Multiscale Change-point Segmentation: Beyond Step Functions

Dissertation

zur Erlangung des mathematisch-naturwissenschaftlichen
Doktorgrades
"Doctor rerum naturalium"
der Georg-August-Universität Göttingen

im Promotionsprogramm
PhD School of Mathematical Sciences (SMS)
der Georg-August University School of Science (GAUSS)

vorgelegt von

Qinghai Guo

aus Jiangxi, China

Göttingen, 2017

Betreuungsausschuss:

Prof. Dr. Axel Munk,
Institut für Mathematische Stochastik, Universität Göttingen

Jun. -Prof. Dr. Andrea Krajina,
Institut für Mathematische Stochastik, Universität Göttingen

Mitglieder der Prüfungskommission:

Referent:
Prof. Dr. Axel Munk,
Institut für Mathematische Stochastik, Universität Göttingen

Korreferent:
Jun. -Prof. Dr. Andrea Krajina,
Institut für Mathematische Stochastik, Universität Göttingen

Weitere Mitglieder der Prüfungskommission:

Dr. Michael Habeck,
Institut für Mathematische Stochastik, Universität Göttingen

Prof. Dr. Stephan Huckemann,
Institut für Mathematische Stochastik, Universität Göttingen

Prof. Dr. Russell Luke,
Institut für Numerische und Angewandte Mathematik, Universität Göttingen

Prof. Dr. Chenchang Zhu,
Mathematisches Institut, Universität Göttingen

Tag der mündlichen Prüfung: 03.02.2017

Acknowledgement

First of all, I would like to express my very great appreciation to my principal supervisor Prof. Axel Munk for introducing me into the research of mathematical statistics, and providing the interesting and challenging topic of this work. His guidance and enthusiasm have always been a great encouragement throughout my work, and his stimulating contributions were also fundamental to this work. I benefit a lot from his great statistical intuition and deep understanding on a wide range of areas of mathematics. Further, I would like to thank my second advisor Jun.-Prof Andrea Krajina, for many assistances during my PhD study, and for proofreading of this work and providing many helpful comments.

Special thanks should be given to Dr. Housen Li, for his extraordinary assistance with this work, as well as his patient help and encouragement from the first day of my PhD study.

I am grateful to Florian Pein, for proofreading of this work and for many helpful discussions.

I wish to express my gratitude to all the members at the IMS, for providing a pleasant working circumstance at the IMS. Special thanks should be given to Merle Behr, for proofreading and comments of this work.

The financial support by the SFB 803 "Functionality controlled by organization in and between membranes" is gratefully acknowledged.

Finally, I would like to express my deep appreciation to my family and my girlfriend, Xiao Yang, for their constant support, understanding, and encouragement.

Summary

Many multiscale segmentation methods have been proven to work successfully for detecting multiple change-points, mainly because they provide faithful statistical statements, while at the same time allowing for efficient computation. Underpinning theory has been studied exclusively for models which assume that the signal is an unknown step function. However, when the signal is only approximately piecewise constant, which often occurs in practical applications, the behavior of multiscale segmentation methods is still not well studied. To narrow this gap, we investigate the asymptotic properties of a certain class of *multiscale change-point segmentation* methods in a general nonparametric regression setting.

The main contribution of this work is the adaptation property of these methods over a wide range of function classes, although they are designed for step functions. On the one hand, this includes the optimal convergence rates (up to log-factor) for step functions with bounded or even increasing to infinite number of jumps. On the other hand, for models beyond step functions, which are characterized by certain approximation spaces, we show the optimal rates (up to log-factor) as well. This includes bounded variation functions and (piecewise) Hölder functions of smoothness order $0 < \alpha \leq 1$. All results are formulated in terms of L^p-loss, $0 < p < \infty$, both almost surely and in expectation. In addition, we show that the convergence rates readily imply accuracy of feature detection, such as change-points, modes, troughs, etc. The practical performance is examined by various numerical simulations.

Contents

List of Symbols

$\#S$	The number of elements in set S
$\mathcal{A}^\gamma$	Certain approximation spaces with order γ
$\|I\|$	The Lebesgue measure of set I
$\mathcal{D}([0,1))$	The class of càdlàg functions on $[0,1)$
$\Gamma(\cdot)$	The approximation error
$\|\cdot\|_{L^p}$	The L^p-norm w.r.t. the Lebesgure measure
$\mathcal{S}([0,1))$	The class of piecewise constant functions on $[0,1)$
$BV([0,1))$	The bounded variation classes on $[0,1)$
$E(X)$	The expectation of X
$H^\alpha([0,1))$	The Hölder function classes with order α on $[0,1)$
$J(f)$	The set of change-points of a step function f

1 Introduction

We assume that the observations are given through the general regression model

$$y_i^n = f\Big(\frac{i}{n}\Big) + \xi_i^n, \qquad i = 0, \dots, n-1, \tag{1.1}$$

where $\xi^n = (\xi_0^n, \dots, \xi_{n-1}^n)$ are independent centered sub-Gaussian random variables.

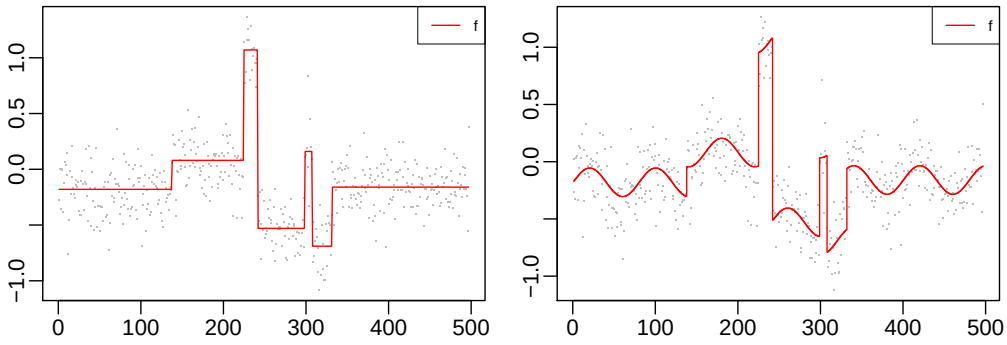

Figure 1.1: Examples of a regression step function (left) and a non-step function (right) with Gaussian noise

When f is a piecewise constant function with an unknown number of change-points (see e.g. Figure 1.1 left), model (1.1) is often referred to as *change-point regression model*, the related (non-parametric) problem turns into estimating the number and locations of change-points, as well as the function value on each constant interval. The corresponding study has a long and rich history in the statistical literature (see Basseville and Nikiforov, 1993; Brodsky and Darkhovsky, 1993; Csörgö and Horváth, 1997; Chen and Gupta, 2000; Lai, 2001; Wu, 2005, for a selective survey). Recent years have witnessed a renaissance in change-point inference motivated by several applications which require fast and efficient finding of many change-points. To this end, many change-point segmentation methods have been lately proposed, which are either based on dynamic programming (Boysen et al., 2009; Killick et al., 2012; Du et al., 2015), local search (Scott and Knott, 1974; Olshen et al., 2004; Fryzlewicz, 2014), or convex optimization (Harchaoui and Lévy-Leduc, 2008; Tibshirani and Wang, 2008; Harchaoui and Lévy-Leduc, 2010). More recently, Frick et al. (2014)

introduced a *multiscale* segmentation approach, Simultaneous MUltiscale Change-point Estimator (SMUCE). SMUCE minimizes the number of change-points under a side constraint based on a simultaneous multiple testing procedure on all scales (length of subsequent observations), see Davies and Kovac (2001), Boysen et al. (2007), Pein et al. (2015) and Li et al. (2016) for related estimators. Implemented by fast dynamic programming algorithms, SMUCE and its variants were found empirically promising in various applications (see e.g. Hotz et al., 2013; Futschik et al., 2014; Behr et al., 2016).

On the other hand, in many applications a piecewise constant function is only an approximation of the underlying signal (see e.g. Figure 1.1 right). For instance, in DNA copy number analysis, a change-point regression model is commonly assumed (see e.g. Olshen et al., 2004; Lai et al., 2005), although a periodic trend distortion (known as genomic waves) exists with biological evidence (Diskin et al., 2008). In this case, i.e., when f is not piecewise constant, motivated by change-point segmentation methods, we are particularly interested in the following problems:

(i) Can we apply segmentation methods for change-point regression settings to model (1.1) when the true signal f is beyond piecewise constant? If so, how robust do these methods perform?

(ii) How well do they recover such functions? More precisely, what are their convergence rates results with respect to L^p-loss, $0 < p < \infty$?

1.1 Methodology

When the underlying signal f is in the space of càdlàg functions (right-continuous with left limits, cf. Section 2.1), following Frick et al. (2014), we introduce *multiscale change-point segmentation estimators* for model (1.1), which approximate f by a step function $\hat{f}_n$, as follows.

For a system of intervals $\mathcal{I}$, we estimate model (1.1) by solving

$$\min_{\hat{f}_n \in \mathcal{S}([0,1))} \#J(\hat{f}_n) \qquad \text{subject to } T_{\mathcal{I}}(y^n; \hat{f}_n) \leq q, \tag{1.2}$$

where $\mathcal{S}([0,1))$ is the space of right-continuous step functions, $J(f)$ is the set of change-points of f, $q \in \mathbb{R}$ is a user-specified threshold, which will be chosen later, and $T_{\mathcal{I}}(y^n; f)$ is a multiscale test statistic, where

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

with s_I a scale penalty to be defined later. Note that the solution to the optimization problem (1.2) might be *non-unique*, in which case one could pick an arbitrary solution. Recall that SMUCE from Frick et al. (2014) is an estimator of the form (1.2), Figure 1.2 shows SMUCE's estimates for some classical testing signals: Blocks, Bumps, Heavisine and Doppler (Donoho and Johnstone, 1994).

The main focus of this work is to investigate convergence rates of the estimator $\hat{f}_n$ in (1.2) with respect to L^p-loss, $0 < p < \infty$. First, we consider the situation when f is a step function but with an increasing number of change-points (probably to infinity). That is, when f is in $\mathcal{S}_L(k_n)$ with

$$\mathcal{S}_L(k_n) := \Big\{ f \in \mathcal{S}([0,1)) : \#J(f) \leq k_n, \text{ and } \|f\|_{L^\infty} \leq L \Big\},$$

for $k_n \in \mathbb{N}$ and $L > 0$.

Then, in order to investigate the convergence behavior of $\hat{f}_n$ for more general functions, we consider functions in certain approximation spaces (c.f. Section 2.3) defined by

$$\mathcal{A}^\gamma := \Big\{ f \in \mathcal{D}([0,1)) : \sup_{k \geq 1} k^\gamma \Gamma_k(f) < \infty \Big\}, \qquad \text{for } \gamma > 0, \tag{1.3}$$

where $\mathcal{D}([0,1))$ is the space of càdlàg functions (cf. Section 2.1) and $\Gamma_k(f)$ is the approximation error (c.f. Section 2.3) defined by

$$\Gamma_k(f) := \inf \Big\{ \|f - g\|_{L^\infty} : g \in \mathcal{S}([0,1)),\, \#J(g) \leq k \Big\}.$$

Furthermore, motivated by Lin et al. (2016), we show how convergence rates yield to accurate feature detection, such as change-points, modes, troughs, etc.

1.2 Related work

Although many segmentation methods have been studied in recent years, most of them require the underlying signal to lie in the step function space and some even need a fixed number of changes. Only a few are studied under slightly more general models, allowing the number of change-points to increase with number of observations, see e.g. (Zhang and Siegmund, 2012; Fryzlewicz, 2014; Li et al., 2016). In general, nothing is known for segmentation methods in the general nonparametric regression setting (1.1). Exceptions include the convergence analysis of the jump-penalized least square estimator in Boysen et al. (2009). There they proved that the Potts minimizer has a convergence rate of $(\log n/n)^{1/2}$ with respect to L^2-loss when f is a step function with bounded number of change-points.

Further, they showed a convergence rate of $(\log n/n)^{\gamma/(2\gamma+1)}$ with respect to L^2-loss when f belongs to aforementioned approximation space (1.3), and as an example, showed a convergence rate of $(\log n/n)^{\alpha/(2\alpha+1)}$ with respect to L^2-loss when f belongs to Hölder class of order α, $0 < \alpha \leq 1$. For the unbalanced Haar wavelets based estimator, Fryzlewicz (2007) proved a convergence rate of $(1/n)^{1/2} \log n$ when f is a step function with bounded number of change-points, and a convergence rate of $(1/n)^{\alpha/(2\alpha+1)} \log n$ when f belongs to Hölder class of order α, $0 < \alpha \leq 1$, both with respect to L^2-loss. Our work extends these results to a class of multiscale change-point segmentation methods.

Besides theoretical interest (cf. Linton and Seo, 2014; Farcomeni, 2014), studying models beyond piecewise constant functions is of particular practical importance (e.g. Olshen et al., 2004; Lai et al., 2005; Diskin et al., 2008). Such a study can be regarded as robustness analysis of segmentation methods against model misspecification. Our viewpoint concerns robustness against a distorted step function. This is different from focusing on locations and magnitudes of jumps for piecewise smooth functions as in Korostelev (1988), Gijbels et al. (1999) and Bigot (2005). It is also in sharp contrast to a recent work by Song et al. (2016) who considered a *reverse* scenario: a sequence of smooth functions approaches a step function in the limit.

1.3 Main results

When f in (1.1) is a step function, the theory behind multiscale segmentation methods in (1.2) is well-understood, including deviation bounds on the number and the location of change-points and optimal detection of vanishing signals. This work derives convergence rates for a sequence of piecewise constant functions with possibly increasing number of changes (see also Frick et al., 2014; Fryzlewicz, 2014). We show that under some general assumptions and an appropriate choice of the threshold q in (1.1), it holds for $0 < r < \infty$ that

$$\|\hat{f}_n - f\|_{L^p}^r = \mathcal{O}\left(\left(\frac{2k_n+1}{n}\right)^{\min\{1/p,1/2\}r} (\log)^{r/2}\right),$$

uniformly for $f \in \mathcal{S}_L(k_n)$, both almost surely and in expectation. Combining this with existing theory on lower bounds (Tsybakov, 2009; Li et al., 2016), yields that the multiscale change-point segmentation estimator is minimax optimal up to a log-factor, see Section 3.1 for details.

Secondly, when f is an arbitrary function in the approximation spaces (1.3) (cf. Section 2.3 and Section 3.2), we also derive a uniform convergence rate of $\hat{f}_n$, both almost surely and in expectation, with respect to L^p-loss for any $0 < p < \infty$. That is,

$$\|\hat{f}_n - f\|_{L^p}^r = \mathcal{O}\left(n^{-\frac{2\gamma}{2\gamma+1}\min\{1/p,1/2\}r} (\log n)^{\frac{\gamma+(1/2-1/p)_+}{2\gamma+1}r}\right),$$

uniformly for f in the approximation space $\mathcal{A}^\gamma$. As special cases we obtain the optimal rates $n^{-2/3 \cdot \min\{1/2,1/p\}}$ and $n^{-2\alpha/(2\alpha+1) \cdot \min\{1/2,1/p\}}$ (up to a log-factor) in terms of the L^p-loss ($0 < p < \infty$), both almost surely and in expectation, for f within bounded variation and (piecewise) Hölder continuous of order $0 < \alpha \leq 1$, respectively.

Thirdly, the convergence rates imply accuracy of feature detection, such as deviation bounds on the locations of jumps. This again extends existing theory on piecewise constant functions to more general functions (cf. Lin et al., 2016). Moreover, for non-step functions we also get statistical justification on the detection of features, such as modes and troughs, deduced by convergence rates, see Section 3.3. More precisely, under some general assumptions, for an appropriate choice of q, it holds almost surely that

$$d\big(J(\hat{f}_n), J(f_{k_n})\big) := \max_{\tau \in J(f_{k_n})} \min_{\hat{\tau} \in J(\hat{f}_n)} |\tau - \hat{\tau}| = \mathcal{O}\Big(\frac{k_n \log n}{\Delta_n^2 n}\Big), \quad \text{a.s.}$$

where (f_{k_n}) is a sequence of step functions with up to k_n jumps and Δ_n is the smallest jump size of f_{k_n}. For $f \in \mathcal{A}^\gamma$, it holds almost surely that

$$\max\big\{|m_I(\hat{f}_n) - m_I(f)| \,:\, I \in \mathcal{I}_n\big\} = \mathcal{O}\left(\frac{1}{\sqrt{\lambda_n}}\Big(\frac{\log n}{n}\Big)^{\gamma/(2\gamma+1)}\right), \quad \text{a.s.}$$

where $m_I(g) := \int_I g(x)dx/|I|$ is the mean of function g over I and λ_n is the smallest length of intervals in $\mathcal{I}_n$.

In summary, the major finding of this work is that the aforementioned multiscale change-point segmentation methods are *universal*, in the sense that they are completely independent of the unknown true regression function. Hence, they automatically *adapt* to the unknown "smoothness" of the underlying function, no matter whether it is piecewise constant (possibly with unbounded number of change-points) or lies in the approximation spaces (1.3). In other words, the estimators in (1.2) are robust to the misspecification of the true smoothness class, provided the degree of such misspecification is *mild*.

This work is organized as follows. In Chapter 2 we introduce some basic preliminaries and multiscale change-point segmentation methods. Some necessary assumptions are listed as well. In Chapter 3 we derive uniform bounds on the L^p-loss over step functions with possibly increasing number of change-points and over classical approximation spaces. We also present some implications on feature detection from convergence rates. Theoretical findings are supported by simulations in Chapter 4. There, we also outline implementation of multiscale change-point segmentation estimators in (1.2), and compare with other change-point methods. This work ends with conclusion and outlook in Chapter 5.

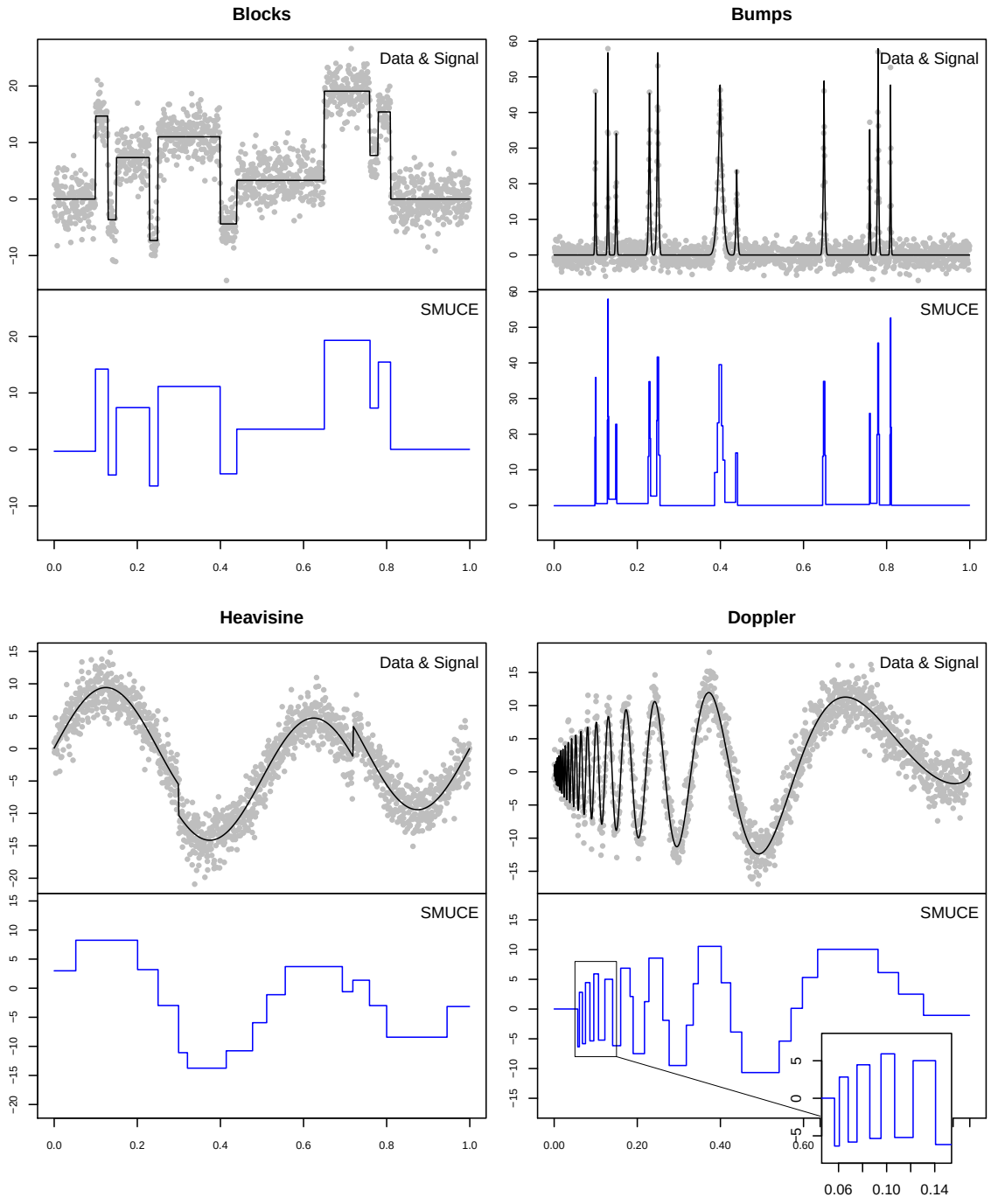

Figure 1.2: Estimation by SMUCE for Blocks, Bumps, Heavisine, and Doppler signals (sample size $n = 1.500$, and SNR $:= \|f\|_{L^2}/\sigma = \sqrt{11}$).

2 Mathematical methodology

In order to state the regression model formally, we will first introduce some notations and terminologies, which will also be used later in this work.

2.1 Model and notation

We begin by recalling the definition of a sub-Gaussian random variable and of the càdlàg functions. We will restrict our consideration on the interval $[0, 1)$ where the regression model is defined.

Definition 2.1.1. A random variable $X \in \mathbb{R}$ is said to be *sub-Gaussian* with variance σ^2 if its moment generating function satisfies

$$\mathbf{E}\left(e^{sX}\right) \leq e^{\sigma^2 s^2/2}, \forall s \in \mathbb{R}, \tag{2.1}$$

in this case we write $X \sim \text{subG}(\sigma^2)$.

Definition 2.1.2 (Billingsley (1999)). Let $\mathcal{D}([0, 1))$ be the space of real functions f on $[0, 1)$ that are right-continuous and have left-hand limits:

(i) For $0 \leq t < 1$, $f(t+) = \lim_{s \downarrow t} f(s)$ exits and $f(t+) = f(t)$.

(ii) For $0 < t < 1$, $f(t-) = \lim_{s \uparrow t} f(s)$ exits.

Functions having these two properties are called *càdlàg functions*.

Remark 2.1.3. A simple example of càdlàg functions is the space of right-continuous change-point functions (*step functions*), which is defined as

$$\mathcal{S}([0, 1)) := \Big\{ f \in \mathcal{D}([0, 1)) : f(t) = \sum_{i=0}^{k} c_i \mathbf{1}_{[\tau_i, \tau_{i+1})}(t), \\ 0 = \tau_0 < \tau_1 < \ldots < \tau_{k+1} = 1, c_i \neq c_{i+1} \Big\}. \tag{2.2}$$

With these preparations, we now state our regression model. Suppose we observe independent random variables $Y = (y_0^n, y_2^n, \ldots, y_{n-1}^n)$ through the regression model

$$y_i^n = f\Big(\frac{i}{n}\Big) + \xi_i^n, \qquad i = 0, \ldots, n-1, \tag{2.3}$$

where $(\xi_i)_{i=0}^{n-1}$ are independent centered sub-Gaussian random variables with the scale parameter σ and the underlying signal f is in the space of càdlàg functions $\mathcal{D}([0,1))$.

If f is an step function in $\mathcal{S}([0,1))$, we denote by $J(f) = (\tau_1, \tau_2, \ldots, \tau_k)$ the increasingly ordered vector of change-points and by $\#J(f) = k$ the number of change-points. Let $\hat{f}_n$ be an estimator of f. If $\hat{f}_n$ lies in $\mathcal{S}([0,1))$ as well, we will denote the estimated number of change-points by $\hat{k}$ and the estimated change-point locations by $J(\hat{f}_n) = (\hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_{\hat{k}})$. By *intervals* we always refer to those of the form $[a,b), 0 \leq a < b \leq 1$. For abbreviation, we write $y^n = (y_i^n)_{i=0}^{n-1}$, $f^n = (f(i/n))_{i=0}^{n-1}$ and $\xi^n = (\xi_i^n)_{i=0}^{n-1}$.

It is useful to introduce a technical concept of *normality*, which describes the richness of a system of intervals.

Definition 2.1.4 (Nemirovski (1985)). A system $\mathcal{I} \equiv \mathcal{I}_n$ of intervals is called *normal* (or *c-normal*) for some constant $c > 1$, if 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 $\{i/n : i = 0, \ldots, n-1\}$.

(iii) The system $\mathcal{I}$ contains at least the intervals $[i/n, (i+1)/n)$, $i = 0, \ldots, n-1$.

Remark 2.1.5. The requirement (i) in the above definition is crucial, while (ii) and (iii) are technical nature aim to the discrete sampling locations $\{i/n\}_{i=0}^{n-1}$. Examples of normal systems include the highly redundant system $\mathcal{I}^0$ of all intervals whose end-points lie on the grid (with the constant $c \geq 2$), used in Siegmund and Yakir (2000), Dümbgen and Spokoiny (2001) and Frick et al. (2014), and less redundant, but still asymptotically efficient systems (see Walther, 2010; Rivera and Walther, 2013). Moreover, there are even normal systems with cardinality of order n, such as the *dyadic partition system*, with the constant $c \geq 4$,

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

see Hotz et al. (2013) and Grasmair et al. (2015) for further information.

2.2 Multiscale change-point segmentation

Given a (normal) system $\mathcal{I}$ of intervals, we introduce a class of *multiscale change-point segmentation* estimators $\hat{f}_n$ (see Frick et al., 2014; Pein et al., 2015; Li et al., 2016) as a solution to

$$\min_{f \in \mathcal{S}([0,1))} \#J(f) \qquad \text{subject to } T_{\mathcal{I}}(y^n; f) \leq q. \tag{2.4}$$

Here $q \in \mathbb{R}$ is a user-specified threshold to be chosen later. The side constraint in (2.4) is defined by a multiscale test statistic

$$T_{\mathcal{I}}(y^n; f) := \sup_{\substack{I \in \mathcal{I} \\ f \equiv c_I \text{ on } I}} \left\{ \frac{1}{\sqrt{n|I|}} \Big| \sum_{i/n \in I} (y_i^n - c_I) \Big| - s_I \right\}, \tag{2.5}$$

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 side constraint in (2.4) originates from a simultaneous multiple testing procedure on various scales, which combines all the local likelihood ratio tests for whether the local mean $\bar{f}_I$ on I of f equals to a given c_I for every $I \in \mathcal{I}$. This provides a criterion for testing the constancy of f on each of the segments $I \in \mathcal{I}$. The scale penalty s_I is introduced to balance the detection power over different scales, see Dümbgen and Spokoiny (2001), Walther (2010), and Frick et al. (2014) for several choices, and see Boysen et al. (2009) and Davies et al. (2012) for the unpenalized estimators, where $s_I \equiv 0$. Thus, the multiscale change-point segmentation estimator yields the most parsimonious candidate over the acceptance region of the multiple tests. The threshold q in (2.4), as a trade-off between data-fit and parsimony, can be chosen such that the truth f is admissible to the side constraint at least with a pre-specified probability $1 - \beta$, i.e., q is the $1 - \beta$ quantile of the distribution of $T_{\mathcal{I}}(\xi^n; 0)$ which can be determined by Monte-Carlo simulations or based on asymptotic considerations (Frick et al., 2014). Here,

$$q \equiv q(\beta) := \inf\Big\{ u \in \mathbb{R} : \mathbf{P}\Big\{ \sup_{I \in \mathcal{I}} \frac{1}{\sqrt{n|I|}} \Big| \sum_{i/n \in I} \xi_i \Big| - s_I \le u \Big\} \ge 1 - \beta \Big\}. \tag{2.6}$$

The choice of significance level β provides an upper bound on the family-wise error rate of the aforementioned multiple testing. It immediately leads to a control of overestimating the number of jumps $\#J(f)$ of f, i.e.,

$$\mathbf{P}\Big\{ \#J(\hat{f}_n) \le \#J(f) \Big\} \ge 1 - \beta.$$

Also, it is possible to control instead the false discovery rate by means of *local* quantiles, see Li et al. (2016) for details. For asymptotic analysis, it will be sufficient to assume $q \asymp \sqrt{\log n}$, as we will see in Chapter 3.

Note that the solution to the optimization problem (2.4) might be *non-unique*. In this case, one can pick any solution of (2.4). In practice often the *constrained maximal likelihood estimator* is used, see Frick et al. (2014) for details. In fact, if we denote the minimal value of $\#J(f)$ in problem (2.4) by $\hat{k}$, then we actually get a confidence set for the true regression function f given by the function class $\mathcal{C}(q)$ (Frick et al., 2014):

$$\mathcal{C}(q) = \Big\{ g \in \mathcal{S} : \#J(g) = \hat{k}, T_{\mathcal{I}}(y^n; g) \le q \Big\}. \tag{2.7}$$

We emphasize that our results hold not only for one particular solution (e.g. maximal likelihood estimator) picked from $\mathcal{C}(q)$. The results hold for all the possible solution from the class $\mathcal{C}(q)$ as well.

In the following, whenever referring to the multiscale change-point segmentation methods by (2.4), we always assume the following.

Assumption 1. (i) The interval system $\mathcal{I}$ is c-normal for some constant $c > 1$.

(ii) The scale penalty s_I satisfies almost surely that

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

Remark 2.2.1. We note that

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

is at most of order $\sqrt{\log n}$ (Shao, 1995), so Assumption 1 is quite natural. In particular, it allows for many common scale penalties (Dümbgen and Spokoiny, 2001; Schmidt-Hieber et al., 2013; Frick et al., 2014), and even includes a no scale penalty (Davies et al., 2012). Thus, Assumption 1 is rather weak, which in turn makes the approach (2.4) rather general. For instance, this includes SMUCE (Frick et al., 2014) and FDRSeg (Li et al., 2016) as special cases. More precisely, for SMUCE we have $\mathcal{I} = \mathcal{I}^0$ and $s_I = \sqrt{2 \log(e/|I|)}$, and for FDRSeg we have again the same system $\mathcal{I} = \mathcal{I}^0$, but the scale penalty $s_I = \sqrt{2 \log(e|\tilde{I}|/|I|)}$, with $\tilde{I}$ the constant segment of the candidate solution, which contains I.

For simplicity, we also assume that the scale parameter (i.e. noise level) σ in model (2.3) is known. In practice, it can be easily pre-estimated, see Dette et al. (1998) for instance.

2.3 Approximation space

The idea for our estimator comes from the setting that the underlying function f in (2.3) is a step function. However, for practical applications, it often occurs that f is only approximately piecewise constant (cf. Chapter 1). It is quite natural to consider extending this method and related results to more general function spaces. When trying to do this extension, the question arises, which properties of the underlying function f determine the convergence and asymptotic properties. It turns out that the speed of approximation speed of f by step functions is crucial. In order to figure this question out precisely, we now introduce the so called *approximation error* and *approximation space* (cf. Pietsch, 1981; DeVore and Lorentz, 1993; DeVore, 1998).

Definition 2.3.1 (Quasi-norm). A quasi-norm is a non-negative function $\|\cdot\|_X$ defined on a (real or complex) linear space X for which the following conditions are satisfied.

(i) If $\|f\|_X = 0$ for some $f \in X$, then $f = 0$.

(ii) $\|\lambda f\|_X = |\lambda|\|f\|_X$ for $f \in X$ and all scalars λ.

(iii) There exists a constant $c_X \geq 1$ such that

$$\|f + g\|_X \leq c_X(\|f\|_X + \|g\|_X) \qquad \text{for } f, g \in X.$$

A quasi-Banach space $(X, \|\cdot\|_X)$ is a linear space X equipped with a quasi-norm $\|\cdot\|_X$ such that every Cauchy sequence is convergent.

A quasi-norm $\|\cdot\|_X$ is called a p-norm $(0 < p \leq 1)$ if

$$\|f + g\|_X^p \leq \|f\|_X^p + \|g\|_X^p \qquad \text{for } f, g \in X.$$

Definition 2.3.2 (Approximation Schemes). An approximation scheme (X, A_n) is a quasi-Banach space X together with a sequence of subsets A_n such that the following conditions are satisfied.

(i) $A_1 \subseteq A_2 \subseteq \ldots \subseteq X$.

(ii) $\lambda A_n \subseteq A_n$ for all scalars λ and $n \in \mathbb{N}$.

(iii) $A_m \cup A_n \subseteq A_{m+n}$ for $m, n \in \mathbb{N}$.

Let (X, A_n) be an approximation scheme. For $f \in X$ and $n \in \mathbb{N}$ the nth approximation number (error) is defined by

$$\Gamma_n(f, X) := \inf\{\|f - a\|_X : a \in A_n\}.$$

Definition 2.3.3 (Approximation Spaces). Let $0 < \rho < \infty$ and $0 < u \leq \infty$. Let l^u be the space of all the bounded sequences of real numbers $(x_n)_{n=1}^\infty$, with the l^u-norm

$$\|(x_n)_{n=1}^\infty\|_{l^u} := \left(\sum_{n=1}^\infty |x_n|^u\right)^{1/u} \qquad \text{for } u < \infty,$$

and

$$\|(x_n)_{n=1}^\infty\|_{l^\infty} := \sup_{n=1,\ldots,\infty} |x_n| \qquad \text{for } u = \infty.$$

Then the approximation space X_u^ρ, or more precisely $(X, A_n)_u^\rho$, consists of all elements $f \in X$ such that $(n^{\rho - 1/u}\Gamma_n(f, X)) \in l^u$, where $n \in \mathbb{N}$.

Example 2.3.4. Consider the space of càdlàg functions $\mathcal{D}([0, 1))$, equipped with the L^∞-norm. For $k \in \mathbb{N}$, let A_k be the space of step functions with no more than k number of change-points, that is,

$$A_k = \mathcal{S}(k) := \Big\{f \in \mathcal{S}([0, 1)) : \#J(f) \leq k\Big\}.$$

It is easy to see that $(\mathcal{D}([0,1)), A_k)$ is an approximation scheme. For $f \in \mathcal{D}([0,1))$, the approximation error is then defined by

$$\Gamma_k(f) := \Gamma_k(f, \mathcal{D}([0,1))) := \inf\Big\{ \|f - g\|_{L^\infty} : g \in \mathcal{S}([0,1)),\ \#J(g) \le k \Big\}.$$

Thus for any $0 < \gamma < \infty$, we have the following approximation space $(\mathcal{D}([0,1)), \mathcal{S}(k))^\gamma_\infty$,

$$(\mathcal{D}([0,1)), \mathcal{S}(k))^\gamma_\infty = \Big\{ f \in \mathcal{D}([0,1)) : \sup_{k \ge 1} k^\gamma \Gamma_k(f) < \infty \Big\}.$$

For abbreviation, we will write $\mathcal{A}^\gamma := (\mathcal{D}([0,1)), \mathcal{S}(k))^\gamma_\infty$ in the following.

3 Theory

In this chapter, we show convergence rates of the multiscale change-point segmentation methods for the model in (2.3) with equidistant sampling points. We stress, that the subsequent results can be easily generalized to non-equidistant (and random) sample points $x_{i,n}$ under appropriate conditions on the design (see Munk and Dette, 1998). This is, however, suppressed to ease presentation.

3.1 Convergence rates for step functions

Consider first the locally constant change-point regression, i.e., the underlying signal f in model (2.3) is piecewise constant. We introduce the class of uniformly bounded piecewise constant functions (recall (2.2)) with up to k jumps

$$\mathcal{S}_L(k) := \Big\{ f \in \mathcal{S}([0,1)) : \#J(f) \leq k, \text{ and } \|f\|_{L^\infty} \leq L \Big\},$$

for $k \in \mathbb{N}$ and $L > 0$. For a step function $f \in \mathcal{S}_L(k)$, let λ_f be the smallest interval length of f, and let Δ_f and $\tilde{\Delta}_f$ be the smallest and the largest jump size of f, respectively.

Now we consider a specific multiscale change-point segmentation estimator, SMUCE (Frick et al., 2014), and derive its convergence rate with respect to L^2-loss, which has not shown in the original paper. To finish the story, we assume the underlying signal f belongs to the following slightly constrained uniformly bounded piecewise constant functions:

$$B_{\nu,\epsilon,H}(L,K) := \{ f \in \mathcal{S}_L(K) | \lambda_f \geq \nu, \epsilon \leq \Delta_f \leq \tilde{\Delta}_f \leq H \}, \tag{3.1}$$

where $0 < \nu < 1/2$ and $0 < \epsilon < H < \infty$. Denote by $\hat{f}_n$ the SMUCE of f in model (2.3), we deduce the uniform upper bound of L^2-loss for SMUCE $\hat{f}_n$.

Theorem 3.1.1. *Under the assumptions above, if we choose $\beta = o(\sqrt{\log n}/n)$ and $\beta \geq n^{-r}, r \geq 1$, in SMUCE (Frick et al., 2014), then*

$$\limsup_{n\to\infty} \sup_{f \in B_{\nu,\epsilon,H}(L,K)} \mathbf{E}\left(\|\hat{f}_n - f\|_{L^2} \right) \left(\frac{\log n}{n} \right)^{\frac{1}{2}} \leq C, \tag{3.2}$$

where C is a constant only depending on $\nu, \epsilon, H, r, \sigma$ and K.

Proof. Assume $\#J(f) = K_f \leq K$, and define the following sets:

$$A := \Big\{\vartheta \in \mathcal{S} : \#J(\vartheta) \leq K_f\Big\},$$

for a given c_n with $c_n \to 0$,

$$B_n := \Big\{\vartheta \in \mathcal{S} : d(J(\vartheta), J(f)) < c_n\Big\},$$

where $d(J(\vartheta), J(f)) := \max_{\tau \in J(\vartheta)} \min_{\hat{\tau} \in J(f)} |\tau - \hat{\tau}|$. Note that

$$\mathbf{E}\left(\|\hat{f}_n - f\|_{L^2}\right) = \int_0^{\sqrt{n}} \mathbf{P}\left\{\|\hat{f}_n - f\|_{L^2} \geq t\right\} dt + \int_{\sqrt{n}}^{\infty} \mathbf{P}\left\{\|\hat{f}_n - f\|_{L^2} \geq t\right\} dt$$

In the following, we will show as $n \to \infty$,

$$\sup_{f \in B_{\nu,\epsilon,H}(L,K)} \int_0^{\sqrt{n}} \mathbf{P}\{\|\hat{f}_n - f\|_{L^2} \geq t\} dt \sqrt{\frac{n}{\log n}} \leq C \tag{3.3}$$

and

$$\sup_{f \in B_{\nu,\epsilon,H}(L,K)} \int_{\sqrt{n}}^{\infty} \mathbf{P}\{\|\hat{f}_n - f\|_{L^2} \geq t\} dt \sqrt{\frac{n}{\log n}} \to 0. \tag{3.4}$$

For (3.3), we have

$$\begin{aligned}
&\int_0^{\sqrt{n}} \mathbf{P}\left\{\|\hat{f}_n - f\|_{L^2} \geq t\right\} dt \sqrt{\frac{n}{\log n}} \\
&\leq \int_0^{\infty} \mathbf{P}\left\{\|\hat{f}_n - f\|_{L^2} \geq t, A \cap B_n\right\} dt \sqrt{\frac{n}{\log n}} \\
&+ \int_0^{\sqrt{n}} \mathbf{P}\left\{\|\hat{f}_n - f\|_{L^2} \geq t, A^c \cup B_n^c\right\} dt \sqrt{\frac{n}{\log n}} \\
&\leq \int_0^{\infty} \mathbf{P}\left\{\|\hat{f}_n - f\|_{L^2} \geq t, A \cap B_n\right\} dt \sqrt{\frac{n}{\log n}} \qquad (3.5) \\
&+ \sqrt{n}(\mathbf{P}\{A^c\} + \mathbf{P}\{B_n^c\}) \sqrt{\frac{n}{\log n}} \qquad (3.6)
\end{aligned}$$

For the first part of (3.6), since $\beta = o(\frac{\sqrt{\log n}}{n})$, it follows from $P(A^c) < \beta$ (c.f. (Frick et al., 2014)) that

$$\limsup_{n \to \infty} \sup_{f \in B_{\nu,\epsilon,H}(L,k)} \frac{n}{\sqrt{\log n}} P(A^c) = 0.$$

For the second part of (3.6), if we take $c_n = \frac{48r\log n}{\Delta_f^2 n} \le \lambda_f/8$ and $\beta \ge 1/n^r$, from the Theorem 7 in (Frick et al., 2014), we have

$$\begin{aligned}
P(B_n^c) &\le 2K_f\big\{\exp(-\frac{1}{16}nc_n\Delta_f^2)\exp(\frac{1}{2}(q+\sqrt{2\log(e/c_n)})^2) + \exp(-\frac{1}{4}nc_n - \Delta_f^2)\big\} \\
&\le 2K_f\big\{\exp(q^2 + 2\log(e/c_n) - \frac{1}{16}nc_n\Delta_f^2) + \exp(-\frac{1}{4}nc_n\Delta_f^2)\big\} \\
&\le 2K_f(e^{-r\log n} + e^{-12r\log n}) \\
&\le \frac{14K_f}{n^r}
\end{aligned}$$

where the third inequality comes from $q \le \sqrt{8\log\frac{2}{\beta}}$. Thus

$$\limsup_{n\to\infty} \sup_{f\in B_{\nu,\epsilon,H}(L,K)} \frac{n}{\sqrt{\log n}} P(B_n^c) = 0$$

On the other hand, it is easy to see that if $\vartheta \in A \cap B_n$, then $\#J(\vartheta) = \#J(f)$. Thus, for (3.5),

$$\begin{aligned}
&\int_0^\infty \mathbf{P}\left\{\|\hat{f}_n - f\|_{L^2} \ge t, A\cap B_n\right\} dt \\
=&\mathbf{E}\left(\|\hat{f}_n - f\|_{L^2}; \mathbf{1}_{\{A\cap B_n\}}\right) \\
\le&\mathbf{E}\left(\|\hat{f}_n - f\|_{L^2}^2; \mathbf{1}_{\{A\cap B_n\}}\right)^{1/2}. \qquad (3.7)
\end{aligned}$$

Let $\tau_i^- = \min\{\tau_i, \hat{\tau}_i\}, \tau_i^+ = \max\{\tau_i, \hat{\tau}_i\}, I_i = [\tau_{i-1}^+, \tau_i^-), \eta_i = |\tau_{i+1} - \tau_i|$, and denote by θ_i and $\hat{\theta}_i$ the value of f and $\hat{f}_n$ on I_i, respectively, then the square of (3.7) is bounded from above by

$$\begin{aligned}
&\mathbf{E}\left(\sum_{i=0}^{K_f} |\hat{\theta}_i - \theta_i|^2(\tau_{i+1}^- - \tau_i^+) + \sum_{i=1}^{K_f} \max\{|\hat{\theta}_{i+1} - \theta_i|^2, |\hat{\theta}_i - \theta_{i+1}|^2\}(\tau_i^+ - \tau_i^-)\right) \\
\le&\sum_{i=0}^{K_f} \eta_i \mathbf{E}\left(|\hat{\theta}_i - \theta_i|^2\right) + \sum_{i=1}^{K_f}(2|\theta_{i+1} - \theta_i|^2 + 2|\hat{\theta}_i - \theta_i|^2)c_n \\
\le&\sum_{i=0}^{K_f}(\eta_i + 2c_n)\mathbf{E}\left(|\hat{\theta}_i - \theta_i|^2\right) + 2K_f\tilde{\Delta}_f^2 c_n
\end{aligned}$$

Note that by the construction of SMUCE, we have for any interval I_i,

$$T_n(Y, \hat{\theta}_i) = |\bar{Y}_{I_i} - \hat{\theta}_i|\sqrt{|I_i|n} - \sqrt{2\log\frac{e}{|I_i|}} \le q,$$

here T_n is the multiscale statistic in SMUCE, $\bar{Y}_{I_i}$ is the average value of Y_i in the interval I_i, and $|I_i|$ is the length of I_i. This implies $\sqrt{n|I_i|}|\bar{Y}_{I_i}-\theta_i-t| \le q+\sqrt{2\log\frac{e}{|I_i|}}$, if $\bar{Y}_{I_i}-\theta_i \le t$ and $\hat{\theta}_i-\theta_i > t$. Then,

$$\begin{aligned}
\mathbf{P}\left\{\hat{\theta}_i-\theta\ge t\right\} &\le \mathbf{P}\left\{\bar{Y}_{I_i}-\theta\le s, \hat{\theta}_i-\theta_i>s\right\}+\mathbf{P}\left\{\bar{Y}_{I_i}-\theta_i>t\right\}\\
&\le \mathbf{P}\left\{\sqrt{n|I_i|}|\bar{Y}_{I_i}-\theta_i-t|\le q+\sqrt{2\log\frac{e}{|I_i|}}\right\}+\mathbf{P}\left\{\bar{Y}_{I_i}>\theta_i+t\right\}\\
&\le \exp\big(-\frac{1}{8}(t\sqrt{n|I_i|}-q-\sqrt{2\log\frac{e}{|I_i|}})_+^2\big)+\exp(-\frac{n|I_i|t^2}{2})\\
&\le 2\exp\big(-\frac{1}{8}(t\sqrt{n|I_i|}-q-\sqrt{2\log\frac{e}{|I_i|}})_+^2\big)
\end{aligned}$$

Since we already know $|I_i| \ge \eta_i - 2c_n > 0$, by monotonicity of $\sqrt{2\log\frac{e}{|I_i|}}$ and symmetry of the gaussian distribution, we have

$$\mathbf{P}\left\{|\hat{\theta}_i-\theta|\ge t\right\} \le 4\exp\big(-\frac{1}{8}(t\sqrt{n(\eta_i-2c_n)}-q-\sqrt{2\log\frac{e}{\eta_i-2c_n}})_+^2\big) \tag{3.8}$$

Using (3.8) to estimate

$$\begin{aligned}
&\mathbf{E}\left(|\hat{\theta}_i-\theta_i|^2\right)\\
=&\int_0^\infty \mathbf{P}\left\{\hat{\theta}_i-\theta_i|^2>t\right\}dt\\
=&\int_0^\infty \mathbf{P}\left\{|\hat{\theta}_i-\theta_i|>t^{1/2}\right\}dt\\
=&\int_0^{(\frac{q+2\sqrt{2\log\frac{e}{\eta_i-2c_n}}}{\sqrt{n(\eta_i-2c_n)}})^2} \mathbf{P}\left\{|\hat{\theta}_i-\theta_i|>t^{1/2}\right\}dt+\int_{(\frac{q+2\sqrt{2\log\frac{e}{\eta_i-2c_n}}}{\sqrt{n(\eta_i-2c_n)}})^2}^\infty \mathbf{P}\left\{|\hat{\theta}_i-\theta_i|>t^{1/2}\right\}dt\\
\le&(\frac{q+2\sqrt{2\log\frac{e}{\eta_i-2c_n}}}{\sqrt{n(\eta_i-2c_n)}})^2\\
&+\int_{(\frac{q+2\sqrt{2\log\frac{e}{\eta_i-2c_n}}}{\sqrt{n(\eta_i-2c_n)}})^2}^\infty 2\exp(-\frac{1}{8}(t^{1/2}\sqrt{n(\eta_i-2c_n)}-q-2\sqrt{2\log\frac{e}{\eta_i-2c_n}})^2)dt.
\end{aligned}$$

It remains to calculate the latter term. Let $a=\frac{q+2\sqrt{2\log\frac{e}{\eta_i-2c_n}}}{\sqrt{n(\eta_i-2c_n)}}, b=\sqrt{n(\eta_i-2c_n)}$, then

$$\int_{a^2}^\infty \exp(-\frac{1}{8}(t^{1/2}\sqrt{n(\eta_i-2c_n)}-q-2\sqrt{2\log\frac{e}{\eta_i-2c_n}})^2)dt$$

$$= \int_0^\infty e^{-\frac{1}{8}x^2} \frac{2}{b^2}(x+ab)dx \tag{3.9}$$

$$\begin{aligned}(3.9) =& \frac{2}{b^2}\int_0^\infty e^{-\frac{1}{8}x^2}(x+ab)dx \\ \leq& \frac{2}{b^2}\int_0^\infty e^{-\frac{1}{8}x^2}(x+ab)dx \\ =& \frac{8}{b^2} + \frac{2a}{b}\sqrt{2\pi}.\end{aligned}$$

Hence,

$$\begin{aligned}&\mathbf{E}\left(|\hat{\theta}_i - \theta_i|^2\right) \\ \leq& \frac{(q+2\sqrt{2\log\frac{e}{\eta_i - 2c_n}})^2}{n(\eta_i - 2c_n)} + \frac{32}{n(\eta_i - 2c_n)} + 8\sqrt{2\pi}\frac{(q+2\sqrt{2\log\frac{e}{\eta_i-2c_n}})}{n(\eta_i-2c_n)},\end{aligned}$$

which implies that the square of (3.5) is bounded by

$$\begin{aligned}&\sum_{i=0}^{K_f}\frac{\eta_i+2c_n}{n(\eta_i-2c_n)}\{(q+2\sqrt{2\log\frac{e}{\lambda_i-2c_n}}+4\sqrt{2\pi})^2+(32-32\pi)\}+2K_f\tilde{Delta}^2c_n \\ \leq& \frac{2}{n}(K_f+1)((q+2\sqrt{2\log\frac{e}{6c_n}}+4\sqrt{2\pi})^2+(32-32\pi))+2K_f\tilde{\Delta}_f^2c_n.\end{aligned}$$

If we take $c_n = \frac{48r\log n}{\Delta_f^2 n}$, and $\beta \geq \frac{1}{n^r}$, then

$$\limsup_{n\to\infty}\sup_{f\in B_{\nu,\epsilon,H}(L,k)}\int_0^{\sqrt{n}}\mathbf{P}\left\{\|\hat{f}_n - f\|_{L^2}\geq t\right\}dt\sqrt{\frac{n}{\log n}} \leq \sqrt{r((96\frac{H^2}{\epsilon^2}+16)K+32)}. \tag{3.10}$$

(3.4) follows by the same method as in (Li et al., 2016). That is, by construction, we have

$$\|\hat{f}_n - \sum_{i=0}^{n-1}Y_i\mathbf{1}_{[\frac{i}{n},\frac{i+1}{n})}\|_{L^2} \leq \max_{0\leq i\leq n-1}|\hat{f}_n(\frac{i}{n}) - Y_i)| \leq q + \sqrt{2\log en}.$$

On the other hand, for $f = \sum_{k=0}^{K_f}\theta_k\mathbf{1}_{[\tau_k,\tau_{k+1})}$, we have

$$\|\sum_{i=0}^{n-1}Y_i\mathbf{1}_{[\frac{i}{n},\frac{i+1}{n})} - f\|_{L^2} \leq \|\sum_{i=0}^{n-1}Y_i\mathbf{1}_{[\frac{i}{n},\frac{i+1}{n})} - \sum_{k=0}^{K_f}\theta_k\mathbf{1}_{[\frac{\lceil n\tau_k\rceil}{n},\frac{\lceil n\tau_{k+1}\rceil}{n})}\|_{L^2}$$

$$
\begin{aligned}
&+ \| \sum_{k=0}^{K_f} \theta_k \mathbf{1}_{[\frac{\lceil n\tau_k \rceil}{n}, \frac{\lceil n\tau_{k+1} \rceil}{n})} - f \|_{L^2} \\
&\leq (\frac{1}{n} \sum_{i=0}^{n-1} |\varepsilon_i|^2)^{1/2} + \tilde{\Delta}_f (\frac{K_f}{n})^{1/2}
\end{aligned}
$$

If n is chosen large enough such that $\sqrt{n}/2 > \tilde{\Delta}_f (\frac{K_f}{n})^{1/2} + q + \sqrt{2 \log en}$, then

$$
\begin{aligned}
&\int_{\sqrt{n}}^{\infty} \mathbf{P} \left\{ \| \hat{f}_n - f \|_{L^2} \geq t \right\} dt \\
\leq &\int_{\sqrt{n}}^{\infty} \mathbf{P} \left\{ \| \hat{f}_n - \sum_{i=0}^{n-1} Y_i \mathbf{1}_{[\frac{i}{n}, \frac{i+1}{n})} \|_{L^2} + \| \sum_{i=0}^{n-1} Y_i \mathbf{1}_{[\frac{i}{n}, \frac{i+1}{n})} - f \|_{L^2} \geq t \right\} dt \\
\leq &\int_{\sqrt{n}}^{\infty} \mathbf{P} \left\{ q + \sqrt{2 \log en} + \tilde{\Delta}_f (\frac{K_f}{n})^{1/2} + (\frac{1}{n} \sum_{i=0}^{n-1} |\varepsilon_i|^2)^{1/2} \geq t \right\} dt \\
\leq &\int_{\sqrt{n}}^{\infty} \mathbf{P} \left\{ \frac{1}{n} \sum_{i=0}^{n-1} |\varepsilon_i|^2)^{1/2} \geq \frac{t}{2} \right\} dt \\
\leq &\int_{\sqrt{n}}^{\infty} \frac{4}{t^2} dt \mathbf{P} \left\{ \frac{1}{n} \sum_{i=0}^{n-1} |\varepsilon_i|^2 \right\} \\
\leq &\frac{4}{\sqrt{n}}.
\end{aligned}
$$

This implies (3.4). Thus Theorem 3.1.1 is proved. □

Remark 3.1.2. The above theorem gives an upper bound for the SMUCE, combined with the following theorem from Li et al. (2016) we show that the SMUCE is minimax optimal up to a log-factor, with respect to L^2-loss.

Theorem 3.1.3 (Li et al. (2016), Theorem 3.4). *There exists a positive constant C, such that*

$$
\inf_{\hat{f}_n \in \mathcal{S}([0,1))} \sup_{f \in B_{\nu,\epsilon,H}(L,K)} \mathbf{E} \left(\| \hat{f}_n - f \|_{L^2} \right) \geq C \left(\frac{\sigma^2}{n} \right)^{1/2}
$$

for any $\sigma > 0, 0 < \nu < 1/2$ and $0 < \epsilon < H > \infty$,

In fact, if the number of change-points is bounded, the estimation problem is, roughly speaking, parametric, by interpreting the 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) or from a likelihood viewpoint (see e.g. Yao and Au, 1989; Siegmund and Yakir, 2000). However, in order to understand the nonparametric nature of the change-point regression, we now allow

the number of change-points to increase as the number of observations tends to infinity, and we get a much more general result for the convergence rate with respect to L^p-loss, $0 < p < \infty$.

Theorem 3.1.4. *Assume model* (2.3) *and that Assumption 1 holds with constants* $c > 1$ *and* $\delta > 0$. *Let* $0 < p, r < \infty$, *and let* $\hat{f}_n$ *be the multiscale change-point segmentation estimator from* (2.4) *with threshold*

$$\begin{aligned} & q = a\sqrt{\log n} && \text{for some } a \geq \delta + \sigma\sqrt{2r+4}, \\ \text{or} \quad & q = q(\beta) && \text{as in (2.6) with } \beta = \mathcal{O}(n^{-r}). \end{aligned}$$

Let k_n *be a sequence of non-negative integers such that* $k_n = o(n)$. *Then it holds that*

$$\|\hat{f}_n - f\|_{L^p} = \mathcal{O}\left(\left(\frac{2k_n+1}{n}\right)^{\min\{1/2,1/p\}} (\log n)^{1/2}\right), \quad a.s.$$

uniformly for $f \in \mathcal{S}_L(k_n)$. *Furthermore, the same result also holds in expectation,*

$$\mathbf{E}\left(\|\hat{f}_n - f\|_{L^p}^r\right) = \mathcal{O}\left(\left(\frac{2k_n+1}{n}\right)^{\min\{1/2,1/p\}r} (\log n)^{r/2}\right),$$

uniformly for $f \in \mathcal{S}_L(k_n)$.

Proof. We first consider the choice of threshold $q = a\sqrt{\log n}$, and structure the proof into three parts.

(i) *Good noise case.* Assume that the true signal f lies in the multiscale constraint, i.e.

$$T_{\mathcal{I}}(y^n; f) \leq a\sqrt{\log n}.$$

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

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

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 $|\tilde{I}_i| \geq |I_i|/c$. It follows that

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

which, together with $|\tilde{I}_i| \geq |I_i|/c$, implies

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

If $|I_i| \le c/n$, then we have for some i_0

$$|\hat{\theta}_i - \theta_i| \le |\hat{\theta}_i - y^n_{i_0}| + \left|y^n_{i_0} - f\big(\frac{i_0}{n}\big)\right| + 2\|f\|_{L^\infty} \le 2(a+\delta)\sqrt{\log n} + 2L.$$

Thus, by combining these two situations, we obtain that

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

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

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

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

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

Therefore, as $n \to \infty$,

$$\|\hat{f}_n - f\|^r_{L^p} \le 2^{r/p}\Big(\frac{(2k_n+1)c}{n}\Big)^{\min\{r/2, r/p\}} \big(4(a+\delta)^2\log n\big)^{r/2}\big(1+o(1)\big). \tag{3.11}$$

(ii) *Almost sure convergence.* Noting that $(n|I|)^{-1/2}\sum_{i/n\in I}\xi^n_i$ is again sub-Gaussian with scale parameter σ for $I \in \mathcal{I}$, we obtain by Boole's inequality that

$$\begin{aligned}\mathbf{P}\left\{T_{\mathcal{I}}(y^n; f) > a\sqrt{\log n}\right\} \le& \mathbf{P}\left\{\sup_{I\in\mathcal{I}} \frac{1}{\sqrt{n|I|}}\Big|\sum_{i/n\in I}\xi^n_i\Big| > (a-\delta)\sqrt{\log n}\right\} \\ \le& 2n^{-\frac{(a-\delta)^2}{2\sigma^2}+2} \le 2n^{-r} \to 0 \qquad \text{as } n\to\infty.\end{aligned} \tag{3.12}$$

This together with (3.11) implies the almost sure convergence assertion for $q = a\sqrt{\log n}$.
(iii) *Convergence in expectation.* It follows from (3.11) that

$$\mathbf{E}\left(\|\hat{f}_n - f\|^r_{L^p}\right) = \mathbf{E}\left(\|\hat{f}_n - f\|^r_{L^p}; T_{\mathcal{I}}(y^n; f) \le a\sqrt{\log n}\right)$$

$$
\begin{aligned}
&\quad + \mathbf{E}\left(\|\hat{f}_n - f\|_{L^p}^r; T_{\mathcal{I}}(y^n; f) > a\sqrt{\log n}\right) \\
&\leq 2^{r/p}\Big(\frac{(2k_n+1)c}{n}\Big)^{\min\{r/2, r/p\}}\big(4(a+\delta)^2 \log n\big)^{r/2}\big(1+o(1)\big) \\
&\quad + \mathbf{E}\left(\|\hat{f}_n - f\|_{L^p}^r; T_{\mathcal{I}}(y^n; f) > a\sqrt{\log n}\right).
\end{aligned}
$$

We next show the second term above asymptotically vanishes faster than the first one. Note that

$$
\begin{aligned}
&\mathbf{E}\left(\|\hat{f}_n - f\|_{L^p}^r; T_{\mathcal{I}}(y^n; f) > a\sqrt{\log n}\right) \\
=&\int_0^{2n^{p/2}} \mathbf{P}\left\{\|\hat{f}_n - f\|_{L^p}^p \geq u; T_{\mathcal{I}}(y^n; f) > a\sqrt{\log n}\right\}\frac{r}{p}u^{r/p-1}du \\
&\quad + \int_{2n^{p/2}}^{\infty} \mathbf{P}\left\{\|\hat{f}_n - f\|_{L^p}^p \geq u; T_{\mathcal{I}}(y^n; f) > a\sqrt{\log n}\right\}\frac{r}{p}u^{r/p-1}du \\
\leq& 2^{r/p} n^{r/2} \mathbf{P}\left\{T_{\mathcal{I}}(y^n; f) > a\sqrt{\log n}\right\} + \int_{2n^{p/2}}^{\infty} \mathbf{P}\left\{\|\hat{f}_n - f\|_{L^p}^p \geq u\right\}\frac{r}{p}u^{r/p-1}du \\
\leq& 2^{r/p+1} n^{-r/2} + \int_{2n^{p/2}}^{\infty} \mathbf{P}\left\{\|\hat{f}_n - f\|_{L^p}^p \geq u\right\}\frac{r}{p}u^{r/p-1}du, \qquad (3.13)
\end{aligned}
$$

where the last inequality is due to (3.12). Introduce functions $g = \sum_{i=0}^{n-1} y_i^n \mathbf{1}_{[i/n,(i+1)/n)}$ and $h = \sum_{i=0}^{n-1} f(i/n)\mathbf{1}_{[i/n,(i+1)/n)}$. Then, with notation $\xi^n := \{\xi_i^n\}_{i=0}^{n-1}$, $(x)_+ := \max\{x, 0\}$ and $s := (2r-p)_+$, it holds that

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

Thus, for large enough n we have

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

where the last inequality holds by the fact $s \geq 2r - p$. Combining this with (3.13) leads to

$$\begin{aligned}\mathbf{E}\left(\|\hat{f}_n - f\|_{L^p}^r; T_{\mathcal{I}}(y^n; f) > a\sqrt{\log n}\right) &= \mathcal{O}(n^{-r/2}) \\ &= o\Big(\big(n^{-1}(2k_n+1)\big)^{\min\{r/p, r/2\}} (\log n)^{r/2}\Big).\end{aligned}$$

This concludes the proof for $q = a\sqrt{\log n}$.

Finally, we consider the choice of threshold $q = q(\beta)$. The corresponding assertions follow readily from the proof above, by noting the facts that $q(\beta) \leq a\sqrt{\log n}$ for some constant a, due to (3.12), and that $\mathbf{P}\{T_{\mathcal{I}}(y^n; f) > q(\beta)\} = \mathcal{O}(n^{-r})$ by the choice of $\beta = \mathcal{O}(n^{-r})$. □

Remark 3.1.5. In the above theorem, we note that the choice of the only tuning parameter q is universal, i.e., completely independent of the (unknown) true regression function. One can easily obtain a lower bound of order $(k_n/n)^{\min\{1/2, 1/p\}}$ on the best possible rate in terms of L^p-loss, $0 < p < \infty$, by standard arguments based on testing many hypotheses and information inequalities (cf. Tsybakov, 2009; Li et al., 2016). Thus, the multiscale change-point segmentation method adapts to the underlying complexity of the truth, and is up to a log-factor minimax optimal over classes $\mathcal{S}_L(k_n)$ with different choices of k_n, such that $k_n = o(n)$, in particular, $k_n \asymp n^\theta$, $0 \leq \theta < 1$. This includes the case $\theta = 0$, where, by convention, k_n is finite. Moreover, we point out that the choice of threshold q is independent of the specific loss function, but depends on the order r of the moments of the loss.

3.2 Robustness to model misspecification

In practical applications the underlying function in model (2.3) is usually not a precise step function. As a robustness study, we next consider the convergence behavior of the multiscale change-point segmentation methods for more general functions. Using the terminology introduced in Section 2.3, we consider the following approximation space.

$$\mathcal{A}^\gamma := \Big\{ f \in \mathcal{D}([0,1)) : \sup_{k \geq 1} k^\gamma \Gamma_k(f) < \infty \Big\}, \quad \text{for } \gamma > 0,$$

and the subclasses

$$\mathcal{A}_L^\gamma := \Big\{ f \in \mathcal{D}([0,1)) : \sup_{k \geq 1} k^\gamma \Gamma_k(f) \leq L, \text{ and } \|f\|_{L^\infty} \leq L \Big\}, \quad \text{for } \gamma > 0 \text{ and } L > 0,$$

where $\Gamma_k(f)$ is the approximation error defined by

$$\Gamma_k(f) := \inf\Big\{ \|f - g\|_{L^\infty} : g \in \mathcal{S}([0,1)),\, \#J(g) \leq k \Big\}. \tag{3.14}$$

Note that $\mathcal{A}^\gamma = \bigcup_{L>0} \mathcal{A}^\gamma_L$. 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. In addition, it is worth noting that if we consider instead the L^q-loss only for a fixed q, then we can replace $\|f-g\|_{L^\infty}$ by $\|f-g\|_{L^q}$ in the definition (3.14) of the approximation error Δ_k. This will slightly enlarge the approximation spaces.

The rates of convergence for these spaces (or classes) are provided in the following theorem.

Theorem 3.2.1. *Assume model* (2.3) *and that Assumption 1 holds with constants $c > 1$ and $\delta > 0$. Let $0 < p, r < \infty$, and let $\hat{f}_n$ be the multiscale change-point segmentation estimator from* (2.4) *with threshold*

$$\begin{aligned} & q = a\sqrt{\log n} && \text{for some } a > \delta + \sigma\sqrt{2r+4}, \\ \text{or} \quad & q = q(\beta) && \text{as in (2.6) with } \beta = \mathcal{O}(n^{-r}). \end{aligned}$$

Then it holds that

$$\|\hat{f}_n - f\|^r_{L^p} = \mathcal{O}\left(n^{-\frac{2\gamma}{2\gamma+1}\min\{1/p,1/2\}r}(\log n)^{\frac{\gamma+(1/2-1/p)_+}{2\gamma+1}r}\right), \quad a.s.,$$

uniformly for $f \in \mathcal{A}^\gamma_L$. Furthermore, the same result also holds in expectation,

$$\mathbf{E}\left(\|\hat{f}_n - f\|^r_{L^p}\right) = \mathcal{O}\left(n^{-\frac{2\gamma}{2\gamma+1}\min\{1/p,1/2\}r}(\log n)^{\frac{\gamma+(1/2-1/p)_+}{2\gamma+1}r}\right),$$

uniformly for $f \in \mathcal{A}^\gamma_L$.

Proof. The idea behind is that we first approximate the truth f 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 (2.3) (with additional approximation error). In this way, it allows us to employ similar techniques as in the proof of Theorem 3.1.4. To be rigorous, we give a detailed proof as follows.

Firstly, we consider the choice of threshold $q = a\sqrt{\log n}$.
(i) *Good noise case.* Assume for the moment that the observations $y^n = \{y^n_i\}_{i=0}^{n-1}$ from model (2.3) are close to the truth f in the sense that the event

$$\mathcal{G}_n := \left\{ y^n \,:\, \sup_{I\in\mathcal{I}} \frac{1}{\sqrt{n|I|}} \Big| \sum_{i/n\in I} y^n_i - f\big(\frac{i}{n}\big) \Big| - s_I \le a_0\sqrt{\log n} \right\}$$

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

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

Note that $f \in \mathcal{A}^\gamma_L$, so for every n, by means of compactness argument, there exists a step function $\tilde{f}_{k_n} \in \mathcal{S}([0,1))$ with $\#J(\tilde{f}_{k_n}) \le k_n$ such that $\|f - \tilde{f}_{k_n}\|_{L^\infty} \le k_n^{-\gamma} L$. By

introducing additional change-points at $\{i/k_n\}_{i=1}^{k_n-1}$ into $\tilde{f}_{k_n}$, one can construct another step function f_{k_n} with $\#J(f_{k_n}) \le 2k_n$ such that its largest segment length $\le 1/k_n$ and $\|f - f_{k_n}\|_{L^\infty} \le 2k_n^{-\gamma} L$. Then

$$\begin{aligned} T_{\mathcal{I}}(y^n; f_{k_n}) \le & \sup_{\substack{I\in\mathcal{I} \\ f_{k_n}\equiv c_I \text{ on } I}} \frac{1}{\sqrt{n|I|}} \Big| \sum_{i/n\in I} \Big(f\big(\frac{i}{n}\big) - c_I \Big) \Big| + \sup_{I\in\mathcal{I}} \frac{1}{\sqrt{n|I|}} \Big| \sum_{i/n\in I} y_i^n - f\big(\frac{i}{n}\big) \Big| - s_I \\ \le & 2n^{1/2} k_n^{-\gamma-1/2} L + a_0 \sqrt{\log n} \le a\sqrt{\log n}. \end{aligned}$$

That is, f_{k_n} lies in the constraint of (2.4). Thus, by definition, $\#J(\hat{f}_n) \le \#J(f_{k_n}) \le 2k_n$. Let intervals $\{I_i\}_{i=0}^m$ be the partition of $[0,1)$ by $J(\hat{f}_n) \cup J(f_{k_n})$ with $m \le 3k_n$. Then

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

If $|I_i| > c/n$, there is $\tilde{I}_i \in \mathcal{I}$ such that $\tilde{I}_i \subseteq I_i$ and $|\tilde{I}_i| \ge |I_i|/c$. Then,

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

which, together with $|\tilde{I}_i| \ge |I_i|/c$, implies

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

If $|I_i| \le c/n$, then we have for some i_0, j_0

$$\begin{aligned} & |\hat{\theta}_i| \le |\hat{\theta}_i - y_{i_0}^n| + \Big| y_{i_0}^n - f\big(\frac{i_0}{n}\big) \Big| + \|f\|_{L^\infty} \le 2(a+\delta)\sqrt{\log n} + L, \\ \text{and} \quad & |\theta_i| = |f_{k_n}\big(\frac{j_0}{n}\big)| \le \|f - f_{k_n}\|_{L^\infty} + \|f\|_{L^\infty} \le (2k_n^{-\gamma} + 1)L, \end{aligned}$$

which lead to

$$|\hat{\theta}_i - \theta_i| \le |\hat{\theta}_i| + |\theta_i| \le 2(a+\delta)\sqrt{\frac{\log n}{n}} + 2(k_n^{-\gamma} + 1)L.$$

Thus, by combining these two situations, we obtain that

$$\begin{aligned} \|\hat{f}_n - f_{k_n}\|_{L^p}^p \le & \sum_{i:|I_i|>c/n} \Big(2(a+\delta)\sqrt{\frac{c\log n}{n|I_i|}} \Big)^p |I_i| \\ & + \sum_{i:|I_i|\le c/n} \Big(2(a+\delta)\sqrt{\log n} + 2(k_n^{-\gamma}+1)L \Big)^p \frac{c}{n}. \end{aligned}$$

Then, with a similar argument as for (3.11), we obtain as $n \to \infty$

$$\|\hat{f}_n - f_{k_n}\|_{L^p}^p \leq 2\Big(4(a+\delta)^2 \log n\Big)^{p/2} \Big(\frac{(3k_n+1)c}{n}\Big)^{\min\{1,p/2\}} \big(1+o(1)\big),$$

which together with a triangular inequality leads to

$$\|\hat{f}_n - f\|_{L^p}^r \leq 2^{(2/p+1)r}\Big(4(a+\delta)^2 \log n\Big)^{r/2} \Big(\frac{(3k_n+1)c}{n}\Big)^{\min\{r/p,r/2\}} \big(1+o(1)\big). \tag{3.15}$$

(ii) *Rates of convergence.* The rate of almost convergence is a consequence of (3.15) and the fact that, due to (3.12),

$$\limsup_{n\to\infty} \mathbf{P}\left\{\mathcal{G}_n^c\right\} \leq \limsup_{n\to\infty} \mathbf{P}\left\{\sup_{I\in\mathcal{I}} \frac{1}{\sqrt{n|I|}} \Big|\sum_{i/n\in I} \xi_i^n\Big| > (a_0-\delta)\sqrt{\log n}\right\} = 0.$$

Similar to the proof step (iii) of Theorem 3.1.4, we drive from (3.15) that, as $n \to \infty$,

$$\begin{aligned}
&\mathbf{E}\left(\|\hat{f}_n - f\|_{L^p}^r\right)\\
=&\mathbf{E}\left(\|\hat{f}_n - f\|_{L^p}^r; \mathcal{G}_n\right) + \mathbf{E}\left(\|\hat{f}_n - f\|_{L^p}^r; \mathcal{G}_n^c\right)\\
\leq&\mathbf{E}\left(\|\hat{f}_n - f\|_{L^p}^r; \mathcal{G}_n\right) + 2^{r/p} n^{r/2} \mathbf{P}\left\{\mathcal{G}_n^c\right\} + \int_{2n^{p/2}}^{\infty} \mathbf{P}\left\{\|\hat{f}_n - f\|_{L^p}^p \geq u\right\} \frac{r}{p} u^{r/p-1} du\\
\leq&\mathcal{O}\Big((\log n)^{r/2}\big(n^{-1}k_n\big)^{\min\{r/p,r/2\}}\Big) + \mathcal{O}\big(n^{-r/2}\big)\\
=&\mathcal{O}\Big((\log n)^{r/2}\big(n^{-1}k_n\big)^{\min\{r/p,r/2\}}\Big),
\end{aligned}$$

which shows the rate of convergence in expectation.

Lastly, for the choice of threshold $q = q(\beta)$, the proof follows in the same way as above, based on the facts that $q(\beta) \leq a\sqrt{\log n}$ for some constant a, due to (3.12), and that $\mathbf{P}\left\{\mathcal{G}_n^c\right\} = \mathcal{O}(n^{-r})$ by the choice of $\beta = \mathcal{O}(n^{-r})$. □

Remark 3.2.2. Similar to Theorem 3.1.4, the above theorem shows that the multiscale change-point segmentation method with a universal threshold automatically adapts to the smoothness of the approximation spaces, in the sense that it has a faster rate for larger order γ. However, unlike in Theorem 3.1.4, we assume the constant a in the definition of the threshold q should be *strictly* greater than $\delta + \sigma\sqrt{2r+4}$, which is necessary since the constant hidden in the $\mathcal{O}$ notation tends to infinity as $a \to \delta + \sigma\sqrt{2r+4}$.

Example 3.2.3. (i) (Piecewise) Hölder functions. For $0 < \alpha \leq 1$ and $L > 0$, we consider the Hölder function classes

$$H_L^\alpha([0,1)) := \big\{f \in \mathcal{D}([0,1)) : \|f\|_{L^\infty} \leq L \text{ and}$$

$$|f(x_1) - f(x_2)| \leq L|x_1 - x_2|^\alpha \text{ for all } x_1, x_2 \in [0,1)\Big\},$$

and the piecewise Hölder function classes with at most κ jumps

$$H^\alpha_{\kappa,L}([0,1)) := \Big\{ f \in \mathcal{D}([0,1) : \text{there is a partition } \{I_i\}_{i=0}^l, \text{ with } l \leq \kappa, \text{ of } [0,1) \\ \text{such that } f\big|_{I_i} \in H^\alpha_L(I_i) \text{ for all possible } i \Big\}.$$

Obviously, the latter one contains the former as a special case when $\kappa = 0$, that is, $H^\alpha_{0,L}([0,1)) \equiv H^\alpha_L([0,1))$. It is easy to see that $H^\alpha_L([0,1)) \subseteq \mathcal{A}^\alpha_{L'}$ with $L' \geq L$, and $H^\alpha_{\kappa,L}([0,1)) \subseteq \mathcal{A}^\alpha_{L'}$ with $L' \geq L(\kappa+1)^{\alpha+1/2}$ (cf. Boysen et al., 2009).

It is known that the fastest possible rate over $H^\alpha_L([0,1))$, $0 < \alpha \leq 1$ with respect to the L^p-loss, $0 < p < \infty$, is at most of order $n^{-2\alpha/(2\alpha+1)\min\{1/2,1/p\}}$ (see eg. Ibragimov and Has'minskiĭ, 1981; Ibragimov and Khas'minskiĭ, 1982). Thus, as a consequence of Theorem 3.2.1, the multiscale change-point segmentation method with a universal threshold is simultaneously minimax optimal (up to a log-factor) over $H^\alpha_L([0,1))$ and $H^\alpha_{\kappa,L}([0,1))$ for every $\kappa \in \mathbb{N}_0$, $0 < \alpha \leq 1$ and $L > 0$, that is, adaptive to the smoothness order α of the underlying function.

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

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

We introduce the càdlàg-bounded variation classes

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

Elementary calculation, together with Jordan decomposition, implies that

$$\mathrm{BV}_L([0,1)) \subseteq \mathcal{A}^1_{L'} \qquad \text{for } L' \geq L.$$

Since the Hölder class $H^1_L([0,1)) \subseteq \mathrm{BV}_L([0,1))$, the best possible rate for $\mathrm{BV}_L([0,1))$ cannot be faster than that for $H^1_L([0,1))$, which is of order $n^{-2/3\min\{1/2,1/p\}}$. Then, Theorem 3.2.1 implies that the multiscale change-point segmentation method attains the minimax optimal rate (up to a log-factor) over the bounded variation classes $\mathrm{BV}_L([0,1))$ for $L > 0$.

We point out that the convergence rates of the multiscale change-point segmentation methods in the examples above coincide with the rates reported in Boysen et al. (2009) for jump-penalized least square estimators, while they are faster than the rates reported in Fryzlewicz (2007) for the unbalanced Haar wavelets based estimator, with the difference being in log-factors. All these examples concern the approximation spaces $\mathcal{A}^\gamma$ for $\gamma \leq 1$. Note, however,

that those for $\gamma > 1$ do not contain functions of higher smoothness, such as Hölder functions with smoothness order $\alpha > 1$, since it is known (Burchard and Hale, 1975) that if f is piecewise continuously differentiable (i.e. piecewise $\mathcal{C}^1$), and if f lies in $\mathcal{A}^\gamma$ for some $\gamma > 1$, then f is piecewise constant, see also (Boysen et al., 2009, Remark 1).

3.3 Implications of the convergence rates

The convergence rates not only reflect the average performance in recovering the truth over its domain, but also, as a byproduct, lead to statistical justification on the detection of features, such as change-points, modes, and troughs, etc. (see also Lin et al., 2016).

We begin with a theorem which was shown in Lin et al. (2016).

Theorem 3.3.1 (Lin et al. (2016), Theorem 8). *Let f be a piecewise constant function, and $\hat{f}$ be an estimator that satisfied the error bound $\|\hat{f} - f\|_{L^2}^2 = \mathcal{O}_{\mathbb{P}}(R_n)$. Assume that $nR_n/\lambda_f^2 = o(\Delta_f)$. Then,*

$$d(J(\hat{f}), J(f)) := \max_{\tau \in J(f_{k_n})} \min_{\hat{\tau} \in J(\hat{f}_n)} |\tau - \hat{\tau}| = \mathcal{O}_{\mathbb{P}}\left(\frac{nR_n}{\lambda_f^2}\right). \tag{3.16}$$

The above theorem shows the approximate recovery in change-point problems from L^2-loss. With the help of this theorem, combining the convergence results we obtained, we show the following theorem.

Theorem 3.3.2. *Assume model* (2.3) *and that Assumption 1 holds with constants $c > 1$ and $\delta > 0$, and that $\hat{f}_n$ is the multiscale change-point segmentation estimator from* (2.4) *with threshold*

$$\begin{aligned} & q = a\sqrt{\log n} && \text{for some } a > \delta + \sigma\sqrt{2r+4}, \\ \text{or} \quad & q = q(\beta) && \text{as in (2.6) with } \beta = \mathcal{O}(n^{-r}), \end{aligned}$$

for some $0 < r < \infty$. Then

(i) Let f_{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. If

$$\frac{k_n \log n}{n\lambda_n \Delta_n^2} \to 0 \qquad \text{as } n \to \infty$$

it holds almost surely that

$$d\big(J(\hat{f}_n), J(f_{k_n})\big) := \max_{\tau \in J(f_{k_n})} \min_{\hat{\tau} \in J(\hat{f}_n)} |\tau - \hat{\tau}| = \mathcal{O}\Big(\frac{k_n \log n}{\Delta_n^2 n}\Big).$$

(ii) For $n \in \mathbb{N}$, let $\mathcal{I}_n$ be a collection of intervals, and λ_n the smallest length of intervals in $\mathcal{I}_n$, i.e. $\lambda_n := \min\{|I| : I \in \mathcal{I}_n\}$. Then for any $f \in \mathcal{A}^\gamma$

$$\max\big\{|m_I(\hat{f}_n) - m_I(f)| \,:\, I \in \mathcal{I}_n\big\} = \mathcal{O}\left(\frac{1}{\sqrt{\lambda_n}}\Big(\frac{\log n}{n}\Big)^{\gamma/(2\gamma+1)}\right), \qquad a.s.$$

where $m_I(g) := \int_I g(x)dx/|I|$ is the mean of function g over I.

Proof. From Theorem 3.1.4, take $p = 2$ and $r = 2$, we have $\|\hat{f} - f\|_{L^2}^2 = \mathcal{O}(\frac{k_n \log n}{n})$ almost surely. By the assumptions, we have

$$\frac{k_n \log n}{\lambda_n \delta_n^2} \to 0 \qquad as \qquad n \to \infty,$$

take R_n in Theorem 3.3.1 as $(k_n \log n)/n$, we get the first result.

For the second part, similarly, taking $p = 2$ and $r = 2$ in Theorem 3.2.1, we have

$$\|\hat{f}_n - f\|_{L^2} = \mathcal{O}(\frac{\log n}{n})^{\gamma/(2\gamma+1)}.$$

Combining with the fact

$$\|\hat{f}_n - f\|_{L^2} \geq \sqrt{\lambda_n} \max\big\{|m_I(\hat{f}_n) - m_I(f)| \,:\, I \in \mathcal{I}_n\big\},$$

we get the result. □

Remark 3.3.3. The rate in Theorem 3.3.2 (i) in particular applies to SMUCE (Frick et al., 2014) and FDRSeg (Li et al., 2016), where rates of the same order are reported, and it is of the fastest order known until now (Fryzlewicz, 2014). It, moreover, implies that $\mathbf{P}\left\{\#J(\hat{f}_n) \geq \#J(f_{k_n})\right\} \to 1$, which together with the fact (by the choice of threshold q) that

$$\mathbf{P}\left\{\#J(\hat{f}_n) > \#J(f_{k_n})\right\} \leq \mathcal{O}(n^{-r}) \to 0$$

leads to the consistency of the multiscale change-point segmentation methods in estimating the number of jumps, that is,

$$\mathbf{P}\left\{\#J(\hat{f}_n) = \#J(f_{k_n})\right\} \to 1,$$

under the setting specified in Theorem 3.3.2 (i). This coincides with the consistency result in Frick et al. (2014), where they induced it by a different approach.

The local means of $m_I(f)$ over a collection of intervals I actually shed light on the shape of f, such as increases and decreases (thus modes and troughs). In fact, for disjoint intervals I_1, I_2 such that $m_{I_1}(f) < m_{I_2}(f)$, if $\lambda_n^{-1/2}(\log n/n)^{\gamma/(2\gamma+1)} \to 0$, then it holds for sufficiently large n that

$$m_{I_1}(\hat{f}_n) < m_{I_2}(\hat{f}_n),$$

by Theorem 3.3.2 (ii). In other words, an increase (or decrease) of f on the convex hull $\mathrm{conv}(I_1 \cup I_2)$ of the union of I_1 and I_2 eventually leads to an increase (or decrease) of $\hat{f}$ on $\mathrm{conv}(I_1 \cup I_2)$. Further, by selecting $\mathcal{I}_n$ as a fixed partition that captures the modes and the troughs of f, one can show that

$$\mathbf{P}\left\{\#\mathrm{modes}(\hat{f}_n) \geq \#\mathrm{modes}(f), \#\mathrm{troughs}(\hat{f}_n) \geq \#\mathrm{troughs}(f)\right\} \to 1, \qquad \text{as } n \to \infty.$$

Another consequence of Theorem 3.3.2 (ii) is a control of the estimation accuracy of jump locations for general functions f in $\mathcal{A}^\gamma$. Define for $f \in \mathcal{A}^\gamma$ the jump locations of f as $J(f) := \{x : f(x) \neq f(x-)\}$, and the smallest jump size as $\hat{\Delta}_f := \min\{|f(x) - f(x-)| : x \in J(f)\}$. By setting $\mathcal{I}_n := \{[x, x+\lambda_n) \text{ or } [x-\lambda_n, x) : x \in J(f)\}$, with $\lambda_n \asymp d\big(J(\hat{f}_n), J(f)\big)$, one can easily obtain from Theorem 3.3.2 (ii) that

$$\frac{\hat{\Delta}_f}{4} \leq |m_{I_n}(f) - m_{I_n}(\hat{f}_n)| = \mathcal{O}\left(\frac{1}{\sqrt{\lambda_n}}\Big(\frac{\log n}{n}\Big)^{\gamma/(2\gamma+1)}\right) \qquad \text{a.s.}$$

for some appropriately chosen $I_n \in \mathcal{I}_n$. This further implies that

$$d\big(J(\hat{f}_n), J(f)\big) = \mathcal{O}\left(\frac{1}{\hat{\Delta}_f}\Big(\frac{\log n}{n}\Big)^{2\gamma/(2\gamma+1)}\right) \qquad \text{a.s.} \qquad \text{for every } f \in \mathcal{A}^\gamma.$$

As step functions lie in $\mathcal{A}^\gamma$ for all $\gamma > 0$, the above result "formally" reproduces Theorem 3.3.2 (i) for the case that the step function f is fixed, by letting γ tend to infinity.

4 Implementation and Simulation

We first give a brief outline of the implementation of multiscale change-point segmentation methods, then investigate the performance of these methods by a couple of simulations.

4.1 Implementation

Note that in the definition of multiscale change-point segmentation method we consider only the local constraints on the intervals where candidate functions are constant. This ensures the structure of the corresponding optimization problem (2.4) to be a directed acyclic graph, which makes *dynamic programming* algorithms (cf. Bellman, 1957) applicable to such a problem. Moreover, the computation can be substantially accelerated by incorporating *pruning* ideas as recently developed in Frick et al. (2014), Pein et al. (2015) and Li et al. (2016). As a consequence, the computation complexity of multiscale change-point segmentation methods can be even almost *linear* in terms of the number of observations, in case that there are many change-points, see Frick et al. (2014), Pein et al. (2015) and Li et al. (2016) for further details.

The multiscale change-point segmentation estimator can be computed by a pruned dynamic programing algorithm followed in Frick et al. (2012) and Futschik et al. (2014). We will briefly outline the algorithm in the following, see Frick et al. (2014) for details. Note that a change-point segmentation estimator $\hat{f}_n$ can be identified with the vector $(\hat{\theta}_1, ..., \hat{\theta}_n) \in \mathbb{R}^n$ where $\hat{\theta}_i = \hat{f}_n(i/n)$. Next, for a given $c \in \mathbb{N}$ on an interval $\{k, ..., l\}$, we define the local cost of c on $\{k, ..., l\}$ as

$$d_{k,l}(y^n, c) = \begin{cases} h(k, l, c, y^n) & \text{if} \max_{I \subset [k/n, l/n)} \frac{1}{\sqrt{n|I|}} \left| \sum_{i/n \in I} (y_i^n - c) \right| - s_I \leq q \\ \infty & \text{else,} \end{cases}$$

where $h(k, l, c, y^n)$ is a cost function whose definition is based on the selection method of the solutions to (2.4). For instance, if we use the maximum likelihood method, as the authors did in SMUCE, then $h(k, l, c, y^n)$ can be defined as the negative log-likelihood. The optimal costs on the interval $\{k, ..., l\}$ are then defined by $d_{k,l} = \min_{c \in \mathbb{R}} d_{k,l}(y^n, c)$. If

$d_{k,l} < \infty$, we say that $c_{k,l}$ is the optimal parameter if $d_{k,l} = d_{k,l}(y^n, c_{k,l})$. If $d_{k,l} = \infty$ then no $c \in \mathbb{R}$ exits such that the multiscale constraint is satisfied on $\{k, .., l\}$.

Now we outline the dynamic programing approach to solve (2.4). To do this, we first compute the optimal costs $d_p := d_{1,p}$ for $p = 1, 2, \ldots$, and the corresponding parameter values $c_{1,p}$ if $d_p < \infty$. If $d_{1,p+1} = \infty$ then we save the latest feasible index by $R_0 = p$. For all $p > R_0$ at least one new change-point has to be added in order to satisfy the multiscale constraint. Note that for $1 \leq l \leq R_0$ we can always find an estimator $\hat{f}(l, p) = c_{1,l}\mathbf{1}_{\{1,\ldots,l\}} + c_{l+1,p}\mathbf{1}_{\{l+1,\ldots,p\}}$ which has the lowest costs on its constant pieces given the jump location l.

By setting

$$l(p) = \operatorname*{arg\,min}_{1 \leq l \leq R_0} d_{1,l} + d_{l+1,p}$$

we find that $\hat{f}(p) = \hat{f}(l(p), p)$ is the estimator on the interval $\{1, ..., p\}$ with the lowest cumulative costs $d_p := d_{1,l(p)} + d_{l(p)+1,p}$ among all the piecewise constant functions with one change-point. Iterate this procedure until $d_{l+1,p+1} = \infty$ for all $1 \leq l \leq R_0$ and then set $R_1 = p$.

Now assume for $k \geq 1$, we already know R_{k-1} and R_k, and for $R_{k-1} < l \leq R_k$ the estimator $\hat{f}(l)$ has the lowest cumulative costs d_l with k change-points on the interval $\{1, ..., l\}$. Then for $p > R_0$,

$$\hat{f}(l, p) = \hat{f}(l)\mathbf{1}_{\{1,\ldots,l\}} + c_{l+1,p}\mathbf{1}_{\{l+1,\ldots,p\}}$$

is an estimator with $k+1$ change-points on the interval $\{1, ..., p\}$ with the lowest cumulative costs given that the last change-point is at l. Again, by setting $l(p) = \arg\min_{R_{k-1} < l \leq R_k} d_l + d_{l+1,p}$ we obtain the estimator $\hat{f}(p) = \hat{f}(l(p), p)$ with the lowest cumulative costs $d_p = d_{1,l(p)} + d_{l(p)+1,p}$. Proceed these procedures until $d_{l+1,p+1} = \infty$ for all $R_{k-1} < l \leq R_k$ (then define $R_{k+1} = p$ and iterate) or until $p = n$ (then we get our estimator $\hat{f}_n = \hat{f}(n)$).

4.2 Simulation by SMUCE

We now investigate the performance of multiscale change-point segmentation methods from various perspectives. For brevity, we only consider a particular multiscale change-point segmentation method, SMUCE (Frick et al., 2014), and stress that the results are similar for other multiscale change-point segmentation methods (which are not shown here), see e.g. Li et al. (2016) for a further simulation study. For SMUCE, we use the implementation of an efficient pruned dynamic program from the CRAN R-package "stepR" (version 1.0), we select the system of all intervals with dyadic lengths for the multiscale constraint, and choose $q(\beta)$ by (2.6) as the threshold, which is simulated 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 $\|f\|_{L^2}/\sigma$.

4.2.1 Stability

We first examine the stability of multiscale change-point segmentation methods with respect to the significance level β in (2.6). The test signal f (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 4.1 presents the behavior of SMUCE with threshold $q = q(\beta)$ for different choices of significance level β. 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 $\beta \in [0.06, 0.94]$. For smaller β (< 0.06), SMUCE tends to underestimate the number of change-points (see the second panel of Figure 4.1 for example, where the missing change-point is marked by a vertical line), while, for larger β (> 0.94), it is inclined to recover false change points (as shown in the last panel of Figure 4.1). 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.

4.2.2 Different noise backgrounds

We next investigate the impact of noise level (or equivalently, of the SNR) on multiscale change-point segmentation methods. We consider the recovery of the Blocks signal (Donoho and Johnstone, 1994) for different noise levels. The estimation by SMUCE with significance level $\beta = 0.1$ is summarized in Figures 4.2. It shows that SMUCE recovers the signal rather well for low and medium noise levels, while misses one or two tiny features for high noise level.

4.2.3 Robustness

Furthermore, we study the robustness of multiscale change-point segmentation methods in case of model misspecification. To that end, we introduce a local trend component as in Olshen et al. (2004) and Zhang and Siegmund (2007) to the test signal f in Section 4.2.1, which leads to the model

$$y_i^n = \left(f\left(\frac{i}{n}\right) + 0.25b\sin(a\pi i)\right) + \xi_i^n, \qquad i = 0, \ldots, n-1. \tag{4.1}$$

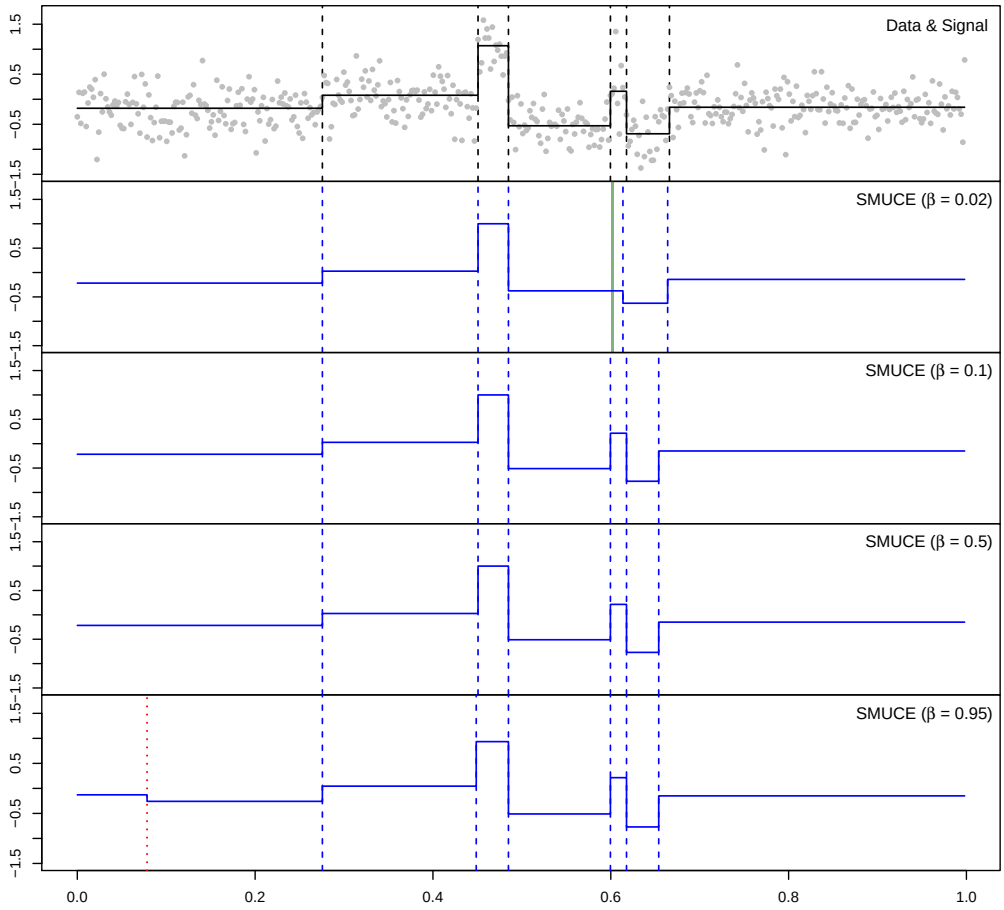

Figure 4.1: Estimation of step functions by SMUCE with $q = q(\beta)$ by (2.6) for different β (sample size $n = 497$, and SNR $= 1$).

In particular, we simulate the data using $a = 0.025$ and $b = 0.3$, and apply SMUCE again with various choices of β, see Figure 4.3. Similar phenomenon as in Figure 4.1 is viewed. For instance, SMUCE captures all relevant features of the signal still for a wide range of β ($0.08 \leq \beta \leq 0.29$, for the data in Figure 4.3).

On the other hand, when the parameter b is large enough, i.e., the fluctuation is strong enough, SMUCE is able to capture the fluctuation by inducing additional change-points. Figure 4.4 shows how SMUCE performs for the signal above with $b = 1.0$ and $b = 1.2$ under different noise levels.

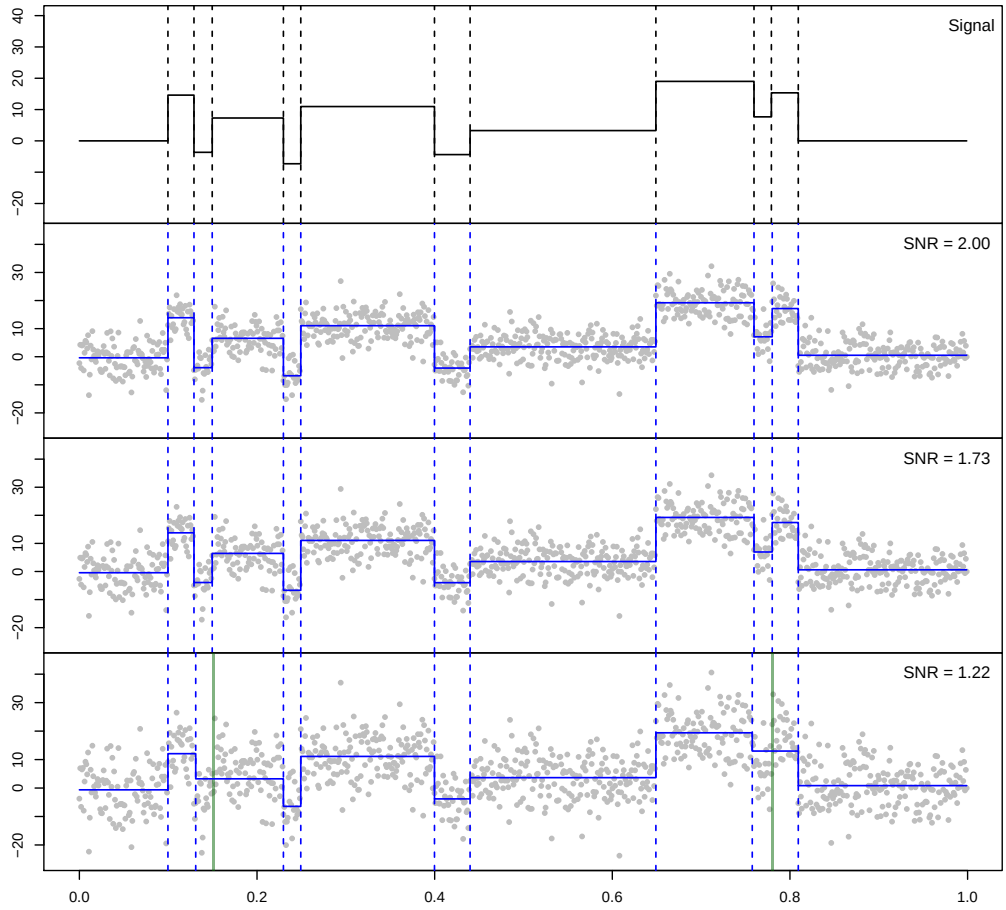

Figure 4.2: Estimation of the Blocks signal for various noise levels (sample size $n = 1{,}023$).

4.2.4 Empirical convergence rates

Lastly, we empirically explore the asymptotic behavior of multiscale change-point methods. The test signals are Blocks and Heavisine (Donoho and Johnstone, 1994). In Figure 4.5, we show the average of L^2-loss of SMUCE with significance level $\beta = 0.1$ over 20 repetitions for a range of sample sizes from 1,023 to 10,230. Note that the empirical convergence rates are quite close to the minimax optimal rates (indicated by slopes of the red straight lines), which confirms our theoretical findings in Theorems 3.1.4 and 3.2.1.

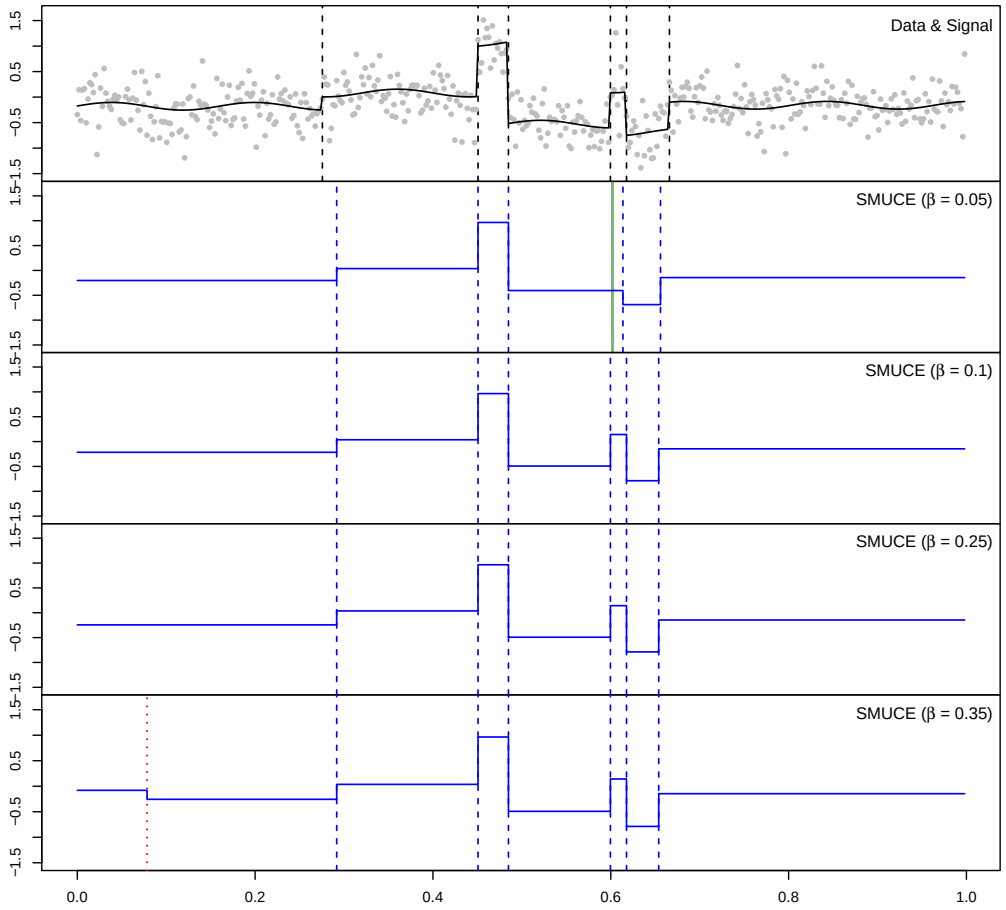

Figure 4.3: Estimation of step functions with local trends by SMUCE with $q = q(\beta)$ by (2.6) for different β (sample size $n = 497$, and SNR $= 1$).

4.3 Comparison

Next we compare the SMUCE and another multiscale change-point segmentation method, FDRSeg (Li et al., 2016) with some other change-points methods, PELT (Killick et al., 2012), CBS (Olshen et al., 2004; Venkatraman and Olshen, 2007) and WBS (Fryzlewicz, 2007). As mentioned in the Introduction, these methods work well in change-point problems either because they use exact and fast global optimization methods based on dynamic programming (PELT, SMUCE and FDRSeg) or because they use fast greedy methods based on local single change-point detection (CBS and WBS). Although all of these methods are developed under the setting of change-point models, we can still apply them

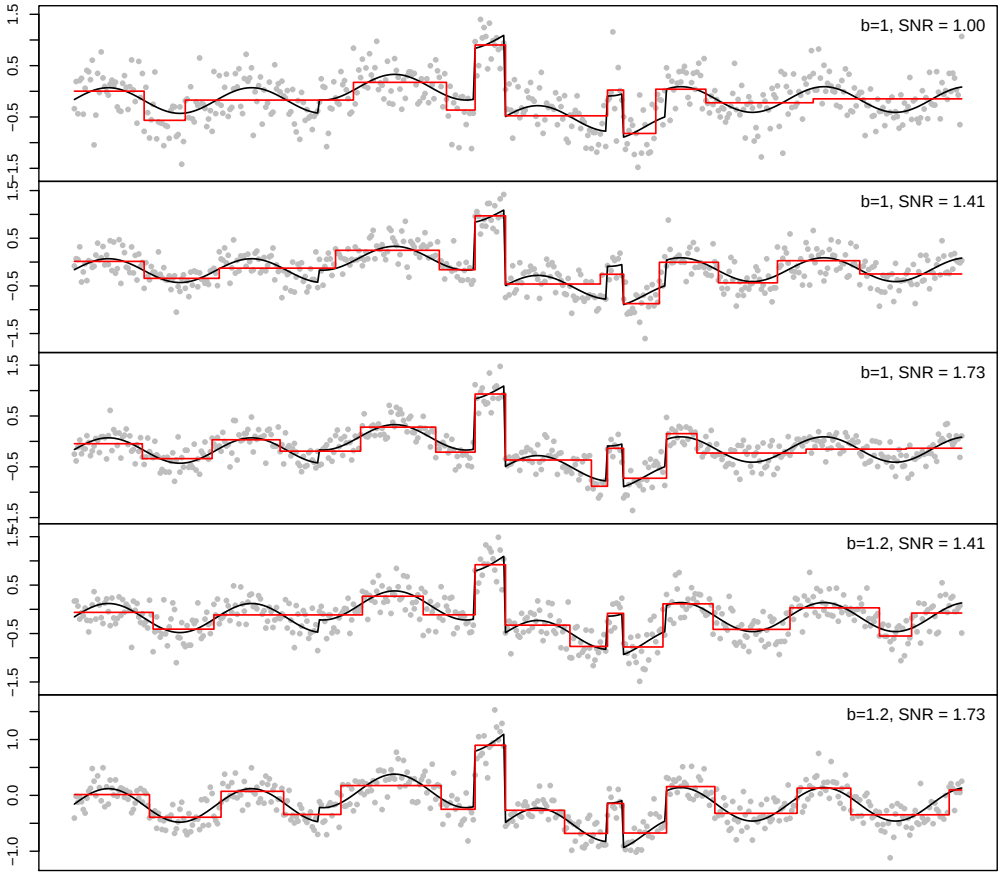

Figure 4.4: Estimation of the signal in (4.1) with $a = 0.025$ for various b and noise levels (sample size $n = 497$).

in non-step function models. Concerning implementation, we use CRAN R-packages "FDRSeg"(version $1.0 - 1$) for FDRSeg, "PSCBS" (version 0.61.0) for CBS, "wbs" for WBS, "changepoint" (version 2.2.2) for PELT.

4.3.1 Overview

Figure 4.6 shows the five methods' behavior in the estimation of the classical testing signals: Blocks, Bumps, Heavisine and Doppler (Donoho and Johnstone, 1994). We see they capture nearly all the main features of each signal, except that CBS misestimates some features for Bumps and Heavisine signals. Note that, compared to the other four methods, SMUCE

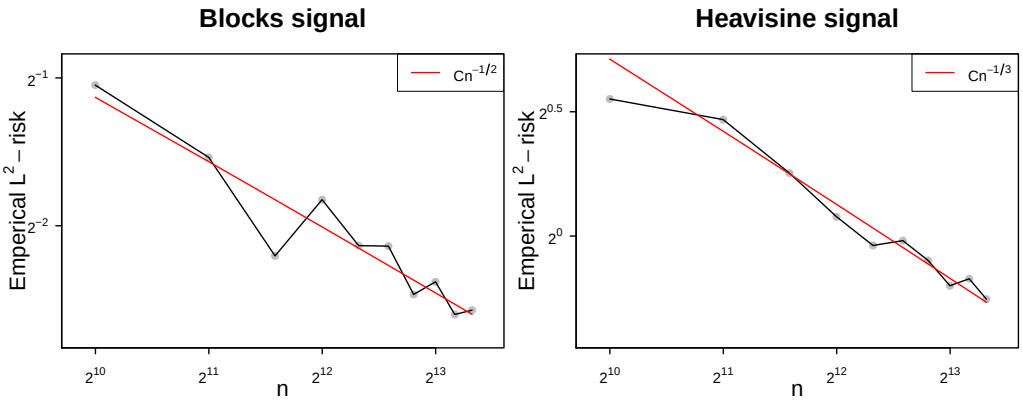

Figure 4.5: Empirical convergence rates of SMUCE for Blocks and Heavisine signals (SNR $= \sqrt{6}$).

captures all the main features and simultaneously controls the parsimony, i.e. the number of change-points.

4.3.2 Robustness

We also do the robustness comparison over these five methods. We use the same test signal with trend components as from (4.1). We see in Figure 4.7, SMUCE, FDRSeg, WBS, PELT obtain the right number of change-points, while CBS overestimate one change-point. On the other hand, Figure 4.8 shows when the trend parameter is large, these methods are also able to capture the fluctuant features by inducing additional change-points, where SMUCE and FDRSeg are with lower numbers of change-points, which is obvious due to their construction.

4.3.3 Empirical convergence rates

We end the simulation study by comparing the asymptotic behaviors of these methods. The test signals are Blocks and Heavisine (Donoho and Johnstone, 1994). Figure 4.9 shows the outcome of the five methods over 20 repetitions for a range of sample sizes from 1.023 to 10.230. It shows that for step-function signals, SMUCE and FDRSeg get a smaller L^2-loss while for the heavisine signal they get a larger L^2-loss. This is also due to the minimization of the number of change-points among the candidate functions.

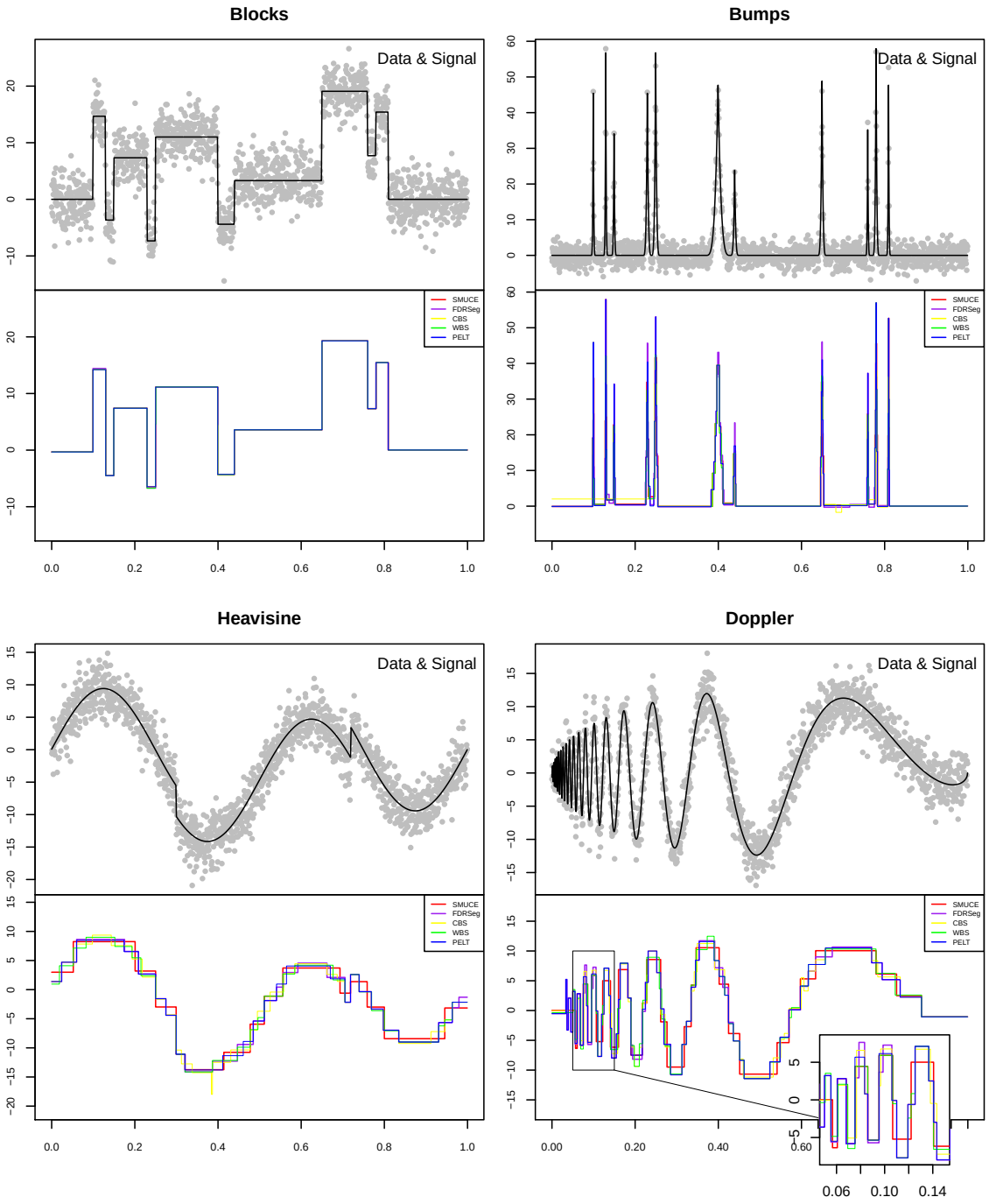

Figure 4.6: Estimation by different methods for Blocks, Bumps, Heavisine, and Doppler signals (sample size $n = 1.500$, and SNR $= \sqrt{11}$).

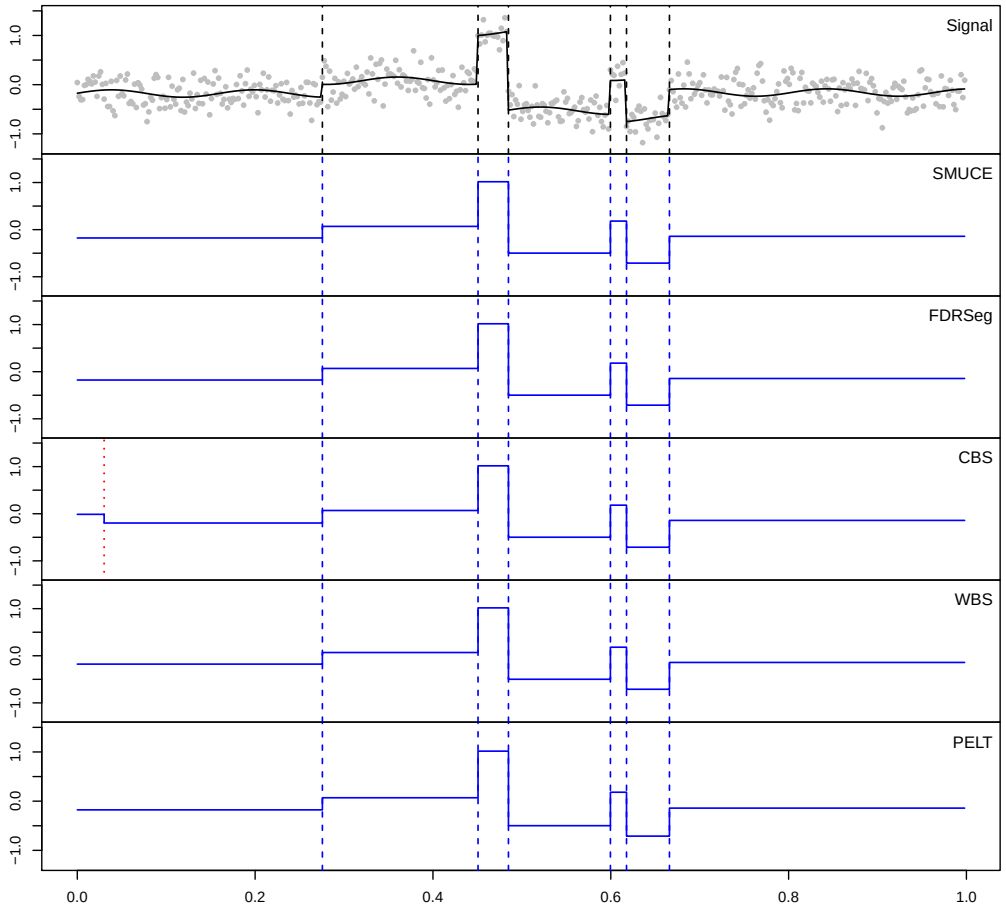

Figure 4.7: Estimation of the signal in (4.1) with $a = 0.025$ and $b = 0.3$ by different methods (sample size $n = 497$, and SNR $= \sqrt{2}$).

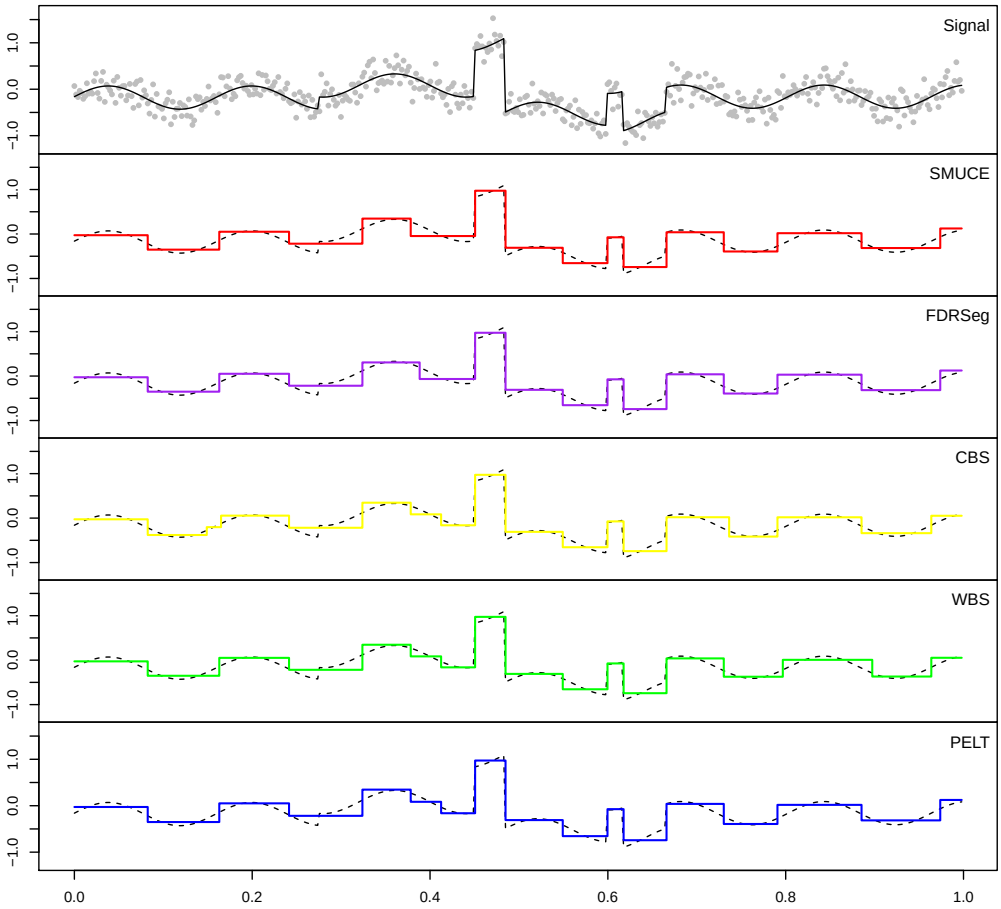

Figure 4.8: Estimation of the signal in (4.1) with $a = 0.025$ and $b = 1$ by different methods (sample size $n = 497$ and SNR $= \sqrt{3}$).

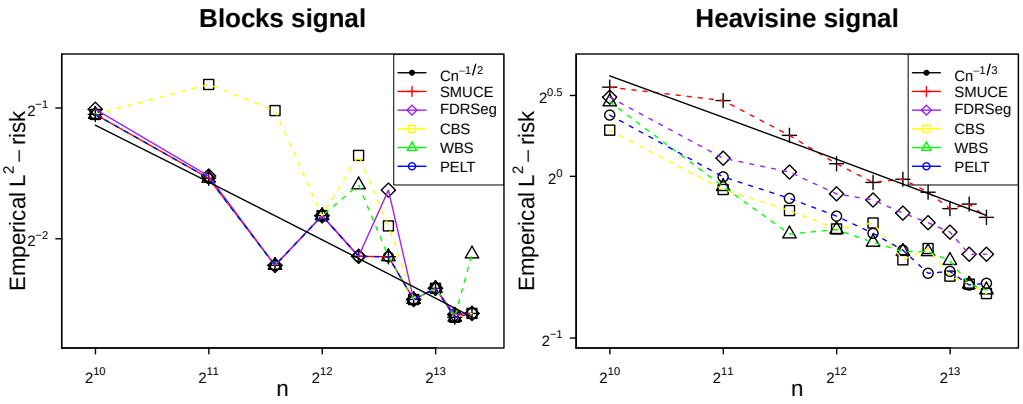

Figure 4.9: Empirical convergence rates of different methods for Blocks and Heavisine signals (SNR $= \sqrt{6}$).

5 Discussion and outlook

In this work we focused on convergence analysis for multiscale change-point segmentation methods, a general family of change-point estimators based on combination of variational estimation and multiple testing over different scales, in a nonparametric regression setting. Special emphasis was put on step functions while allowing for various distortions, where we found that estimation difficulty for is mainly determined by its number of jumps. We showed that multiscale change-point segmentation methods attain nearly optimal convergence rates for step functions with asymptotically bounded or even varying number of jumps.

As a robustness study, we also examined convergence behavior of these methods for more general functions, which are viewed as distorted jump functions. Such distortion is precisely characterized by classical approximation spaces. In particular, we derived nearly optimal convergence rates for multiscale change-point segmentation methods in case the regression function is either a (piecewise) Hölder function or a bounded variation function. Furthermore, it was shown that these methods automatically adapt to the unknown smoothness of the corresponding function classes, as the only tuning parameter can be selected in a universal way. The convergence rates also provide statistical justification with respect to detection of features, such as change-points, modes and troughs.

Finally, we collect some possible extensions of our methodology and theoretical analysis, which we plan to explore in future.

(a) Multiscale change-point segmentation methods cannot attain faster convergence rates for functions of stronger smoothness than above, since these estimators are piecewise constant. This can be improved by considering piecewise polynomial estimators (see e.g. Spokoiny, 1998). However, proper combination with multiscale methodology needs further investigation (see the rejoinder by Frick et al., 2014, for a first attempt). Alternatively, certain smoothness penalties can be selected instead of the number of jumps in the formulation of multiscale change-point segmentation, see e.g. Grasmair et al. (2015), where nearly optimal rates are shown for higher order Sobolev/Besov classes.

(b) Recall from Section 3.2, that by the classical approximation theory (c.f. DeVore, 1998), for any $f \in \mathcal{A}^\gamma$ and any number K of change-points, there always exists a best approximation of f by a step function $\tilde{f}_K$ with K number of change-points. It

is natural to ask how different the multiscale change-point segmentation estimator $\hat{f}_n$ is from the best approximation $\tilde{f}_{\#J(\hat{f}_n)}$. Although Theorem 3.2.1 already gives an asymptotic answer, the non-asymptotic case needs further research. In addition, recall Figure 4.3 and 4.4 where the estimator $\hat{f}_n$ detects no change-points on a constant interval with small fluctuations, but detects a change-point when the fluctuation is strong enough. Another issue is to further analyze precisely under which conditions $\hat{f}_n$ induces a change-point. It is clear that the fluctuation and noise level play an important role, as we can see from Figure 4.4.

(c) Our theory and analysis assumes the noise to be sub-Gaussian. Extension of our results to models with general errors beyond sub-Gaussian would be interesting as well. For instance, if considering exponential families, the corresponding regression model becomes

$$y_i^n \sim F_{f(i/n)}, \quad \textit{for } i = 0, \ldots, n-1,$$

where $\{F_\theta\}_{\theta \in \mathbb{R}}$ is a regular and minimal one-dimensional exponential family of distributions and $f \in \mathcal{D}([0,1))$. See Frick et al. (2014) for the case when f is a step function.

(d) Boysen et al. (2009) show that jump-penalized least squares estimators have a convergence rate of $(\log n/n)^{\alpha/(2\alpha+1)}$ with respect to L^2-loss when the underlying function is in Hölder class of order α, $0 < \alpha \leq 1$. Moreover, Fryzlewicz (2007) shows that unbalanced Haar wavelets based estimator has a convergence rate of $(1/n)^{\alpha/(2\alpha+1)} \log n$ with respect to L^2-loss when the underlying function is in Hölder class of order α, $0 < \alpha \leq 1$. Note that the empirical convergence rates of several change-point estimators in Figure 4.9 are all approximately of order $n^{1/3}$ for the Heavisine signal. We conjecture that these estimators also have a certain convergence rate when the underlying signal f is in a certain approximation space. Moreover, it is not clear how differences of estimation methods relate the differences of convergence rates. This raises challenging issues, which we plan to address in future.

Bibliography

Basseville, M. and Nikiforov, I. V. (1993). *Detection of abrupt changes: theory and application*. Prentice Hall Information and System Sciences Series. Prentice Hall, Inc., Englewood Cliffs, NJ.

Behr, M., Holmes, C., and Munk, A. (2016). Multiscale blind source separation. *arXiv preprint arXiv:1608.07173*.

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

Bigot, J. (2005). A scale-space approach with wavelets to singularity estimation. *ESAIM Probab. Stat.*, 9:143–164.

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

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(1):157–183.

Boysen, L., Liebscher, V., Munk, A., and Wittich, O. (2007). Scale space consistency of piecewise constant least squares estimators: Another look at the regressogram. *Lecture Notes-Monograph Series*, pages 65–84.

Brodsky, B. E. and Darkhovsky, B. S. (1993). *Nonparametric methods in change-point problems*, volume 243 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht.

Burchard, H. G. and Hale, D. F. (1975). Piecewise polynomial approximation on optimal meshes. *J. Approximation Theory*, 14(2):128–147.

Chen, J. and Gupta, A. K. (2000). *Parametric statistical change point analysis*. Birkhäuser Boston Inc., Boston, MA.

Csörgö, M. and Horváth, L. (1997). *Limit Theorems in Change-point Analysis*. John Wiley & Sons Ltd., Chichester.

Davies, L., Höhenrieder, C., and Krämer, W. (2012). Recursive computation of piecewise constant volatilities. *Comput. Stat. Data Anal.*, 56(11):3623 – 3631.

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

Dette, H., Munk, A., and Wagner, T. (1998). Estimating the variance in nonparametric regression—what is a reasonable choice? *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 60(4):751–764.

DeVore, R. A. (1998). Nonlinear approximation. In *Acta numerica, 1998*, volume 7 of *Acta Numer.*, pages 51–150. Cambridge Univ. Press, Cambridge.

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

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(19):e126.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455.

Du, C., Kao, C.-L. M., and Kou, S. C. (2015). Stepwise signal extraction via marginal likelihood. *J. Amer. Statist. Assoc.*, in press.

Dümbgen, L. and Spokoiny, V. G. (2001). Multiscale testing of qualitative hypotheses. *Ann. Statist.*, 29(1):124–152.

Farcomeni, A. (2014). Discussion of "multiscale change-point inference". *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 76(3):546–547.

Frick, K., Marnitz, P., and Munk, A. (2012). Statistical multiresolution Dantzig estimation in imaging: Fundamental concepts and algorithmic framework. *Electron. J. Stat.*, 6:231–268.

Frick, K., Munk, A., and Sieling, H. (2014). Multiscale change-point inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol., with discussion and rejoinder by the authors*, 76:495–580.

Fryzlewicz, P. (2007). Unbalanced Haar technique for nonparametric function estimation. *J. Amer. Statist. Assoc.*, 102(480):1318–1327.

Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *Ann. Statist.*, 42(6):2243–2281.

Futschik, A., Hotz, T., Munk, A., and Sieling, H. (2014). Multiresolution DNA partitioning: statistical evidence for segments. *Bioinformatics*, 30(16):2255–2262.

Gijbels, I., Hall, P., and Kneip, A. (1999). On the estimation of jump points in smooth curves. *Ann. Inst. Statist. Math.*, 51(2):231–251.

Grasmair, M., Li, H., and Munk, A. (2015). Variational multiscale nonparametric regression: smooth functions. *arXiv:1512.01068*.

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(492):1480–1493.

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(part II):201–215. Probastat '02. Part II.

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

Ibragimov, I. A. and Khas'minskiĭ, R. Z. (1982). Bounds for the risks of non-parametric regression estimates. *Theory Probab. Appl.*, 27(1):84–99.

Killick, R., Fearnhead, P., and Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *J. Amer. Statist. Assoc.*, 107(500):1590–1598.

Korostelev, A. P. (1988). On minimax estimation of a discontinuous signal. *Theory Probab. Appl.*, 32(4):727–730.

Lai, T. L. (2001). Sequential analysis: some classical problems and new challenges. *Statist. Sinica*, 11(2):303–408. With comments and a rejoinder by the author.

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(19):3763–3770.

Li, H., Munk, A., and Sieling, H. (2016). FDR-control in multiscale change-point segmentation. *Electron. J. Stat.*, 10(1):918–959.

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*, page 42.

Linton, O. and Seo, M. H. (2014). Discussion of "multiscale change-point inference". *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 76(3):548.

Munk, A. and Dette, H. (1998). Nonparametric comparison of several regression functions: exact and asymptotic theory. *Ann. Statist.*, 26(6):2339–2368.

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).

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(4):557–572.

Pein, F., Sieling, H., and Munk, A. (2015). Heterogeneous change point inference. *arXiv preprint arXiv:1505.04898*.

Pietsch, A. (1981). Approximation spaces. *Journal of Approximation Theory*, 32(2):115–134.

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.

Schmidt-Hieber, J., Munk, A., and Dümbgen, L. (2013). Multiscale methods for shape constraints in deconvolution: confidence statements for qualitative features. *Ann. Statist.*, 41(3):1299–1328.

Scott, A. J. and Knott, M. (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, 30(3):pp. 507–512.

Shao, Q. M. (1995). On a conjecture of Révész. *Proc. Amer. Math. Soc.*, 123(2):575–582.

Siegmund, D. and Yakir, B. (2000). Tail probabilities for the null distribution of scanning statistics. *Bernoulli*, 6(2):191–213.

Song, R., Banerjee, M., and Kosorok, M. R. (2016). Asymptotics for change-point models under varying degrees of mis-specification. *Ann. Statist.*, 44(1):153–182.

Spokoiny, V. G. (1998). Estimation of a function with discontinuities via local polynomial fit with an adaptive window choice. *Ann. Statist.*, 26(4):1356–1378.

Tibshirani, R. and Wang, P. (2008). Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics*, 9(1):18–29.

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

Venkatraman, E. S. and Olshen, A. B. (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, 23(6):657–663.

Walther, G. (2010). Optimal and fast detection of spatial clusters with scan statistics. *Ann. Statist.*, 38(2):1010–1033.

Wu, Y. (2005). *Inference for change-point and post-change means after a CUSUM test*, volume 180 of *Lecture Notes in Statistics*. Springer, New York.

Yao, Y.-C. and Au, S. T. (1989). Least-squares estimation of a step function. *Sankhyā Ser. A*, 51(3):370–381.

Zhang, N. and Siegmund, D. (2007). A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63(1):22–32.

Zhang, N. and Siegmund, D. (2012). Model selection for high-dimensional, multi-sequence change-point problems. *Statist. Sinica*, 22:1507–1538.

Curriculum Vitae

Name	Qinghai Guo
Address	Goldschmidtstrasse 7 37077 Göttingen, Germany
Email:	qguo@gwdg.de

Personal Details

Gender	Male
Date of birth	November 2, 1989
Place of birth	Jiangxi, China
Citizenship	Chinese

Education

Since 04/2013	Ph.D. student of mathematics at the University of Göttingen, Germany Supervisor: Prof. Dr. Axel Munk, Jun.-Prof. Dr. Andrea Krajina
04/2011-03/2013	Master student of mathematical science at Yamagata University, Japan Supervisor: Prof. Dr. Hiroyuki Matsumoto
09/2006-07/2010	Student of mathematics and applied mathematics at Beijing Forestry University, China
09/2003-06/2006	Secondary school "Nankang middle school" in Jiangxi, China
09/2000-06/2003	Middle school "Rongjiang middle school" in Jiangxi, China

---

Research Experience

Since 04/2013	Member of the Research Group SFB 803 "Functionality controlled by organization in and between membranes"