Siegmund and Yakir (2000); Boysen et al. (2009) and Korostelev and Korosteleva (2011, Chapter 5)). However, in order to understand the increasing difficulty of change-point estimation as the number of change-points gets larger, i.e. the nonparametric nature of change-point regression, we allow now the number of change-points to increase as the number of observations tends to infinity.

Theorem 1 (Adaptation I)**.** *Assume model* (1)*. Let* $0 < p < \infty$*, and* $k_n \in \mathbb{N}_0$ *be such that* $k_n = o(n)$ *as* $n \to \infty$*. Then:*

(i) *For any multiscale change-point segmentation estimator* $\hat{f}_n$ *in Definition 2 with some* $r_0 \in (0, \infty)$*, the following upper bound holds for each* $r \in (0, r_0]$

$$\limsup_{n\to\infty} \frac{1}{\sqrt{\log n}} \Big(\frac{n}{k_n + 1}\Big)^{1/2 \wedge 1/p} \sup_{f_0 \in \mathcal{S}_L(k_n)} \mathbb{E}\left[\|\hat{f}_n - f_0\|_{L^p}^r\right]^{1/r} < \infty.$$

The same result also holds almost surely if we drop the expectation $\mathbb{E}[\cdot]$*.*

(ii) *If noise* ξ_i^n *in model* (1) *has a density* $\varphi_{i,n}$ *such that for some constants* σ_0 *and* z_0

$$\max_{i,n} \int \varphi_{i,n}(x) \log \frac{\varphi_{i,n}(x)}{\varphi_{i,n}(x+z)} dx \le \frac{z^2}{\sigma_0^2} \qquad \text{for } |z| \le z_0 \tag{10}$$

then the following lower bound holds for each $r \in (0, \infty)$

$$\liminf_{n\to\infty} \Big(\frac{n}{k_n+1} \Big)^{1/2 \wedge 1/p} \inf_{\hat{g}_n} \sup_{f_0 \in \mathcal{S}_L(k_n)} \mathbb{E}\left[\|\hat{g}_n - f_0\|_{L^p}^r \right]^{1/r} > 0,$$

where the infimum is taken over all estimators $\hat{g}_n$.

Proof. See Appendix A.1. □

Remark 4.

(i) The condition (10) is a typical assumption for establishing lower bounds (see e.g. Tsybakov, 2009, Section 2.5). If $\exp(-c_1 x^2) \lesssim \varphi_{i,n}(x) \lesssim \exp(-c_2 x^2)$ with constants c_1, c_2, then (10) holds for any $z_0 > 0$, e.g. a Gaussian density. Note that the universal threshold η in (6) is independent of the truth f_0 and the specific loss function $\|\cdot\|_{L^p}$ for $0 < p < \infty$. The restriction of $r \le r_0$ is mainly due to control the r-th moment of the noise, which is quite natural. In most cases, one is interested in $r = 1$ or 2, for which it is sufficient to set $r_0 = 2$. Thus, Theorem 1 states that the MCPS estimators are up to a log-factor adaptively minimax optimal over sequences of classes $\mathcal{S}_L(k_n)$ for all possible k_n and L.

(ii) Theorem 1 also reveals that the underlying difficulty in estimation of step functions with respect to L^p-loss is actually determined by the number of change-points. A common choice of k_n is $k_n \asymp n^\theta$, $0 \le \theta < 1$, which in particular reproduces the convergence results in Li, Munk and Sieling (2016, Theorem 3.4) but now under weaker assumptions (here no assumption on the minimal segment length and the minimal jump size is made). It also includes the case $\theta = 0$, where, by convention, $k_n \equiv k$ is bounded.

(iii) Note further that the restriction $p < \infty$ in Theorem 1 is necessary and natural, because L^∞-loss is not reasonable in change-point estimation problems (as no estimator can detect change-point locations at a rate faster than $\mathcal{O}(1/n)$, see Chan and Walther, 2013, which leads to inconsistency of any estimator with respect to L^∞-loss).

(iv) In general, it is not clear whether the lower bound in Theorem 1 (ii) is sharp or not. However, in the particular case that $f_0 \in \mathcal{S}_L(k_n)$ is isotonic, it has recently been shown that the minimax rate in terms of squared L^2 risk is exactly of order $n^{-1} k_n \log\log n$, see Gao, Han and Zhang (2019).

3.2. Robustness to model misspecification

As discussed in Section 1, in practical applications, it often occurs that the underlying signal f_0 in model (1) is only approximately piecewise constant. To

address this issue, we next consider the L^p-loss of the MCPS methods for more general functions. In order to characterize the degree of model misspecification, we adopt from nonlinear approximation theory (cf. DeVore and Lorentz, 1993; DeVore, 1998) the *approximation spaces* as

$$\mathcal{A}_q^\gamma := \Big\{ f \in \mathcal{D} : \sup_{k\in\mathbb{N}} k^\gamma \Delta_{q,k}(f) < \infty \Big\}, \quad \text{for } 0 < q \le \infty,\ 0 < \gamma < \infty,$$

where the approximation error $\Delta_{q,k}$ is defined as

$$\Delta_{q,k}(f) := \inf\Big\{ \|f - g\|_{L^q} : g \in \mathcal{S},\ \#J(g) \le k \Big\}. \tag{11}$$

Introduce the subclasses

$$\mathcal{A}_{q,L}^\gamma := \Big\{ f \in \mathcal{D} : \sup_{k\ge 1} k^\gamma \Delta_{q,k}(f) \le L, \text{ and } \|f\|_{L^\infty} \le L \Big\},$$

for $0 < q \le \infty$, and $0 < \gamma, L < \infty$. The best approximant in (11) exists, but is in general non-unique, see e.g., DeVore and Lorentz (1993, Chapter 12). It follows readily from definition that $\mathcal{A}_q^\gamma = \bigcup_{L>0} \mathcal{A}_{q,L}^\gamma$ and that $\mathcal{A}_{q_1,L}^\gamma \subseteq \mathcal{A}_{q_2,L}^\gamma$ for all $q_1 \ge q_2$. Note that $\mathcal{A}_q^\gamma$ is actually an interpolation space between L^q and some Besov space (see Petrushev, 1988, Corollary 2.2). The order γ of these spaces (or classes) reflects the speed of approximation of f by step functions as the number of change-points increases. It is further known that if f lies in $\mathcal{A}_q^\gamma$ for some $\gamma > 1$ and if f is piecewise continuous, then f is piecewise constant, see Burchard and Hale (1975) (which is often referred to as a *saturation* result in the approximation theory community). Thus, it is custom to consider $\mathcal{A}_q^\gamma$ with $0 < \gamma \le 1$.

The rates of convergence for approximation classes are provided below.

Theorem 2 (Adaptation II)**.** *Let $0 < p < \infty$, $p \vee 2 \le q \le \infty$, and assume that $\hat{f}_n$ is an MCPS estimator in Definition 2 with some $r_0 \in (0, \infty)$. Then, for $0 < r \le r_0$ and $0 < \gamma, L < \infty$,*

$$\limsup_{n\to\infty} (\log n)^{-\frac{\gamma+(1/2-1/p)_+}{2\gamma+1}} n^{\frac{2\gamma}{2\gamma+1}(1/2\wedge 1/p)} \sup_{f_0\in\mathcal{A}_{q,L}^\gamma} \mathbb{E}\Big[\|\hat{f}_n - f_0\|_{L^p}^r\Big]^{1/r} < \infty.$$

The same result also holds almost surely if we drop the expectation $\mathbb{E}[\cdot]$.

Proof. See Appendix A.2. □

Remark 5. Similar to Theorem 1, the above theorem shows that any MCPS method automatically adapts to the smoothness of the approximation spaces, in the sense that it has a faster rate for larger γ. Note that such convergence rates in L^p-loss, $0 < p \le 2$, are nearly (i.e. up to a log-factor) minimax optimal over $\mathcal{A}_{q,L}^\gamma$ for every $0 < \gamma \le 1$, $2 \le q \le \infty$ and $L > 0$, since $n^{-\gamma/(2\gamma+1)}$ are known to be minimax rates for a smaller class H_L^γ, see Example 1 (i) below. We conjecture that the convergence rates in Theorem 2 are also nearly minimax

optimal for $\mathcal{A}^{\gamma}_{q,L}$ with respect to L^p-loss when $2 < p \le q \le \infty$, because this is indeed the case for $\gamma = 1$, as shown later in Example 1 (ii).

Moreover, note that the convergence rates of the MCPS methods above generalize the rates reported in Boysen et al. (2009) for jump-penalized least square estimators, and are faster than the rates reported in Fryzlewicz (2007) for the unbalanced Haar wavelets based estimator, with the difference being in log-factors.

Example 1. (i) *(Piecewise) Hölder functions.* For $0 < \alpha \le 1$ and $L \in (0, \infty)$, we consider the Hölder function classes

$$H^{\alpha}_{L} \equiv H^{\alpha}_{L}([0,1)) \coloneqq \big\{ f \in \mathcal{D} : \|f\|_{L^\infty} \le L, \text{ and } |f(x_1) - f(x_2)| \le L|x_1 - x_2|^{\alpha} \text{ for all } x_1, x_2 \in [0,1) \big\},$$

and the piecewise Hölder function classes with at most κ jumps, $\kappa \in \mathbb{N}_0$

$$\begin{aligned} H^{\alpha}_{\kappa,L} &\equiv H^{\alpha}_{\kappa,L}([0,1)) \\ &\coloneqq \Big\{ f \in \mathcal{D} : \text{there is a partition } \{I_i\}_{i=0}^{l}, \text{ with } l \le \kappa, \text{ of } [0,1) \\ &\qquad \text{such that } f\big|_{I_i} \in H^{\alpha}_{L}(I_i) \text{ for all possible } i \Big\}. \end{aligned}$$

Obviously, the latter one contains the former as a special case when $\kappa = 0$, that is, $H^{\alpha}_{0,L} \equiv H^{\alpha}_{L}$. By considering step functions with segments of equal length, one can easily show that $H^{\alpha}_{L} \subseteq \mathcal{A}^{\alpha}_{q,L'}$ with finite $L' \ge L$ and $0 < q \le \infty$, and in a similar way that $H^{\alpha}_{\kappa,L} \subseteq \mathcal{A}^{\alpha}_{q,L'}$ with finite $L' \ge L(\kappa+1)^{\alpha+1/2}$ and $0 < q \le \infty$.

It is known that the fastest possible rate over H^{α}_{L}, $0 < \alpha \le 1$, is of order $n^{-\alpha/(2\alpha+1)}$ with respect to the L^p-loss, $0 < p < \infty$, see e.g. Nemirovski (2000, Theorem 3.1). Thus, as a consequence of Theorem 2, the MCPS methods are simultaneously minimax optimal (up to a log-factor) over $\mathcal{A}^{\alpha}_{q,L}$, H^{α}_{L} and $H^{\alpha}_{\kappa,L}$ for every $\kappa \in \mathbb{N}_0$, $0 < p \le 2 \le q \le \infty$, $0 < \alpha \le 1$ and $L \in (0, \infty)$, that is, adaptive to the smoothness order α of the underlying function. The difference in convergence rates for L^p-loss, $2 < p < \infty$, is mainly because $\mathcal{A}^{\alpha}_{q,L}$ is strictly larger than H^{α}_{L}, see the next example for $\alpha = 1$.

(ii) *Bounded variation functions.* Recall that the (total) variation $\|\cdot\|_{\mathrm{TV}}$ of a function f is defined as

$$\|f\|_{\mathrm{TV}} \coloneqq \sup\Big\{ \sum_{i=0}^{m} |f(x_{i+1}) - f(x_i)| : 0 = x_0 < \cdots < x_{m+1} = 1,\ m \in \mathbb{N} \Big\}.$$

We introduce the càdlàg bounded variation classes

$$\mathrm{BV}_L \equiv \mathrm{BV}_L([0,1)) \coloneqq \big\{ f \in \mathcal{D} : \|f\|_{L^\infty} \le L, \text{ and } \|f\|_{\mathrm{TV}} \le L \big\} \quad \text{for } L \in (0, \infty).$$

Elementary calculation, together with Jordan decomposition, implies that

$$\mathrm{BV}_L \subseteq \mathcal{A}^{1}_{q,L'} \qquad \text{for finite } L' \ge L \text{ and } 0 < q \le \infty.$$

The best possible rate for BV_L are of order $n^{-2/3\min\{1/2,1/p\}}$ (see e.g. del Alamo, Li and Munk, 2018). Then, Theorem 2 implies that the MCPS methods attain the minimax optimal rate (up to a log-factor) over the bounded variation classes BV_L and $\mathcal{A}^1_{q,L}$ for $L \in (0,\infty)$, with respect to L^p-loss, $0 < p < \infty$.

All the examples above concern functions of smoothness order ≤ 1. For smoother functions, say H^α_L with $\alpha > 1$, which is defined as

$$H^\alpha_L \equiv H^\alpha([0,1)) \coloneqq \big\{ f \in \mathcal{D} : \|f\|_{L^\infty} \leq L, \text{ and } |f^{(\lfloor\alpha\rfloor)}(x_1) - f^{(\lfloor\alpha\rfloor)}(x_2)| \leq L|x_1 - x_2|^{\alpha - \lfloor\alpha\rfloor} \text{ for all } x_1, x_2 \in [0,1) \big\},$$

with $\lfloor\alpha\rfloor \coloneqq \max\{k \in \mathbb{N} : k < \alpha\}$, it holds that $H^\alpha_L \subseteq \mathcal{A}^1_q$ but $H^\alpha_L \not\subseteq \mathcal{A}^\gamma_q$ for any $\gamma > 1$. Thus, by Theorem 2, we obtain that the MCPS estimators attain (up to a log-factor) the rates of order $n^{-1/3}$ for H^α_L with $\alpha > 1$ in terms of L^2-loss. Note that such rates are suboptimal, but turn out to be the saturation barrier for every piecewise constant segmentation estimator; As we will see in Example 2 in Section 5.1, piecewise constant segmentation estimators even with the oracle choice of change-points cannot attain faster rates for functions of smoothness order > 1.

In summary, in the particular case of L^2-loss, we find that the MCPS methods are minimax optimal (up to log factors) simultaneously over sequences of step function classes $\mathcal{S}_L(k_n)$ ($k_n = o(n)$, $0 < L < \infty$), and over approximation spaces $\mathcal{A}^\gamma_{q,L}$ ($0 < \gamma \leq 1$, $2 \leq q \leq \infty$, $0 < L < \infty$). This especially includes sequences of step function classes $\mathcal{S}_L(n^\theta)$ ($0 \leq \theta < 1$, $0 < L < \infty$), Hölder classes H^α_L and $H^\alpha_{\kappa,L}$ ($0 < \alpha \leq 1$, $\kappa \in \mathbb{N}_0$, $0 < L < \infty$), and bounded variation classes BV_L ($0 < L < \infty$).

4. Feature detection

The convergence rates in Theorems 1 and 2 not only reflect the average performance in recovering the truth over its domain, but also, as a byproduct, lead to further statistical justifications on detection of features, such as change-points, modes and troughs.

Proposition 1. *Assume model (1) and let the truth $f_0 \equiv f_{k_n} \in \mathcal{S}_L(k_n)$ be a sequence of step functions with up to k_n jumps. By Δ_n and λ_n denote the smallest jump size, and the smallest segment length of f_{k_n}, respectively. Let $\hat{f}_n$ be an MCPS method in Definition 2. If*

$$\lim_{n\to\infty} \frac{k_n \log n}{\lambda_n \Delta_n^2 n} = 0,$$

then there is a constant C independent of f_{k_n} such that

$$\lim_{n\to\infty} \mathbb{P}\Big\{ \#J(\hat{f}_n) = \#J(f_{k_n}),\ d\big(J(\hat{f}_n); J(f_{k_n})\big) \leq C\frac{k_n \log n}{\Delta_n^2 n} \Big\} = 1,$$

with $d\big(J(\hat{f}_n); J(f_{k_n})\big) \coloneqq \max_{\tau \in J(f_{k_n})} \min_{\hat{\tau} \in J(\hat{f}_n)} |\tau - \hat{\tau}|$.

Proof. By Theorem 1 and Lin et al. (2016, Theorem 8) it holds almost surely that $d\big(J(\hat{f}_n); J(f_{k_n})\big) \leq C_1 k_n \log n/(\Delta_n^2 n)$, so $\mathbb{P}\big\{\#J(\hat{f}_n) \geq \#J(f_{k_n})\big\} \to 1$. This, together with the fact that $\mathbb{P}\big\{\#J(\hat{f}_n) > \#J(f_{k_n})\big\} \leq \mathcal{O}(n^{-r}) \to 0$ (see (18) in Appendix A.1) completes the proof. □

Remark 6. Proposition 1 concerns step functions, and is a typical consistency result in change-point literature (e.g. Boysen et al., 2009; Harchaoui and Lévy-Leduc, 2010; Chan and Chen, 2017). It in particular applies to SMUCE (Frick, Munk and Sieling, 2014) and FDRSeg (Li, Munk and Sieling, 2016), where the same error rate on the accuracy of estimated change-points is reported, and is of the fastest order known up to now (see also Fryzlewicz, 2014).

Assume now $f \in \mathcal{D}$, an arbitrary (not necessarily piecewise constant) function. We consider a similar concept of change-points as for step functions. To this end, we define, for any $\varepsilon > 0$, the set of ε-jump locations of f as

$$J_\varepsilon(f) \coloneqq \{x : |f(x) - f(x-0)| > \varepsilon\} ,$$

and the smallest ε-jump size as $\Delta_f^\varepsilon \coloneqq \min\{|f(x) - f(x-0)| : x \in J_\varepsilon(f)\}$. By Billingsley (1999, Lemma 1 in Section 12), the above concepts are well-defined, and satisfy that $\#J_\varepsilon(f) < \infty$ and $\Delta_f^\varepsilon \geq \varepsilon > 0$. Note that, in the particular case of step functions f, we always have $J_\varepsilon(f) \subseteq J(f)$ and $\Delta_f^\varepsilon \geq \Delta_f$, with equality holding for both if ε is smaller than the smallest jump size Δ_f of f. Moreover, if there exist $x_0 < x_1 < x_2 \in [a,b) \subseteq [0,1)$ such that $f(x_1) > f(x_0) \vee f(x_2)$ or $f(x_1) < f(x_0) \vee f(x_2)$, we say that there is a mode or a trough of f on $[a,b)$, respectively. We further define the number of modes of $f \in \mathcal{D}$ as

$$\begin{aligned}\#\mathrm{mode}(f) \coloneqq \big\{k : \text{there exist } x_0 < x_1 < \cdots < x_{2k} \in [0,1) \text{ such that} \\ f(x_{2i-1}) > f(x_{2i-1}) \vee f(x_{2i}) \text{ for each } i = 1, \ldots, k\big\} ,\end{aligned}$$

and the number of troughs of f as $\#\mathrm{trough}(f) \coloneqq \#\mathrm{mode}(-f)$. In order to investigate the shape of f, we introduce the local mean of f over an interval I as $m_I(f) \coloneqq \int_I f(x)dx/|I|$.

Theorem 3 (Feature recovery)**.** *Assume model (1) with the truth $f_0 \in \mathcal{A}_{2,L}^\gamma$ with $\gamma, L \in (0,\infty)$. Let $\hat{f}_n$ be an MCPS method in Definition 2. Then:*

(i) *If $\#\mathrm{mode}(f_0) \vee \#\mathrm{trough}(f_0) < \infty$, then*

$$\lim_{n\to\infty} \mathbb{P}\Big\{\#\mathrm{mode}(\hat{f}_n) \geq \#\mathrm{mode}(f_0);\ \#\mathrm{trough}(\hat{f}_n) \geq \#\mathrm{trough}(f_0)\Big\} = 1;$$

(ii) *There is a constant C independent of f_0 such that for every $\varepsilon > 0$*

$$\lim_{n\to\infty} \mathbb{P}\Big\{ d\big(J(\hat{f}_n), J_\varepsilon(f_0)\big) \leq \frac{C}{(\Delta_{f_0}^\varepsilon)^2}\Big(\frac{\log n}{n}\Big)^{\frac{2\gamma}{2\gamma+1}};$$
$$\text{and} \quad \#J(\hat{f}_n) \geq \#J_\varepsilon(f_0)\Big\} = 1.$$

Proof. See Appendix A.3. □

Remark 7. Since step functions lie in $\mathcal{A}_2^\gamma$ for all $\gamma > 0$, Theorem 3 (ii) "formally" reproduces Proposition 1 for the case that the step function f is fixed, by letting γ tend to infinity.

The statistical justifications of Theorem 3 are of one-sided nature, in the sense that an MCPS method $\hat{f}_n$ reproduces the features of f_0. Note that statistical guarantees for the reverse order are in general not possible, as long as an arbitrary number of jumps/features on small scales cannot be excluded, see e.g. Donoho (1988). However, the MCPS methods $\hat{f}_n$ will not include too many artificial features (e.g., jumps, modes or troughs), due to their parsimony nature by construction, namely, minimization of the number of jumps, see (4). Further, we can, to some extent, tell whether a feature reported by $\hat{f}_n$ is genuine or false, as follows.

Theorem 4 (Feature inference)**.** *Assume model* (1) *with the truth* $f_0 \in \mathcal{D}$*. Let* $\hat{f}_n$ *be an MCPS method in Definition 2 with interval system* $\mathcal{I}$ *and threshold* $\eta = \eta(\beta)$*,* $\beta \in (0,1)$*, in* (7)*. Define* $r_I = 2\big(\eta(\beta) + s_I\big)/\sqrt{n|I|}$ *for* $I \in \mathcal{I}$*. Then*

$$m_{I_1}(\hat{f}_n) > m_{I_2}(\hat{f}_n) + r_{I_1} + r_{I_2} \quad \text{for some } I_1, I_2 \in \mathcal{I} \text{ where } \hat{f}_n \text{ is constant,} \tag{12}$$

implies $m_{I_1}(f_0) > m_{I_2}(f_0)$*, simultaneously over all such pairs of* I_1 *and* I_2*, with probability at least* $1 - \beta$*.*

Proof. See Appendix A.3. □

Remark 8. Theorem 4 states that large increases (or decreases) of MCPS estimators imply increases (or decreases) of the true signal. This is actually a finite-sample inference guarantee, and holds simultaneously for many intervals, which thus provides inference guarantee on modes and troughs. In this way, we can discern a collection of genuine features among all the detected features, with controllable confidence. To be precise, let $\hat{f}_n = \sum_{i=1}^{\hat{k}} \hat{c}_i \mathbf{1}_{[\hat{\tau}_{i-1},\, \hat{\tau}_i)}$ with $0 = \hat{\tau}_0 < \cdots < \hat{\tau}_{\hat{k}} = 1$ and $\hat{c}_i \neq \hat{c}_{i+1}$ be an MCPS estimator with threshold $\eta(\beta)$.

(i) *Increase or decrease.* Let $\hat{\tau}_{i+1/2} = (\hat{\tau}_i + \hat{\tau}_{i+1})/2$. Define

$$u_i^R = \min_{I \in \mathcal{I},\, I \subseteq [\hat{\tau}_i,\, \hat{\tau}_{i+1/2})} (\hat{c}_{i+1} + r_I), \quad l_i^R = \max_{I \in \mathcal{I},\, I \subseteq [\hat{\tau}_i,\, \hat{\tau}_{i+1/2})} (\hat{c}_{i+1} - r_I),$$
$$\text{and} \quad u_i^L = \min_{I \in \mathcal{I},\, I \subseteq [\hat{\tau}_{i-1/2},\, \hat{\tau}_i)} (\hat{c}_i + r_I), \quad l_i^L = \max_{I \in \mathcal{I},\, I \subseteq [\hat{\tau}_{i-1/2},\, \hat{\tau}_i)} (\hat{c}_i - r_I).$$

Then, by Theorem 4, there is at least an increase (or a decrease) of f_0 on interval $[\hat{\tau}_{i-1/2}, \hat{\tau}_{i+1/2})$ if $u_i^L < l_i^R$ (or if $l_i^L > u_i^R$) with confidence level no less than $1 - \beta$. Further, because of the simultaneous confidence control, the inferred increases and decreases on non-overlapped intervals $[\hat{\tau}_{i-1/2}, \hat{\tau}_{i+1/2})$ leads naturally to inference on modes and troughs.

(ii) *Change-point.* Assume the true signal f_0 is piecewise Lipschitz continuous, namely, $f_0 \in H^1_{\kappa,L}$ with $\kappa \in \mathbb{N}_0$ and $L \in (0,\infty)$, see Example 1 (i). If for some ω and i such that $\hat\tau_{i-1} \le \hat\tau_i - \omega \le \hat\tau_i + \omega \le \hat\tau_{i+1}$,

$$\left|m_{[\hat\tau_i-\omega,\,\hat\tau_i)}(\hat f_n) - m_{[\hat\tau_i,\,\hat\tau_i+\omega)}(\hat f_n)\right| > r_{[\hat\tau_i-\omega,\,\hat\tau_i)} + r_{[\hat\tau_i,\,\hat\tau_i+\omega)} + \omega L\,, \quad (13)$$

then similar to Theorem 4 (ii), see Appendix A.3, we have

$$\left|m_{[\hat\tau_i-\omega,\,\hat\tau_i)}(f_0) - m_{[\hat\tau_i,\,\hat\tau_i+\omega)}(f_0)\right| > \omega L\,,$$

with confidence level no less than $1-\beta$. Note that if $f_0 \in H^1_{\kappa,L}$ is Lipschitz continuous on $[\hat\tau_i-\omega,\,\hat\tau_i+\omega)$, then

$$\begin{aligned}&\left|m_{[\hat\tau_i-\omega,\,\hat\tau_i)}(f_0) - m_{[\hat\tau_i,\,\hat\tau_i+\omega)}(f_0)\right| \\ &\quad\le \omega^{-1}\int_{[\hat\tau_i-\omega,\,\hat\tau_i)} |f_0(x) - f_0(x+\omega)|dx \le \omega L\,.\end{aligned}$$

Thus, condition (13) implies that there is at least a change-point of f_0 in $[\hat\tau_i-\omega,\,\hat\tau_i+\omega)$ with confidence level no less than $1-\beta$. That is, a significant change-point in most cases leads to a true change-point.

See Figure 2 (in Section 1) for an illustration. The SMUCE has detected 3 change-points, 1 mode and 1 trough. By the method described above, we can claim that the truth has at least 1 mode (in region [0.36, 0.88)), 1 trough (in region [0.1, 0.63)) and 2 change-points (around 0.5 and 0.75, if we assume $f_0 \in H^1_{\kappa,L}$ with $L \le 10$; note that the smallest Lipschitz constant of f_0 on its continuous parts is 2π in this example), with probability at least 90%. Such inference is nicely confirmed by the underlying truth.

5. Oracle properties

This section focuses on the oracle properties of MCPS methods. For simplicity, we restrict ourselves to $\mathcal{A}_2^\gamma$ and L^2-topology.

5.1. Oracle segmentation

It is well-known that the crucial difficulty in change-point segmentation problems is to infer the locations of change-points; Once the change-point locations are detected, the height of each segment can easily be determined via any reasonable estimator, e.g. a maximum likelihood estimator, locally on each segment (see e.g. Killick, Fearnhead and Eckley, 2012; Fryzlewicz, 2014). In line of this thought, we define

$$\Pi_n \coloneqq \left\{(\tau_0,\tau_1,\ldots,\tau_k) : \tau_0 = 0 < \tau_1 < \cdots < \tau_k = 1,\, k \in \mathbb{N},\ \text{and } \{n\tau_i\}_{i=0}^k \subseteq \mathbb{N}\right\}.$$

For each $\tau \equiv (\tau_0, \ldots, \tau_k) \in \Pi_n$, we introduce the piecewise constant segmentation estimator $\hat{f}_{\tau,n}$, conditioned on τ, for model (1) as

$$\hat{f}_{\tau,n} \coloneqq \sum_{i=1}^{k} \hat{c}_i \mathbf{1}_{[\tau_{i-1},\tau_i)} \qquad \text{with } \hat{c}_i = \frac{1}{n(\tau_i - \tau_{i-1})} \sum_{j \in [n\tau_{i-1}, n\tau_i)} y_j^n \,.$$

Theorem 5. *Assume model (1), and sub-Gaussian noises satisfy* $\mathbb{E}[(\xi_i^n)^2] \asymp \sigma_0^2$, *i.e., for some constants* c_1, c_2 *it holds that* $c_1\sigma_0^2 \le \mathbb{E}[(\xi_i^n)^2] \le c_2\sigma_0^2$ *for every possible* i *and* n. *Let* $\hat{f}_n$ *be an MCPS method in Definition 2. Then, there is a constant* C *such that for every* f_0 *in* $\cup_{\gamma>0}\mathcal{A}_2^\gamma \cap L^\infty$

$$\mathbb{E}[\|\hat{f}_n - f_0\|_{L^2}^2] \le C \log n \min_{\tau \in \Pi_n} \mathbb{E}[\|\hat{f}_{\tau,n} - f_0\|_{L^2}^2] \qquad \textit{for sufficiently large } n.$$

Proof. See Appendix A.4. □

Remark 9. Theorem 5 states that the MCPS methods perform nearly (up to a log-factor) as well as the piecewise constant segmentation estimator using an oracle for the change-point locations.

We next consider a *saturation phenomenon* of piecewise constant segmentation estimators via a simple example.

Example 2. Assume model (1) with the truth $f_0(x) \equiv x$ and the noise ξ_i^n being standard Gaussian. For simplicity, let $n = 6m^3$ with $m \in \mathbb{N}$. Elementary calculation shows that

$$\mathbb{E}[\|\hat{f}_{\tau_*,n} - f_0\|_{L^2}^2] = \min_{\tau \in \Pi_n} \mathbb{E}[\|\hat{f}_{\tau,n} - f_0\|_{L^2}^2] = \frac{6^{2/3} + 6^{-1/3}}{12} n^{-2/3}$$

and $\tau_* = \big(0, 1/m, \ldots, (m-1)/m, 1\big)$. Note that $f_0(x) \equiv x$ lies in every Hölder class H_L^α with $0 < \alpha < \infty$ and $L \ge 1$, and that the minimax optimal rates in terms of squared L^2-risk for H_L^α is of order $n^{-2\alpha/(2\alpha+1)}$. Thus, it indicates that the piecewise segmentation estimator even with the oracle choice of change-points saturates at smoothness order $\alpha = 1$. This in turn explains why MCPS methods cannot achieve faster rates for functions of smoothness order ≥ 1.

Note that such a saturation phenomenon for piecewise constant segmentation estimators is by no means due to the discontinuity of the estimator. In fact, one could discretize a smooth estimator (i.e., wavelet shrinkage estimators, Donoho et al., 1995) on the sample grids $\{i/n\}_{i=0}^n$ into a piecewise constant one: the discretized version performs equally well as the original estimator in asymptotical sense, since the discretization error vanishes faster than statistical estimation error. In contrast, the underlying reason for the aforementioned saturation is because piecewise constant segmentation estimators aim to segment data into constant pieces with the best possible recovery of change-point locations, rather than approximate the truth as well as possible. The purpose of segmentation into constant pieces provides an easy interpretation of the data, but it turns out to be less sufficient if the complete recovery of the function is the statistical

task. To overcome this saturation barrier, one could smoothen each segment based on detected change-point locations (see Boneva, Kendall and Stefanov, 1971). For instance, one could modify the MCPS estimators in (4) by considering polynomials or splines in each segment instead, which would lead to a procedure that detects sharp changes and meanwhile fits smooth pieces. Alternatively, in a similar spirit as Abramovich, Antoniadis and Pensky (2007), one could develop a two-step procedure: applying the MCPS estimators in the first step to estimate change-points, and in the second step fitting smooth pieces between change-points by spline or local polynomial estimators. The detailed study is, however, beyond the scope of this paper, and will be part of our future work.

5.2. Oracle approximant

Here we examine the performance of MCPS methods $\hat{f}_n$ by comparing it with the best piecewise constant approximants of f_0 with up to $\#J(\hat{f}_n)$ jumps. By means of compactness arguments and the convexity of L^2-norm, we can define

$$f_k^{\mathrm{app}} \in \operatorname*{argmin}_{f\in\mathcal{S},\, \#J(f)\le k} \|f_0 - f\|_{L^2} \qquad \text{for } k \in \mathbb{N}, \tag{14}$$

which always exists, but might be non-unique, as mentioned earlier in Section 3.2.

Proposition 2. *Assume model* (1). *Let* $\hat{f}_n$ *be an MCPS method in Definition 2, and* $\hat{K}_n := \#J(\hat{f}_n)$. *Then*

$$\lim_{n\to\infty} \sup_{f_0\in\mathcal{A}^\gamma_{2,L}} \mathbb{P}\Big\{ \|f_0 - f^{\mathrm{app}}_{\hat{K}_n}\|_{L^2} \ge C\|f_0 - \hat{f}_n\|_{L^2} \Big\} = 1 \qquad \textit{for some constant } C.$$

Proof. Following the proof of Theorem 2 in Appendix A.2, one can see that

$$\lim_{n\to\infty} \mathbb{P}\big\{A_n\big\} = 1,$$

where the event A_n is defined as

$$A_n := \left\{ \hat{K}_n \le k_n,\ \sup_{f_0\in\mathcal{A}^\gamma_{2,L}} \|f_0 - \hat{f}_n\|_{L^2} \le C_1 \left(\frac{\log n}{n}\right)^{\frac{\gamma}{2\gamma+1}} \right\},$$

with $k_n = C_2(n/\log n)^{1/(2\gamma+1)}$. Note that there is a sequence of $f_n \in \mathcal{A}^\gamma_{2,L}$ such that $\|f_n - f^{\mathrm{app}}_{k_n}\|_{L^2} \ge C_3 k_n^{-\gamma}$. Then, on the event A_n,

$$\|f_n - f^{\mathrm{app}}_{\hat{K}_n}\|_{L^2} \ge \|f_n - f^{\mathrm{app}}_{k_n}\|_{L^2} \ge C_3 k_n^{-\gamma} \ge C_4 \left(\frac{\log n}{n}\right)^{\frac{\gamma}{2\gamma+1}} \ge C_5 \|f_n - \hat{f}_n\|_{L^2}.$$

As a consequence, we have

$$\sup_{f_0\in\mathcal{A}^\gamma_{2,L}} \mathbb{P}\Big\{ \|f_0 - f^{\mathrm{app}}_{\hat{K}_n}\|_{L^2} \ge C_5\|f_0 - \hat{f}_n\|_{L^2} \Big\}$$

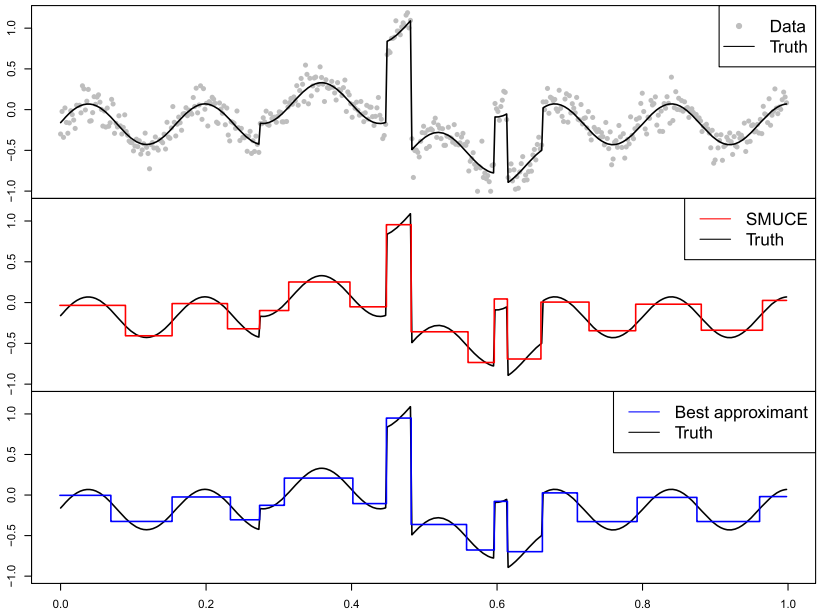


FIG 4. *Performance of SMUCE $\hat{f}_n$ with threshold $\eta(0.1)$ as oracle approximants for the signal in Olshen et al. (2004) and Zhang and Siegmund (2007). The bottom panel shows the best approximant $f^{\mathrm{app}}_{\hat{K}_n}$, defined in (14), of the truth with up to $\hat{K}_n$ jumps. Here SNR $= 3$ and $\|f - \hat{f}_n\|_{L^2} = 1.3\|f - f^{\mathrm{app}}_{\hat{K}_n}\|_{L^2}$.*

$$\begin{aligned} &\geq \mathbb{P}\Big\{\|f_n - f^{\mathrm{app}}_{\hat{K}_n}\|_{L^2} \geq C_5\|f_n - \hat{f}_n\|_{L^2}\Big\} \\ &\geq \mathbb{P}\{A_n\} \to 1 \qquad \text{as } n \to \infty, \end{aligned}$$

which concludes the proof. □

Remark 10. Proposition 2 indicates that $\hat{f}_n$ performs almost (up to a constant) as well as the best approximants $f^{\mathrm{app}}_{\hat{K}_n}$ for "complicated" functions f_0 in $\mathcal{A}^{\gamma}_{2,L}$, see Figure 4 for a visual illustration. For simpler functions f_0 in $\mathcal{A}^{\gamma}_{2,L}$, e.g., f_0 is piecewise constant, note that $\|f_0 - f^{\mathrm{app}}_{\hat{K}_n}\|_{L^2}$ can be zero. Thus, in this sense, the result in Proposition 2 cannot be improved by replacing $\sup_{f_0 \in \mathcal{A}^{\gamma}_{2,L}}$ by $\inf_{f_0 \in \mathcal{A}^{\gamma}_{2,L}}$.

6. Simulation study

Note that in the optimization problem (4) for MCPS methods, we optimize only over the local intervals, where the candidate function is constant, cf. (5). This leads to independence of values of candidate function among different segments, and thus ensures the structure of (4) to be a directed acyclic graph, which makes

dynamic programming algorithms (cf. Bellman, 1957) applicable to such a problem, see also Korte and Vygen (2012, Chapter 7). Moreover, the computation can be substantially accelerated by incorporating *pruning* ideas as e.g. recently developed in Killick, Fearnhead and Eckley (2012), Frick, Munk and Sieling (2014) and Li, Munk and Sieling (2016). As a consequence, the computational complexity of MCPS methods can be even *linear* in terms of the number of observations, in case that there are many change-points, see Frick, Munk and Sieling (2014) and Li, Munk and Sieling (2016) for further details.

We now investigate the finite sample performance of MCPS methods from the previously discussed perspectives. For brevity, we only consider a particular MCPS method, SMUCE (Frick, Munk and Sieling, 2014), and stress that the results are similar for other methods of type (4) (which are not shown here), see e.g. Frick, Munk and Sieling (2014) and Li, Munk and Sieling (2016) for an extensive simulation study. For SMUCE, we use the implementation of a pruned dynamic program from the CRAN R-package "stepR", select the system of all intervals with dyadic lengths for the multiscale constraint, and choose $\eta = \eta(\beta)$ in (7) as the threshold, which is estimated by 10,000 Monte-Carlo simulations. In what follows, the noise is assumed to be Gaussian with a known noise level σ, and SNR denotes the signal-to-noise ratio $\|f\|_{L^2}/\sigma$.

6.1. Stability

We first examine the stability of MCPS methods with respect to the significance level β, i.e. to the threshold η. The test signal f_0 (adopted from Olshen et al., 2004; Zhang and Siegmund, 2007) has 6 change points at 138, 225, 242, 299, 308, 332, and its values on each segment are -0.18, 0.08, 1.07, -0.53, 0.16, -0.69, -0.16, respectively. Figure 5 presents the behavior of SMUCE with threshold $\eta = \eta(\beta)$ for different choices of significance level β, on some specific data (see Table 1 in Frick, Munk and Sieling (2014) for the performance over many random repetitions). In fact, for the shown data, SMUCE detects the correct number of change-points, and recovers the location and the height of each segment in high accuracy, for the whole range of $0.06 \le \beta \le 0.94$ (i.e. $0.47\sqrt{\log n} \ge \eta \ge -0.04\sqrt{\log n}$). Only for smaller β (< 0.06, i.e. $\eta > 0.47\sqrt{\log n}$), SMUCE tends to underestimate the number of change-points (see the second panel of Figure 5 for example, where the missing change-point is marked by a vertical solid line), while, for larger β (> 0.94, i.e. $\eta < -0.04\sqrt{\log n}$), it is inclined to recover false change points (as shown in the last panel of Figure 5). Note that in either case the estimated locations and heights of the remaining segments (away from the missing/spurious jumps) are fairly accurate. This reveals that SMUCE is remarkably stable with respect to the choice of β (or η), in accordance with Theorem 1, Remark 3 and Proposition 1.

6.2. Different noise levels

We next investigate the impact of the noise level (or equivalently SNR) on MCPS methods. We consider the recovery of the Blocks signal (Donoho and

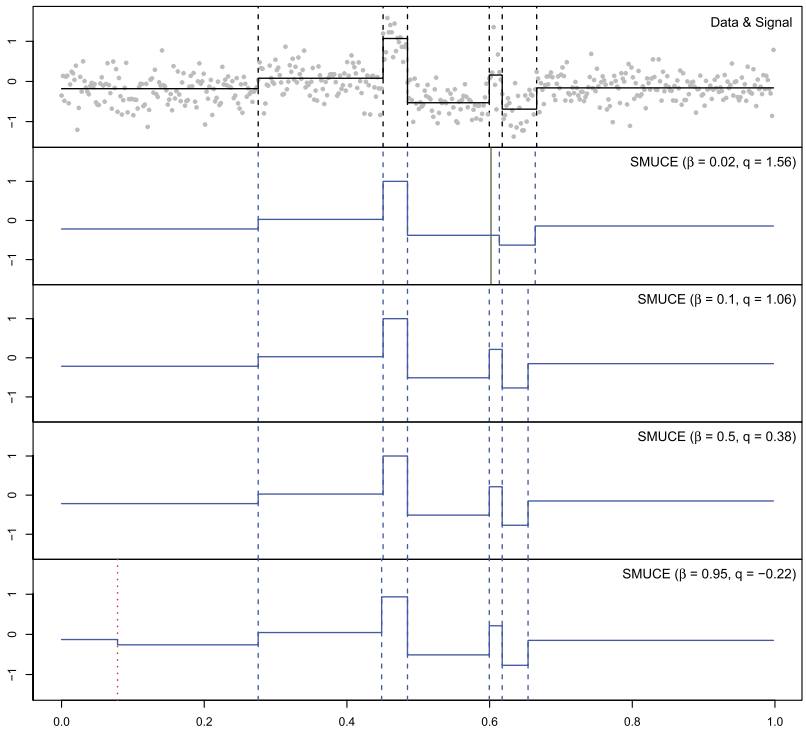


FIG 5. *Estimation of the step signal in Olshen et al. (2004) and Zhang and Siegmund (2007) by SMUCE with* $\eta = \eta(\beta)$ *for different* β *(sample size* $n = 497$*, and SNR* $= 1$*).*

Johnstone, 1994) for different noise levels. The result for SMUCE at significance level $\beta = 0.1$ is summarized in Figure 6 and Table 1. It shows that SMUCE recovers the truth rather well, in terms of e.g. change-point locations and L^2-loss, for the low and medium noise levels (SNR = 2.5 or 2), while it tends to miss one or two small scale features for the high noise level (SNR = 1.5).

6.3. Robustness and feature detection

In order to investigate the robustness of MCPS methods with respect to model misspecification, we introduce a local trend component as in Olshen et al. (2004) and Zhang and Siegmund (2007) to the test signal f_0 in Section 6.1, which leads to the model (with $n = 497$)

$$y_i^n = \big(\bar{f}_i^n + 0.25b\sin(a\pi i)\big) + \xi_i^n, \qquad i = 0, \dots, n-1. \tag{15}$$

We consider two scenarios separately.

(i) *Weak background waves.* We simulate data for $a = 0.025$ and $b = 0.3$, and apply SMUCE again with various choice of β, see Figure 7, with

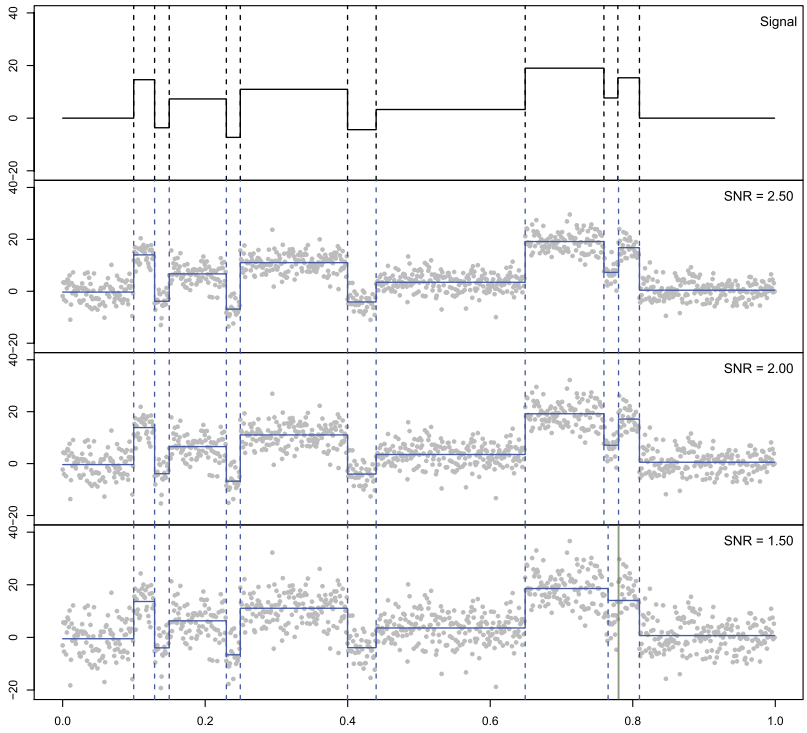


FIG 6. *Blocks signal: SMUCE for various noise levels (sample size $n = 1{,}023$).*

the average performance given in the top part of Table 2. It shows that SMUCE captures all relevant features, e.g., change-points and modes, of the piecewise constant part (cf. Figure 7) of the true signal, and is stable with respect to the choice of β. This is in accordance with the previous simulations and Theorems 3 and 5.

(ii) *Strong background waves.* When the scale b of the background wave becomes larger, i.e., the fluctuation is stronger, SMUCE captures the fluctuation by including additional change-points according to Theorems 2, 3, 5 and Proposition 2. Figure 8, as well as Table 2, illustrates the performance of SMUCE with $\beta = 0.1$ for the signal in (15) with $b = 1.0$ and $b = 1.2$

TABLE 1
Performance of SMUCE ($\beta = 0.1$) on the Blocks signal ($n = 1{,}023$, cf. Figure 6) for various noise levels over 100 random repetitions.

SNR	Counts of $\#J(\hat{f}_n) - \#J(f_0)$				Average of	
	≤ -2	-1	0	≥ 1	$n \cdot d(J(\hat{f}_n), J(f_0))$	$\|\hat{f}_n - f_0\|_{L^2}/\|f_0\|_{L^2}$
2.5	0	1	99	0	1	0.046
2	0	16	84	0	4.2	0.071
1.5	2	71	27	0	15	0.12

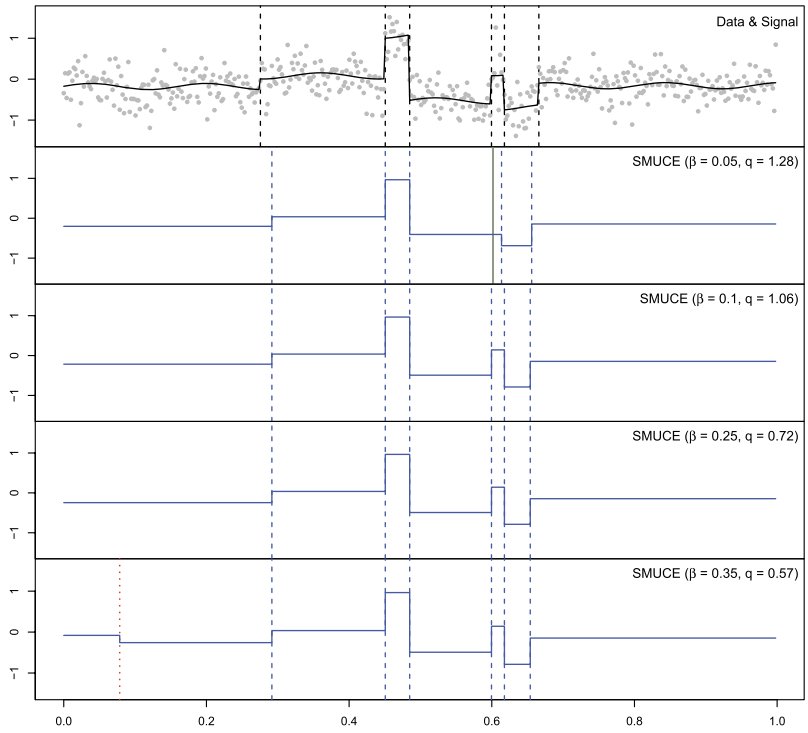


FIG 7. *Estimation of the signal in* (15) *($a = 0.025, b = 0.3$) by SMUCE with $\eta = \eta(\beta)$ for different β (SNR $= 1$).*

under different noise levels. It can be seen that SMUCE recovers the shape and modes of the whole true signal (which has 8 modes in total).

We stress, moreover, that by Theorem 4 it is possible to screen whether the recovered features are genuine or not with pre-specified confidence level β. This can be done by the procedure detailed in Remark 8. From Table 2, we observe that nearly all the recovered features are genuine when the noise level is low (e.g., SNR $= 2.5$), while only large features are guaranteed to be there for the medium and high noise levels (e.g., SNR $= 2$ or 1.5), with probability at least 90%. This is mainly due to the built-in parsimony of the method, namely, the minimization of number of change-points, see (4).

6.4. Empirical convergence rates

Finally, we empirically explore how well the finite sample risk is approximated by our asymptotic analysis. The test signals are the Blocks and the Heavisine from Donoho and Johnstone (1994). Note that the Blocks signal is a piecewise constant function with a fixed number of change-points, so the convergence rates

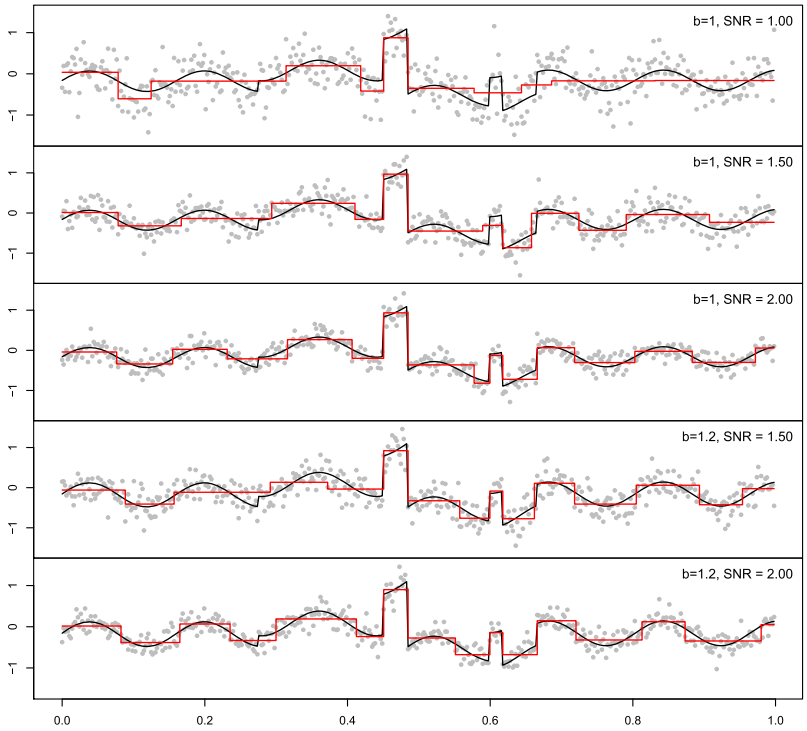


FIG 8. *Estimation of the signal in* (15) *with* $a = 0.025$ *and* $b = 1$ *or* 1.2 *by SMUCE for various noise levels.*

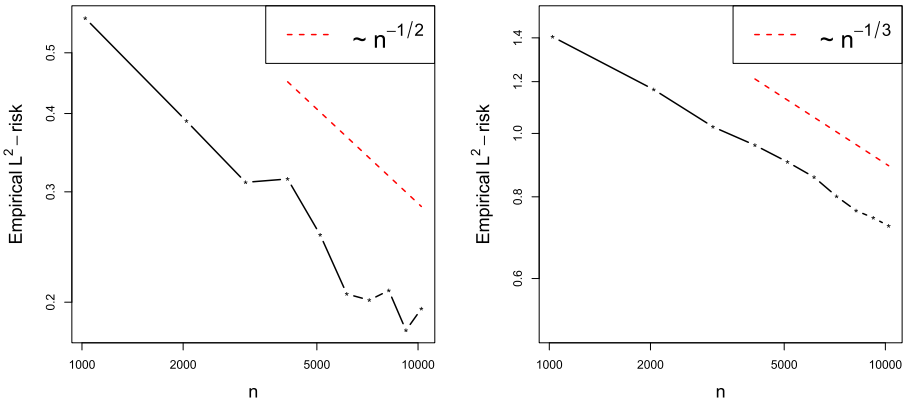
TABLE 2

Average performance of SMUCE ($\beta = 0.1$) *on the signal in* (15) ($a = 0.025$, *cf. Figures 7 and 8) over 100 random repetitions. The number of inferred modes is computed according to the procedure in Remark 8.*

b	SNR	$\#J(\hat{f}_n) - \#J(f_0)$	$\#\text{mode}(\hat{f}_n)$	# inferred modes	$\|\hat{f}_n - f_0\|_{L^2}/\|f_0\|_{L^2}$
	2.5	0.51	2.2	2.1	0.17
0.3	2	0.14	2	2	0.17
	1.5	−0.09	1.9	1.5	0.21
	2.5	9.1	6	5.5	0.27
1	2	8.5	5.8	3.7	0.31
	1.5	4.8	3.5	1.7	0.45
	2.5	9.2	6	5.8	0.3
1.2	2	8.9	5.9	4.9	0.32
	1.5	6.3	4.4	2.1	0.45

in L^2-risk is of order $n^{-1/2}$ (up to a log-factor) by Theorem 1. For the Heavisine signal, the convergence rate is of order $n^{-1/3}$ (up to a log-factor), since it lies in $H^1_{1,L}$ and BV_L for some L, see Theorem 2 and Example 1. Although the Heavisine also lies in $H^\alpha_{1,L}$, $\alpha \geq 1$, this will not lead to a faster rate for the MCPS methods due to the saturation phenomenon, see Example 2.

In Figure 9, we display the average of L^2-loss of SMUCE with significance



(a) Blocks signal (b) Heavisine signal

FIG 9. *Convergence rates of SMUCE averaged over 100 random repetitions (SNR* $= 2.5$*).*

level $\beta = 0.1$ for a range of sample sizes from 1,023 to 10,230. Note that both axes are in log-scale, so the slopes reflect the order of rates. By linear regression, the estimated order of rates are $n^{-0.48}$ and $n^{-0.29}$ for the Blocks and the Heavisine, respectively. There is only a little difference to the optimal order of rates, which are indicated by the slopes of dashed lines. It is partially due to the log-factors. Thus, this confirms our theoretical findings in Theorems 1 and 2.

7. Conclusion

In this paper we focus on the convergence analysis for MCPS methods, a general family of change-point estimators based on the combination of variational estimation and multiple testing over different scales, in a nonparametric regression setting with special emphasis on step functions while allowing for various distortions. We found that the estimation difficulty for a step function is mainly determined by its number of jumps, and shown that the MCPS methods attain the nearly optimal convergence rates for step functions with asymptotically bounded or even increasing number of jumps. As a robustness study, we also examined the convergence behavior of these methods for more general functions, which are viewed as distorted jump functions. Such distortion is precisely characterized by certain approximation spaces. In particular, we have derived nearly optimal convergence rates for MCPS methods in case that the regression function is either a (piecewise) Hölder function or a bounded variation function. Remarkably, these methods automatically adapt to the unknown smoothness for all aforementioned function classes, as the only tuning parameter can be selected in a universal way. The convergence rates also provide statistical justification with respect to the detection of features, such as change-points and modes (or troughs). In addition, the MCPS methods $\hat{f}_n$ are shown perform nearly as well as the oracle piecewise constant segmentation estimators, and the best piece-

wise constant (oracle) approximants of the truth with less or the same number of jumps as $\hat{f}_n$.

The MCPS methods, however, cannot attain faster rates for functions of stronger smoothness than above, which is indeed a common saturation shared by all piecewise constant segment estimators. This can be improved by considering piecewise polynomial or spline estimators (see e.g. Spokoiny, 1998; Abramovich, Antoniadis and Pensky, 2007), but the proper combination with multiscale methodology needs further investigation (see the rejoinder by Frick, Munk and Sieling, 2014, for a first attempt). Alternatively, certain smoothness penalty can be employed instead of the number of jumps in the formulation of MCPS, see e.g. Grasmair, Li and Munk (2018), where the nearly optimal rates are shown for higher order Sobolev/Besov classes. In addition, extension of our results to models with general errors beyond sub-Gaussian, such as heavy tailed distributions (see e.g. Han and Wellner, 2019), and stationery Gaussian processes (see e.g. Schwartzman, Gavrilov and Adler, 2011), would be interesting for future research.

Appendix A: Proofs

A.1. Proof of Theorem 1

Part (i): We structure the proof into three steps. We shift the change-points of the truth f_0 to their nearest points in the grid $\{0, 1/n, \dots, (n-1)/n\}$, while keeping the heights of segments unchanged, and denote the resulting function by $\tilde{f}_0$. Then $\#J(\tilde{f}_0) \le k_n$ and

$$\|f_0 - \tilde{f}_0\|_{L^p} \le 2\|f\|_{L^\infty} \left(\frac{k_n}{2n}\right)^{1/p} \le 2L \left(\frac{k_n}{2n}\right)^{1/p}. \tag{16}$$

a) *Good noise case.* Assume that $\tilde{f}_0$ lies in the multiscale constraint, i.e.,

$$T_{\mathcal{I}}(y^n; \tilde{f}_0) \le \eta = a\sqrt{\log n}.$$

By (4), it holds that $\#J(\hat{f}_n) \le \#J(\tilde{f}_0) \le k_n$. Let intervals $\{I_i\}_{i=0}^m$ be the partition of $[0,1)$ by $J(\hat{f}_n) \cup J(\tilde{f}_0)$ with $m \le 2k_n$. Then

$$\|\hat{f}_n - \tilde{f}_0\|_{L^p}^p = \sum_{i=0}^{m} |\hat{\theta}_i - \theta_i|^p |I_i| \qquad \text{with } \hat{f}_n|_{I_i} \equiv \hat{\theta}_i \text{ and } \tilde{f}_0|_{I_i} \equiv \theta_i.$$

If $|I_i| > c/n$, then by c-normality of $\mathcal{I}$, there is $\tilde{I}_i \in \mathcal{I}$ such that $\tilde{I}_i \subseteq I_i$ and $|\tilde{I}_i| \ge |I_i|/c$. It follows that

$$\left|\tilde{I}_i\right|^{1/2} \left|\theta - \frac{1}{n|\tilde{I}_i|} \sum_{j/n \in \tilde{I}_i} y_j^n\right| \le (a+\delta)\sqrt{\frac{\log n}{n}} \qquad \text{for } \theta = \theta_i \text{ or } \hat{\theta}_i.$$

By a triangular inequality and $|\tilde{I}_i| \geq |I_i|/c$, we obtain

$$|I_i|^{1/2}|\hat{\theta}_i - \theta_i| \leq 2(a+\delta)\sqrt{\frac{c\log n}{n}}.$$

If $|I_i| \leq c/n$, by Definition 1, we have $[i_0/n, (i_0+1)/n) \subseteq I_i$ for some i_0. Then

$$|\hat{\theta}_i - \theta_i| \leq |\hat{\theta}_i - y^n_{i_0}| + |y^n_{i_0} - \theta_i| \leq 2(a+\delta)\sqrt{\log n}\,.$$

Thus, by combining these two situations, we obtain

$$\|\hat{f}_n - \tilde{f}_0\|^p_{L^p} \leq \sum_{i:|I_i|>c/n} |I_i| \left(2(a+\delta)\sqrt{\frac{c\log n}{n|I_i|}}\right)^p + \sum_{i:|I_i|\leq c/n} \frac{c}{n}\left(2(a+\delta)\sqrt{\log n}\right)^p.$$

Note that for $0 < p < 2$, by the Hölder's inequality,

$$\begin{aligned}
&\sum_{i:|I_i|>c/n} |I_i| \left(2(a+\delta)\sqrt{\frac{c\log n}{n|I_i|}}\right)^p \\
&= \sum_{i:|I_i|>c/n} |I_i|^{1-p/2} \left(2(a+\delta)^2\frac{c\log n}{n}\right)^{p/2} \\
&\leq \Big(\sum_{i:|I_i|>c/n} |I_i|\Big)^{1-p/2} \Big(\sum_{i:|I_i|>c/n} 4(a+\delta)^2\frac{c\log n}{n}\Big)^{p/2} \\
&\leq \left(4(2k_n+1)(a+\delta)^2\frac{c\log n}{n}\right)^{p/2},
\end{aligned}$$

and for $2 \leq p < \infty$,

$$\begin{aligned}
\sum_{i:|I_i|>c/n} |I_i| \left(2(a+\delta)\sqrt{\frac{c\log n}{n|I_i|}}\right)^p &\leq \sum_{i:|I_i|>c/n} \left(4(a+\delta)^2\frac{c\log n}{n}\right)^{p/2} \left(\frac{c}{n}\right)^{1-p/2} \\
&\leq \frac{(2k_n+1)c}{n}\big(4(a+\delta)^2\log n\big)^{p/2}.
\end{aligned}$$

Since $k_n = o(n)$, we have for large enough n

$$\begin{aligned}
\|\hat{f}_n - \tilde{f}_0\|^p_{L^p} &\leq \left(\left(\frac{(2k_n+1)c}{n}\right)^{p/2\wedge 1} + \frac{(2k_n+1)c}{n}\right)\big(4(a+\delta)^2\log n\big)^{p/2} \\
&\leq 2\left(\frac{(2k_n+1)c}{n}\right)^{p/2\wedge 1}\big(4(a+\delta)^2\log n\big)^{p/2}.
\end{aligned}$$

Then, together with (16), for large enough n,

$$\|\hat{f}_n - f_0\|^r_{L^p} \leq 2^{(r-1)_+}\left(\|\hat{f}_n - \tilde{f}_0\|^r_{L^p} + \|\tilde{f}_0 - f_0\|^r_{L^p}\right)$$

$$\leq 2^{r(1+1/p)} \left(\frac{(2k_n+1)c}{n} \right)^{r/2 \wedge r/p} \left(4(a+\delta)^2 \log n\right)^{r/2}. \tag{17}$$

b) *Almost sure convergence.* For each $I \in \mathcal{I}$, note that $(n|I|)^{-1/2} \sum_{i/n \in I} \xi_i^n$ is again sub-Gaussian with scale parameter σ, so

$$\mathbb{P}\left\{ \frac{1}{\sqrt{n|I|}} \Big| \sum_{i/n \in I} \xi_i^n \Big| > x \right\} \leq 2\exp(-x^2/2\sigma^2)$$

for any $x > 0$. Note that, on every $I \in \mathcal{I}$ where $\tilde{f}_0$ is constant,

$$\frac{1}{\sqrt{n|I|}} \left| \sum_{i/n \in I} \left(\bar{f}_i^n - \tilde{f}_0(\frac{i}{n}) \right) \right| \leq \frac{4L}{\sqrt{n|I|}} \leq 4L\,,$$

with $\bar{f}_i^n$ in (1). Then, by the Boole's inequality, it holds that for large enough n

$$\begin{aligned}
&\mathbb{P}\left\{ T_{\mathcal{I}}(y^n; \tilde{f}_0) > a\sqrt{\log n} \right\} \\
&\leq \mathbb{P}\left\{ \max_{I \in \mathcal{I}} \frac{1}{\sqrt{n|I|}} \Big| \sum_{i/n \in I} \xi_i^n \Big| > (a-\delta)\sqrt{\log n} - 4L \right\} \\
&\leq 2\exp\left(-\frac{(a-\delta)^2 \log n}{2\sigma^2} (1+o(1)) \right) \frac{n^2}{2} \\
&= n^{-\frac{(a-\delta)^2}{2\sigma^2}(1+o(1))+2} \leq n^{-r},
\end{aligned} \tag{18}$$

where the last equality is due to (6) and $r \leq r_0$. This and (17) imply the almost sure convergence assertion.

c) *Convergence in expectation.* It follows from (17) that for large enough n

$$\begin{aligned}
\mathbb{E}\left[\|\hat{f}_n - f_0\|_{L^p}^r \right] &= \mathbb{E}\left[\|\hat{f}_n - f_0\|_{L^p}^r;\, T_{\mathcal{I}}(y^n; \tilde{f}_0) \leq a\sqrt{\log n} \right] \\
&\quad + \mathbb{E}\left[\|\hat{f}_n - f_0\|_{L^p}^r;\, T_{\mathcal{I}}(y^n; \tilde{f}_0) > a\sqrt{\log n} \right] \\
&\leq 2^{r(1+1/p)} \left(\frac{(2k_n+1)c}{n} \right)^{r/2 \wedge r/p} \left(4(a+\delta)^2 \log n\right)^{r/2} \\
&\quad + \mathbb{E}\left[\|\hat{f}_n - f_0\|_{L^p}^r;\, T_{\mathcal{I}}(y^n; \tilde{f}_0) > a\sqrt{\log n} \right].
\end{aligned}$$

We next show the second term above asymptotically vanishes faster.

$$\begin{aligned}
&\mathbb{E}\left[\|\hat{f}_n - f_0\|_{L^p}^r;\, T_{\mathcal{I}}(y^n; \tilde{f}_0) > a\sqrt{\log n} \right] \\
&= \int_0^{2n^{p/2}} \mathbb{P}\left\{ \|\hat{f}_n - f_0\|_{L^p}^p \geq u;\, T_{\mathcal{I}}(y^n; \tilde{f}_0) > a\sqrt{\log n} \right\} \frac{r}{p} u^{r/p-1} du \\
&\qquad + \int_{2n^{p/2}}^{\infty} \mathbb{P}\left\{ \|\hat{f}_n - f_0\|_{L^p}^p \geq u;\, T_{\mathcal{I}}(y^n; \tilde{f}_0) > a\sqrt{\log n} \right\} \frac{r}{p} u^{r/p-1} du
\end{aligned}$$

$$\leq 2^{r/p} n^{r/2} \mathbb{P}\left\{T_{\mathcal{I}}(y^n; \tilde{f}_0) > a\sqrt{\log n}\right\} + \int_{2n^{p/2}}^{\infty} \mathbb{P}\left\{\|\hat{f}_n - f_0\|_{L^p}^p \geq u\right\} \frac{r}{p} u^{r/p-1} du$$

$$\leq 2^{r/p} n^{-r/2} + \int_{2n^{p/2}}^{\infty} \mathbb{P}\left\{\|\hat{f}_n - f_0\|_{L^p}^p \geq u\right\} \frac{r}{p} u^{r/p-1} du, \tag{19}$$

where the last inequality is due to (18). Define functions $g : [0,1) \to \mathbb{R}$ and $h : [0,1) \to \mathbb{R}$ as

$$g \coloneqq \sum_{i=0}^{n-1} y_i^n \mathbf{1}_{[i/n,(i+1)/n)}. \quad \text{and} \quad h \coloneqq \sum_{i=0}^{n-1} \bar{f}_i^n \mathbf{1}_{[i/n,(i+1)/n)}.$$

Let $\xi^n \coloneqq \{\xi_i^n\}_{i=0}^{n-1}$ and $s \coloneqq (2r-p)_+$. Then,

$$\begin{aligned}
\|\hat{f}_n - f_0\|_{L^p}^p \leq & 3^{(p-1)_+} \left(\|\hat{f}_n - g\|_{L^p}^p + \|g - h\|_{L^p}^p + \|h - f_0\|_{L^p}^p\right) \\
\leq & 3^{(p-1)_+} \left((a+\delta)^p (\log n)^{p/2} + n^{-1} \|\xi^n\|_{\ell^p}^p + (2L)^p\right) \\
\leq & 3^{(p-1)_+} \left((a+\delta)^p (\log n)^{p/2} + n^{-p/(p+s)} \|\xi^n\|_{\ell^{p+s}}^p + (2L)^p\right).
\end{aligned}$$

Thus, for large enough n, i.e. if $n^{p/2} \geq 3^{(p-1)_+}\big((a+\delta)^p (\log n)^{p/2} + (2L)^p\big)$,

$$\begin{aligned}
& \int_{2n^{p/2}}^{\infty} \mathbb{P}\left\{\|\hat{f}_n - f_0\|_{L^p}^p \geq u\right\} \frac{r}{p} u^{r/p-1} du \\
\leq & \int_{2n^{p/2}}^{\infty} \mathbb{P}\left\{3^{(p-1)_+} \left((a+\delta)^p (\log n)^{p/2} + n^{-p/(p+s)} \|\xi^n\|_{\ell^{p+s}}^p + (2L)^p\right) \geq u\right\} \\
& \times \frac{r}{p} u^{r/p-1} du \\
\leq & \int_{n^{p/2}}^{\infty} \mathbb{P}\left\{3^{(p-1)_+} \left((a+\delta)^p (\log n)^{p/2} + n^{-p/(p+s)} \|\xi^n\|_{\ell^{p+s}}^p + (2L)^p\right) \geq u + n^{p/2}\right\} \\
& \times \frac{r}{p} (2u)^{r/p-1} du \\
\leq & \int_{n^{p/2}}^{\infty} \mathbb{P}\left\{3^{(1+s/p)(p-1)_+} \frac{1}{n} \sum_{i=0}^{n-1} |\xi_i^n|^{p+s} \geq u^{1+s/p}\right\} \frac{r}{p} (2u)^{r/p-1} du \\
\leq & 2^{r/p-1} 3^{(1+s/p)(p-1)_+} \mathbb{E}\left[\frac{1}{n} \sum_{i=0}^{n-1} |\xi_i^n|^{p+s}\right] \int_{n^{p/2}}^{\infty} \frac{r}{p} u^{-(s-r)/p-2} du = \mathcal{O}(n^{-r/2}),
\end{aligned}$$

where the last inequality holds by the fact $s \geq 2r - p$ and

$$\mathbb{E}[|\xi_i^n|^{p+s}] \leq (p+s) 2^{(p+s)/2} \sigma^{p+s} \Gamma\big(\frac{p+s}{2}\big) = \mathcal{O}(1) \qquad \text{for each } i\,.$$

Thus, by (19) it holds that

$$\mathbb{E}\left[\|\hat{f}_n - f_0\|_{L^p}^r;\, T_{\mathcal{I}}(y^n; \tilde{f}_0) > a\sqrt{\log n}\right] = \mathcal{O}(n^{-r/2})$$

$$= o\Big(\big(n^{-1}(k_n+1)\big)^{r/p \wedge r/2} (\log n)^{r/2} \Big).$$

This concludes the assertion in expectation.

Part (ii): The lower bound can be proven similarly as Li, Munk and Sieling (2016, Theorem 3.4), by means of standard arguments based on testing many hypotheses (pioneered by Ibragimov and Has'minskiĭ, 1977; Has'minskiĭ, 1978). More precisely, we consider two collections of hypotheses

$$\Big\{ \sum_{i=1}^{2k_n+2} \frac{(-1)^i \tilde{z}_0}{2} \mathbf{1}_{[\frac{i-1}{2k_n+2}+c_{i-1}, \frac{i}{2k_n+2}+c_i)} : c_i = \pm \frac{\sigma_0^2 \log 2}{32 n \tilde{z}_0^2}, c_0 = c_{2k_n+2} = 0 \Big\} \subseteq \mathcal{S}_L(k_n)$$

with $\tilde{z}_0 \coloneqq z_0 \wedge L$, and

$$\Big\{ \sum_{i=1}^{k_n+1} \frac{(-1)^i L + c_i}{2} \mathbf{1}_{[\frac{i-1}{k_n+1}, \frac{i}{k_n+1})} : c_i = \pm \frac{\sigma_0}{4} \sqrt{\frac{k_n \log 2}{2n}} \Big\} \subseteq \mathcal{S}_L(k_n).$$

Elementary calculation together with Fano's lemma (cf. Tsybakov, 2009, Corollary 2.6) concludes the proof. □

A.2. Proof of Theorem 2

The idea behind is that we first approximate the truth f_0 by a step function f_{k_n} with $\mathcal{O}(k_n)$ jumps, and then treat f_{k_n} as the underlying "true" signal in model (1) (with additional approximation error). In this way, it allows us to employ similar techniques as in the proof of Theorem 1. To be rigorous, we give a detailed proof as follows.

Since $\mathcal{A}^\gamma_{q_1,L} \subseteq \mathcal{A}^\gamma_{q_2,L}$ for $q_1 \geq q_2$, it is sufficient to consider $q = p \vee 2 < \infty$.

a) *Good noise case.* Assume that the observations $y^n = \{y_i^n\}_{i=0}^{n-1}$ from model (1) are close to the truth f_0 in the sense that the event

$$\mathcal{G}_n \coloneqq \Big\{ y^n : \max_{I \in \mathcal{I}} \frac{1}{\sqrt{n|I|}} \Big| \sum_{i/n \in I} \big(y_i^n - \bar{f}_i^n\big) \Big| - s_I \leq a_0 \sqrt{\log n} \Big\} \tag{20}$$

holds with $a_0 = \delta + \sigma\sqrt{2r_0+4}$. Now let

$$k_n \coloneqq \Big\lceil \Big(\frac{4L}{a-a_0}\Big)^{2/(2\gamma+1)} \Big(\frac{n}{\log n}\Big)^{1/(2\gamma+1)} \Big\rceil .$$

Since $f_0 \in \mathcal{A}^\gamma_{q,L}$, for every n there is a step function $\tilde{f}_{k_n} \in \mathcal{S}$ such that

$$\#J(\tilde{f}_{k_n}) \leq k_n, \quad \|\tilde{f}_{k_n}\|_{L^\infty} \leq L, \quad \text{and } \|f_0 - \tilde{f}_{k_n}\|_{L^q} \leq L k_n^{-\gamma},$$

by means of compactness argument. Based on the continuity of

$$x \mapsto \int_{[0,x)} |f_0(t) - \tilde{f}_{k_n}(t)|^2 dt,$$

one can find $\tau_0 = 0 < \tau_1 < \cdots < \tau_{k_n} = 1$ satisfying

$$\int_{[\tau_{i-1},\tau_i)} |f_0(t) - \tilde{f}_{k_n}(t)|^2 dt = \frac{1}{k_n} \|f_0 - \tilde{f}_{k_n}\|_{L^2}^2 \quad \text{for each } i\,.$$

Let $\{\tilde{I}_i\}_{i=0}^{\ell}$ be the partition of $[0,1)$ by $J(\tilde{f}_{k_n}) \cup \{\tau_1, \ldots, \tau_{k_n-1}\}$. Then $\ell \le 2k_n$. Fix $\varepsilon_n > 0$ such that $\varepsilon_n/2$ is smaller than the smallest jump size of $\tilde{f}_{k_n}$ and $\varepsilon_n \le L k_n^{-\gamma-1/2}$. Define a step function $\breve{f}_{k_n} : [0,1) \to \mathbb{R}$ as

$$\breve{f}_{k_n}(x) = \tilde{f}_{k_n}(x) + (-1)^i \varepsilon_n \quad \text{for } x \in \tilde{I}_i\,.$$

Then $J(\breve{f}_{k_n}) = J(\tilde{f}_{k_n}) \cup \{\tau_1, \ldots, \tau_{k_n-1}\}$ and

$$\begin{aligned}
&\|f_0 - \breve{f}_{k_n}\|_{L^q} \le \|f_0 - \tilde{f}_{k_n}\|_{L^q} + \|\tilde{f}_{k_n} - \breve{f}_{k_n}\|_{L^q} \le 2Lk_n^{-\gamma}\,, \quad \text{and} \\
&\|(f_0 - \breve{f}_{k_n})\mathbf{1}_{\tilde{I}_i}\|_{L^2} \le \|(f_0 - \tilde{f}_{k_n})\mathbf{1}_{\tilde{I}_i}\|_{L^2} + \|(\tilde{f}_{k_n} - \breve{f}_{k_n})\mathbf{1}_{\tilde{I}_i}\|_{L^2} \le 2Lk_n^{-\gamma-1/2}\,,
\end{aligned}$$

for every $i = 0, \ldots, \ell$. Moving each change-point of $\breve{f}_{k_n}$ to the closest point in $\{0, 1/n, \ldots, (n-1)/n\}$ but leaving the heights of segments unchanged, we obtain a step function f_{k_n} such that $\#J(f_{k_n}) \le \#J(\breve{f}_{k_n}) \le 2k_n$ and

$$\|f_0 - f_{k_n}\|_{L^q} \le \|f_0 - \breve{f}_{k_n}\|_{L^q} + \|\breve{f}_{k_n} - f_{k_n}\|_{L^q} \le 2Lk_n^{-\gamma} + 4L\left(\frac{k_n}{n}\right)^{1/q}.$$

Note that $\|f_{k_n}\|_{L^\infty} = \|\breve{f}_{k_n}\|_\infty \le L + \varepsilon_n \le 2L$. Then

$$\begin{aligned}
\|(f_0 - f_{k_n})\mathbf{1}_I\|_{L^2} &\le \|(f_0 - \breve{f}_{k_n})\mathbf{1}_I\|_{L^2} + \|(\breve{f}_{k_n} - f_{k_n})\mathbf{1}_I\|_{L^2} \\
&\le 2Lk_n^{-\gamma-1/2} + 3Ln^{-1/2} + 4Ln^{-1/2} \\
&= 2Lk_n^{-\gamma-1/2} + 7Ln^{-1/2}
\end{aligned}$$

for every segment I of f_{k_n}. Thus, for sufficiently large n

$$\begin{aligned}
&T_{\mathcal{I}}(y^n; f_{k_n}) \\
&\quad\le \max_{\substack{I\in\mathcal{I} \\ f_{k_n}\equiv c_I \text{ on } I}} \frac{1}{\sqrt{n|I|}} \Big| \sum_{i/n\in I} (\bar{f}_i^n - c_I) \Big| + \max_{I\in\mathcal{I}} \frac{1}{\sqrt{n|I|}} \Big| \sum_{i/n\in I} (y_i^n - \bar{f}_i^n) \Big| - s_I \\
&\quad\le \max_{\substack{I\in\mathcal{I} \\ f_{k_n}\equiv c_I \text{ on } I}} \sqrt{\frac{n}{|I|}} \int_I |f_0(t) - f_{k_n}(t)| dt + a_0\sqrt{\log n} \\
&\quad\le \max_{\substack{I\in\mathcal{I} \\ f_{k_n}\equiv c_I \text{ on } I}} \sqrt{n} \|(f_0 - f_{k_n})\mathbf{1}_I\|_{L^2} + a_0\sqrt{\log n} \\
&\quad\le 2n^{1/2}k_n^{-\gamma-1/2}L + 7L + a_0\sqrt{\log n} \le \eta = a\sqrt{\log n}.
\end{aligned}$$

That is, f_{k_n} lies in the constraint of (4). Thus, $\#J(\hat{f}_n) \le \#J(f_{k_n}) \le 2k_n$. Then, with the same argument as in the proof of Theorem 1 for (17), we obtain

$$\|\hat{f}_n - f_{k_n}\|_{L^p}^r \le 2^{r(1+1/p)} \left(\frac{(2k_n+1)c}{n}\right)^{r/2 \wedge r/p} \left(4(a+\delta)^2 \log n\right)^{r/2}.$$

Then, as $n \to \infty$

$$\begin{aligned} \|\hat{f}_n - f_0\|_{L^p}^r &\le 2^{(r-1)_+} \left(\|\hat{f}_n - f_{k_n}\|_{L^p}^r + \|f_{k_n} - f_0\|_{L^p}^r\right) \\ &\le \mathcal{O}\left((\log n)^{r/2} \left(n^{-1}k_n\right)^{r/p \wedge r/2}\right) \\ &= \mathcal{O}\left((\log n)^{\frac{\gamma+(1/2-1/p)_+}{2\gamma+1}} n^{-\frac{2\gamma}{2\gamma+1}(1/2 \wedge 1/p)}\right). \end{aligned} \tag{21}$$

b) *Rates of convergence.* As in (18), we have

$$\begin{aligned} \mathbb{P}\{\mathcal{G}_n^c\} &\le \mathbb{P}\left\{\max_{I \in \mathcal{I}} \frac{1}{\sqrt{n|I|}} \Big| \sum_{i/n \in I} \xi_i^n \Big| > (a_0 - \delta)\sqrt{\log n}\right\} \\ &\le 2\exp\left(-\frac{(a_0-\delta)^2 \log n}{2\sigma^2}\right) \frac{n^2}{2} \\ &= n^{-\frac{(a_0-\delta)^2}{2\sigma^2}+2} = n^{-r_0} \le n^{-r}. \end{aligned}$$

This together with (21) implies the rate of almost sure convergence.

As in part (i) c) of the proof of Theorem 1, we drive from (21) that

$$\begin{aligned} &\mathbb{E}\left[\|\hat{f}_n - f_0\|_{L^p}^r\right] \\ &= \mathbb{E}\left[\|\hat{f}_n - f_0\|_{L^p}^r; \mathcal{G}_n\right] + \mathbb{E}\left[\|\hat{f}_n - f_0\|_{L^p}^r; \mathcal{G}_n^c\right] \\ &\le \mathbb{E}\left[\|\hat{f}_n - f_0\|_{L^p}^r; \mathcal{G}_n\right] + 2^{r/p} n^{r/2} \mathbb{P}\{\mathcal{G}_n^c\} \\ &\quad + \int_{2n^{p/2}}^{\infty} \mathbb{P}\left\{\|\hat{f}_n - f_0\|_{L^p}^p \ge u\right\} \frac{r}{p} u^{r/p-1} du \\ &\le \mathcal{O}\Big((\log n)^{\frac{\gamma+(1/2-1/p)_+}{2\gamma+1}} n^{-\frac{2\gamma}{2\gamma+1}(1/2 \wedge 1/p)}\Big) + \mathcal{O}\big(n^{-r/2}\big) \\ &= \mathcal{O}\Big((\log n)^{\frac{\gamma+(1/2-1/p)_+}{2\gamma+1}} n^{-\frac{2\gamma}{2\gamma+1}(1/2 \wedge 1/p)}\Big), \end{aligned}$$

as $n \to \infty$, which shows the rate of convergence in expectation. □

A.3. Feature detection

The proofs of Theorems 3 and 4 rely on the following lemma.

Lemma 1. *Under model (1) with the truth $f_0 \in \mathcal{D}$, let $\hat{f}_n$ be an MCPS method in Definition 2 with interval system $\mathcal{I}$, and $\mathcal{J}_n$ be an arbitrary collection of (possibly random) intervals.*

(i) *If* $f_0 \in \mathcal{A}^{\gamma}_{2,L}$ *for some finite* $\gamma, L > 0$, *then*

$$\lim_{n\to\infty} \mathbb{P}\left\{\max\left\{|I|^{1/2}|m_I(\hat{f}_n) - m_I(f_0)| : I \in \mathcal{J}_n\right\} \le C\left(\frac{\log n}{n}\right)^{\gamma/(2\gamma+1)}\right\} = 1,$$

where C *is a constant independent of* f_0.

(ii) *If* $\mathcal{J}_n \subseteq \mathcal{I}$, *and on each* $I \in \mathcal{J}_n$ *we have* $\hat{f}_n$ *is constant, then*

$$\mathbb{P}\left\{|I|^{1/2}|m_I(\hat{f}_n) - m_I(f_0)| \le \frac{2(\eta + s_I)}{n^{1/2}} \quad \text{for all } I \in \mathcal{J}_n\right\} \ge \mathbb{P}\left\{T_{\mathcal{I}}(\xi^n; 0) \le \eta\right\},$$

where the right hand side converges to 1 *as* $n \to \infty$.

Proof. Part (i): Note that for each $I \in \mathcal{J}_n$,

$$|I|^{1/2}|m_I(\hat{f}_n) - m_I(f_0)| \le \frac{1}{|I|^{1/2}} \int_I |\hat{f}_n(x) - f_0(x)|dx \le \frac{1}{|I|^{1/2}} |I|^{1/2} \left(\int_I |\hat{f}_n(x) - f_0(x)|^2\right)^{1/2} \le \|\hat{f}_n - f_0\|_{L^2}.$$

Then, the assertion follows from Theorem 2.

Part (ii): Assume $T_{\mathcal{I}}(\xi^n; 0) \le \eta$. Recall $\bar{f}^n_i$ in (1). Then,

$$T_{\mathcal{I}}\left(y^n; \sum_{i=0}^{n-1} \bar{f}^n_i \mathbf{1}_{[i/n,(i+1)/n)}\right) \le \eta.$$

Since $T_{\mathcal{I}}(y^n; \hat{f}_n) \le \eta$ by definition, we obtain for either $g = f_0$ or $g = \hat{f}_n$

$$|I|^{1/2}\left|m_I(g) - \frac{1}{n|I|} \sum_{j/n \in I} y^n_j\right| \le s_I + \eta \qquad \text{for any } I \in \mathcal{J}_n \subseteq \mathcal{I}.$$

Then, by a triangular inequality, $|I|^{1/2}\left|m_I(\hat{f}_n) - m_I(f_0)\right| \le 2(s_I + \eta)$. It implies

$$\{T_{\mathcal{I}}(\xi^n; 0) \le \eta\} \subseteq \left\{|I|^{1/2}|m_I(\hat{f}_n) - m_I(f_0)| \le \frac{2(\eta + s_I)}{n^{1/2}} \quad \text{for all } I \in \mathcal{J}_n\right\},$$

which shows the assertion. By (18), it holds that $\lim_{n\to\infty} \mathbb{P}\{T_{\mathcal{I}}(\xi^n; 0)\} = 1$. □

Proof of Theorem 3. Part (i): We consider only the case of modes, since the case of troughs follows if we replace f_0 by $-f_0$. By definition of modes and the right continuity of f_0, we can select $\mathcal{J}_n$ as a fixed collection of intervals that capture the modes of f_0. That is, $\mathcal{J}_n := \{I_0, I_1, \ldots, I_{2k}\}$ with $k = \#\text{mode}(f_0)$ such that $I_0 < I_1 < \cdots < I_{2k}$ and $m_{I_{2i-1}}(f_0) > m_{I_{2i-2}}(f_0) \vee m_{I_{2i}}(f_0)$ for each $i = 1, \ldots, k$. By Lemma 1 (i), we have

$$\mathbb{P}\left\{\max\left\{|I|^{1/2}|m_I(\hat{f}_n) - m_I(f)| : I \in \mathcal{J}_n\right\} \to 0\right\}$$

$$\geq \limsup_{n\to\infty} \mathbb{P}\left\{\max\left\{|I|^{1/2}|m_I(\hat{f}_n)-m_I(f)| : I\in\mathcal{J}_n\right\} \leq C\Big(\frac{\log n}{n}\Big)^{\gamma/(2\gamma+1)}\right\} = 1.$$

It implies $m_{I_{2i-1}}(\hat{f}_n) > m_{I_{2i-2}}(\hat{f}_n)\vee m_{I_{2i}}(\hat{f}_n)$, $i=1,\dots,k$, for sufficiently large n, which shows the assertion.

Part (ii): Now we set $\mathcal{J}_n := \big\{[x,x+\lambda_n^\varepsilon), [x-\lambda_n^\varepsilon, x) : x\in J_\varepsilon(f_0)\big\}$ with

$$\lambda_n^\varepsilon := \min\big\{d\big(J(\hat{f}_n), J_\varepsilon(f_0)\big),\, \delta_n\big\}$$

for some positive $\delta_n\to 0$ arbitrarily slow. For $x\in J_\varepsilon(f_0)$, note that $\hat{f}_n$ is constant on $[x-\lambda_n^\varepsilon, x+\lambda_n^\varepsilon)$, since $d\big(J(\hat{f}_n), J_\varepsilon(f_0)\big) \geq \lambda_n^\varepsilon$. This in particular implies $m_{[x-\lambda_n^\varepsilon,x)}(\hat{f}_n) = m_{[x,x+\lambda_n^\varepsilon)}(\hat{f}_n)$. Moreover, as $\lambda_n^\varepsilon\to 0$, from the definition of $\Delta_{f_0}^\varepsilon$ and $f_0\in\mathcal{D}$ it follows for sufficiently large n

$$\big|m_{[x-\lambda_n^\varepsilon,x)}(f_0) - m_{[x,x+\lambda_n^\varepsilon)}(f_0)\big| \geq \frac{1}{2}\Delta_{f_0}^\varepsilon \qquad \text{for all } x\in J_\varepsilon(f_0). \tag{22}$$

We claim that for each $x\in J_\varepsilon(f_0)$ there exists $I_x = [x,x+\lambda_n^\varepsilon)$ or $[x-\lambda_n^\varepsilon, x)$ such that $|m_{I_x}(f_0) - m_{I_x}(\hat{f}_n)| \geq \Delta_{f_0}^\varepsilon/4$. Otherwise, if $|m_{I_x}(f_0) - m_{I_x}(\hat{f}_n)| < \Delta_{f_0}^\varepsilon/4$ holds for both $I_x = [x,x+\lambda_n^\varepsilon)$ and $[x-\lambda_n^\varepsilon, x)$, then it leads to

$$\big|m_{[x-\lambda_n^\varepsilon,x)}(f_0) - m_{[x,x+\lambda_n^\varepsilon)}(f_0)\big| < \Delta_{f_0}^\varepsilon/2\,,$$

which contradicts with (22). Thus, by Lemma 1 (i), it holds, as $n\to\infty$,

$$\mathbb{P}\left\{\frac{\Delta_{f_0}^\varepsilon}{4} \leq |m_{I_x}(f_0)-m_{I_x}(\hat{f}_n)| \leq \frac{C}{\sqrt{\lambda_n^\varepsilon}}\Big(\frac{\log n}{n}\Big)^{\gamma/(2\gamma+1)} \quad \text{for all } x\in J_\varepsilon(f_0)\right\} \to 1.$$

It implies $\lambda_n^\varepsilon \leq 16C^2(\Delta_{f_0}^\varepsilon)^{-2}(\log n/n)^{2\gamma/(2\gamma+1)}$ almost surely, as $n\to\infty$. By letting $\delta_n\to 0$ slower than $(\log n/n)^{2\gamma/(2\gamma+1)}$, we obtain

$$\lim_{n\to\infty}\mathbb{P}\left\{d\big(J(\hat{f}_n), J_\varepsilon(f_0)\big) \leq \frac{16C^2}{(\Delta_{f_0}^\varepsilon)^2}\Big(\frac{\log n}{n}\Big)^{2\gamma/(2\gamma+1)}\right\} = 1,$$

and then $\lim_{n\to\infty}\mathbb{P}\big\{\#J(\hat{f}_n) \geq \#J_\varepsilon(f_0)\big\} = 1$. Thus, the assertion holds. □

Proof of Theorem 4. By Lemma 1 (ii), we have

$$\mathbb{P}\Big\{|I_i|^{1/2}|m_{I_i}(\hat{f}_n) - m_{I_i}(f_0)| \leq \frac{2(\eta+s_{I_i})}{n^{1/2}} \text{ for } i=1,2\Big\} \geq \mathbb{P}\left\{T_{\mathcal{I}}(\xi^n;0) \leq \eta(\beta)\right\} \geq 1-\beta.$$

Note that $|I_i|^{1/2}|m_{I_i}(\hat{f}_n) - m_{I_i}(f_0)| \leq \frac{2(\eta+s_{I_i})}{n^{1/2}}$ for $i=1,2$ and (12) imply

$$m_{I_1}(f_0) - m_{I_2}(f_0)$$

$$\geq m_{I_1}(\hat{f}_n) - m_{I_2}(\hat{f}_n) - \sum_{i=1}^{2} |m_{I_i}(\hat{f}_n) - m_{I_i}(f_0)|$$
$$> \frac{2\big(\eta(\beta) + s_{I_1}\big)}{\sqrt{n|I_1|}} + \frac{2\big(\eta(\beta) + s_{I_2}\big)}{\sqrt{n|I_2|}} - \sum_{i=1}^{2} |m_{I_i}(\hat{f}_n) - m_{I_i}(f_0)| \geq 0.$$

This concludes the proof. □

Remark 11. From the proof above, one can easily see that if $m_{I_1}(\hat{f}_n) - m_{I_2}(\hat{f}_n) > r_{I_1} + r_{I_2} + \theta$ for some $\theta \in \mathbb{R}$, then it holds with probability $\geq 1 - \beta$ that $m_{I_1}(f_0) - m_{I_2}(f_0) > \theta$.

A.4. Proof of Theorem 5

For simplicity, we assume that the noises ξ_i^n have homogeneous variance σ_0^2, since for the general case it is obvious to modify the following proof accordingly. For every $\tau \equiv (\tau_0, \tau_1, \ldots, \tau_k) \in \Pi_n$, we define $\#\tau := k$, and by elementary calculation obtain

$$\mathbb{E}[\|\hat{f}_{\tau,n} - f_0\|_{L^2}^2] = \|s_\tau - f_0\|_{L^2}^2 + \frac{\#\tau}{n}\sigma_0^2$$

where s_τ is the best L^2-approximant of f with change-points specified by τ. Define $\tau_* \equiv \tau_*(n) \in \Pi_n$ such that $\mathbb{E}[\|\hat{f}_{\tau_*,n} - f_0\|_{L^2}^2] = \min_{\tau\in\Pi_n} \mathbb{E}[\|\hat{f}_{\tau,n} - f_0\|_{L^2}^2]$. Now we claim that there exists a constant C satisfying

$$\|s_{\tau_*} - f_0\|_{L^2}^2 \leq C\frac{\#\tau_*}{n}\sigma_0^2 \qquad \text{for sufficiently large } n. \tag{23}$$

To prove the claim (23), we, anticipating contradiction, assume that

$$\limsup_{n\to\infty} \frac{n\|s_{\tau_*} - f_0\|_{L^2}^2}{\#\tau_*\sigma_0^2} = \infty.$$

One can choose $m \equiv m(n)$ such that $\limsup_{n\to\infty} n\|s_{\tau_*} - f_0\|_{L^2}^2 (m\#\tau_*\sigma_0^2)^{-1} = \infty$, and $\lim_{n\to\infty} m = \infty$. Define υ_* as $\|s_{\upsilon_*} - f_0\|_{L^2} = \min_{\upsilon\in U_{\tau_*,m}} \|s_\upsilon - f_0\|_{L^2}$ with

$$U_{\tau_*,m} := \left\{\upsilon \in \Pi_n : \upsilon \equiv (0, \upsilon_1^1, \ldots, \upsilon_m^1 \equiv \tau_*^1, \ldots, \upsilon_1^k, \ldots, \upsilon_m^k \equiv \tau_*^k)\right\},$$

where $\tau_* \equiv (0, \tau_*^1, \ldots, \tau_*^k)$. It follows from $m \to \infty$ and $f_0 \in \mathcal{A}_2^\gamma \cap L^\infty$ for some γ that $\|s_{\upsilon_*} - f_0\|_{L^2} / \|s_{\tau_*} - f_0\|_{L^2} \to 0$. Then we obtain

$$\limsup_{n\to\infty} \frac{\mathbb{E}[\|\hat{f}_{\tau_*,n} - f_0\|_{L^2}^2]}{\mathbb{E}[\|\hat{f}_{\upsilon_*,n} - f_0\|_{L^2}^2]} \geq \limsup_{n\to\infty} \frac{\|s_{\tau_*} - f_0\|_{L^2}^2}{\|s_{\upsilon_*} - f_0\|_{L^2}^2 + m\#\tau_*\sigma_0^2/n} = \infty,$$

which contradicts the definition of τ_*.

Denote $L := \|f_0\|_{L^\infty}$. Similar to part a) in the proof of Theorem 2, one can construct a step function $\tilde{s}_{\tau_*}$, by adding another $\#\tau_*$ change-points to s_{τ_*} and

later shifting all the change-points to the grid points i/n, such that $\#J(\tilde{s}_{\tau_*}) \leq 2(\#\tau_* - 1)$, $\|\tilde{s}_{\tau_*} - f_0\|_{L^2}^2 \leq 2\|s_{\tau_*} - f_0\|_{L^2}^2 + 2n^{-1}\#\tau_* L^2$, and $\|(\tilde{s}_{\tau_*} - f_0)\mathbf{1}_I\|_{L^2}^2 \leq 2(\#\tau_*)^{-1}\|s_{\tau_*} - f_0\|_{L^2}^2 + 2n^{-1}L^2$ for each segment I of $\tilde{s}_{\tau_*}$.

Assume now the "good noise" case, namely, event $\mathcal{G}_n$ in (20) holds true. Then we have for sufficiently large n,

$$\begin{aligned}
&T_{\mathcal{I}}(y^n; \tilde{s}_{\tau_*}) \\
&\quad \leq \max_{\substack{I\in\mathcal{I} \\ \tilde{s}_{\tau_*}\equiv c_I \text{ on } I}} \frac{1}{\sqrt{n|I|}}\Big|\sum_{i/n\in I}(\bar{f}_i^n - c_I)\Big| + \max_{I\in\mathcal{I}} \frac{1}{\sqrt{n|I|}}\Big|\sum_{i/n\in I}(y_i^n - \bar{f}_i^n)\Big| - s_I \\
&\quad \leq \max_{\substack{I\in\mathcal{I} \\ \tilde{s}_{\tau_*}\equiv c_I \text{ on } I}} \sqrt{n}\|(f_0 - \tilde{s}_{\tau_*})\mathbf{1}_I\|_{L^2} + a_0\sqrt{\log n} \\
&\quad \leq \sqrt{2n(\#\tau_*)^{-1}\|s_{\tau_*} - f_0\|_{L^2}^2 + 2L^2} + a_0\sqrt{\log n} \\
&\quad \leq \sqrt{2C\sigma_0^2 + 2L^2} + a_0\sqrt{\log n} \leq a\sqrt{\log n},
\end{aligned}$$

where C is the constant in (23), and $\bar{f}_i^n$ in (1). Again following similar lines of part a) in the proof of Theorem 2, one can obtain

$$\|\hat{f}_n - \tilde{s}_{\tau_*}\|_{L^2}^2 \leq 32(a+\delta)^2 c\log n\frac{\#\tau_*}{n}(1+o(1)).$$

It further follows that

$$\begin{aligned}
\|\hat{f}_n - f_0\|_{L2}^2 &\leq 2\|f_0 - \tilde{s}_{\tau_*}\|_{L^2}^2 + 2\|\hat{f}_n - \tilde{s}_{\tau_*}\|_{L^2}^2 \\
&\leq 4\|s_{\tau_*} - f_0\|_{L^2}^2 + 4L^2\frac{\tau_*}{n} + 64(a+\delta)^2 c\log n\frac{\#\tau_*}{n}(1+o(1)).
\end{aligned}$$

Thus, under event $\mathcal{G}_n$, we obtain for large enough n

$$\|\hat{f}_n - f_0\|_{L^2}^2 \leq \tilde{C}\log n\big(\|s_{\tau_*} - f_0\|_{L^2}^2 + \frac{\#\tau_*}{n}\sigma_0^2\big) \leq \tilde{C}\log n\mathbb{E}[\|\hat{f}_{\tau_*,n} - f_0\|_{L^2}^2],$$

where $\tilde{C}$ is a constant independent of f_0. □

Acknowledgements

Li is supported by RTG 2088 subproject B2, and CRC 937 subproject A10, and Guo is supported by CRC 803 subproject Z02. Munk and Li are supported by the DFG Cluster of Excellence, EXC 2067/1 - 390729940. The authors would like to thank an associate editor and two reviewers for helpful comments.

References

Abramovich, F., Antoniadis, A. and Pensky, M. (2007). Estimation of piecewise-smooth functions by amalgamated bridge regression splines. *Sankhyā* **69** 1–27. MR2385276

Antoch, J. and Hušková, M. (2000). Bayesian-type estimators of change points. *J. Statist. Plann. Inference* **91** 195–208. Prague Workshop on Perspectives in Modern Statistical Inference: Parametrics, Semi-parametrics, Non-parametrics (1998). MR1814780

Aue, A., Cheung, R. C. Y., Lee, T. C. M. and Zhong, M. (2014). Segmented model selection in quantile regression using the minimum description length principle. *J. Amer. Statist. Assoc.* **109** 1241–1256. MR3265694

Behr, M., Holmes, C. and Munk, A. (2018). Multiscale blind source separation. *Ann. Statist.* **46** 711–744. MR3782382

Behr, M. and Munk, A. (2017). Identifiability for blind source separation of multiple finite alphabet linear mixtures. *IEEE Trans. Inform. Theory* **63** 5506–5517. MR3688042

Bellman, R. (1957). *Dynamic Programming*. Princeton University Press, Princeton, NJ, USA. MR0090477

Billingsley, P. (1999). *Convergence of Probability Measures*, second ed. *Wiley Series in Probability and Statistics: Probability and Statistics*. John Wiley & Sons, Inc., New York. A Wiley-Interscience Publication. MR1700749

Boneva, L. I., Kendall, D. and Stefanov, I. (1971). Spline transformations: Three new diagnostic aids for the statistical data-analyst. *J. Roy. Statist. Soc. Ser. B* **33** 1–70. MR0288888

Boysen, L., Kempe, A., Liebscher, V., Munk, A. and Wittich, O. (2009). Consistencies and rates of convergence of jump-penalized least squares estimators. *Ann. Statist.* **37** 157–183. MR2488348

Braun, J. V., Braun, R. K. and Mueller, H. G. (2000). Multiple change-point fitting via quasilikelihood, with application to DNA sequence segmentation. *Biometrika* **87** 301–314. MR1782480

Burchard, H. G. and Hale, D. F. (1975). Piecewise polynomial approximation on optimal meshes. *J. Approximation Theory* **14** 128–147. MR0374761

Cai, T. T., Jeng, X. J. and Li, H. (2012). Robust detection and identification of sparse segments in ultrahigh dimensional data analysis. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 773–797. MR2988906

Chan, H.-P. and Chen, H. (2017). Multi-sequence segmentation via score and higher-criticism tests. arXiv:1706.07586.

Chan, H. P. and Walther, G. (2013). Detection with the scan and the average likelihood ratio. *Statist. Sinica* **23** 409–428. MR3076173

Chen, H. and Zhang, N. (2015). Graph-based change-point detection. *Ann. Statist.* **43** 139–176. MR3285603

Davies, L., Höhenrieder, C. and Krämer, W. (2012). Recursive computation of piecewise constant volatilities. *Comput. Stat. Data Anal.* **56** 3623–3631. MR2943916

Davies, P. L. and Kovac, A. (2001). Local extremes, runs, strings and multiresolution. *Ann. Statist.* **29** 1–65. With discussion and rejoinder by the authors. MR1833958

del Alamo, M., Li, H. and Munk, A. (2018). Frame-constrained total variation regularization for white noise regression. arXiv:1807.02038.

DeVore, R. A. (1998). Nonlinear approximation. In *Acta Numerica, 1998*.

Acta Numer. **7** 51–150. Cambridge Univ. Press, Cambridge. MR1689432

DeVore, R. A. and Lorentz, G. G. (1993). *Constructive Approximation. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]* **303**. Springer-Verlag, Berlin. MR1261635

Diskin, S. J., Li, M., Hou, C., Yang, S., Glessner, J., Hakonarson, H., Bucan, M., Maris, J. M. and Wang, K. (2008). Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res.* **36** e126.

Donoho, D. L. (1988). One-Sided inference about functionals of a density. *Ann. Statist.* **16** 1390–1420. MR0964930

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455. MR1311089

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: asymptopia? *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **57** 301–369. With discussion and a reply by the authors. MR1323344

Du, C., Kao, C.-L. M. and Kou, S. C. (2016). Stepwise signal extraction via marginal likelihood. *J. Amer. Statist. Assoc.* **111** 314–330. MR3494662

Dümbgen, L. and Spokoiny, V. G. (2001). Multiscale testing of qualitative hypotheses. *Ann. Statist.* **29** 124–152. MR1833961

Fang, X., Li, J. and Siegmund, D. (2019). Segmentation and estimation of change-point models: false positive control and confidence regions. *Ann. Statist.* To appear.

Farcomeni, A. (2014). Discussion of "Multiscale change-point inference". *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 546–547.

Frick, K., Munk, A. and Sieling, H. (2014). Multiscale change point inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 495–580. With 32 discussions by 47 authors and a rejoinder by the authors. MR3210728

Fryzlewicz, P. (2007). Unbalanced Haar technique for nonparametric function estimation. *J. Amer. Statist. Assoc.* **102** 1318–1327. MR2412552

Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *Ann. Statist.* **42** 2243–2281. MR3269979

Futschik, A., Hotz, T., Munk, A. and Sieling, H. (2014). Multiresolution DNA partitioning: statistical evidence for segments. *Bioinformatics* **30** 2255–2262.

Gao, C., Han, F. and Zhang, C.-H. (2019). On estimation of isotonic piecewise constant signals. *Ann. Statist.* To appear.

Grasmair, M., Li, H. and Munk, A. (2018). Variational multiscale nonparametric regression: Smooth functions. *Ann. Inst. Henri Poincaré Probab. Stat.* **54** 1058–1097. MR3795077

Hall, P. and Marron, J. S. (1990). On variance estimation in nonparametric regression. *Biometrika* **77** 415–419. MR1064818

Han, Q. and Wellner, J. A. (2019). Convergence rates of least squares regression estimators with heavy-tailed errors. *Ann. Statist.* **47** 2286–2319. MR3953452

Harchaoui, Z. and Lévy-Leduc, C. (2008). Catching change-points with

lasso. *Adv. in Neur. Inform. Processing Syst.* **20** 161–168.

Harchaoui, Z. and Lévy-Leduc, C. (2010). Multiple change-point estimation with a total variation penalty. *J. Amer. Statist. Assoc.* **105** 1480–1493. MR2796565

Has'minskiĭ, R. Z. (1978). A lower bound on the risks of nonparametric estimates of densities in the uniform metric. *Theory Probab. Appl.* **23** 794–798.

Haynes, K., Eckley, I. A. and Fearnhead, P. (2017). Computationally efficient changepoint detection for a range of penalties. *J. Comput. Graph. Statist.* **26** 134–143. MR3610414

Hotz, T., Schütte, O. M., Sieling, H., Polupanow, T., Diederichsen, U., Steinem, C. and Munk, A. (2013). Idealizing ion channel recordings by jump segmentation and statistical multiresolution analysis. *IEEE Trans. Nanobiosci.* **12** 376–386.

Hušková, M. and Antoch, J. (2003). Detection of structural changes in regression. *Tatra Mt. Math. Publ.* **26** 201–215. Probastat '02. Part II. MR2055177

Ibragimov, I. A. and Has'minskiĭ, R. Z. (1977). On the estimation of an infinite-dimensional parameter in Gaussian white noise. *Sov. Math. Dokl.* **18** 1307–1309.

Ibragimov, I. A. and Has'minskiĭ, R. Z. (1981). *Statistical Estimation. Applications of Mathematics* **16**. Springer-Verlag, New York-Berlin Asymptotic theory, Translated from the Russian by Samuel Kotz. MR0620321

Kabluchko, Z. (2007). Extreme-value analysis of standardized Gaussian increments. arXiv:0706.1849.

Killick, R., Fearnhead, P. and Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *J. Amer. Statist. Assoc.* **107** 1590–1598. MR3036418

Korostelev, A. and Korosteleva, O. (2011). *Mathematical Statistics. Graduate Studies in Mathematics* **119**. American Mathematical Society, Providence, RI Asymptotic minimax theory. MR2767163

Korte, B. and Vygen, J. (2012). *Combinatorial Optimization*, fifth ed. *Algorithms and Combinatorics* **21**. Springer, Heidelberg. Theory and algorithms. MR2850465

Lai, W. R., Johnson, M. D., Kucherlapati, R. and Park, P. J. (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* **21** 3763–3770.

Li, H., Munk, A. and Sieling, H. (2016). FDR-control in multiscale changepoint segmentation. *Electron. J. Stat.* **10** 918–959. MR3486421

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

Linton, O. and Seo, M. H. (2014). Discussion of "Multiscale change-point inference". *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 548. MR3210728

Maidstone, R., Hocking, T., Rigaill, G. and Fearnhead, P. (2016). On optimal multiple changepoint algorithms for large data. *Stat. Comput.* 1–15. MR3599687

MAMMEN, E. and VAN DE GEER, S. (1997). Locally adaptive regression splines. *Ann. Statist.* **25** 387–413. MR1429931

MÜLLER, H.-G. and STADTMÜLLER, U. (1987). Estimation of heteroscedasticity in regression analysis. *Ann. Statist.* **15** 610–625. MR0888429

MUNK, A., BISSANTZ, N., WAGNER, T. and FREITAG, G. (2005). On difference-based variance estimation in nonparametric regression when the covariate is high dimensional. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 19–41. MR2136637

NEMIROVSKI, A. (1985). Nonparametric estimation of smooth regression functions. *Izv. Akad. Nauk. SSR Teckhn. Kibernet. (in Russian)* **3** 50–60. *J. Comput. System Sci.*, 23:1–11, 1986 (in English). MR0844292

NEMIROVSKI, A. (2000). Topics in non-parametric statistics. In *Lectures on Probability Theory and Statistics (Saint-Flour, 1998). Lecture Notes in Math.* **1738** 85–277. Springer, Berlin. MR1775640

OLSHEN, A. B., VENKATRAMAN, E. S., LUCITO, R. and WIGLER, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5** 557–572.

PEIN, F., SIELING, H. and MUNK, A. (2017). Heterogeneous change point inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1207–1227. MR3689315

PETRUSHEV, P. P. (1988). Direct and converse theorems for spline and rational approximation and Besov spaces. In *Function Spaces and Applications (Lund, 1986). Lecture Notes in Math.* **1302** 363–377. Springer, Berlin. MR0942281

RIVERA, C. and WALTHER, G. (2013). Optimal detection of a jump in the intensity of a Poisson process or in a density with likelihood ratio statistics. *Scand. J. Stat.* **40** 752–769. MR3145116

SCHWARTZMAN, A., GAVRILOV, Y. and ADLER, R. J. (2011). Multiple testing of local maxima for detection of peaks in 1D. *Ann. Statist.* **39** 3290–3319. MR3012409

SCOTT, A. J. and KNOTT, M. (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics* **30** 507–512.

SHAO, Q. M. (1995). On a conjecture of Révész. *Proc. Amer. Math. Soc.* **123** 575–582. MR1231304 (95c:60031)

SIEGMUND, D. (2013). Change-points: from sequential detection to biology and back. *Sequential Anal.* **32** 2–14. MR3023983

SIEGMUND, D. and VENKATRAMAN, E. S. (1995). Using the generalized likelihood ratio statistic for sequential detection of a change-point. *Ann. Statist.* **23** 255–271. MR1331667

SIEGMUND, D. and YAKIR, B. (2000). Tail probabilities for the null distribution of scanning statistics. *Bernoulli* **6** 191–213. MR1748719 (2001e:62036)

SONG, R., BANERJEE, M. and KOSOROK, M. R. (2016). Asymptotics for change-point models under varying degrees of mis-specification. *Ann. Statist.* **44** 153–182. MR3449765

SPOKOINY, V. G. (1998). Estimation of a function with discontinuities via local polynomial fit with an adaptive window choice. *Ann. Statist.* **26** 1356–1378. MR1647669

SPRAUL, M., NEIDIG, P., KLAUCK, U., KESSLER, P., HOLMES, E., NICHOL-

SON, J. K., SWEATMAN, B. C., SALMAN, S. R., FARRANT, R. D., RAHR, E., BEDDELL, C. R. and LINDON, J. C. (1994). Automatic reduction of NMR spectroscopic data for statistical and pattern recognition classification of samples. *J. Pharm. Biomed. Anal.* **12** 1215–1225.

TECUAPETLA-GÓMEZ, I. and MUNK, A. (2017). Autocovariance estimation in regression with a discontinuous signal and m-dependent errors: a difference-based approach. *Scand. J. Stat.* **44** 346–368. MR3658518

TIBSHIRANI, R. and WANG, P. (2008). Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics* **9** 18–29.

TSYBAKOV, A. (2009). *Introduction to Nonparametric Estimation*. Springer-Verlag, New York. MR2724359

TUKEY, J. W. (1961). Curves as parameters, and touch estimation. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I* 681–694. Univ. California Press, Berkeley, Calif. MR0132677

WALTHER, G. (2010). Optimal and fast detection of spatial clusters with scan statistics. *Ann. Statist.* **38** 1010–1033. MR2604703

YAO, Y.-C. and AU, S. T. (1989). Least-squares estimation of a step function. *Sankhyā Ser. A* **51** 370–381. MR1175613

ZHANG, N. R. and SIEGMUND, D. O. (2007). A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics* **63** 22–32. MR2345571

ZHANG, N. R. and SIEGMUND, D. O. (2012). Model selection for high-dimensional, multi-sequence change-point problems. *Statist. Sinica* **22** 1507–1538. MR3027097