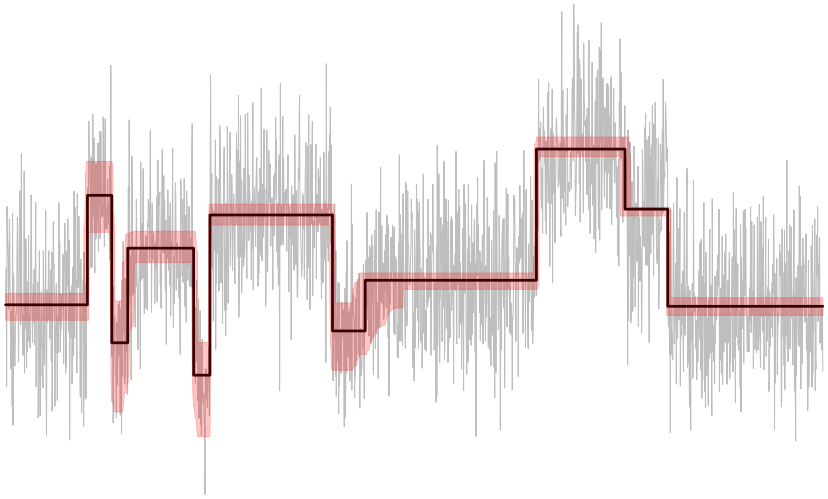
Statistical Multiscale Segmentation: Inference, Algorithms and Applications



Dissertation zur Erlangung

des mathematisch-naturwissenschaftlichen Doktorgrades

"Doctor rerum naturalium"
der Georg-August-Universität zu Göttingen

im Promotionsprogramm

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

vorgelegt von
Hannes Sieling
aus Oldenburg (Oldb)

Göttingen, 2013

Betreuungsausschuss:

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

Prof. Dr. Anja Sturm,
Institut für Mathematische Stochastik, Universität Göttingen

Mitglieder der Prüfungskommission:

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

Korreferent:
Prof. Dr. Dominic Schuhmacher,
Institut für Mathematische Stochastik, Universität Göttingen

Weitere Mitglieder der Prüfungskommission:

Prof. Dr. Dorothea Bahns,
Mathematisches Institut, Universität Göttingen

Prof. Dr. Tatyana Krivobokova,
Institut für Mathematische Stochastik, Universität Göttingen

Prof. Dr. Stephan Waack,
Institut fr Informatik, Universität Göttingen

Prof. Dr. Max Wardetzki,
Institut fr Numerische und Angewandte Mathematik, Universität Göttingen

Tag der mündlichen Püfung: 22.01.2014

Acknowledgment

First and foremost, I would like to express my very great appreciation to my advisor Prof. Axel Munk for providing the interesting and challenging topic of this thesis. His guidance and enthusiasm have been a great encouragement throughout my work. He has always kept an open mind to my ideas during various vivid and interesting discussions, and his stimulating contributions were fundamental to the completion of this thesis. His statistical intuition was a great inspiration and essentially formed my statistical understanding.
Further, I would like to thank Prof. Dominic Schuhmacher for taking on the Korreferat.
Special thanks should be given to Klaus Frick for his extraordinary assistance with this work, as well as for his companionship and encouragement throughout the time spent together at the IMS.
Thomas Hotz expertise in statistical computing has been vital for this thesis and related topics.
I am grateful to Prof. Lutz Dümbgen from the University Bern for sharing fruitful comments and ideas during my visit.
I would like to offer my special thanks to Prof. Günther Walther from Stanford University for enlightening discussions on this work and related topics. I deeply appreciate him sharing his ideas with me and providing a very pleasant stay in Stanford.
I wish to acknowledge the help of Prof. Chris Holmes from Oxford University, who introduced me to some interesting challenges in statistical genomics.
Support provided by the DFG-SNF research group 916 "Statistical Regularization and Qualitative Constraints" was greatly appreciated.
I am particularly grateful to my colleagues for providing a pleasant experience at the IMS. Special thanks should be given to Till Sabel, Rebekka Brink-Spalink and Ina Schachtschneider for generating a supportive, creative and very enjoyable office environment.
Finally, I would like to express my heartfelt thanks to my parents and my girlfriend, Birte Dunker, for their constant support and encouragement.

Preface

Piecewise constant step functions with a finite number of change-points provide a suitable regression model in many situations. Estimation of such change-point functions is deemed to be a classical problem in statistics, which experienced a revival with applications in various interdisciplinary fields in recent years. Two examples that received particular attention are the detection of gene copy number aberrations in genomics and the unveiling of changes in the volatility of time series in financial econometrics.

This thesis mainly concerns change-point models with independent observations from an exponential family, with constant mean in between change-points. An inferential scheme for estimation and confidence statements based on a multiscale statistic is provided, which allows for efficient and accurate detection of multiple change-points. A universal bound for the asymptotic null-distribution of the considered multiscale statistic is derived. Based on this, the probability of over- and underestimation of change-points is bounded explicitly. From these bounds, model consistency is obtained and (asymptotically) honest confidence sets for the unknown change-point function and its change-points are constructed. It is shown that the change-point locations are estimated at the minimax rate $\mathcal{O}(n^{-1})$ up to a logarithmic term. Moreover, the optimal detection rate of vanishing signals as $n \to \infty$ is attained.

The general methodology, as in Section 1 and Section 2, and large parts of the theory in Section 3 have been published in Frick et al. (2013). However, several theoretical findings are extended and refined, as described precisely at the beginning of Section 3.

It is shown how dynamic programming can be used for efficient computation of estimators, confidence intervals and confidence bands for the change-point function.

The performance and robustness of the approach are illustrated in various simulations. The proposed estimate has been applied to DNA segmentation (Futschik et al., 2013) and with some modifications to idealization of ion-channel recordings (Hotz et al., 2012). Both papers are not part of this thesis, yet, the application in Futschik et al. (2013) is illustrated by means of a data set from the literature in Section 6.7.

This thesis extends the work of Frick et al. (2013) by including two generalizations beyond exponential families (Section 5). In addition, an approach is derived, which is tailor-suited for applications in which the change-point function is known to have few different values (Section 7). Finally, extensions and modifications that give motivation for future work are discussed in Section 8.

Contents

List of Symbols

$\mathbf{E}[X], \mathbf{Var}[X], \mathbf{med}[X]$	expected value, variance, median of X
$\mathcal{N}(\mu, \sigma^2)$	normal distribution with mean μ and variance σ^2
$U[a, b]$	uniform distribution on $[a, b]$
$C(l, s)$	Cauchy distribution with location l and scale s
χ^2_k	chi-squared distribution with k degrees of freedom
$\xrightarrow{\mathcal{D}}$	convergence in distribution
$\overset{\mathcal{D}}{=}$	equality in distribution
$\overset{\mathcal{D}}{\leq}, \overset{\mathcal{D}}{\geq}$	bounded in distribution
$\mathcal{O}_{\mathbf{P}}()$	a sequence of random variables $(X_n)_{n\in\mathbb{N}}$ is $\mathcal{O}_{\mathbf{P}}(a_n)$, if X_n/a_n is bounded in probability.
$\mathbb{N}_0$	set of non-negative integers
$\#I$	number of observation in the interval $I \subset [0, 1]$
$\lvert I\rvert$	Lebesgue measure of the interval $I \subset [0, 1]$
$\mathbf{1}$	indicator function
$\dot{\psi}, \ddot{\psi}$	first and second derivative of ψ
$\lVert x\rVert_p$, $\lVert x\rVert_{\mathrm{TV}}$	l_p norm and total variation semi-norm of x

SECTION 1

Introduction

We assume that independent random variables $Y = (Y_1, \ldots, Y_n)$ are given by the regression model

$$Y_i \sim F_{\vartheta(i/n)}, \qquad \text{for } i = 1, \ldots, n. \tag{1.1}$$

Here, $\{F_\theta\}_{\theta \in \Theta}$ is a one-dimensional exponential family with densities f_θ and the regression function $\vartheta : [0, 1) \to \Theta \subseteq \mathbb{R}$ is a right-continuous change-point function with an unknown number K of change-points. The change-points locations will be denoted by $(\tau_1, \ldots, \tau_K)$ and the value of the function by $(\theta_1, \ldots, \theta_K)$. Figure 1 depicts such a step function with $K = 11$ change-points and corresponding data Y for the Gaussian family $F_\theta = \mathcal{N}(\theta, 1)$. A formal definition of the model is given in Section 2.1. The statistical problem related with this model is often referred to as the *change-point problem* (Carlstein et al., 1994) and consists in estimating

(i) the number of change-points K,

(ii) the change-point locations $(\tau_1, \ldots, \tau_K)$ and

(iii) the function values $(\theta_1, \ldots, \theta_K)$.

Additionally, we address the more involved issue

(iv) of constructing simultaneous asymptotic confidence statements for the function ϑ, its number of change-points and its change-point locations.

Within this work we present an approach to the change-point problem based on a multiscale test statistic. In general, the problem of detecting changes in the characteristics of a sequence of observations has a long history in statistics and related fields, dating back to the 1950's (see e.g. Page (1955)). For a selective survey, we refer the reader also to the books of Basseville and Nikiforov (1993), Brodsky and Darkhovsky (1993), Csörgö and Horváth (1997), Chen and Gupta (2000), Wu (2005) and the extensive list in Khodadadi and Asgharian (2008).

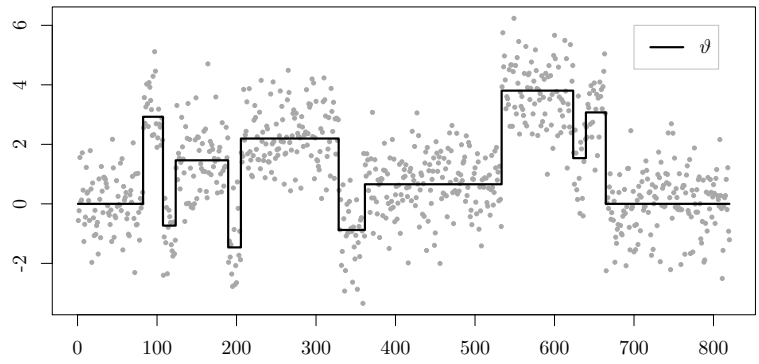


Figure 1: Example of a regression function $\vartheta \in S$ with Gaussian observations Y and variance $\sigma^2 = 1$.

In recent years, the change-point problem experienced a renaissance in the context of regression analysis due to novel applications that mainly came along with the rapid development in genetic engineering, looking at detection of changes in gene copy numbers in the genome (Jeng et al., 2010; Lai et al., 2005; Olshen et al., 2004; Zhang and Siegmund, 2007). Also in the context of detecting changes in the volatility of time series in financial econometrics much research has been done (Davies et al., 2012; Inclán and Tiao, 1994; Lavielle and Teyssière, 2007; Spokoiny, 2009). Motivated by these applications for large data sets, fast computation of estimates is crucial and a lot of work on efficient algorithms was carried out recently (see e.g. Friedrich et al. (2008), Killick et al. (2011) and Venkatraman and Olshen (2007)).

1.1 Method

In order to address the points (i) − (iv), we propose a methodology, which can be considered as a hybrid method of two well-established approaches to the change-point problem.
Likelihood ratio statistics are frequently employed to test for a change in the parameter of the distribution family and to construct confidence regions for change-point locations. Approaches of this type date back as far as Chernoff and Zacks (1964), Kander and Zacks (1966) and have gained considerable attention afterwards (Dümbgen, 1991; Hinkley, 1970; Hinkley and Hinkley, 1970; Hušková and Antoch, 2003; Siegmund, 1988; Worsley, 1983, 1986). The likelihood-ratio test was also extensively studied for sequential change-point analysis (Siegmund, 1986; Siegmund and Venkatraman, 1995; Yakir and Pollak, 1998). These methods are primarily designed to detect a predefined maximal number (mostly one) of change-points. A generalization of this approach towards testing of multiple (i.e. an unknown number of) change-points yields a multiple testing problem. Such problems have e.g. been addressed by multiscale (scanning) statistics, see Dümbgen and Spokoiny (2001), Dümbgen and Walther

(2008) and in the context of change-point regression Siegmund and Yakir (2000). In this work we employ a multiscale statistic which will be derived in detail in Section 2.2 and is based on results from Dümbgen and Spokoiny (2001). By these approaches simultaneous confidence statements about multiple qualitative features are obtained, which makes this approach particular suitable for the problem raised in (iv). Moreover, it was shown in Chan and Walther (2013) that statistics of this kind achieve optimality in detection of signals on segments of any lengths simultaneously.

Another popular approach in change-point regression is based on minimizing a *penalized cost function*, i.e. solving an optimization problem of the form

$$\inf_{\vartheta \in \mathcal{S}} c(Y, \vartheta) + \mathrm{pen}(\vartheta). \tag{1.2}$$

Here the cost function $c(Y, \vartheta)$ serves as a goodness-of-fit measure and the penalty term $\mathrm{pen}(\vartheta)$, which may e.g. depend on the number of change-points, penalizes the complexity of ϑ and prevents over-fitting. It increases with the dimension of the model and provides a *model selection criterion*. A minimizer of the optimization problem (1.2) naturally provides solutions for (i)-(iii).

A special case of (1.2) is linear penalization of the number of change-points, more precisely $\mathrm{pen}(\vartheta) = \omega \# J(\vartheta)$, which has been considered in Yao (1988) and Yao and Au (1989) with a BIC type weight $\omega \sim \log n$. Model selection based ℓ_0-penalized functionals, which are nonlinear in $\# J(\vartheta)$ have been investigated in Birgé and Massart (2001) for change-point regression. Furthermore, Zhang and Siegmund (2007) introduced a penalty, which depends on the number of change-points and additionally on its locations. Various methods based on weighted l_0-penalties have since been developed in Braun et al. (2000), Winkler and Liebscher (2002), Wittich et al. (2008) and Boysen et al. (2009). As an eligible property of l_0-penalization, it was shown that exact solutions of such optimization problems can often be computed efficiently by dynamic programing (see the literature in Section 4 for a selective overview on the literature).

In many situations the optimization problem in (1.2) may equivalently be written as

$$\inf_{\vartheta \in \mathcal{S}} \mathrm{pen}(\vartheta) \quad \text{s.t.} \quad c(Y, \vartheta) \leq q, \tag{1.3}$$

for some (unknown) threshold $q > 0$. In this work, we combine these two ideas and propose to solve an optimization problem of the type (1.3), where the goodness-of-fit measure c is chosen to be a multiscale statistic. This statistic will be restricted to constant parts of ϑ, which makes dynamic programing applicable while maintaining optimal detection properties of the multiscale statistic. By this the above mentioned advantages of both approaches are combined, as we will point out in this work: on the one hand, we obtain confidence statements

for the estimate originating from the multiscale statistic (see Section 3) and on the other hand we show that it can be implemented with worst case complexity $\mathcal{O}(n^2)$ by dynamic programing (see Section 4).

In order to outline the estimation procedure, let $T_n(Y, \vartheta)$ denote a (later specified) multiscale statistic. The goals (i)-(iv) will then be achieved based on an estimation and inference method for the change-point problem in exponential families: the **S**imultaneous **MU**ltiscale **C**hange-point **E**stimator (SMUCE). For $\vartheta \in \mathcal{S}$ we denote by $J(\vartheta)$ the ordered vector of change-points and by $\#J(\vartheta)$ its length, i.e. the number of change-points. We consider the optimization problem

$$\inf_{\vartheta \in \mathcal{S}} \#J(\vartheta) \quad \text{s.t.} \quad T_n(Y, \vartheta) \leq q. \tag{1.4}$$

SMUCE addresses change-point regression in two simultaneously combined estimation steps: model selection (estimation of K) and estimation of ϑ given K. The minimal value of $\#J(\vartheta)$ in (1.4) gives the estimated number of change-points, denoted by $\hat{K}(q)$. To obtain an estimator for ϑ first consider the set of all solutions of (1.4) given by

$$\mathcal{H}(q) = \left\{ \vartheta \in \mathcal{S} \ : \ \#J(\vartheta) = \hat{K}(q) \text{ and } T_n(Y, \vartheta) \leq q \right\}. \tag{1.5}$$

We will show in Section 3.6 that $\mathcal{H}(q)$ constitutes a confidence set for the true regression function ϑ. Based on this confidence set, we address (iv) and derive confidence bands for ϑ and confidence intervals for the change-point locations. As the final estimate $\hat{\vartheta}(q)$ for ϑ we propose the *constrained maximum likelihood estimator* within this confidence set, i.e.

$$\hat{\vartheta}(q) = \operatorname*{argmax}_{\vartheta \in \mathcal{H}(q)} \sum_{i=1}^{n} \log \left(f_{\vartheta(i/n)}(Y_i) \right). \tag{1.6}$$

Since $\hat{\vartheta}(q)$ implies an estimate of the change-point locations and function values, this provides a solution to (ii)-(iii). Figure 2 shows the SMUCE (red solid line) for the data example in Figure 1. As stressed above, the multiscale constraint on the r.h.s. of (1.4) renders SMUCE sensitive to the multiscale nature of the signal ϑ. The signal in Figure 2 illustrates this as the signal is recovered on large and small scales equally well.

1.2 Related work

Estimates, which minimize target functionals under a statistical multiscale constraint have been already considered in Nemirovski (1985), Donoho (1995) and more recently in Davies and Kovac (2001), Candès and Tao (2007), Davies et al. (2009) and Frick et al. (2012). To piecewise constant regression this idea was first applied in Höhenrieder (2008) for approximation of financial data in a Gaussian model, see also Davies et al. (2012). There it was also shown that

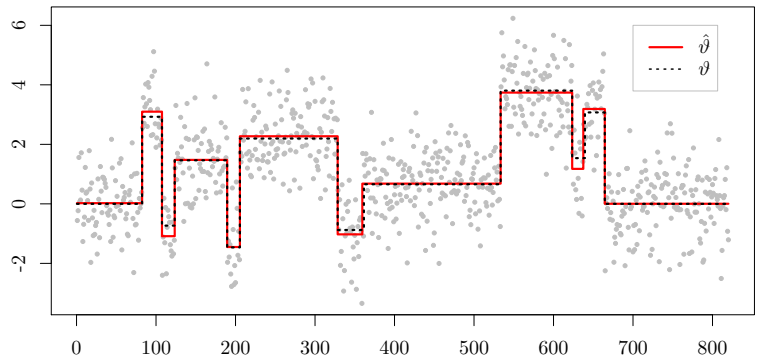


Figure 2: Example of a regression function $\vartheta \in S$ (black, dotted line) with Gaussian observations Y and variance $\sigma^2 = 1$ and SMUCE (solid, red line).

the reduction to a multiscale statistic acting on constant parts makes dynamic programing applicable (see Section 4 for more details).

The literature in Section 1.1 can be complemented by further prominent penalization approaches of the type (1.2) including the fused lasso procedure (Friedman et al., 2007; Tibshirani et al., 2005) and Harchaoui and Lévy-Leduc (2010) that use a linear combination of the total-variation and the ℓ^1-norm to penalize complexity. Multiscale based partitioning methods include binary segmentation in Sen and Srivastava (1975), Vostrikova (1981), Olshen et al. (2004) and Fryzlewicz (2012). Besides the already mentioned frequentists work, there are also several Bayesian approaches to the change-point problem. For some recent literature, we refer to Du and Kou (2012), Fearnhead (2006), Luong et al. (2012), Rigaill et al. (2012) and the references therein.

1.3 Main results

1.3.1 Deviation bounds and confidence sets

The parameter $q \in \mathbb{R}$ in (1.4) plays a crucial role for estimation as it governs the trade-off between data-fit and parsimony, represented by the number of change-points. It has an immediate statistical interpretation. From (1.4) it follows that

$$\mathbf{P}\left(\hat{K}(q) > K\right) \leq \mathbf{P}(T_n(Y, \vartheta) > q). \tag{1.7}$$

Hence, by choosing $q = q(\alpha)$ to be the $(1-\alpha)$-quantile of the (asymptotic) null-distribution of $T_n(Y, \vartheta)$, we can (asymptotically) control the *probability of overestimating* the number of change-points by α. In fact, we show that the null-distribution of $T_n(Y, \vartheta)$ can be bounded asymptotically by a distribution which is independent of ϑ (see Section 3.1). In addition,

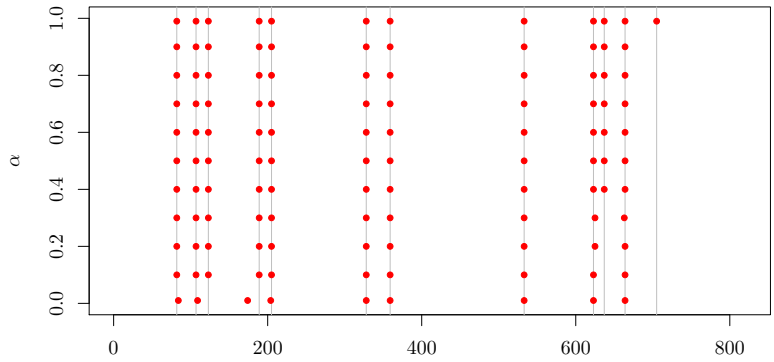


Figure 3: Estimated change-points (red dots) for the signal in Figure 1 and different values of α. The true change-point locations are shown grey vertical lines.

in Theorem 37 we provide an estimate for the tails of this limit distribution, which yields explicit bounds. It is noteworthy that for Gaussian observations these bounds are even non-asymptotic (see Section 3.5). In Figure 3 we reconsider the previous example and show for different choices of α (y-axis) the corresponding estimates for the change-point locations (red dots). The vertical ticks mark the true change-point locations. The number of estimated change-points is monotonically increasing in α in accordance with (1.7), which guarantees at error level α that SMUCE has not more change-points than the true signal ϑ.

As mentioned before, the threshold $q(\alpha)$ for SMUCE automatically controls the probability of overestimating the number of change-points. In Section 3.2 we prove a refinement (Theorem 5) which actually shows that for any $k \in \mathbb{N}_0$

$$\mathbf{P}\left(\hat{K}(q(\alpha)) - K > 2k\right) \leq \alpha^{k+1}.$$

Based on this bound we will derive an upper bound for the *expected number* of overestimated change-points (Corollary 6). This bound in turn opens the opportunity for a data-driven choice of q, based on controlling the false discovery rate (FDR), as we will show in Section 8.1.

In addition, we prove an upper bound for the *probability of underestimating* the number of change-points. Any such bound necessarily depends on characteristics of the signal ϑ, as no method can recover arbitrary fine features for given sample size n, see Donoho (1988) for a rigorous argument in the context of density estimation. Our bound (see Theorem 7) reflects this fact and is given in terms of the length of segments of ϑ and the height of its jumps. A simplified version, which only depends on the smallest interval length Λ, the smallest absolute

jump size Δ and the number of change-points K of the true regression function ϑ reads as

$$\mathbf{P}\left(\hat{K}(q) < K\right) \leq 2Ke^{-Cn\Lambda\Delta^2}\left[e^{\left(q+\sqrt{2\log(e/\Lambda)}\right)^2}+1\right]. \tag{1.8}$$

Here, $C > 0$ is some known universal constant only depending on the family of distributions (see Section 3.3). While the bounds for overestimation are essentially build on the control of the null-distribution of T_n, these bounds rely on power approximations for the *local* test statistics. For the case of Gaussian observations we derive the detection power of the multiscale statistic T_n, i.e. we determine the rate and constants at which a signal may vanish with increasing n but still can be detected with probability 1, asymptotically. For the task of detecting a single constant signal against a noisy background, we prove that the obtained rate is optimal (cf. Dümbgen and Spokoiny (2001), Dümbgen and Walther (2008) and Chan and Walther (2013)). Further, we extend this result to the case of an arbitrary number of change-points, retrieving the same optimal rate but different constants (Section 3.5.1).
As a consequence of the bounds for over- and underestimation, $\mathcal{H}(q(\alpha))$ in (1.5) constitutes an asymptotic confidence set at level $1-\alpha$ and we will explain in Section 4.5 how confidence bands for the graph of ϑ and confidence intervals for its change-points can be obtained from this. Of course, honest (i.e. uniform) confidence sets cannot be obtained on the entire set of step functions $\mathcal{S}$, as Δ and Λ can become arbitrarily small. Nevertheless, we can show that simultaneously, confidence bands for ϑ and intervals for the change-points are both *asymptotically honest with respect to to a sequence of nested models* $\mathcal{S}^{(n)} \subset \mathcal{S}$ that satisfy

$$\frac{n}{\log n}\Delta_n^2\Lambda_n \to \infty, \qquad \text{as } n \to \infty. \tag{1.9}$$

In other words, the confidence level α is kept uniformly over $\mathcal{S}^{(n)}$ as $n \to \infty$ (c.f. Section 3.6). Here Λ_n and Δ_n represent the smallest interval length and smallest absolute jump size in $\mathcal{S}^{(n)}$, respectively.

1.4 Beyond exponential families

Even though the results in Section 3 generally rely on the restriction to exponential families, the SMUCE methodology can be applied to other distributions. Extending the results from Section 3.1, we show that the null-distribution of the multiscale statistic with Gaussian likelihoods converges to the same limit distribution for any sub-Gaussian additive noise. This makes the procedure applicable in this more general model (Section 5.1). These findings may also be understood as a certain robustness property of the SMUCE with Gaussian likelihood, which is confirmed by simulations in Section 6.6 for uniformly distributed noise.
Moreover, we provide a modification of SMUCE for quantile regression. The approach is based

on a multiscale analysis of the signs of residuals, and is hence applicable to any distributions (Section 5.2).

1.5 Implementation and Software

The applicability of dynamic programming to the change-point problem has been subject of research recently (Auger and Lawrence, 1989; Fearnhead, 2006; Friedrich et al., 2008; Harchaoui and Lévy-Leduc, 2010; Jackson et al., 2005). The SMUCE $\hat{\vartheta}(q)$ can also be computed by a dynamic program due to the restriction of the local likelihoods to the constant parts of candidate functions. This was shown in Höhenrieder (2008) for the multiscale constraint considered there.

Much in the spirit of the dynamic program suggested in Killick et al. (2011), our implementation exploits the structure of the constraint set in (1.6) to include pruning steps. These reduce the worst case computation time $\mathcal{O}(n^2)$ considerably in practice. Simultaneously, the algorithm returns a confidence band for the graph of ϑ as well as confidence intervals for the location of the change-points (Section 4.5), the latter without any additional cost. A complete pseudo-code of the algorithm is given and complexity and computation time are discussed. An R-package (stepR) including an implementation of the pruned dynamic program for SMUCE is available (Hotz and Sieling, 2013)[1].

1.6 Choice of q, simulations and applications

We investigate the performance of our approach in simulations and real world data examples. For this purpose, we first discuss the choice of the threshold parameter q. As pointed out above, q can be chosen such that the probability of overestimation is controlled. Moreover, balancing the probabilities for over- and underestimation gives an upper bound on $\mathbf{P}(\hat{K}(q) \neq K)$, i.e. the probability that the number of change-points is misspecified. This bound depends on n, q, Λ and Δ in an explicit way and opens the door for several strategies to select q, e.g. such that $\mathbf{P}(\hat{K}(q) = K)$ is maximized if prior information on Δ and Λ is incorporated. We discuss different approaches and suggest a simple way how to do this in Section 6.1. Additionally, we relate our findings to false and true discoveries in Section 8.1. From this in turn we derive an alternative, data-driven parameter choice, designed to control the false discovery rate.

Extensive simulations reveal that SMUCE is competitive with state-of-the-art methods for the change-point problem. Our simulation study includes the CBS method (Olshen et al., 2004), the fused lasso (Tibshirani et al., 2005) and the modified BIC (Zhang and Siegmund, 2007) for Gaussian regression, the multiscale estimator in Davies et al. (2012) for piecewise

[1] R package available at `http://www.stochastik.math.uni-goettingen.de/smuce`

constant volatility estimation and the extended taut string method for quantile regression in Dümbgen and Kovac (2009). In our simulations we consider several risk measures, including the mean integrated squared error (MISE), the mean integrated absolute error (MIAE) and the model selection error $\mathbf{P}(\hat{K} \neq K)$. Within these simulations the robustness to violations of the assumption of a piecewise constant function is investigated.
As stressed before the applications for change-point models are vast. Besides the data examples in Frick et al. (2013) the procedure underlying SMUCE has been applied to idealization of ion channels recordings (Hotz et al., 2012) and to segmentation of DNA-sequences (Futschik et al., 2013). In extension to the results in Futschik et al. (2013) we illustrate the capacity of SMUCE by means of a data example from the literature.

1.7 Multiscale segmentation with few levels

A modification of SMUCE is presented, which is designed for applications in which it is known that the signal only takes few different values. The application, which we bear in mind is the analysis of array CGH data. It is shown how the prior information of few different values can be incorporated into the estimation procedure underlying SMUCE. The superiority of the modified approach is illustrated in simulations and it is applied to an array CGH data set, which has been considered in Snijders et al. (2001) and Olshen et al. (2004).

1.8 Discussion

In this section possible extensions and modifications of the proposed methodology are discussed. Motivated by the bounds for the expected number of overestimated change-points in Section 3.2 we relate our findings to false discoveries. From this in turn we derive a data-driven choice of q and show promising results in simulations.
Moreover, we investigate possibilities to further reduce the computation time of SMUCE by considering fewer intervals in the multiscale statistic T_n. This is reduction is based on ideas in Walther (2010) and makes SMUCE applicable to large data sets.
In addition we outline how SMUCE can be applied to dependent data in certain situation, where the dependence structure is known. The ideas, which have been elaborated in detail for an applications in Hotz et al. (2012), are shown at a simple example.
Finally, the scale-calibrated penalization chosen for T_n is discussed and a possible extension of SMUCE to more general piecewise parametric models is outlined.

SECTION 2

Statistical methodology

2.1 Model and notation

Before we can formally state the regression model, some definitions have to be introduced. We recall the definition of exponential families and define the space of right-continuous change-point functions.

Definition 1. Let ν be a σ-finite measure on the Borel set of $\mathbb{R}$. Let $\mathcal{F}$ be the family of distributions with ν-densities

$$f_\theta(x) = \exp\left(\theta x - \psi(\theta)\right), \quad x \in \mathbb{R}, \tag{2.1}$$

and with natural parameter space

$$\Theta = \left\{ \theta \in \mathbb{R} \ : \ \int_{\mathbb{R}} \exp(\theta x)\, \mathrm{d}\nu(x) < \infty \right\}.$$

The family $\mathcal{F}$ is called a *natural exponential family* and is said to be *regular* and *minimal* if Θ is an open interval and the cumulant transform ψ is strictly convex on Θ.

Some well-known examples of exponential families are Gaussian distributions with fixed variance σ^2, Poisson distributions and Bernoulli distributions.

Definition 2. The class of right-continuous change-point functions is defined as

$$\mathcal{S} := \left\{ \vartheta : \vartheta(t) = \sum_{k=0}^{K} \theta_k \mathbf{1}_{[\tau_k, \tau_{k+1})}(t), \theta_k \in \Theta, 0 = \tau_0 < \tau_1 < \ldots < \tau_K < \tau_{K+1} = 1, K < \infty \right\}.$$

With these preparations, we now state the regression model.

Model 1. *Suppose we observe the independent random variables* $Y = (Y_1, \dots, Y_n)$ *from*

$$Y_i \sim F_{\vartheta(i/n)}, \qquad \text{for } i = 1, \dots, n, \tag{2.2}$$

where $\{F_\theta\}_{\theta\in\Theta}$ *is a regular and minimal one-dimensional exponential family of distributions and* $\vartheta \in \mathcal{S}$ *a right-continuous change-point function.*

It will be useful to define the functions

$$m(\theta) := \dot{\psi}(\theta) = \mathbf{E}\,[X] \quad \text{and} \quad v(\theta) := \ddot{\psi}(\theta) = \mathbf{Var}\,[X]\,, \tag{2.3}$$

for $X \sim F_\theta$. Note that m is strictly increasing and v is positive on Θ. In Definition 2 the values τ_k are the *change-point locations* and $\theta_k \in \Theta$ the corresponding *intensities* of ϑ. We will assume that $\theta_k \neq \theta_{k+1}$ for $k = 0, \dots, K$ to ensure identifiability. To ease presentation we also use the notation $I_k = [\tau_k, \tau_{k+1})$ for the k-th *segment* of ϑ.

Also, it turns out to be useful to consider the mean-value parameterization of ϑ and θ_k given by

$$\mu(x) = m(\vartheta(x)) \quad \text{and} \quad \mathfrak{m}_k = m(\theta_k). \tag{2.4}$$

Due to the monotonicity of m, the mapping $\mu \mapsto \vartheta$ is one-to-one and hence inference on ϑ and μ are equivalent. Clearly, the same is true for any strictly monotone transformation of ϑ. For $\vartheta \in \mathcal{S}$ as in Definition 2 we denote by $J(\vartheta) = (\tau_1, \dots, \tau_K)$ the increasingly ordered vector of change-points and by $\#J(\vartheta) = K$ its length.

For any estimator $\hat{\vartheta}$ of $\vartheta \in \mathcal{S}$, the estimated number of change-points will be denoted by $\#J(\hat{\vartheta}) = \hat{K}$, the change-point locations by $J(\hat{\vartheta}) = (\hat{\tau}_1, \dots, \hat{\tau}_{\hat{K}})$. Further, we set $\hat{\theta}_k = \hat{\vartheta}(t)$ for $t \in [\hat{\tau}_k, \hat{\tau}_{k+1})$, i.e. $\hat{\theta}_k$ is the value of $\hat{\vartheta}$ on the k-th segments $\hat{I}_k$. Analogously we set $\hat{\mu} = m(\hat{\vartheta})$ and $\hat{\mathfrak{m}}_k = m(\hat{\theta}_k)$.

Let $S[k]$ denote the class of all functions in $\mathcal{S}$ which number of change-points is less or equal to k. For simplicity, for each $n \in \mathbb{N}$ we restrict ourselves to estimators which have change-points only at sampling points, i.e. $\hat{\vartheta} \in \mathcal{S}_n[K]$ with $\hat{\tau}_k = \hat{l}_k/n$ for some $1 \leq \hat{l}_k \leq n$. For a simple presentation, we consider an equidistant sampling scheme as in Model 1. However, extensions to more general designs are straightforward.

2.2 A Multiscale test for change-point regression

In this section we derive the multiscale statistic, which we employ for change-point inference throughout this work. We will first consider local likelihood-ratio tests for local intensities of ϑ (Subsection 2.2.1) and then combine these into a multiscale statistic (Subsection 2.2.2).

2.2.1 Local likelihood-ratio tests

Given a candidate function $\hat{\vartheta} \in \mathcal{S}$ we want to decide whether or not $\hat{\vartheta}$ is a good reconstruction of ϑ. With a slight abuse of notation, $\hat{\vartheta}$ is considered as a fixed non-random function at this point. To begin with, we fix some $1 \leq k \leq K$ and consider one fixed interval $[i/n, j/n] \subset \hat{I}_k$, i.e. which $\hat{\vartheta}$ is constant on with value $\hat{\theta}_k$. Then, consider the *local* test problem

$$\begin{aligned} H_{i,j} &: Y_i, \ldots, Y_j \sim F_{\hat{\theta}_k} \quad \text{vs.} \\ K_{i,j} &: Y_i, \ldots, Y_j \sim F_{\tilde{\theta}} \ \text{ for some } \ \tilde{\theta} \in \{\Theta \setminus \hat{\theta}_k\}. \end{aligned} \tag{2.5}$$

For i.i.d. observations $Y_i, \ldots, Y_j$, the *local* likelihood-ratio statistic for this test is given by

$$T_i^j(Y, \hat{\theta}_k) = \log \left(\frac{\sup_{\tilde{\theta} \in \Theta} \prod_{l=i}^{j} f_{\tilde{\theta}}(Y_l)}{\prod_{l=i}^{j} f_{\hat{\theta}_k}(Y_l)} \right). \tag{2.6}$$

Introducing the notation $\phi(x) = \sup_{\theta \in \Theta}(\theta x - \psi(\theta))$ and $J(x, \theta) = \phi(x) - (\theta x - \psi(\theta))$ we find

$$T_i^j(Y, \hat{\theta}_k) = (j - i + 1) J(\overline{Y}_i^j, \hat{\theta}_k) \geq 0, \tag{2.7}$$

where $\overline{Y}_i^j = (\sum_{i \leq l \leq j} Y_l)/(j - i + 1)$. This reveals the property of the likelihood-ratio test to achieve reduction of the data by sufficiency, as the local test statistic T_i^j depends on the minimal sufficient statistic $\overline{Y}_i^j$ only. The resulting test at level $\alpha \in (0, 1)$ is of the form

$$\phi(Y) = \begin{cases} 1 & \text{if} \quad T_i^j(Y, \hat{\theta}_k) \leq q_{i,j}(\alpha) \quad \text{and} \\ 0 & \text{otherwise,} \end{cases} \tag{2.8}$$

for some constant $q_{i,j}(\alpha)$, determined by the level of significance $\alpha \in (0, 1)$ of the test. Hence, $H_{i,j}$ is rejected if T_i^j exceeds the threshold $q_{i,j}(\alpha)$. Given the observations $Y_i, \ldots, Y_j$, there exist constants $\underline{b}_{i,j}$ and $\overline{b}_{i,j}$ such that $\hat{\theta}_k$ is accepted if and only if

$$\underline{b}_{i,j} \leq \hat{\theta}_k \leq \overline{b}_{i,j}. \tag{2.9}$$

This follows from the strict convexity of T_i^j, as we will show in Section 4.3. In summary, any function $\hat{\vartheta}$ which is constant on $[i/n, j/n]$ is rejected if its value on $[i/n, j/n]$ is not in the interval $[\underline{b}_{i,j}, \overline{b}_{i,j}]$.

Our goal is to decide if $\hat{\vartheta}$ is a good reconstruction of the entire signal, i.e. on all intervals simultaneously. For $\hat{\vartheta} \in \mathcal{S}$ with $\hat{K}$ segments $\hat{I}_1, \ldots, \hat{I}_{\hat{K}}$ and values $\hat{\theta}_1, \ldots, \hat{\theta}_{\hat{K}}$ we therefore

consider the following multiple testing problem

$$\bigcap_{k=1}^{\hat{K}} \bigcap_{[i/n,j/n]\subset\hat{I}_k} H_{i,j} \quad \text{vs.} \quad \bigcup_{k=1}^{\hat{K}} \bigcup_{[i/n,j/n]\subset\hat{I}_k} K_{i,j}.$$

In other words, $\hat{\vartheta}$ is rejected, whenever *any* of the local hypotheses in (2.5) is rejected on an interval, which $\hat{\vartheta}$ is constant on. In the upcoming section we discuss how the local test statistics in (2.7) can be combined into a multiscale statistic.

2.2.2 Combing local tests

Recall that given a candidate function $\hat{\vartheta} \in \mathcal{S}$, we perform the *local test* in (2.7) on any interval, which $\hat{\vartheta}$ is constant on. We aim for finding a testing procedure which will not reject the true signal ϑ with a specified probability $\alpha \in (0,1)$. In the theory of multiple testing this corresponds to controlling the *family wise error (FWE)*. By this approach the error of first type is controlled uniformly over all local tests. Assuming the values $q_{i,j}$ in (2.8) could be chosen such that

$$\mathbf{P}\left(\max_{k=1\ldots,K} \max_{[i/n,j/n]\subset I_k} T_i^j(Y,\theta_k) - q_{i,j}(\alpha) > 0\right) \leq \alpha, \tag{2.10}$$

for the true signal $\vartheta \in S$, one can guarantee that the true function ϑ is not rejected with probability greater than $1-\alpha$ by any of the local tests. Following the argumentation in (2.9), we can construct the acceptance region for the multiple test:

$$\max_{k=1\ldots,\hat{K}} \max_{[i/n,j/n]\subset\hat{I}_k} T_i^j(Y,\hat{\theta}_k) - q_{i,j}(\alpha) \leq 0$$

is satisfied if and only if for all $k = 1,\ldots,\hat{K}$

$$\underline{b}_{i,j} \leq \hat{\theta}_k \leq \overline{b}_{i,j} \quad \text{for all} \quad [i/n, j/n] \subset \hat{I}_k. \tag{2.11}$$

Here, the bounds $\underline{b}_{i,j}$ and $\overline{b}_{i,j}$ depend on Y and $q_{i,j}(\alpha)$. The computation of these bounds is crucial for an efficient implementation of our approach (see Section 4.3). For the moment, however, we focus on the statistical problem to find constants $q_{i,j}$ that satisfy condition (2.10). Clearly, this problem has no unique solution. The particular choice we make enables us to prove optimal detection of segments an all scales simultaneously. For this purpose, it puts different scales on equal footing by penalization of small intervals. This becomes advantageous, since there are many more small than large intervals. Without a scale-calibration the null-distribution would hence be dominated by the small scales. We use an additive penalization introduced in Dümbgen and Spokoiny (2001) and consider the ***penalized*** *multiscale*

statistic

$$T_n(Y,\hat{\vartheta}) = \max_{0\le k\le \hat{K}} \max_{[i/n,j/n]\in\hat{I}_k} \left(\sqrt{2T_i^j(Y,\hat{\theta}_k)} - p\left(\frac{j-i+1}{n}\right)\right) \tag{2.12}$$

with penalties $p(x) = \sqrt{2\log(e/x)}$. We use a penalization of the square root of the likelihood-ratios instead of the likelihood-ratios. As it was argued in Rivera and Walther (2012) this allows for optimal detection with a simple additive penalty term. The same is not true if the likelihood-ratios were penalized instead. In Section 8.4 we will briefly discuss different penalizations. Assume that $q(\alpha)$ is the $(1-\alpha)$-quantile of the null-distribution of $T_n(Y,\vartheta)$, i.e. the distribution of $T_n(Y,\vartheta)$ for the true signal $\vartheta \in \mathcal{S}$. Then we easily find that

$$q_{i,j} = q(\alpha) + p\left(\frac{j-i+1}{n}\right) \tag{2.13}$$

satisfies (2.10). We will investigate the null-distribution $T_n(Y,\vartheta)$ (asymptotically) in Section 3.1. In the further course of this thesis, we will consider the *multiscale constraint* $T_n(Y,\vartheta) \le q$ for the multiscale statistic T_n in (2.12) and a threshold $q \in \mathbb{R}$.

2.3 Statistical multiscale change-point inference

With the definition of the multiscale statistic T_n in (2.12), we formally state the inference scheme, which we employ in this thesis. For $q \in \mathbb{R}$ the set of function, that fulfill the multiscale constraint, will be denoted by

$$\mathcal{C}(q) := \{\vartheta \in \mathcal{S} : T_n(Y,\vartheta) \le q\}. \tag{2.14}$$

We then consider the multiscale constraint optimization problem

$$\inf_{\vartheta\in\mathcal{S}} \#J(\vartheta) \quad \text{s.t.} \quad \vartheta \in \mathcal{C}(q). \tag{2.15}$$

Let the estimate $\hat{K}(q)$ for K be given by be the minimal value $\#J(\vartheta)$ of (2.15), i.e.

$$\hat{K}(q) = \min\{k \in \mathbb{N} \ : \ \exists\ \vartheta \in \mathcal{S}_n[k] : T_n(Y,\vartheta) \le q\}. \tag{2.16}$$

Further, define the set of all solutions of (2.15) as

$$\mathcal{H}(q) := \left\{\vartheta \in \mathcal{S} : T_n(Y,\vartheta) \le q \quad \text{and} \quad \#J(\vartheta) = \hat{K}(q)\right\}. \tag{2.17}$$

Finally, let the estimate $\hat{\vartheta}(q)$ for ϑ be the maximum likelihood estimator among all functions in $\mathcal{H}(q)$, i.e.

$$\hat{\vartheta}(q) := \underset{\vartheta \in \mathcal{H}(q)}{\operatorname{argmax}} \sum_{i=1}^{n} \log \left(f_{\vartheta(i/n)}(Y_i) \right). \tag{2.18}$$

Clearly, $\hat{\vartheta}(q)$ implicitly defines estimates for the change-points locations by

$$\left(\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{K}(q)} \right) := J(\hat{\vartheta}(q)). \tag{2.19}$$

In the upcoming section we develop a theory for these estimates and show that $\mathcal{H}(q)$ constitutes an asymptotic confidence set. Further, we will show in Section 4 that an efficient computation of a solution of (2.18) relies crucially on the equivalence in (2.11).

SECTION 3

Theory

In this section asymptotic and non-asymptotic properties of SMUCE are shown. Parts of these results have appeared in Frick et al. (2013). In Section 3.1, we prove convergence of the null-distribution of the statistic T_n. These findings from Frick et al. (2013) are complemented by explicit bounds for the tails of the limit distribution. Based on these results, the probability of overestimating the number of change-points and the expected number of overestimated change-points is bounded. This extends the results in Frick et al. (2013) and opens the door to a data-driven threshold selection as we show in Section 8.1. Additionally, bounds for the probability of underestimation are shown in the spirit of Frick et al. (2013). Here, a refined version is derived, which yields sharper finite bounds. Finally, we prove asymptotic confidence statements for the set $\mathcal{H}(q)$ as in (2.17). We stress that non-asymptotic versions of these results exists in the Gaussian case (Section 3.5).

3.1 Asymptotic null-distribution

We now investigate the null-distribution of T_n as in (2.12). It is well known that in exponential families the null-distribution of the local likelihood-ratio tests T_i^j are χ_1^2-distributed asymptotically (i.e. as $n \to \infty$, s.t. $(j - i + 1) \to \infty$), see e.g. the book of van der Vaart (1998)[Chapter 16]. Put differently, this says that the asymptotic null-distribution of the local tests is the same as in the Gaussian case and depends neither on the specific exponential family nor on the regression function ϑ.

We will prove a result in that spirit for the multiscale statistic T_n, i.e. for the scale-calibrated maximum of the local tests. For Gaussian observations, it follows from Dümbgen and Spokoiny (2001) and Dümbgen et al. (2006) that under the null-hypothesis T_n converges to a random variable, concentrated on the positive reals, which is finite almost surely. Moreover, it has sub-exponential tails, as we will prove in Section A.1.2. In this section we show weak convergence of the null-distribution of T_n to the Gaussian limit distribution under Model 1.

For the proof we bound the smallest size of intervals and consider a modified version of (2.12), which reads as

$$T_n(Y,\vartheta;c_n) = \max_{0\le k\le K} \max_{\substack{\tau_k\le i/n\le j/n<\tau_{k+1}\\ (j-i+1)/n\ge c_n}} \left(\sqrt{2T_i^j(Y,\theta_k)} - \sqrt{2\log\frac{en}{j-i+1}}\right), \tag{3.1}$$

where it is assumed that

$$c_n^{-1}\log^3(n)/n \to 0. \tag{3.2}$$

This lower bound is necessary by technical reasons. We use strong approximations of partial sum processes (see Lemma 41), which require $c_n\log^2(n)/n \to 0$. Furthermore, Taylor expansion of the local likelihood-ratios T_i^j (see Lemma 40) is used to show convergence to a Gaussian limit law. These rely on the assumption that $c_n^{-1}\log^3(n)/n \to 0$.
The representation of the asymptotic null-distribution is given in terms of the random variable

$$M := \sup_{0\le s<t\le 1}\left(\frac{|B(t)-B(s)|}{\sqrt{t-s}} - \sqrt{2\log\frac{e}{t-s}}\right), \tag{3.3}$$

where $(B(t))_{t\ge 0}$ denotes the standard Brownian motion. After these preparations we can state the main theorem on the null-distribution.

Theorem 3 (Asymptotic null-distribution)**.** *Let Y be given by Model 1 and assume $(c_n)_{n\in\mathbb{N}}$ satisfies* (3.2)*. Then,*

$$T_n(Y,\vartheta;c_n) \xrightarrow{D} \max_{0\le k\le K}\sup_{\tau_k\le s<t\le\tau_{k+1}}\left(\frac{|B(t)-B(s)|}{\sqrt{t-s}} - \sqrt{2\log\frac{e}{t-s}}\right). \tag{3.4}$$

Further, let $M_0,\dots,M_K$ be independent copies of M as in (3.3)*. Then, the r.h.s. in* (3.4) *is stochastically bounded from above by M and from below by*

$$\max_{0\le k\le K}\left(M_k - \sqrt{2\log\frac{1}{\tau_{k+1}-\tau_k}}\right). \tag{3.5}$$

We emphasize that the limit distribution in (3.4) (as well as the lower bound in (3.5)) depends on the unknown regression function ϑ only through the change-point locations $\tau_1,\dots,\tau_K$. Whereas the function values of ϑ do not influence the limit law. The upper bound M is independent of ϑ, i.e. for any $x>0$

$$\lim_{n\to\infty}\sup_{\vartheta\in\mathcal{S}}\mathbf{P}\left(T_n(Y,\vartheta,c_n)>x\right) \le \mathbf{P}\left(M>x\right). \tag{3.6}$$

We will show in Section A.1.2 that M has sub-Gaussian tails (see Theorem 37). Together with Theorem 3 this yields the following corollary.

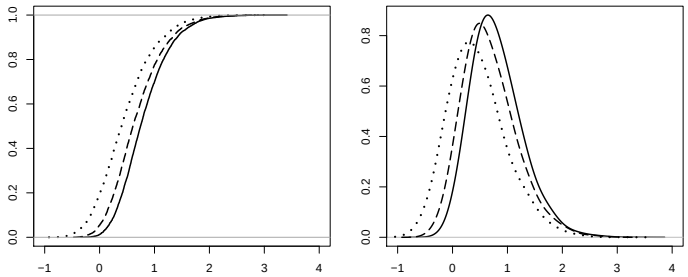


Figure 4: Simulations of the cumulative distribution function (left) and density (right) of M as in (3.3) for $n = 50$(dotted line), $n = 500$(dashed line) and $n = 5{,}000$(solid line) equidistant discretization points.

Corollary 4. *Let Y be given by Model 1. For all $x > 2\mathbf{E}[M]$ it holds that*

$$\lim_{n\to\infty} \mathbf{P}\left(T_n(Y, \vartheta; c_n) > x\right) \leq 2\exp(-x^2/8).$$

This bound on the tails of the null-distribution turns out to be useful throughout this thesis. For example it allows us to prove almost sure consistency for the estimated number of change-points $\hat{K}$ (see Corollary 15) in the Gaussian setting. In addition, the result can be employed in order to approximate quantiles of M in the tails.

In Section 3.5 we will show that for the Gaussian case even non-asymptotic versions of the latter results can be obtained, which allows for finite sample refinement of the null-distribution of T_n. More precisely, in (3.6) the random variable M can be replaced by

$$M^{(n)} = \max_{0\leq i<j\leq n} \frac{|B(j/n) - B(i/n)|}{\sqrt{(j-i)/n}} - \sqrt{2\frac{en}{j-i}}.$$

As the convergence in Theorem 3 is rather slow, this finite sample correction is helpful even for relatively large samples, say if n is of the order of a few thousands. This is highlighted in Figure 4 where it also becomes apparent that the empirical null-distributions for finite samples, obtained from simulations, is in general not supported in $[0, \infty)$.

Hence, it is advantageous for Gaussian data to use finite sample simulations from $M^{(n)}$. For non-Gaussian data the bound is valid asymptotically only. Empirically, however, we found that the approximation of the likelihood-ratios by the Gaussian version is very accurate, even for small sample sizes. This is illustrated in Figure 6, which shows probability-probability plots of $M^{(n)}$ against the null-distribution of T_n for Poisson observations with constant mean 3 (first row) and Bernoulli observations with constant mean 0.8 (second row) for sample size $n = 100$ (left), $n = 500$ (middle) and $n = 1{,}000$ (right). Even for the smallest sample size $n = 100$ we find that $M^{(n)}$ approximates the null-distributions quite well in both cases.

The inequality in (3.6) is not sharp, if the true function has at least one change-point. For an

illustration of this, Figure 5 shows probability-probability plots of the exact null-distribution of signals with two, four and ten equidistant change-points against the null-distribution of a signal without change-points for sample size $n = 500$. Clearly, further information on the number and location of change-points could be used to improve the distributional bound.

3.2 Overestimation of change-points

We first note that with the additional constraint in (3.1) on the minimal interval length, the estimated number of change-points is given by

$$\hat{K}(q) = \min\left\{k \in \mathbb{N} \;:\; \exists\, \vartheta \in \mathcal{S}_n[k] : T_n(Y, \vartheta; c_n) \leq q\right\}, \quad q \in \mathbb{R}. \tag{3.7}$$

From the construction of SMUCE, it is immediate that if $q = q(\alpha)$ is chosen to be the $(1-\alpha)$-quantile of M, then

$$\limsup_{n\to\infty} \mathbf{P}(\hat{K}(q(\alpha)) > K) \leq \alpha. \tag{3.8}$$

This holds since the number of change-points is minimized among all functions in $\mathcal{C}(q)$ and $\mathbf{P}(\vartheta \in \mathcal{C}(q)) \geq 1-\alpha$. However, in (3.8) we only distinguish between the events that the number of change-points was overestimated or not. In many applications as well as from a theoretical point of view, it is certainly of interest to quantify the number of overestimated change-points. For this purpose, we extend the latter result in the following theorem.

Theorem 5 (Overestimation bound)**.** *Let Y be given by Model 1, $\hat{K}(q)$ as in* (3.7)*, $q = q(\alpha)$ be the $(1-\alpha)$-quantile of M and $k \in \mathbb{N}_0$. Then,*

$$\limsup_{n\to\infty} \mathbf{P}\left(\hat{K}(q(\alpha)) > K + 2k\right) \leq \alpha^{k+1}. \tag{3.9}$$

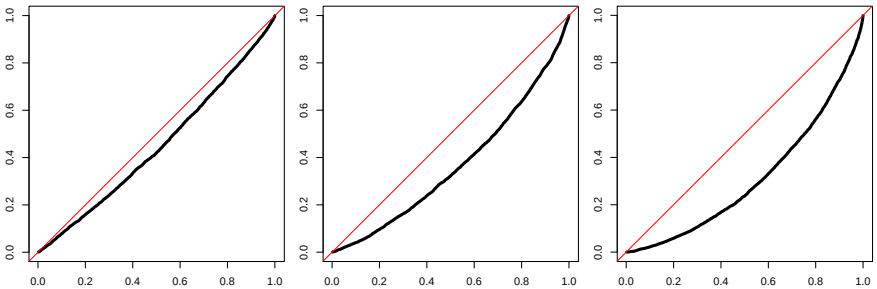


Figure 5: Probability-probability plots of the empirical null-distribution of a signal without change-points (x-axis) against signals with 2(left), 5 (center) and 10 (right) equidistant change-points (y-axis) for $n = 500$.

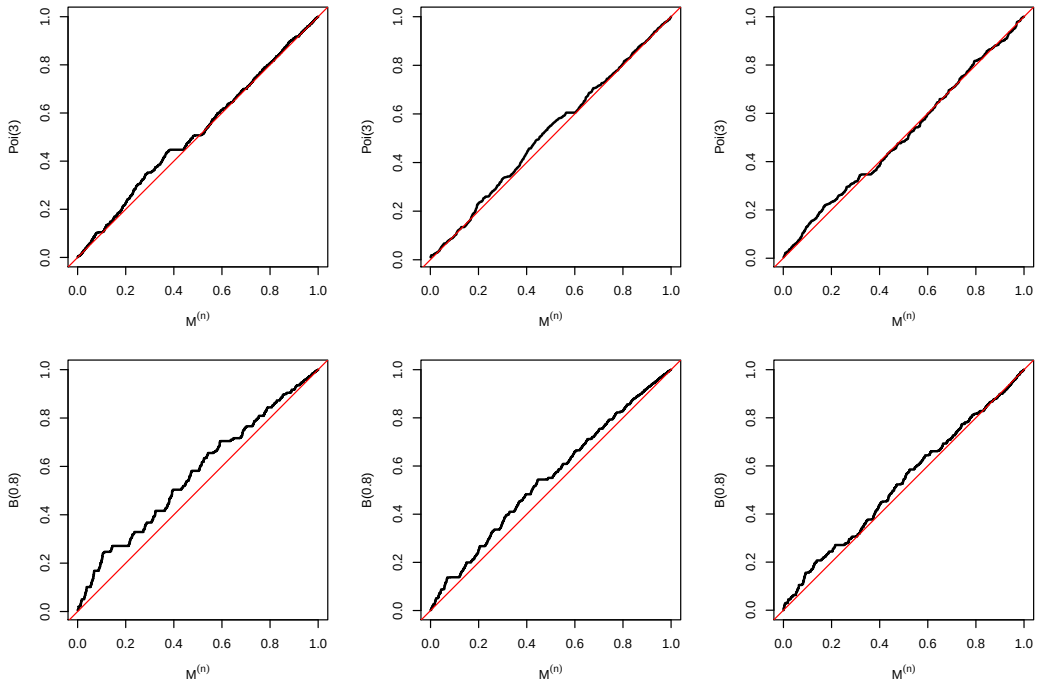


Figure 6: Probability-probability plots (black line) of $M^{(n)}$ against the null-distribution of T_n for Poisson observations with $\mu \equiv 3$ (first row) and Bernoulli observations with $\mu \equiv 0.8$ (second row) for sample size $n = 100$ (left), $n = 500$ (middle) and $n = 1{,}000$ (right).

First, we observe that for $k = 0$, (3.9) boils down to (3.8). For general $k \geq 1$, the theorem reveals that we cannot only control the probability of overestimation but, moreover, give confidence statements about the number of overestimated change-points. As an application, this allows to control the expected value of overestimated change-points, as shown in the following corollary.

Corollary 6. *Let Y be given by Model 1, $\hat{K}(q)$ as in (3.7), $q = q(\alpha)$ be the $(1-\alpha)$-quantile of M. Then,*

$$\lim_{n\to\infty} \mathbf{E}\left[\left(\hat{K}\left(q(\alpha)\right) - K\right)_+\right] \leq 2\frac{\alpha}{1-\alpha},$$

where $(x)_+ = \max(x, 0)$.

This shows that even for rather large values of α, the expected value of overestimated change-points is relatively small, see also Figure 7 for an illustration. Hence, SMUCE is a method, which first of all guarantees the error of overestimation to be small.

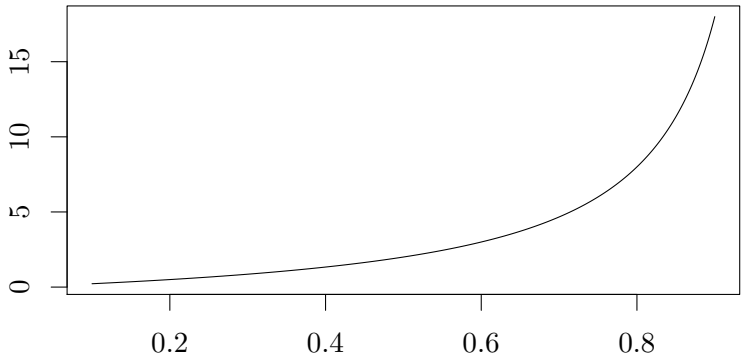


Figure 7: Bounds for the expected value of $(\hat{K}(q(\alpha)) - K)_+$ in Corollary 6 in dependence of $\alpha \in (0, 1)$ (x-axis).

3.3 Underestimation of change-points

In this section we derive explicit bounds for the probability that $\hat{K}(q)$ as defined in (2.15) underestimates the true number of change-points K. For these bounds it is not necessary to impose a lower bound on the lengths of the considered intervals. Bounds for the probability of underestimation necessarily have to depend on the true signal ϑ, as no method can recover changes of arbitrarily small height or on arbitrarily small segments for a given sample size n. For a similar argument in the context of density estimation we refer to the work of Donoho (1988). Under assumptions on the true signal ϑ such two-sided inference can be achieved.

We begin with a general result that bounds the probability of missing change-points given some characteristics of the regression function ϑ. This needs some preparations. First, define for $k = 1, \ldots, K$ the height of the k-th change-point δ_k and as a measure for the lengths of the corresponding segments λ_k as

$$\delta_k = |\theta_{k+1} - \theta_k| \quad \text{and} \quad \lambda_k = \min\left\{\frac{\tau_k - \tau_{k-1}}{2}, \frac{\tau_{k+1} - \tau_k}{2}\right\}.$$

We will also frequently use the notations

$$\Delta = \min_{1 \le k \le K} \delta_k \quad \text{and} \quad \Lambda = 2 \min_{1 \le k \le K} \lambda_k \tag{3.10}$$

for the smallest jump and smallest segment of ϑ, respectively. By $D(\theta||\tilde{\theta})$ we will denote the *Kullback-Leibler divergence* of F_θ and $F_{\tilde{\theta}}$, i.e.

$$D(\theta||\tilde{\theta}) = \int_{\mathbb{R}} f_\theta(x) \log \frac{f_\theta(x)}{f_{\tilde{\theta}}(x)} \, \mathrm{d}\nu(x) = \psi(\tilde{\theta}) - \psi(\theta) - (\tilde{\theta} - \theta) m(\theta). \tag{3.11}$$

To state bounds for the probability of underestimation we further require the functions

$$\kappa_1^{\pm}(v,w,x,y) = \inf_{\substack{v\le\theta\le w \\ \theta\pm x\in[v,w]}} \sup_{\varepsilon\in[0,x]} \left[\frac{\varepsilon}{x}\left(D(\theta||\theta\pm x)-y\right) - D(\theta||\theta\pm\varepsilon)\right], \tag{3.12}$$

$$\kappa_2^{\pm}(v,w,x) = \inf_{\substack{v\le\theta\le w \\ \theta\pm x\in[v,w]}} D(\theta\pm x||\theta). \tag{3.13}$$

Finally, we define

$$\kappa_1^k = \min\left\{\kappa_1^+\left(\underline{\theta},\overline{\theta},\frac{\delta_k}{2},\frac{\left(q+\sqrt{2\log\frac{e}{\lambda_k}}\right)^2}{n\lambda_k}\right),\kappa_1^-\left(\underline{\theta},\overline{\theta},\frac{\delta_k}{2},\frac{\left(q+\sqrt{2\log\frac{e}{\lambda_k}}\right)^2}{n\lambda_k}\right)\right\}, \tag{3.14}$$

$$\kappa_2^k = \min\left\{\kappa_2^+\left(\underline{\theta},\overline{\theta},\frac{\delta_k}{2}\right),\kappa_2^-\left(\underline{\theta},\overline{\theta},\frac{\delta_k}{2}\right)\right\}. \tag{3.15}$$

After these preparations we can now give an explicit bound on the probability of underestimating the number of change-points.

Theorem 7 (Underestimation bound)**.** *Let Y be given by Model 1, $q>0$ and $\hat{K}(q)$ be defined by* (2.16) *and let*

$$\beta_{nk}(q) = \left[1-e^{-\kappa_1^k n\lambda_k}-e^{-\kappa_2^k n\lambda_k}\right]^2. \tag{3.16}$$

Then,

$$\mathbf{P}\left(\hat{K}(q)\ge K\right) \ge \prod_{k=1}^{K}\beta_{nk}(q)$$

and moreover

$$\mathbf{E}\left[\left(K-\hat{K}(q)\right)_+\right] \le \sum_{k=1}^{K}(1-\beta_{nk}(q)).$$

As it becomes clear in the proofs, $\beta_{nk}(q)$ is a lower bound for the probability of detecting the k-th change-point. Let

$$\beta_n(q) = \min_{1\le k\le K}\beta_{nk}(q), \tag{3.17}$$

which bounds the probability of detecting the change-point, which is hardest to detect. As a direct consequence of Theorem 7, we obtain from the inequality $(1-x)^m\ge 1-mx$ (for all $x\in(0,1)$ and $m\in\mathbb{N}_0$) that

$$\mathbf{P}\left(\hat{K}(q)\ge K\right) \ge \beta_n(q)^K \ge 1-K\left(1-\beta_n(q)\right). \tag{3.18}$$

Furthermore, it holds that

$$\mathbf{E}\left[(K - \hat{K}(q))_+\right] \leq K(1 - \beta_n(q)). \tag{3.19}$$

The parameters $\beta_{nk}(q)$ depend not only on the true function ϑ but also on the family of distribution $\mathcal{F}$. Their explicit computation can be rather tedious and has to be done for each exponential family separately (for the Gaussian case see Section 3.5). Therefore, it is useful to have a lower bound for these constants, which is given in the following.

Lemma 8. *Let v be as in* (2.3) *and $\kappa_1^{\pm}$ and $\kappa_2^{\pm}$ be defined as in* (3.12) *and* (3.13)*, respectively. Then,*

$$\kappa_1^{\pm}(v, w, x, y) \geq \frac{x^2}{8} \frac{\inf_{v \leq t \leq w} v(t)^2}{\sup_{v \leq t \leq w} v(t)} - y \quad \text{and} \quad \kappa_2^{\pm}(v, w, x) \geq \frac{x^2}{2} \inf_{v \leq t \leq w} v(t).$$

Clearly, Lemma 8 can be used to bound the results in Theorem 7 further. In particular, combination with (3.18) yields a simplified version, which only depends on Λ and Δ as in (3.10). For this purpose, we assume that $\vartheta \in \mathcal{S}$ is so that $\underline{\theta} \leq \vartheta(t) \leq \overline{\theta}$ for all $t \in [0, 1]$. Then,

$$\mathbf{P}\left(\hat{K}(q) < K\right) \leq 2Ke^{-Cn\Lambda\Delta^2/2}\left[e^{\left(q+\sqrt{2\log(2e/\Lambda)}\right)^2} + 1\right], \tag{3.20}$$

where

$$C = C(\mathcal{F}, \underline{\theta}, \overline{\theta}) = \frac{1}{32} \frac{\inf_{\underline{\theta} \leq \theta \leq \overline{\theta}} v(\theta)^2}{\sup_{\underline{\theta} \leq \theta \leq \overline{\theta}} v(\theta)}. \tag{3.21}$$

Such simplified bounds were also derived in Frick et al. (2013). We stress that the refined version in Theorem 7 is sharper, since the height and length corresponding to the same change-point are taken into account, which is reflected in the definition of β_{nk} in (3.16).

3.4 Consistency and locations of estimated change-points

We will employ the latter results, in order to investigate the asymptotic behavior of SMUCE for a fixed signal $\vartheta \in \mathcal{S}$ as $n \to \infty$. Under rather mild assumption on q_n the estimate $\hat{K}(q_n)$ converges to the true number of change-points K in probability. This is made precise in the following corollary.

Corollary 9 (Model selection consistency)**.** *Let $\vartheta \in \mathcal{S}$ be fixed and $\hat{K}(q)$ be as in* (3.7)*. Further, assume that $q_n/\sqrt{n} \to 0$ and $q_n \to \infty$. Then,*

$$\mathbf{P}\left(\hat{K}(q_n) = K\right) \to 1.$$

We will show in Section 3.5 that this result can be extended to a.s. convergence for Gaussian observations.

Remark 10. In Corollary 9 we found that $q_n/\sqrt{n} \to 0$ is sufficient to ensure consistency of SMUCE. Since e.g. $q_n = \mathcal{O}(\sqrt{n/\log n})$ fulfills these assumptions, we find from Corollary 4 that $\mathbf{P}(\hat{K}(q_n) > K) \leq 2\exp(-n/\log n)$ can be achieved. Hence, the error of overestimation can be controlled at an (almost) exponential rate while ensuring consistency.

Next, we investigate the localization of estimated change-points as in (2.19). Any candidate in $\mathcal{H}(q)$ recovers the change-point locations of the true regression function ϑ with the same convergence rate. In other words, the maximum likelihood step in (1.6) is not needed for these results. To ease notations we nevertheless state the result only for $(\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{K}})$, estimated by SMUCE as in (2.19). The results are proved similarly as the bounds for underestimation. Here, we focus on asymptotic rates here and not on finite bounds. For this reason we give a simpler bound which is less sharp for small n, similar to (3.20).

Theorem 11. *Let $q \in \mathbb{R}$ and $(\epsilon_n)_{n\in\mathbb{N}}$ a sequence in $(0, 1]$. Further let $\vartheta \in \mathcal{S}$ be bounded by $\underline{\theta} \leq \vartheta \leq \overline{\theta}$ and let $C = C(\mathcal{F}, \underline{\theta}, \overline{\theta})$ be as in (3.21). Then, for all $n \in \mathbb{N}$*

$$\mathbf{P}\left(\max_{\tau\in J(\vartheta)} \min_{\hat{\tau}\in J(\hat{\vartheta}(q))} |\hat{\tau} - \tau| > \epsilon_n\right) \leq 2Ke^{-Cn\epsilon_n\Delta^2}\left[e^{-\left(q+\sqrt{2\log(2e/\epsilon_n)}\right)^2} + 1\right].$$

For a fixed signal $\vartheta \in \mathcal{S}$, a sufficient condition for the r.h.s. in Theorem 11 to vanish as $n \to \infty$ is

$$\epsilon_n \geq \frac{1}{\Delta^2 C}\frac{\log n}{n}. \tag{3.22}$$

This improves several results obtained for other methods, e.g. in Harchaoui and Lévy-Leduc (2010) for a total variation penalized estimator a $\log^2(n)/n$ rate has been shown.
In the following we will apply Theorem 11 to determine subclasses of $\mathcal{S}$ in which the change-point locations are reconstructed *uniformly* with rate ϵ_n. These subclasses are delimited by conditions on the smallest absolute jump height Δ_n and on the number of change-points K_n (or the smallest interval lengths Λ_n by using the relation $K_n \leq 1/\Lambda_n$) of its members. For instance, the rate function $\epsilon_n = n^{-\beta}$ with some $\beta \in (0, 1)$ implies the condition

$$\frac{n^\beta \exp(-n^{1-\beta}\Delta_n)}{\Lambda_n} \to 0.$$

A value of β close to 1 gives a small subclass of functions which then can be reconstructed uniformly with convergence rate arbitrarily close to the sampling rate $1/n$. We finally point out that the result in Theorem 11 does not presume the number of change-points to be estimated correctly. If ϵ_n additionally satisfies (3.2) and if $q = q_n \to \infty$ slower than $\sqrt{-\log \epsilon_n}$ in Theorem 11, we find that $\mathbf{P}(\hat{K}(q) = K) \to 1$ and it follows that

$$\mathbf{P}\left(\epsilon_n^{-1}|\tau_k - \hat{\tau}_k| > 1\right) \to 0, \quad \text{for } k = 1, \ldots, K.$$

3.5 Gaussian observations

We now derive explicit results for the case when $\mathcal{F}$ is the Gaussian family of distributions with constant variance σ^2. In this case the regression Model 1 can be rewritten to

$$Y_i = \mu(i/n) + \sigma\varepsilon_i, \quad i = 1, \dots, n, \tag{3.23}$$

where $\varepsilon_1, \dots, \varepsilon_n$ are independent standard Gaussian random variables and $\sigma > 0$. We will refer to the value of the k-th segment as $\mathfrak{m}_k \in \mathbb{R}$ as in (2.4).
The result on the asymptotic null-distribution in Theorem 3 is based on strong approximation of the likelihood-ratios by Gaussian partial sums. Since this step is superfluous for Gaussian observations, it is possible to get rid of the lower bound for the smallest scales c_n as in (3.2). Furthermore, we can bound the null-distribution *non-asymptotically*. To this end, we set (as before)

$$M^{(n)} = \max_{0 \le i < j \le n} \frac{|B(j/n) - B(i/n)|}{\sqrt{(j-i)/n}} - \sqrt{2\frac{en}{j-i}}. \tag{3.24}$$

Clearly, $M^{(n)} \overset{\mathcal{D}}{\le} M$, since the maximum is taken over a subset of $\{[s,t] : 0 \le s < t \le 1\}$.

Corollary 12 (Null-Distribution of T_n)**.** *Let Y be given by* (3.23) *and let $M^{(n)}$ be as in* (3.24)*. Then, for any $n \in \mathbb{N}$*

$$T_n(Y, \mu) \overset{\mathcal{D}}{\le} M^{(n)} \overset{\mathcal{D}}{\le} M.$$

In contrast to Theorem 3, this result is non-asymptotic and we can control the error of overestimation for any finite n, which e.g. enables us to prove almost sure consistency (see Corollary 15). Further, we can also state finite bounds for overestimation.

Corollary 13 (Overestimation bound)**.** *Let Y be given by* (3.23)*, $q(\alpha)$ be the $(1-\alpha)$-quantile of $M^{(n)}$ and $\hat{K}(q)$ be defined as in* (2.16)*. Then, for any $k \in \mathbb{N}_0$ and $n \in \mathbb{N}$*

$$\mathbf{P}\left(\hat{K}(q(\alpha)) > K + 2k\right) \le \alpha^{k+1}$$

and

$$\mathbf{E}\left[(\hat{K}(q(\alpha)) - K)_+\right] \le 2\frac{\alpha}{1-\alpha}. \tag{3.25}$$

We now turn to bound the probability of underestimating the number of change-points. Similar to Section 3.3 we set

$$\delta_k = \frac{|\mathfrak{m}_{k+1} - \mathfrak{m}_k|}{\sigma}, \ \lambda_k = \min\left\{\frac{\tau_k - \tau_{k-1}}{2}, \frac{\tau_{k+1} - \tau_k}{2}\right\} \tag{3.26}$$

and let $\Lambda = 2\min_{1 \le k \le K} \lambda_k$ be the smallest interval length and $\Delta = \min_{1 \le k \le K} \delta_k$ the smallest

normalized jump height. We then define

$$\beta_{nk}(q) := \left[1 - \exp\left(-\frac{\left(\sqrt{n\lambda_k}\delta_k - 2q - \sqrt{8\log\frac{e}{\lambda_k}}\right)_+^2}{8}\right) - \exp\left(-\frac{n\lambda_k\delta_k^2}{8}\right)\right]^2$$

and $\beta_n(q) = \min_{k=1,\dots,K} \beta_{nk}(q)$.

Theorem 14 (Underestimation bound)**.** *Let $\mu \in \mathcal{S}$ and Y be given by* (3.23)*. Let $q > 0$ and $\hat{K}(q)$ be defined as in* (2.16)*. Then,*

$$\mathbf{P}\left(\hat{K}(q) \geq K\right) \geq \prod_{k=1}^{K} \beta_{nk}(q) \tag{3.27}$$

and

$$\mathbf{E}\left[(K - \hat{K}(q))_+\right] \leq \sum_{k=1}^{K}(1 - \beta_{nk}(q)). \tag{3.28}$$

Note that the r.h.s. in (3.27) can be further bounded by $(\beta_n(q))^K$. Similarly the r.h.s. in (3.28) can be bounded from above by $K(1 - \beta_n(q))$. We will employ the latter results, in order to investigate the asymptotic behavior of SMUCE for a fixed signal $\mu \in \mathcal{S}$ as $n \to \infty$.

Corollary 15. *Let $\mu \in \mathcal{S}$ and Y be given by* (3.23)*. Let $\hat{K}(q)$ be defined as in* (2.16)*. Further, set $0 < \zeta < 0.5$, let q_n such that $q_n/\sqrt{\log n} \to \infty$ and $q_n n^{-\zeta} \to 0$. Then,*

$$\lim_{n\to\infty} \hat{K}(q_n) = K \quad a.s.$$

In comparison to Corollary 9 this shows that almost sure consistency can be obtained if the assumption $q_n/\sqrt{n} \to 0$ is replaced by $q_n n^{-\zeta} \to 0$ and it is additionally assumed that $q_n/\sqrt{\log n} \to \infty$.

3.5.1 Detection of vanishing signals

The previous results may also be seen from a different angle. Instead of considering a fixed signal μ as $n \to \infty$, we shall now determine sequences of subclasses of $\mathcal{S}$, among which the number of change-points is estimated correctly with asymptotic probability 1. In other words, we investigate at which rate signals may vanish while still being detected by SMUCE. We begin with signals on a single interval against an unknown background.

Theorem 16. *Let* $\mathfrak{m}_0 \in \mathbb{R}$ *and* $\mu_n(t) = \mathfrak{m}_0 + \Delta_n \mathbf{1}_{I_n}(t)$ *for a sequence of intervals* $I_n \subset [0,1]$ *with lengths* $|I_n|$*. Further, let* Y *be given by* (3.23) *and* $(q_n)_{n\in\mathbb{N}}$ *be bounded away from zero. Assume*

(1.) for signals on a large scale (i.e. $\liminf |I_n| > 0$*), that* $\sqrt{|I_n|\, n}\Delta_n/q_n \to \infty$*,*

(2.) for signals on a small scale (i.e. $|I_n| \to 0$*) that* $\sqrt{|I_n|\, n}\Delta_n \geq (\sqrt{2}+\varepsilon_n)\sqrt{-\log(|I_n|)}$ *with* ε_n*, s.t.* $\varepsilon_n\sqrt{-\log(|I_n|)}/q_n \to \infty$*.*

Then,

$$\mathbf{P}\left(\hat{K}(q_n) > 0\right) \to 1. \tag{3.29}$$

Remark 17. In Theorem 16 it is shown that SMUCE detects at least one change-point with asymptotic probability one. This is due to the fact that the intervals I_n are allowed to be arbitrarily close to the borders of the unit interval. If it is further assumed that for some $\epsilon > 0$ the intervals I_n are in $[\epsilon, 1-\epsilon]$, i.e. I_n is bounded away from the borders, we find $\mathbf{P}\left(\hat{K}(q_n) \geq K\right) \to 1$. This is a direct consequence of the proofs of Theorem 16.

Theorem 16 gives sufficient conditions on the signals μ_n (through the interval length $|I_n|$ and the jump height Δ_n) as well as on the thresholds q_n such that the multiscale statistic T_n detects the signals with probability one asymptotically. The following theorem shows that the result is optimal in the sense that any test with level $\alpha \in (0,1)$ has non-trivial power if ϵ_n in (2.) of Theorem 16 is replaced by $-\epsilon_n$. To state this more precisely, define

$$\tilde{S}_n = \left\{\mu(t) = \Delta_n \mathbf{1}_{I_n}(t) : \sqrt{|I_n|\, n}\Delta_n = (\sqrt{2} - \varepsilon_n)\sqrt{-\log(|I_n|)}\right\}. \tag{3.30}$$

Theorem 18. *Let* ϵ_n *be such that* $\lim_{n\to\infty} \epsilon_n = 0$ *and* $\lim_{n\to\infty} \epsilon_n\sqrt{-\log|I_n|} \to \infty$*. For any test* $\phi_n(Y)$ *with asymptotic level* α *under the null-hypothesis* $\mu \equiv 0$*, it holds that*

$$\inf_{\mu\in\tilde{S}_n} \mathbf{E}_\mu \phi_n(Y) - \alpha \leq \mathrm{o}(1).$$

For the special case, when $q_n(= q(\alpha))$ is a fixed α-quantile of the limiting null-distribution M in (3.3), the result in Theorem 16 boils down to the findings in Chan and Walther (2013). In particular, aside to the optimal asymptotic power (3.29), the error of first kind is bounded by α. The result in Theorem 16 goes beyond that and allows to shrink the error of first kind to zero asymptotically, by choosing $q_n \to \infty$. The rate of q_n then determines explicitly the rate of ϵ_n through the assumption $\varepsilon_n\sqrt{-\log(|I_n|)}/q_n \to \infty$.
We finally generalize the results in Theorem 16 to the case when $\mu \in \mathcal{S}$ has multiple change-points. These results are based on the bound in (3.27). To ease notations, we formulate the result in terms of the smallest interval length Λ_n and the smallest jump height Δ_n of μ_n. We give conditions on Λ_n and Δ_n so that no change-point is missed asymptotically.

Theorem 19. *Let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{S}$ with K_n change-points and denote by Δ_n and Λ_n the smallest absolute jump size and smallest interval of μ_n, respectively. Further, assume that q_n is bounded away from zero and*

(1.) for signals on large scales (i.e. $\liminf \Lambda_n > 0$), that $\sqrt{\Lambda_n n}\Delta_n/q_n \to \infty$.

(2.) for signals on small scales (i.e. $\Lambda_n \to 0$) with K_n bounded, that $\sqrt{\Lambda_n n}\Delta_n \geq (4+\varepsilon_n)\sqrt{-\log(\Lambda_n)}$ with $\varepsilon_n\sqrt{-\log(\Lambda_n)}/q_n \to \infty$.

(3.) the same as in (2), with K_n unbounded and the constant 8 instead of 4.

Then,

$$\mathbf{P}_{\mu_n}\left(\hat{K}(q_n) \geq K_n\right) \to 1.$$

Theorem 19 amounts to say that the statistic T_n is capable of detecting multiple change-points simultaneously *at the same optimal rate* (in terms of the smallest interval and jump) as a single change-point (see Theorem 16). The only difference being the constants that bound the size of the signals that can be detected. These increase with the dimension of the model: $\sqrt{2}$ for a single change against an unknown background, 4 for a bounded (but unknown), and 8 for an unbounded number of change-points. In Jeng et al. (2010) it was shown that for step functions that exhibit certain sparsity patterns the optimal constant $\sqrt{2}$ can be achieved. It is important to note that we do not make any such sparsity assumption on the true signal. It is not clear, if the constants in Theorem 19 are optimal. Finally we mention an analogy to Theorem 4.1. of Dümbgen and Walther (2008) in the context of detecting areas of local increase and decrease of a density. As in Theorem 19 they showed that only the constants and not the detection rates changes for simultaneous detection of infinitely many features.

3.6 Confidence regions

In this Section we discuss how confidence statements can be constructed from the approach in Section 2.3. Using the terminology in Li (1989), we agree upon the following definition.

Definition 20. A set $\mathcal{C} \subset \mathcal{S}$ is called *asymptotically honest* for the class $\mathcal{S}$ at level $1-\alpha$ if

$$\lim_{n\to\infty} \inf_{\vartheta\in\mathcal{S}} \mathbf{P}\left(\vartheta \in \mathcal{C}\right) \geq 1-\alpha.$$

Clearly, from Theorem 5 it follows that for $q(\alpha)$ being the $(1-\alpha)$-quantile of M, the set $\mathcal{C}(q)$ (as in (2.14)) is an *asymptotically honest confidence set* at level $1-\alpha$. This set, however, is large, e.g. any interpolation of the data is in $\mathcal{C}(q(\alpha))$. Recall that SMUCE is the maximum likelihood estimate restricted to $\mathcal{H}(q)$ (as in (2.17)). From the definition of $\mathcal{H}(q)$ we observe

$$\mathbf{P}\left(\vartheta \in \mathcal{H}(q(\alpha))\right) \geq \mathbf{P}\left(\vartheta \in \mathcal{C}(q(\alpha))\right) - \mathbf{P}\left(\hat{K}(q(\alpha)) < K\right). \tag{3.31}$$

Combining Theorem 7 with the latter inequality gives the following corollary, which bounds the coverage of $\mathcal{H}(q)$.

Corollary 21. *Let $\alpha \in (0,1)$ and $q(\alpha)$ be the $(1-\alpha)$-quantile M as in* (3.3)*. Moreover, set β_n as in* (3.17)*. Then, we find from* (3.18) *that*

$$\mathbf{P}\left(\vartheta \in \mathcal{H}(q)\right) \geq 1-\alpha-K(1-\beta_n(q)) + \mathrm{o}(1). \tag{3.32}$$

Since $\beta_n(q(\alpha)) \to 1$ for any $\vartheta \in S$ as $n\to\infty$ it holds for any $\vartheta \in \mathcal{S}$ that

$$\lim_{n\to\infty} \mathbf{P}\left(\vartheta \in \mathcal{H}(q)\right) \geq 1-\alpha.$$

We mention that for the Gaussian family (see Section 3.5) inequality (3.32) even holds for any n, i.e. the o(1) term on the r.h.s. can be omitted. Thus, the r.h.s. of (3.32) gives an explicit and finite lower bound for the true confidence level of $\mathcal{H}(q(\alpha))$.
Being a subset of $\mathcal{S}$, the confidence set $\mathcal{H}(q)$ is hard to visualize in practice. Therefore, in Section 4.5 we compute a confidence band $B(q) \subset [0,1]\times\Theta$ that contains the graphs of all functions in $\mathcal{H}(q)$ as well as disjoint confidence intervals for the change-point locations. These will be denoted by $[\tau_k^l(q), \tau_k^r(q)] \subset [0,1]$ for $k = 1,\dots,\hat{K}(q)$. For the sake of simplicity, we abbreviate the collection $\{\hat{K}(q), B(q), \{[\tau_k^l(q),\tau_k^r(q)]\}_{k=1,\dots,\hat{K}(q)}\}$ by $I(q)$ and agree upon the notation

$$\begin{aligned} &\vartheta \prec I(q) \quad \text{if } \hat{K}(q) = K, (t,\vartheta(t)) \in B(q) \text{ and } \tau_k \in [\tau_k^l(q),\tau_k^r(q)] \text{ for } k = 1,\dots,K, \\ &\vartheta \not\prec I(q) \quad \text{otherwise.} \end{aligned} \tag{3.33}$$

Put differently, $\vartheta \prec I(q)$ implies that *simultaneously*

1. the number of change-points is estimated correctly,
2. the change-points lie within the confidence intervals (*i.e.*$\tau_k \in [\tau_k^l(q), \tau_k^r(q))$ and
3. the graph is contained in the confidence band $B(q)$.

As it becomes clear from the construction of $I(q)$ in Section 4.5, the confidence set $\mathcal{H}(q)$ and $I(q)$ are linked by the relation

$$\vartheta \in \mathcal{H}(q) \Rightarrow \vartheta \prec I(q). \tag{3.34}$$

In the following we will use this result to determine classes of step functions on which these confidence statements hold uniformly. Following the terminology in Definition 20 we call $I(q)$ asymptotically honest for $\mathcal{S}$, if

$$\liminf_{n\to\infty} \inf_{\vartheta\in\mathcal{S}} \mathbf{P}\left(\vartheta \prec I(q)\right) \geq 1-\alpha. \tag{3.35}$$

Such a condition obviously cannot hold over the entire class $\mathcal{S}$, since signals cannot be detected if they vanish too fast as $n \to \infty$. Because of this, assumption (1) cannot be fulfilled uniformly over $\mathcal{S}$ by any statistical procedure. For Gaussian observations this was made precise in Section 3.5.
To overcome this difficulty, we will relax the notion of asymptotic honesty. Let $\mathcal{S}^{(n)} \subset \mathcal{S}$, $n \in \mathbb{N}$ be a sequence of subclasses of $\mathcal{S}$. Then, we call $I(q)$ *sequentially honest with respect to* $S^{(n)}$ at level $1-\alpha$ if

$$\liminf_{n\to\infty} \inf_{\vartheta\in\mathcal{S}^{(n)}} \mathbf{P}\left(\vartheta \prec I(q)\right) \geq 1-\alpha.$$

By combining (3.31), (3.34) and Theorem 7 we obtain the following result about the asymptotic honesty of $I(q(\alpha))$.

Corollary 22. *Let* $\underline{\theta} < \overline{\theta}$, $\alpha \in (0,1)$ *and* $q(\alpha)$ *be the* $(1-\alpha)$*-quantile of* M *as in* (3.3) *and assume that* $(b_n)_{n\in\mathbb{N}} \to \infty$ *is a sequence of positive numbers. Define*

$$\mathcal{S}^{(n)} = \left\{\vartheta \in \mathcal{S} : n\Lambda\Delta^2/\log(1/\Lambda) \geq b_n,\ \underline{\theta} \leq \vartheta \leq \overline{\theta}\right\}.$$

Then $I(q(\alpha))$ *is* sequentially honest with respect to $\mathcal{S}^{(n)}$ *at level* $1-\alpha$, *i.e.*

$$\lim_{n\to\infty} \inf_{\vartheta\in\mathcal{S}^{(n)}} \mathbf{P}\left(\vartheta \prec I(q(\alpha))\right) \geq 1-\alpha.$$

By estimating $1/\Lambda_n \leq n$ we find that the confidence level α is kept uniformly over nested models $\mathcal{S}^{(n)} \subset \mathcal{S}$, as long as $n\Lambda_n\Delta_n^2/\log n \to \infty$. Here Λ_n and Δ_n are again the smallest interval length and smallest absolute jump size in $\mathcal{S}^{(n)}$, respectively.

3.6.1 Empirical coverage of confidence sets $I(q)$

So far we gave asymptotic results on the simultaneous coverage of the confidence sets $I(q)$ as defined in (3.33). We now investigate the simultaneous coverage empirically. To this end, we consider the test signals shown in Figure 8 for Gaussian observations with varying mean, Gaussian observations with varying variance, Poisson observations and Bernoulli observations. In our simulations we choose $q = q(\alpha)$ to be the $(1-\alpha)$-quantile of M as in (3.3). It then follows from Corollary 22 that asymptotically the simultaneous coverage is larger than $1-\alpha$. Table 1 summarizes the empirical coverage (first column) for different values for α and n obtained by 500 simulation runs each and the relative frequencies of correctly estimated change-points (second column). The results show that for $n = 2{,}000$ the empirical coverage exceeds $1-\alpha$ in all scenarios. The same is not true for smaller n (indicated by bold letters), since here the number of change-points is misspecified rather frequently. Given K has been estimated correctly, we find that the empirical coverage of bands and intervals is in fact larger than the nominal $1-\alpha$ for all simulations (see third column). The low coverage for small sample size is hence caused by underestimation of the number of change-points.

n	$1-\alpha$	Gaussian (mean)			Gaussian (variance)			Poisson			Bernoulli		
1,000	0.8	**0.59**	0.64	0.92	**0.66**	0.68	0.97	0.87	0.89	0.98	0.85	0.90	0.94
	0.9	**0.48**	0.49	0.98	**0.39**	0.39	1.00	**0.85**	0.86	0.99	**0.86**	0.86	0.99
	0.95	**0.28**	0.28	1.00	**0.16**	0.18	0.93	**0.71**	0.74	0.96	**0.66**	0.70	0.94
1,500	0.8	0.84	0.90	0.93	0.87	0.88	0.98	0.92	0.95	0.96	0.93	0.97	0.96
	0.9	**0.73**	0.74	0.98	**0.72**	0.74	0.97	**0.95**	0.97	0.98	0.96	0.97	0.99
	0.95	**0.55**	0.56	0.98	**0.45**	0.47	0.98	**0.92**	0.93	0.99	**0.89**	0.90	0.99
2,000	0.8	0.94	0.99	0.95	0.98	1.00	0.98	0.95	0.99	0.95	0.96	0.99	0.97
	0.9	0.98	1.00	0.98	0.99	1.00	0.99	0.96	0.99	0.96	0.97	0.99	0.98
	0.95	0.99	1.00	0.99	0.97	0.99	0.98	1.00	1.00	1.00	0.99	1.00	0.99

Table 1: Empirical coverage obtained from 500 simulations for the signals shown in Figure 8. For each choice of α and n we computed the simultaneous coverage of $I(q(\alpha))$, as in (3.33) (first value), the percentage of correctly estimated number of change-points (second value) and the simultaneous coverage of confidence bands and intervals for the change-points given $\hat{K}(q(\alpha)) = K$ (third value).

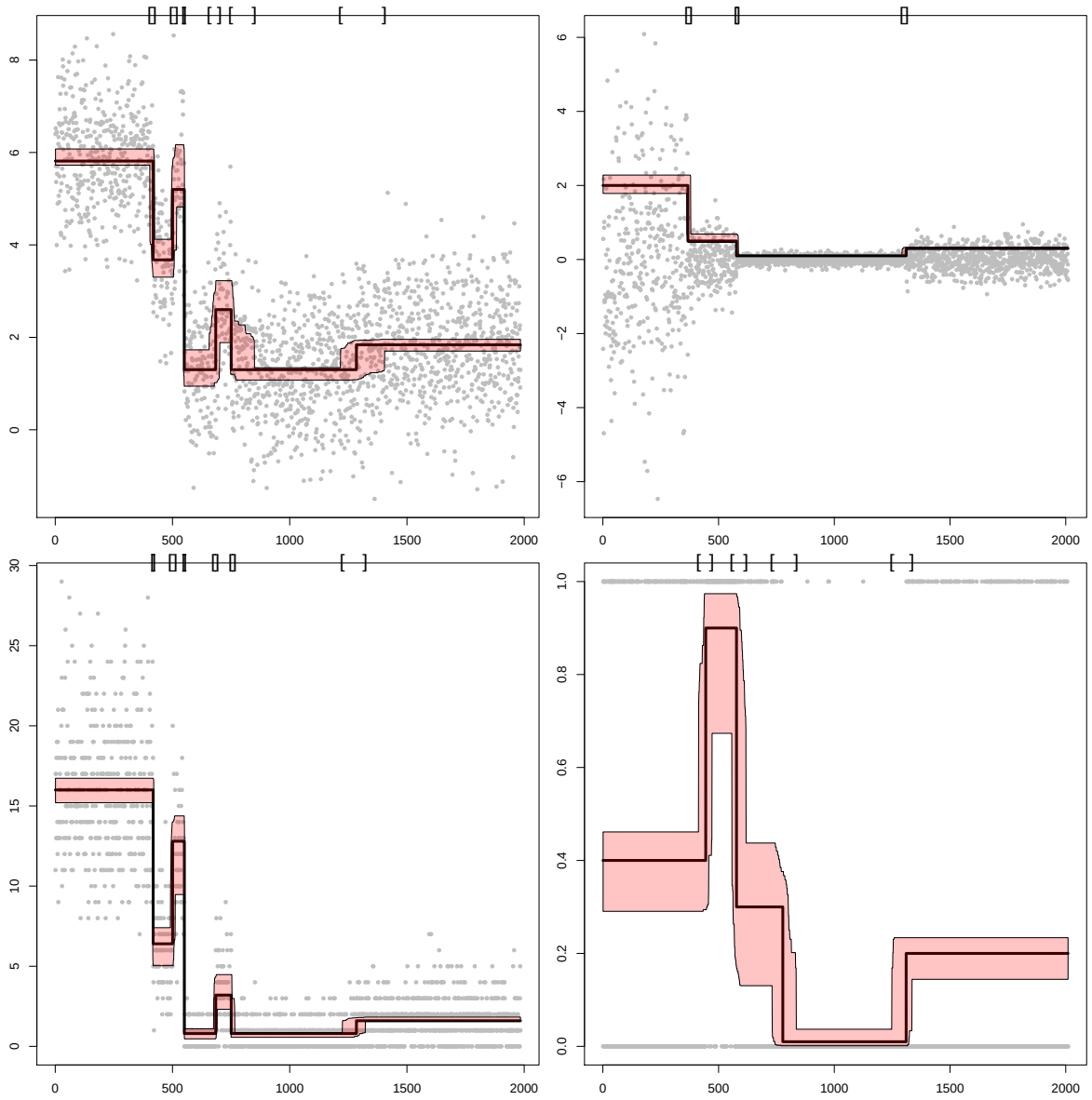


Figure 8: From top left to bottom right: Gaussian observations with varying mean, Gaussian observations with varying variance, Poisson and Bernoulli observations and SMUCE (solid black line) with confidence bands (red hatched) and confidence intervals for change-points (shown by brackets []).

SECTION 4

Implementation

In this section we show how SMUCE can be computed by a pruned dynamic programing algorithm. This approach has been outlined in Frick et al. (2013) and Futschik et al. (2013). We describe the implementation in detail here (Section 4.2). Further, the complexity is discussed and an efficient computation for confidence region is shown (Section 4.5). An R-package of the implementation is available (Hotz and Sieling, 2013)[1].

4.1 Dynamic programming in change-point regression

Dynamic programming algorithms for the change-point problem based on penalized least-square fitting can be traced back to the work of Bellman (1961). In fact, the underlying idea was already introduced in Arrow et al. (1949). All later approaches using dynamic programming are essentially based on *Bellman's Principle of Optimality* (Bellman, 1961):

> *"An optimal policy has the property that whatever the initial state and the initial decision are, the remaining decisions must consist an optimal policy with regard to the state resulting from the first decision."*

In the context of change-point regression this boils down to the observation that a part of the optimal segmentation is optimal itself. We will explain in the following section why such an observation is true for SMUCE. More general, dynamic programming has been used to find exact solutions of penalized cost functionals in the Segment Neighborhood method, which was suggested in Auger and Lawrence (1989). Under the specification of an upper bound on the number of change-points $K_{\max}$, the solution is computed in $\mathcal{O}(K_{\max}n^2)$. Also for penalized cost functionals, dynamic programming has been used more recently in Jackson et al. (2005) and Friedrich et al. (2008) and it was shown that the proposed algorithms are $\mathcal{O}(n^2)$. Killick et al. (2011) developed this approach further by including a pruning step into

[1] R package available at `http://www.stochastik.math.uni-goettingen.de/smuce`

the dynamic program. Their *PELT*-algorithm has a worst case complexity of $\mathcal{O}(n^2)$ but under the assumption that the number of change-points increases in a certain way as more data is collected, the expected computational costs are linearly increasing in n.

Any of the latter approaches is based on the idea of minimizing a *global* cost functional (sometimes referred to as measure-of-fit) with the additional penalization of change-points. In contrast to that, SMUCE requires the final estimate to fulfill a *multiscale* constraint, which acts locally. An adaption of the dynamic program to such multiscale problems has been introduced in Davies et al. (2012) in a similar setting. The authors showed that a solution with minimal number of jumps can be computed in $\mathcal{O}(n^2)$ operations. Moreover, an algorithm which minimizes the empirical quadratic deviations among the solutions with minimal jumps is proposed. This is similar in spirit to the maximum likelihood step in (2.18). The author stated that this algorithm is $\mathcal{O}(n^3)$.

We will exploit the structure of the optimization problem in (2.18) explicitly by including several pruning steps, similar in spirit to Killick et al. (2011). Due to this we can show that the computation of the multiscale restricted maximum likelihood estimator in (2.18) has a worst case complexity of $\mathcal{O}(n^2)$. In addition, we will illustrate that in many situations the computation is much faster than this complexity suggests.

Simultaneously, we derive an efficient way to compute confidence bands for the graph of ϑ (Section 4.5.2) as well as confidence intervals for the location of the change-points (Section 4.5.1).

4.2 A pruned dynamic program for SMUCE

Suppose that $n \in \mathbb{N}$ and $q > 0$ are fixed and that $Y = (Y_1, \ldots, Y_n)$ are observed data. In this section we present a dynamic programming algorithm to compute the estimated number of change-points $\hat{K}(q)$ in (1.4) and the statistical multiscale estimator $\hat{\vartheta}(q)$ as defined in (2.18). To this end, we note that an estimator $\hat{\vartheta}$ can be identified with the vector $(\hat{\vartheta}_1, \ldots, \hat{\vartheta}_n) \in \Theta^n$ where

$$\hat{\vartheta}_i = \hat{\vartheta}(i/n).$$

Next, we note that for a given $\theta \in \Theta$ the log-likelihood on an interval $\{k, \ldots, l\}$ is given by $(k - l + 1)(\theta \overline{Y}_k^l - \psi(\theta))$. With this we define the *local costs* of θ on $\{k, \ldots, l\}$ as

$$c_{k,l}(\theta) = \begin{cases} (k-l+1)(\psi(\theta) - \theta \overline{Y}_k^l) & \text{if } \max_{k \le i \le j \le l} \sqrt{2T_i^j(Y,\theta)} - \sqrt{2 \log \frac{en}{j-i+1}} \le q \\ \infty & \text{else.} \end{cases}$$

In other words, the costs for $\theta \in \Theta$ coincide with the negative log-likelihood if θ satisfies the multiscale constraint on $\{k, \ldots, l\}$ and are infinitely large else. A parameter value that has

finite costs will be referred to as *admissible* on the interval $\{k, \dots, l\}$. The *optimal costs* on the interval $\{k \dots, l\}$ are defined as

$$c_{k,l} = \min_{\theta \in \Theta} c_{k,l}(\theta).$$

If $c_{k,l} < \infty$ we say that $\hat{\theta}_{k,l}$ is the optimal parameter if $c_{k,l} = c_{k,l}(\hat{\theta}_{k,l})$. If $c_{k,l} = \infty$ then there exists no parameter $\theta \in \Theta$ such that the multiscale constraint is satisfied on $\{k, \dots, l\}$.

A detailed pseudocode is given in Algorithm 1. Here, we outline the main idea. To this end, assume the data is not available at once but piece by piece and the optimal estimator is computed iteratively. For $p = 1, 2, \dots$ we compute the optimal costs $c_p := c_{1,p}$ and the corresponding parameter values $\hat{\theta}_{1,p}$ as long as $c_p < \infty$. If $c_{1,p+1} = \infty$ then there are no admissible constant estimators on $\{1, \dots, p+1\}$ and we save the latest feasible index by $R_0 = p$.

For all $p > R_0$ at least one change-point has to be introduced into the reconstruction in order to satisfy the multiscale constraint. Note that for $1 \leq l \leq R_0$ the estimator

$$\hat{\vartheta}(l, p) = \hat{\theta}_{1,l} \mathbf{1}_{\{1,\dots,l\}} + \hat{\theta}_{l+1,p} \mathbf{1}_{\{l+1,\dots,p\}}$$

is the estimator with lowest costs on its constant pieces given the jump location l.

By setting

$$l(p) = \operatorname*{argmin}_{1 \leq l \leq R_0} c_{1,l} + c_{l+1,p}$$

we find that $\hat{\vartheta}(p) = \hat{\vartheta}(l(p), p)$. It is the estimator on the interval $\{1, \dots, p\}$ with lowest cumulative costs $c_p := c_{1,l(p)} + c_{l(p)+1,p}$ among all piecewise constant estimators with one jump. Proceed until $c_{l+1,p+1} = \infty$ for all $1 \leq l \leq R_0$ and then set $R_1 = p$. Put differently, for $p > R_1$ no piecewise constant estimator with one change-point exists on $\{1, \dots, p\}$ that satisfies the multiscale constraint.

Now assume that $k \geq 1$, R_{k-1} and R_k are known and that for $R_{k-1} < l \leq R_k$ the estimator $\hat{\vartheta}(l)$ is the one with lowest cumulative costs c_l with k jumps on the interval $\{1, \dots, l\}$. Then, for $p > R_k$

$$\hat{\vartheta}(l, p) = \hat{\vartheta}(l) \mathbf{1}_{\{1,\dots,l\}} + \hat{\theta}_{l+1,p} \mathbf{1}_{\{l+1,\dots,p\}}$$

is an estimator with $k+1$ jumps on the interval $\{1, \dots, p\}$ with lowest cumulative costs given that the last jump is at l. Again, by setting

$$l(p) = \operatorname*{argmin}_{R_{k-1} < l \leq R_k} c_l + c_{l+1,p}$$

we obtain the estimator $\hat{\vartheta}(p) = \hat{\vartheta}(l(p), p)$ with lowest cumulative costs $c_p = c_{1,l(p)} + c_{l(p)+1,p}$. Such equalities, which constitute the key ingredient for the application of dynamic programing

Data: $Y_1, \ldots, Y_n$
Result: Estimate for number of change-points $\hat{K}$, location of change-points $\hat{l}_1, \ldots, \hat{l}_{\hat{K}}$, values on the segments $\hat{\theta}_0, \ldots, \hat{\theta}_{\hat{K}}$

for $p \leftarrow 1$ **to** n **do**
  Compute $c_p \leftarrow c_{1,p} \leftarrow \inf_{\theta \in \Theta} c_{1,p}(\theta)$;
  if $c_{1,p} = \infty$ **then** // note that $p > 1$ always
    $R_0 \leftarrow p - 1$;
    break
  else
    $\hat{\theta}_{1,p} \leftarrow \operatorname{argmin}_{\theta \in \Theta} c_{1,p}(\theta)$;
  end
end
if $c_{1,n} < \infty$ **then** // there exists a constant feasible estimate
  return $\hat{K} \leftarrow 0$ *and* $\hat{\theta}_0 \leftarrow \hat{\theta}_{1,n}$;
end
$R_{-1} \leftarrow 1$;
for $k \leftarrow 1$ **to** n **do**
  $L_{k-1} \leftarrow R_{k-2} + 1$;
  for $p \leftarrow R_{k-1} + 1$ **to** n **do**
    Initialize $c_p \leftarrow \infty$;
    for $l \leftarrow R_{k-1} - 1$ **to** R_{k-2} **do**
      Compute $c_{l+1,p} \leftarrow \inf_{\theta \in \Theta} c_{l+1,p}(\theta)$ // possibly $c_{l+1,p} = \infty$!;
      if $c_{l+1,p} = \infty$ **then**
        if $p = R_{k-1} + 1$ **then**
          $L_{k-1} \leftarrow l$ // save left bound for confidence intervals;
        end
        break;
      end
      if $c_l + c_{l+1,p} \leq c_p$ **then** // placing a change-point at l reduces overall costs
        $c_p \leftarrow c_l + c_{l+1,p}$, $l(p) \leftarrow l$, $\hat{\theta}_{l+1,p} \leftarrow \operatorname{argmin}_{\theta \in \Theta} c_{l+1,p}(\theta)$;
      end
    end
    if $c_p = \infty$ **then** // there is no feasible estimate with k change-points on $\{1, \ldots, p\}$
      $R_K \leftarrow p - 1$;
      break
    end
  end
  if $c_n < \infty$ **then** // there exists a feasible estimate with k change-points!
    return $\hat{K} \leftarrow k$;
    $\hat{l}_{\hat{K}+1} \leftarrow n$;
    for $k \leftarrow \hat{K}$ **to** 1 **do**
      return $\hat{l}_k \leftarrow l(\hat{l}_{k+1})$ *and* $\hat{\theta}_k \leftarrow \hat{\theta}_{\hat{l}_{k+1}, \hat{l}_{k+1}}$
    end
  end
end

Algorithm 1: Dynamic programming algorithm for the computation of SMUCE

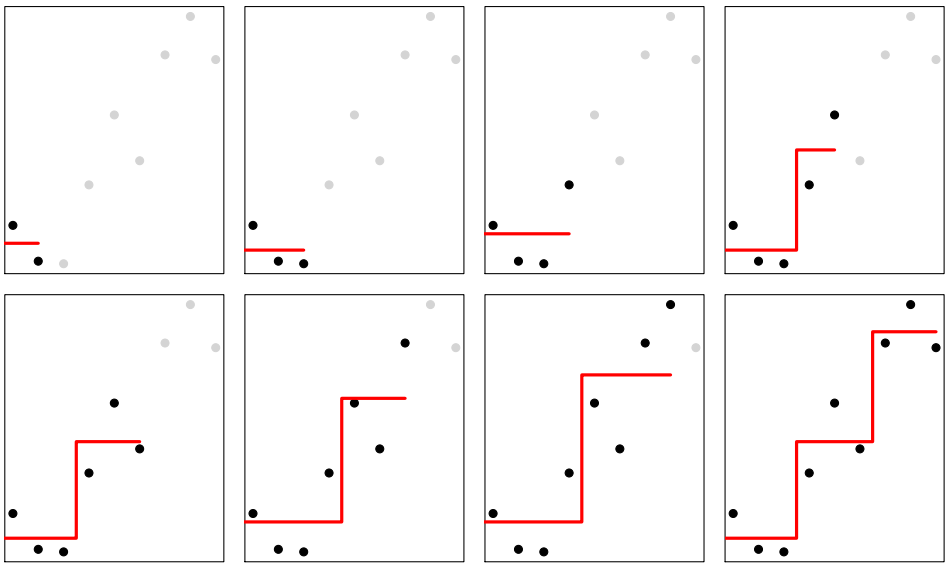


Figure 9: Illustration of the iterative computation for $p = 2, \ldots, 9$; estimate $\hat{\vartheta}(p)$ (red line) and standard Gaussian observations (black dots) for $q = 1$.

are referred to as *Bellman equalities*. This equality shows explicitly why Bellman's principle of optimality holds for the computation of SMUCE.

Proceed until $c_{l+1,p+1} = \infty$ for all $R_{k-1} < l \leq R_k$ (then define $R_{k+1} = p$ and restart the iteration) or until $p = n$ (then set $\hat{\vartheta} = \hat{\vartheta}(n)$ and exit).

In Figure 9 we show the iteratively computed estimators $\hat{\vartheta}(p)$ for a given data set. The example illustrates that the location of detected change-point may alter within the computation: the location of the first change-point changes in the last iteration.

Comparing it to the general dynamic programming approach, it shows important differences which allow for considerable speed-ups. They take into account the specific structure of cost functional by including *pruning steps*, which are founded on the following three observations:

1. Whenever $c_{l,p} = \infty$ for some p, no smaller values for l need to be considered, since $c_{\tilde{l}p} = \infty$ for any $\tilde{l} \leq l$ (see line 24 in Algorithm 1).

2. Whenever there exists no feasible solution with k change-points on $\{1, \ldots, p\}$, then there also exists no feasible solution with k change-points on $\{1, \ldots, \tilde{p}\}$ for any $\tilde{p} > p$ (see line 32 in Algorithm 1).

3. If for a fixed p a feasible solution with k change-points exists, there is no need to consider functions which have more than $k + 1$ change-points up to point p. Consequently, for p in $[R_{k-1}, R_k]$ only estimates whose rightmost change-point is in $[R_{k-2}, R_{k-1})$ have to be computed (see line 18 in Algorithm 1).

Particularly, for signals with many detected change-points these lead to faster computation, as we will discuss in Section 4.4. A crucial part of the algorithm is the computation of the optimal local costs $c_{i,j}$. For the sake of clarity, we gave no details about the computation of these costs in the pseudo-code in Algorithm 1 but postpone this issue to the next subsections.

4.3 Computation of the optimal costs

We will show how the optimal costs $c_{r,p}$ can be computed for some $1 \leq r \leq p \leq n$. To this end, fix some $r \leq i < j \leq p$. Since $\{F_\theta\}_{\theta\in\Theta}$ was assumed to be a regular, one-dimensional exponential family, the natural parameter space Θ is a nonempty, open interval (θ_1, θ_2) with $-\infty \leq \theta_1 < \theta_2 \leq \infty$. Moreover, the mapping $\theta \mapsto J(\overline{Y}_i^j, \theta)$ is strictly convex on Θ and has the unique global minimum at $m^{-1}(\overline{Y}_i^j)$ if and only if $m^{-1}(\overline{Y}_i^j) \in \mathrm{int}(\Theta)$. In this case it follows from Theorem 6.2 in Nielsen (1973) that for all $q > 0$

$$\left\{\theta \in \Theta : \sqrt{2T_i^j(Y,\theta)} - \sqrt{2\log\frac{en}{j-i+1}} \leq q\right\}$$
$$= \left\{\theta \in \Theta : J(\overline{Y}_i^j, \theta) \leq \frac{\left(q + \sqrt{2\log\frac{en}{j-i+1}}\right)^2}{2(j-i+1)}\right\} =: [\underline{b}_{i,j}, \overline{b}_{i,j}],$$

with $-\infty < \underline{b}_{i,j} \leq m^{-1}(\overline{Y}_i^j) \leq \overline{b}_{i,j} < \infty$. In other words, $\underline{b}_{i,j}$ and $\overline{b}_{i,j}$ are the two finite solutions of the equation

$$J(\overline{Y}_i^j, \theta) = \frac{\left(q + \sqrt{2\log\frac{en}{j-i+1}}\right)^2}{2(j-i+1)}. \tag{4.1}$$

If $m^{-1}(\overline{Y}_i^j) \notin \mathrm{int}(\Theta)$, then Nielsen (1973) [Thm. 6.2] implies that either $\underline{b}_{i,j} = -\infty$ or $\overline{b}_{i,j} = \infty$. Let us assume without restriction that $\underline{b}_{i,j} = -\infty$ which in turn shows that $\Theta = (-\infty, \theta_2)$ and $m^{-1}(\overline{Y}_i^j) = -\infty$. In this case, the infimum of $\theta \mapsto J(\overline{Y}_i^j, \theta)$ is not attained and (4.1) has only one finite solution $\overline{b}_{i,j}$. The lower bound $\underline{b}_{i,j} = -\infty$ then is trivial.
After computing $\underline{b}_{i,j}$ and $\overline{b}_{i,j}$ for all $r \leq i \leq j \leq p$, define $\underline{B}_{r,p} = \max_{r\leq i\leq j\leq p} \underline{b}_{i,j}$ and $\overline{B}_{r,p} = \min_{r\leq i\leq j\leq p} \overline{b}_{i,j}$. Hence, if $c_{r,p} < \infty$ we obtain

$$\theta^*_{r,p} = \operatorname*{argmin}_{\theta\in[\underline{B}_{r,p}, \overline{B}_{r,p}]} c_{r,p}(Y,\theta) = \begin{cases} \overline{B}_{r,p} & \text{if} \quad m^{-1}(\overline{Y}_r^p) \geq \overline{B}_{r,p}, \\ \underline{B}_{r,p} & \text{if} \quad m^{-1}(\overline{Y}_r^p) \leq \underline{B}_{r,p}, \\ m^{-1}(\overline{Y}_r^p) & \text{otherwise.} \end{cases} \tag{4.2}$$

Moreover, $c_{r,p} = \infty$ if and only if $\underline{B}_{r,p} > \overline{B}_{r,p}$.
To summarize, the computation of $\theta^*_{r,p}$ (and hence the computation of the minimal costs $c_{r,p}$) reduces to finding the non-trivial solutions of (4.1) for all $r \leq i \leq j \leq p$. This can either be

done explicitly (as for the Gaussian family) or approximately e.g. by Newton's method.

4.4 Complexity and computation times

In this section we investigate the complexity of Algorithm 1. To this end, we will only assume that the local bounds $\underline{b}_{i,j}$ and $\overline{b}_{i,j}$ can be computed in $\mathcal{O}(1)$. Let us for a moment fix $p = p_0$ and $k = k_0$. Observe from Algorithm 1 that l runs from R_{k_0-1} to L_{k_0-1} (see line 16 and line 20). For every value of l the *optimal costs* $c_{l+1,p}$ have to be computed, which requires the computation of $\underline{B}_{l+1,p_0}$ and $\overline{B}_{l+1,p_0}$ as shown in (4.2). We first consider the case $l = R_{k_0-1}$ and observe

$$\underline{B}_{l+1,p_0} = \max\left\{\underline{B}_{l+1,p_0-1}, (\underline{b}_{i,p_0})_{i=l+1,\dots,p_0}\right\} \text{ and } \overline{B}_{l+1,p_0} = \min\left\{\overline{B}_{l+1,p_0-1}, (\overline{b}_{i,p_0})_{i=l+1,\dots,p_0}\right\}.$$

In the previous steps of the algorithm, i.e. for $p = p_0 - 1$, the values $\underline{B}_{l_0+1,p_0-1}$ and $\overline{B}_{l_0+1,p_0-1}$ have already been computed and hence can be recycled. Thus, there are $p_0 - R_{k_0-1}$ values from which the maximum/ minimum have to be determined. Hence, this leads to costs of order $\mathcal{O}(p_0 - R_{k_0-1})$ for $l = R_{k_0-1}$. Also for $l = R_{k_0-1} - 1, \dots, L_{k_0-1}$ we find that previous computations can be employed. This follows from the observation that

$$\underline{B}_{l+1,p_0} = \max\left\{\underline{B}_{l+2,p_0}, \underline{B}_{l+1,p_0-1}, \underline{b}_{l+1,p}\right\} \text{ and } \overline{B}_{l+1,p_0} = \min\left\{\overline{B}_{l+2,p_0}, \overline{B}_{l+1,p_0-1}, \overline{b}_{l+1,p}\right\}.$$

In the previous steps, $\overline{B}_{l+2,p_0}$ and $\underline{B}_{l+2,p_0}$ have been computed as well as $\overline{B}_{l+1,p_0-1}$ and $\underline{B}_{l+1,p_0-1}$ have been computed for $p = p_0 - 1$. Therefore, the computation of $c_{l+1,p}$ essentially reduces to the computation of $\underline{b}_{l_0+1,p}$ and $\overline{b}_{l_0+1,p}$, which shows that for any $l = R_{k_0-1} - 1, \dots, L_{k_0}$ it is $\mathcal{O}(1)$. In summary, the costs for all computations for $p = p_0$ (and $k = k_0$) are of order $\mathcal{O}(p_0 - R_{k_0-1}) + \mathcal{O}(R_{k_0-1} - L_{k_0-1})$.
We now consider the overall complexity of Algorithm 1. Since, for every k, p runs from R_{k-1} to R_k, one finds that for a sufficiently large constant C the complexity can be bounded by

$$\begin{aligned}
&C \sum_{k=1}^{\hat{K}+1} \sum_{p=R_{k-1}}^{R_k} [(p - R_{k-1}) + (R_{k-1} - L_{k-1})] \\
\leq & C \sum_{k=1}^{\hat{K}+1} (R_k - R_{k-1})^2 + ((R_k - R_{k-1})(R_{k-1} - L_{k-1})) \\
= & C \sum_{k=1}^{\hat{K}+1} (R_k - R_{k-1})((R_k - R_{k-1}) + (R_{k-1} - L_{k-1})) \\
\leq & Cn \left[\max_{k=1,\dots,\hat{K}+1} (R_k - R_{k-1}) + \max_{k=1,\dots,\hat{K}+1} (R_k - L_k) \right], \qquad (4.3)
\end{aligned}$$

where we used $\sum_{k=1}^{\hat{K}+1}(R_k - R_{k-1}) = n$ for the last inequality. We stress that this bound for the complexity is *a-posteriori*, since it depends on L_k as well as on R_k, which in turn depend on the data Y. It sheds some light on these bounds to regard two "extreme" scenarios: If $R_k - R_{k-1}$ can be uniformly bounded from above by a constant the bound in (4.3) is linear in n. If, in contrast the estimated signal is constant, which implies $L_0 = R_0 = 1$ and hence $L_1 = R_1 = n$, the bound is quadratic in n.

From Theorem 11 and the subsequent remark, we may (asymptotically) include the uncertainty of estimating the change-points correctly into the bound in (4.3). Roughly speaking, the result says that the true change-points can be localized with asymptotic probability one at a rate of order $\log n$. For a fixed signal $\vartheta \in \mathcal{S}$ there hence exists a constant $C_1 < \infty$, so that with probability tending to one

$$\max_{k=1,\dots,\hat{K}+1}(R_k - R_{k-1}) \leq \max_{k=1,\dots,\hat{K}+1} n(\tau_k - \tau_{k-1}) + C_1 \log(n) \quad \text{and}$$
$$\max_{k=1,\dots,\hat{K}+1}(R_k - L_k) \leq C_1 \log(n).$$

Together with (4.3) this shows that the complexity is bounded by

$$\max_{k=1,\dots,K+1}(\tau_k - \tau_{k-1})n^2 + 2C_1 n \log n \tag{4.4}$$

with probability converging to one. Clearly, this bound is not a-posteriori, as it only depends on the true locations $(\tau_1, \dots, \tau_K)$ and not on the data Y.

We investigate the actual computation time empirically. To this end, we consider two different signals (see Figure 10). The first signal (left) is constant with value zero, while the second signal (right) consists of segments of 50 observations and values alternating between zero and ten. For $n = 100, 500, 1{,}000, 2{,}000, \dots, 8{,}000$ and standard Gaussian noise we compute the average computation time of SMUCE with $\alpha = 0.1$ in 100 runs each. In order to justify the assumptions that $\underline{b}_{i,j}$ and $\overline{b}_{i,j}$ are computed in $\mathcal{O}(1)$, these values are pre-computed in a first step. The results are shown in Figure 10. The increase is approximately quadratical for the second signal (blue line). For the first signal (red) we find that the computation time increases slightly faster than linearly. However, we note that in fact the computation is much faster for the signal with many change-points, in particular for large values of n. This is in accordance with (4.3) and (4.4). All simulations were performed on a single core system with 2.67 GHz and 8 GB RAM in a 64-bit OS.

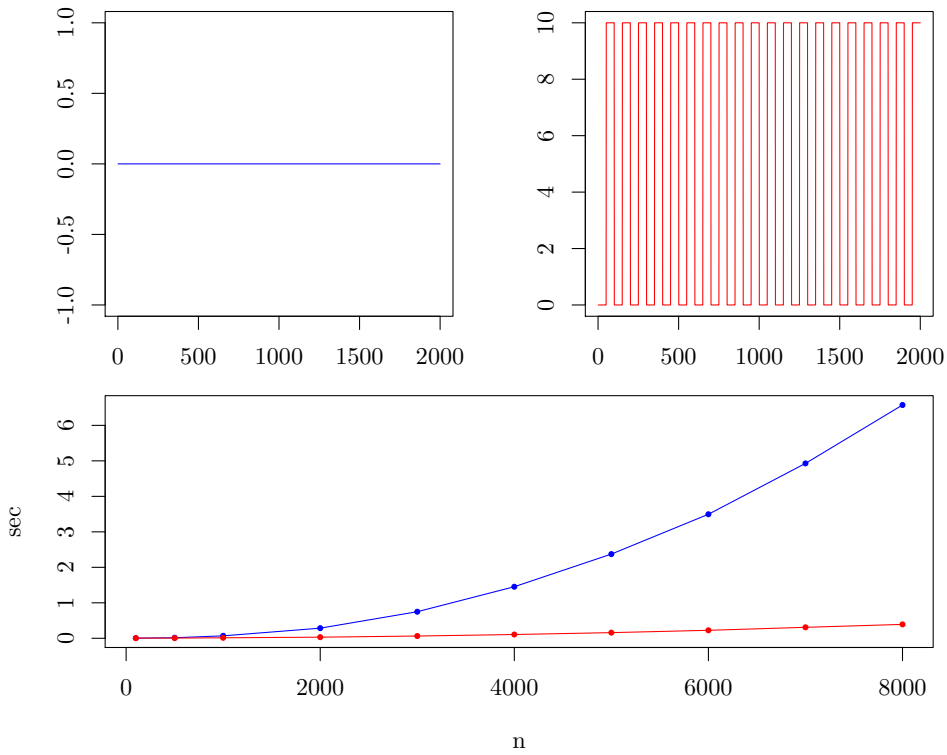


Figure 10: First row: test signals for $n = 2{,}000$ observations; second row: average computation time for SMUCE in dependence of n for both signals.

4.5 Confidence sets

The multiscale constraint underlying SMUCE can be used to obtain confidence intervals $[\tau_k^l, \tau_k^r]$ as well as a confidence band $B(q) \subset [0,1] \times \Theta$ such that for each estimator $\hat{\vartheta} \in \mathcal{H}(q)$

$$\hat{\tau}_k \in [\tau_k^l, \tau_k^r] \text{ for } k = 1, \ldots, \hat{K}(q) \quad \text{and} \quad (t, \hat{\vartheta}(t)) \in B(q), \text{ for all } t \in [0,1].$$

In this section the computational aspects will be discussed in detail. A theoretical foundation for the confidence sets was given in Section 3.6.

4.5.1 Confidence intervals

For the construction of confidence intervals we will consider R_k as in Section 4.2 and further define $L_k = \min\{l : c_{l,R_k} < \infty\}$. Recall that R_k denotes the largest p such that there exists a feasible estimate with k change-points on $\{1, \ldots, p\}$. Then, L_k is the smallest l, such that a feasible constant estimate exists on $\{l, \ldots, R_k\}$. Then, for any estimator $\hat{\vartheta} \in \mathcal{S}[\hat{K}(q)]$ (i.e. with $\hat{K}(q)$ change-points) that satisfies $T_n(Y, \hat{\vartheta}) \leq q$, it holds that $\hat{\tau}_k \in [\tau_k^l, \tau_k^r]$ with $\tau_k^l = n^{-1} L_k$ and $\tau_k^r = n^{-1} R_k$.

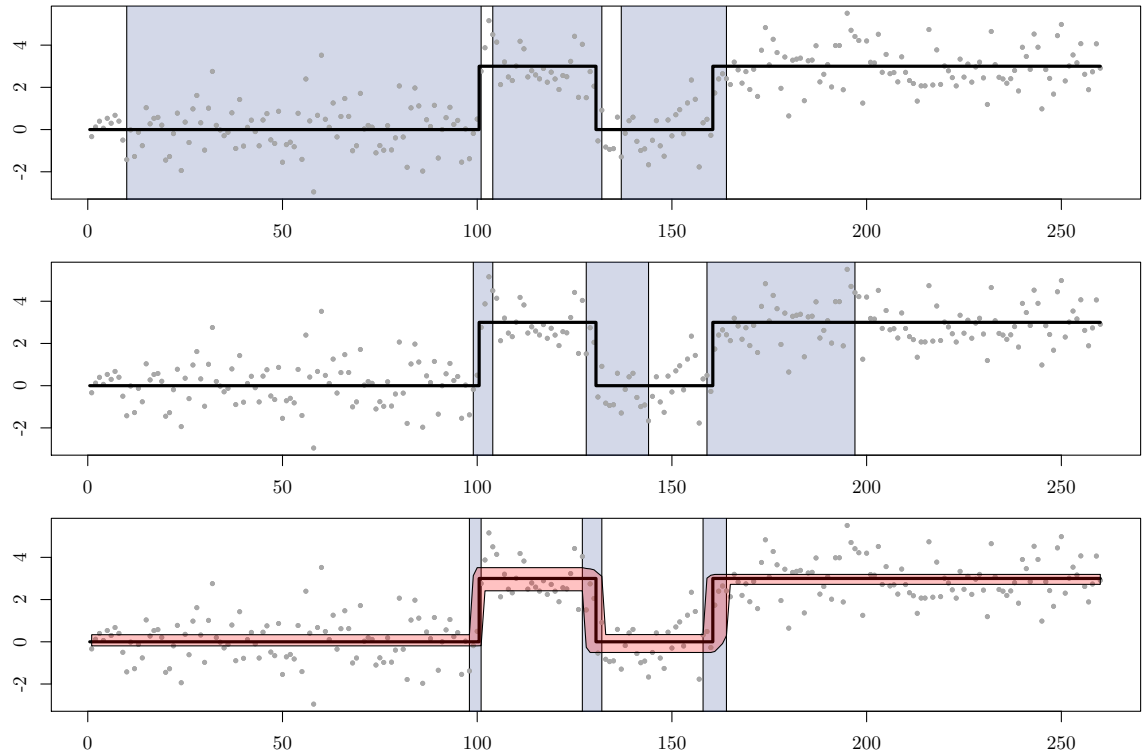


Figure 11: Simulated data and confidence intervals (grey hatched) obtained from applying SMUCE forwards (upper panel), backwards (center panel) and intersection of both (lower panel). The lower panel also shows confidence bands (red hatched).

To show this, assume for a moment that $\hat{\vartheta} \in \mathcal{H}(q)$ is fixed. If for some k, $\hat{\vartheta}$ has no change-point in $[L_k, R_k]$, $\hat{\vartheta}$ must be constant on $[L_k - 1, R_k]$. Since by the definition of L_k no feasible estimate on $[L_k - 1, R_k]$ exists, this contradicts $T_n(Y, \hat{\vartheta}) \leq q$. Since $\hat{\vartheta}$ has exactly $\hat{K}$ change-points, we find that $\hat{\tau}_k \in [\tau_k^l, \tau_k^r]$ for all $k = 1, \ldots, \hat{K}$.

These bounds have been computed within Algorithm 1, and can hence be read of without additional costs. Moreover, the precision of these intervals can be increased by the following idea. Since the optimization problem (2.18) is invariant with respect to reversion of the data, i.e. to consider $(Y_n, \ldots, Y_1)$ instead of $(Y_1, \ldots, Y_n)$, we can also run the dynamic program on the reversed data and compute the corresponding confidence intervals. Taking the intersections of the obtained intervals tightens the confidence intervals considerably in practice, as we illustrated in Figure 11. Clearly, this doubles the overall computation time.

4.5.2 Confidence bands

We construct a confidence band $B(q)$ that contains the graphs of all functions in $\mathcal{H}(q)$. To this end, fix some $\hat{\vartheta} \in \mathcal{H}(q)$ and recall that for $1 \leq k \leq \hat{K}(q)$ there is exactly one change-point in the interval $[\tau_k^l, \tau_k^r]$ and no change-point in (τ_k^r, τ_{k+1}^l). We will consider two cases separately.

First, assume that $t \in (\tau_k^r, \tau_{k+1}^l)$. Then we get a lower and an upper bound for $\hat{\vartheta}(t)$ by $\underline{B}_{R_k+1,L_{k+1}-1}$ and $\overline{B}_{R_{k+1},L_{k+1}-1}$, respectively. Now let $t \in [\tau_k^l, \tau_k^r]$. Then, the k-th change-point is either to the left or to the right of t and hence any feasible estimator is constant either on $[\tau_{k-1}^r, t]$ or on $[t, \tau_{k+1}^l]$. Thus, we obtain a lower bound by $\min\left\{\underline{B}_{R_{k-1},\lfloor tn\rfloor}, \underline{B}_{\lceil nt\rceil,L_{k+1}}\right\}$ and an upper bound by $\max\left\{\overline{B}_{R_{k-1},\lfloor tn\rfloor}, \overline{B}_{\lceil nt\rceil,L_{k+1}}\right\}$, where $\underline{B}_{r,p}$ and $\overline{B}_{r,p}$ are chosen as in (4.2). In the lower panel of Figure 11 we have depicted confidence bands obtained in this manner.

4.6 Software

SMUCE is implemented for the statistical software R (R Core Team, 2013) in the package *stepR* (Hotz and Sieling, 2013)[2]. The SMUCE procedure for Gaussian mean regression, Gaussian variance regression, Binomial regression and Poisson regression is available via the function *smuceR*, which also provides confidence bands and confidence intervals as described in the previous sections.

[2]R package available at `http://www.stochastik.math.uni-goettingen.de/smuce`

SECTION 5

Beyond exponential families

So far we have assumed that the data are given by an exponential family regression model. We extend the methodology to an additive noise model in Section 5.1. Further, we provide a distribution-free approach based on the signs of residuals in Section 5.2, which elaborates the idea as it was outlined in Frick et al. (2013) in detail.

5.1 Sub-Gaussian additive noise

In this section, we show how the methodology and theory underlying SMUCE can be extended to additive noise, different than normal noise. We show that for additive noise with sub-Gaussian tails the limit null-distribution is the same as in the Gaussian case. More precisely, we will consider the following model:

Model 2. *Let $\epsilon_1, \dots \epsilon_n$ be independent and identically distributed observations with $\mathbf{E}[\epsilon_i] = 0$ and $\mathbf{Var}[\epsilon_i] = \sigma^2 > 0$ and assume that there exists a constant $A > 0$ such that*

$$\mathbf{P}(|\epsilon_i| > x\sigma) \leq A \exp(-x^2/2), \quad \text{for all } x > 0. \tag{5.1}$$

Let $\mu \in \mathcal{S}$ be a piecewise constant, right-continuous function with values in $\mathbb{R}$. Further, let the observations $W_1 \dots, W_n$ be given by

$$W_i = \mu(i/n) + \epsilon_i, \quad i = 1, \dots, n.$$

Example 23. Distributions that fulfill the assumptions of Model 2 are e.g. any bounded random variables with mean zero, such as the uniform distribution $U[-u, u]$ for some $u < \infty$ and the Beta$(2, 2)$ distribution.

Since the noise in Model 2 is additive, local tests of the type (2.5) do not depend on the values of the true signal μ. We will consider the multiscale test statistic for Gaussian likelihoods and

investigate the properties of the resulting estimate under Model 2. The multiscale statistic reads as

$$T_n(W,\mu) = \max_{0\le k\le K} \max_{\tau_k \le i/n \le j/n < \tau_{k+1}} \frac{\left|\sum_{l=i}^{j} W_l - \mu(l/n)\right|}{\sqrt{j-i+1}\sigma} - \sqrt{2\log\frac{en}{j-i+1}}. \tag{5.2}$$

We will first show that the null-distribution of T_n converges to the same limit law as for Gaussian data.

As a direct consequence of Dümbgen and Walther (2008)[Theorem 7.1] the asymptotic null-distribution of $T_n(W,\mu)$ is finite almost surely, due to the sub-Gaussian tails of the noise. Moreover, by refined strong Gaussian approximation (see Sakhanenko (1985)), we can prove that under the null-hypothesis the statistic converges weakly to the same distribution as in Theorem 3. Note further that here we do not assume any lower bound on the length of the considered intervals as in Theorem 3.

Theorem 24 (Asymptotic null-distribution)**.** *Set $T_n(W,\mu)$ as in* (5.2)*, let $M_0,\dots,M_K$ be independent copies of M as in* (3.3) *and let W be observations from Model 2. Then,*

$$T_n(W,\mu) \xrightarrow{\mathcal{D}} \max_{0\le k\le K} \sup_{\tau_k \le s < t \le \tau_{k+1}} \left(\frac{|B(t)-B(s)|}{\sqrt{t-s}} - \sqrt{2\log\frac{e}{t-s}} \right) \overset{\mathcal{D}}{\le} M.$$

Empirically, we find that even for relatively small sample sizes the null-distribution of the multiscale statistic is close to the Gaussian version $M^{(n)}$. Figure 12 shows quantile-quantile plots of the null-distribution for Gaussian observations against uniform noise supported on $[-1,1]$ and standardized Beta(2,2) random variables. For the sample size $n = 1{,}000$ we observe only minor differences between the null-distribution of T_n and the Gaussian version. Based on Theorem 24, we can also derive the asymptotic distribution of the extension of T_n on all subintervals.

Corollary 25. *Let $W_1,\dots,W_n$ be observations from Model 2 and let M be as in* (3.3)*. Then,*

$$\max_{1\le i\le j<n} \frac{\left|\sum_{l=i}^{j}(W_l - \mu(l/n))\right|}{\sigma\sqrt{j-i+1}} - \sqrt{2\log\frac{en}{j-i+1}} \xrightarrow{\mathcal{D}} M.$$

Proof. Under the null-hypotheses, the l.h.s. is distributed as $T_n(\epsilon,\mu_o)$ for $\mu_0 \equiv 0$ and $\epsilon_1,\dots,\epsilon_n$ as in Model 2. With this observation the assertion follows directly from Theorem 24. □

Without any further assumptions we can give the rate for multiple detection of change-points, which is purely based on the almost sure finiteness of the asymptotic null-distribution of T_n.

Theorem 26 (Detection rates)**.** *Let $(\mu_n)_{n\in\mathbb{N}} \in \mathcal{S}$ be a sequence of change-point functions with K_n change-points and set Δ_n and Λ_n as in* (3.26)*. Further, let $W_1, \dots, W_n$ be given by Model 2. Assume that q_n is bounded away from zero and*

(1.) for signals on large scales (i.e. $\liminf \Lambda_n > 0$*), that* $\sqrt{\Lambda_n n}\Delta_n/q_n \to \infty$*.*

(2.) for signals on small scales (i.e. $\Lambda_n \to 0$*) and K_n unbounded, that* $\sqrt{\Lambda_n n}\Delta_n \geq (8+\varepsilon_n)\sqrt{-\log(\Lambda_n)}$ *with* $\varepsilon_n\sqrt{-\log(\Lambda_n)}/q_n \to \infty$*.*

Then,

$$\mathbf{P}_{\mu_n}\left(\hat{K}(q_n) < K_n\right) \to 0.$$

Comparing the result to Theorem 19 shows that under either of the assumptions (1.) or (2.) the same constants are obtained as for Gaussian observations. Combination of Theorem 26 and Theorem 24 yields model consistency for any fixed signal $\mu \in \mathcal{S}$ as long as $q_n = \mathrm{o}(1/\sqrt{n})$. In Section 6.6 we empirically assess the performance of the SMUCE with Gaussian likelihoods for uniformly distributed noise.

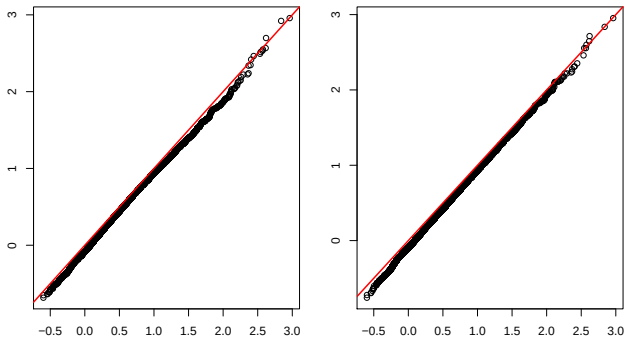


Figure 12: Left: Quantile-quantile plots of the empirical null-distribution of T_n for Gaussian (x-axis) vs. Beta(2,2) data ($y-$axis) with sample size $n = 1{,}000$; right: quantile-quantile plots of the empirical null-distribution of T_n for Gaussian (x-axis) vs. $U(-1,1)$ data ($y-$axis) with sample size $n = 1{,}000$.

5.2 A sign-based version of SMUCE for quantile regression

In this section, we adopt the SMUCE methodology to quantile regression problems. This is of interest e.g. for distributions with heavy tails (e.g. Student's t- or Cauchy distribution) or if the distribution of the observations is unknown. The approach is based on the idea to consider only the signs of residuals. Such sign-based approaches have a long history in non-parametric regression. For example, the *run procedure*, proposed in Davies and Kovac (2001), is build on the construction of "simple" estimates under the constraint that the maximal length of a sequence of residuals with the same sign is below a specified threshold. In contrast, we

suggest to use a multiscale analysis of the residuals' signs in the spirit of the approach in Section 2.3. To state this precisely, let the observations $Z_1, \dots, Z_n$ be given by the following model.

Model 3. *Let $\vartheta_\beta \in \mathcal{S}$ be a piecewise constant, right-continuous function with values in $\mathbb{R}$. Further, let $Z_1, \dots, Z_n$ be independent random variables, such that for some $\beta \in (0,1)$*

$$\mathbf{P}\left(Z_i \leq \vartheta_\beta(i/n)\right) = \beta \quad \textit{for all} \quad i = 1, \dots, n.$$

We now aim for estimation of the piecewise-constant β-quantile function ϑ_β. This can be turned into a Bernoulli regression problem as follows: given the β-quantile function ϑ_β, define the random variables $W(Z, \vartheta_\beta) = (W_1, \dots, W_m)$ as

$$W_i = \begin{cases} 0 & \text{if} \quad \vartheta_\beta(i/n) - Z_i < 0 \quad \text{and} \\ 1 & \text{if} \quad \vartheta_\beta(i/n) - Z_i \geq 0. \end{cases}$$

Under Model 3 we then find that $W_1, \dots, W_n$ are i.i.d. Bernoulli random variables with mean value β. Extending the idea from Section 2.3 we aim for computing the estimate with fewest change-points, such that the signs of the residuals fulfill the multiscale test for Bernoulli observations with mean β. To this end, we consider the multiscale statistic

$$T_n(Z, \vartheta_\beta) = \max_{\substack{1 \leq i \leq j \leq n \\ \vartheta_\beta \text{is constant on } [i/n, j/n]}} \sqrt{2 T_i^j(W(Z, \vartheta_\beta), \beta)} - \sqrt{2 \log \frac{en}{k-j+1}}$$

with the local likelihood-ratio statistic for Bernoulli observations

$$T_i^j(W(Z, \vartheta_\beta)), \beta) = (j-i+1) \left(\overline{W}_i^j \log \left(\frac{\overline{W}_i^j}{\beta} \right) + (1 - \overline{W}_i^j) \log \left(\frac{1 - \overline{W}_i^j}{1 - \hat{\beta}} \right) \right). \tag{5.3}$$

Here, $\overline{W}_i^j = (j-i+1)^{-1} \sum_{l=i}^j W_l$. As before we consider the optimization problem

$$\inf_{\vartheta_\beta \in \mathcal{S}} \# J(\vartheta_\beta) \quad \text{s.t.} \quad T_n(Z, \vartheta_\beta) \leq q. \tag{5.4}$$

Let $\mathcal{H}_\beta(q)$ denote the set of all solutions of (5.4) and define the estimate $\hat{\vartheta}_\beta$ as the restricted maximum likelihood estimate

$$\hat{\vartheta}_\beta(q) = \operatorname*{argmax}_{\vartheta \in \mathcal{H}(q)} \sum_{i=1}^n \log \left(f_\beta(W(Z, \vartheta_\beta)) \right), \tag{5.5}$$

where f_β denotes the density of a Bernoulli distribution with mean β. We will refer to $\hat{\vartheta}_\beta(q)$ as Q-SMUCE in the following. The null-distribution of $T_n(Z, \vartheta_\beta)$ can again be bounded

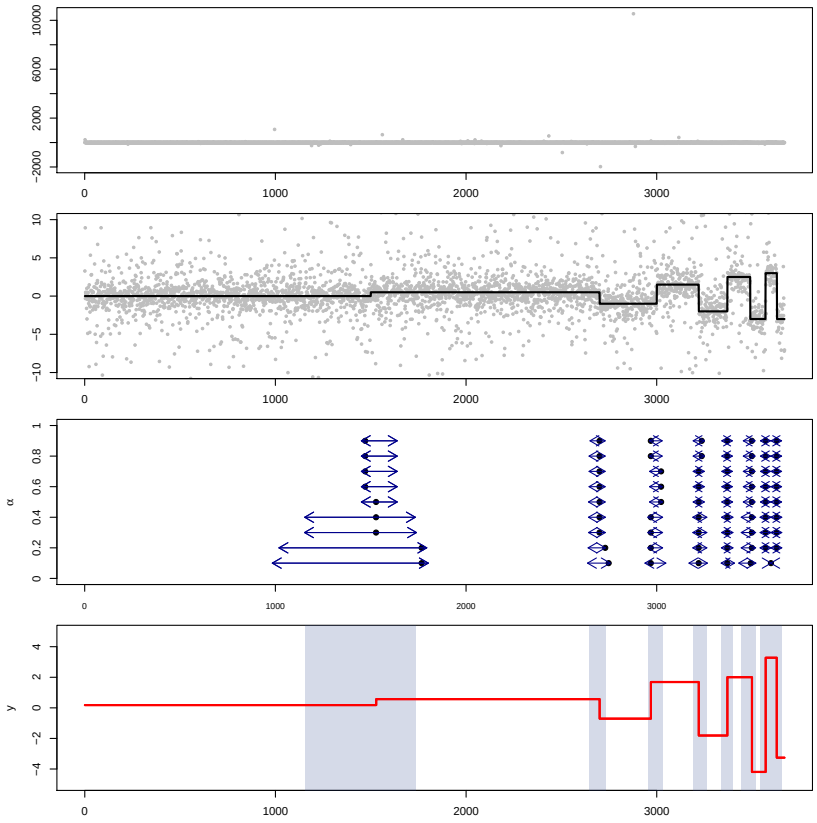


Figure 13: from top to bottom: data Z; data Z (zoomed in); median function $\vartheta_{0.5}$; estimated locations of change-points by Q-SMUCE for different values of alpha with confidence intervals; Q-SMUCE $\hat{\vartheta}_{0.5}$ for $\alpha = 0.3$ with confidence intervals for the change-points location (blue hatched area).

asymptotically by M, as a direct consequence of Theorem 3. Moreover, we emphasize that even non-asymptotically the null-distribution is bounded in probability by $T_n(X, \tilde{\vartheta}_\beta)$ where $X = (X_1, \ldots, X_n)$ are independent Bernoulli random variables with mean β and $\tilde{\vartheta}_\beta \equiv \beta$. Therefore, in practice we can approximate the quantiles by Monte-Carlo simulations with Bernoulli random variables at sample size n.

This idea enables us to control the error of overestimation non-asymptotically as for SMUCE by choosing $q(\alpha)$ as the $(1-\alpha)$ quantile of the null-distribution, such that for all $k \in \mathbb{N}_0$

$$\mathbf{P}\left(\hat{K}(q(\alpha)) > K + k\right) \leq \alpha^{k+1}.$$

Further, confidence bands and intervals can be computed analogously to Section 4.5. An illustration of Q-SMUCE is depicted in 13. A simulation study for an evaluation of the approach can be found in Section 6.5.

5.2.1 Implementation

The Q-SMUCE $\hat{\vartheta}_\beta(q)$ as in (5.5) can be implemented similarly to the general approach in Section 4. However, the computation of the local costs differs from the procedure described in Section 4.3. For this reason, we restrict ourselves to discuss the computation of the local optimal costs. For an interval $[i/n, j/n] \subset [0,1)$ let $\underline{b}_{i,j}$ and $\overline{b}_{i,j}$ denote the two unique solutions of

$$x \log\left(\frac{x}{\beta}\right) + (1-x)\log\left(\frac{1-x}{(1-\beta)}\right) = q/(j-i+1)$$

such that $\underline{b}_{i,j} < \overline{b}_{i,j}$ (see (5.3)). Note that these bounds are independent of the observations Z. Due to the monotonicity of the log-likelihood we then find that $T_j^k(W(\theta_\beta), \beta) \leq q$ if and only if $\overline{W}_i^j(\theta_\beta) \in \left[\underline{b}_{i,j}, \overline{b}_{i,j}\right]$

Moreover, let f_β to be the log-likelihood for a Bernoulli observation with mean β. With this we define the local costs of an estimate θ_β on $[i/n, j/n]$ as

$$c_{i,j}(\theta_\beta) = \begin{cases} -\prod_{t=i}^{j} f_\beta(W_t(\theta_\beta)) & \text{if } \max_{i\leq k\leq l\leq j} T_j^k(W(\theta_\beta),\beta) \leq q, \\ \infty & \text{otherwise.} \end{cases}$$

Following the notation in Section 4.3 let $\underline{B}_{r,p} = \max_{r\leq i\leq j\leq p} \underline{b}_{i,j}$ and $\overline{B}_{r,p} = \min_{r\leq i\leq j\leq p} \overline{b}_{i,j}$. Further, set

$$x^* = \operatorname*{argmax}_{m\in\mathbb{N}:0\leq m\leq j-i+1} \left(f_\beta(1)^m f_\beta(0)^{j-i+1-m}\right).$$

In other words, for the sum of $j-i+1$ independent Bernoulli observations with mean β, x^* is observed with highest probability. Furthermore, let $Z_{[1]}^{ij}, Z_{[2]}^{ij}, \ldots, Z_{[j-i+1]}^{ij}$ be the order statistic of $Z_i, \ldots, Z_j$. Then, $\prod_{t=i}^{j} f_\beta(W_t(\theta_\beta))$ is maximized by $Z_{[x^*]}^{rp}$. Therefore, the optimal local estimate on an interval $[i/n, j/n]$ is given as

$$\theta_{i,j}^* = \begin{cases} \overline{B}_{i,j} & \text{if} \quad Z_{[x^*]}^{ij} \geq \overline{B}_{i,j}, \\ \underline{B}_{i,j} & \text{if} \quad Z_{[x^*]}^{ij} \leq \underline{B}_{i,j}, \\ Z_{[x^*]}^{ij} & \text{otherwise,} \end{cases}$$

by the same convexity argument as in Section 4.3. Therefore, with these modifications, the estimate $\hat{\vartheta}_\beta$ as defined in (5.5) can be computed by dynamic programing as described in Section 4.

SECTION 6

Simulations and applications

We first discuss strategies to choose the threshold parameter q in practice. Then, the performance of SMUCE is investigated in various simulations and the results are compared with state-of-the-art methods for change-point regression. The Sections 6.1-6.4 trace back to Frick et al. (2013) and are complemented in Section 6.5 and Section 6.6 by simulations for the modifications introduced in Section 5. Finally, the application of SMUCE to binary observations in DNA segmentation (Futschik et al., 2013) is illustrated by means of a data set.

6.1 On the choice of q for finite sample size n

The choice of the parameter q in (2.15) is crucial for it balances data fit and parsimony of the estimator. First we discuss a general methodology that takes into account prior information on the true signal ϑ. Based on this, a specific choice is given in the second part which we found particularly suitable for our purposes. Further generalizations are discussed briefly. In addition, a data-driven choice of q based on controlling the false discovery rate is introduced in Section 8.1.

For the Gaussian case we have shown in Section 3.5 that the bound for overestimation is non-asymptotic, i.e.

$$\mathbf{P}(\hat{K}(q) > K) \leq \alpha_n(q), \tag{6.1}$$

where $\alpha_n(q)$ is defined as $\alpha_n(q) = \mathbf{P}(M^{(n)} \geq q)$ with $M^{(n)}$ as in (3.1). This allows to control the probability of overestimating the number of change-points. If the latter is considered as a measure of smoothness, (6.1) can be interpreted as a *minimal smoothness guarantee*. This is similar in spirit to results on other multiscale regularization methods (see Donoho (1995) and Frick et al. (2012)). As argued in Section 3.6, in general it is not possible to bound the minimal number of change-points without further assumptions on the true function ϑ. However, we can draw a bound for the probability of underestimating the number of change-points from Theorem (14) in terms of the minimal interval length Λ and minimal feature size $\eta^2 = n\Lambda\Delta^2$,

which gives

$$\mathbf{P}\left(\hat{K}(q) < K\right) \leq \left[1 - \exp\left(-\frac{\left(\sqrt{n(\Lambda/2)}\Delta - 2q - \sqrt{8\log\frac{2e}{\Lambda}}\right)_+^2}{8}\right) - \exp\left(-\frac{n\Lambda\Delta^2}{16}\right)\right]^{2/\Lambda}$$
$$=: \beta(q, \eta, \Lambda),$$

where we have exploited the fact that $K \leq 1/\Lambda$. By combining (6.1) with the bound above one finds

$$\mathbf{P}\left(\hat{K}(q) = K\right) \geq 1 - \alpha_n(q) - \beta(q, \eta, \Lambda). \tag{6.2}$$

In order to optimize the bound on the probability of estimating the correct number of change-points, one has to balance the error of over- and underestimation. Therefore, we aim for maximizing the r.h.s. over q. Given Λ and $\eta^2 = n\Lambda\Delta^2$ we therefore suggest to choose q as

$$q^*_{\Lambda,\eta} = \operatorname*{argmax}_{q>0} \left\{1 - \alpha_n(q) - \beta(q, \eta, \Lambda)\right\}. \tag{6.3}$$

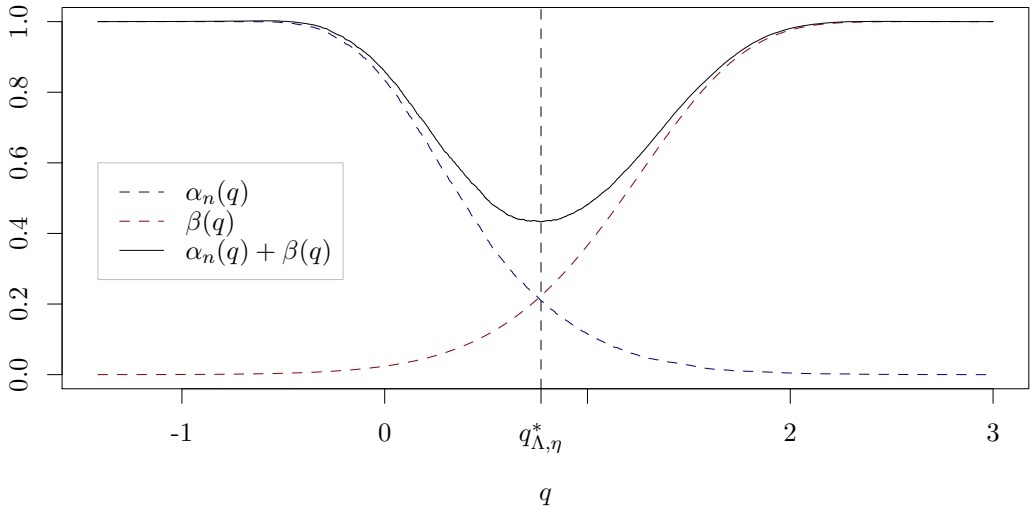


Figure 14: Illustration of $\alpha_n(q)$, $\beta(q, \eta, \Delta)$ and $q^*_{\Lambda\eta}$ as in (6.3) for $n = 500$, $\Delta = 1$ and $\Lambda = 0.25$.

Figure 14 illustrates this balancing of both error terms. The explicit knowledge of the influence of Λ and η in (6.3) paves the way to various strategies for incorporating prior information in order to determine q. One might e.g. use a full prior distribution on (Λ, η) and minimize the posterior model selection error, i.e.

$$\max_{q\in\mathbb{R}} \mathbf{E}\left[1 - \alpha_n(q) - \beta(q, \eta, \Lambda)\right].$$

Here, we suggest a rather simple way to proceed, which we found empirically to perform quite well. We stress that there is certainly room for further improvement. Motivated by the results of Section 3.5.1 we suggest to define Λ and $\eta = \sqrt{n\Lambda}\Delta$ in dependence of n implicitly by the following assumptions

(i) $\eta^* = 8\sqrt{-\log(\Lambda^*)}$ and

(ii) $\sqrt{\Lambda^*} = g(\Delta, n)$,

for some function g with values in $(0, 1]$. According to Theorem 19, the first assumption reflects the worst case scenario among all signals that can be recovered with probability 1 asymptotically. The second assumption corresponds to a prior belief in the true function μ. In the following simulations we always choose $g(\Delta, n) = \Delta$ which puts the decay of Λ and Δ on equal footing. We then come back to the approach in (6.3) and define

$$q_n^* = \max_{q>0} \{1 - \alpha_n(q) - \beta(q, \eta^*, \Lambda^*)\}, \tag{6.4}$$

where λ^* and η^* are defined by (i) and (ii). Consequently, q_n^* is the element that maximizes the probability bound in (6.2). Note that q_n^* does not depend on the true signal μ but only on the number of observations n.

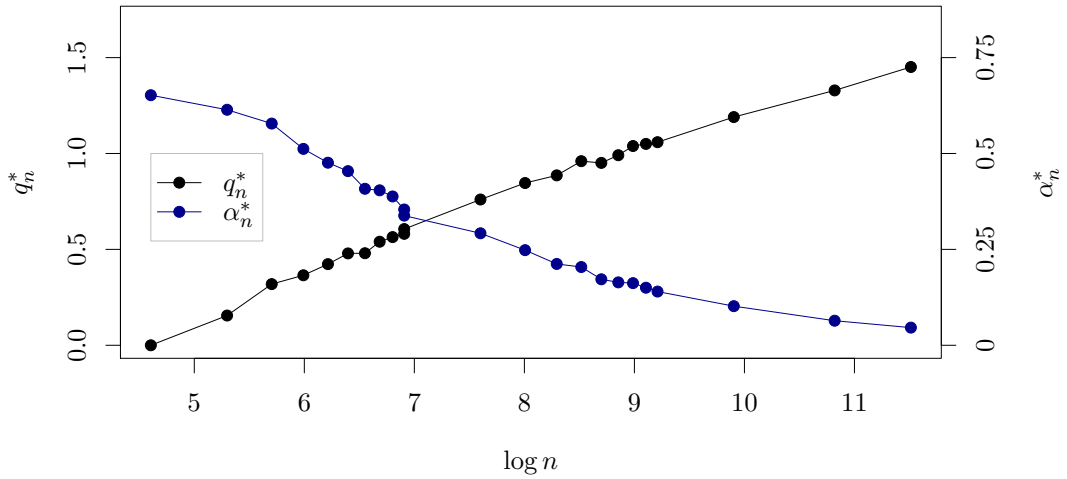


Figure 15: Optimal values q_n^* as in (6.4) obtained from simulations together with the corresponding $\alpha_n^* = \alpha_n(q_n^*)$.

Even though the motivation for q_n^* is build on the assumption of Gaussian observations, simulations indicate that it performs also well for other distributions. That is why we choose $q = q_n^*$, unless stated differently throughout all simulations. To compute q_n^* in practice $\alpha_n(q)$ is estimated by Monte-Carlo simulations. These simulations are rather expensive but only need

to be performed once. For a given n, a solution of (6.4) may then be approximated numerically by computing the r.h.s. for a range of values for q. Figure 15 shows the approximated values of q_n^* for a variety of n, obtained from Monte Carlo simulations of $M^{(n)}$. We stress again that the general concept given by (6.3) can be employed further to incorporate prior knowledge of the signal in applications.

6.2 Gaussian mean regression

Recall model (3.23) in Section 3.5. Throughout this section we assume the variance σ^2 to be known, otherwise one may estimate it by standard methods, see e.g. Davies and Kovac (2001) or Dette et al. (1998). Then, the multiscale statistic (2.12) evaluated at $\hat{\mu} \in \mathcal{S}_n[\hat{K}]$ reads as

$$T_n(Y, \hat{\mu}) = \max_{0 \le k \le \hat{K}} \max_{\hat{l}_k < i \le j \le \hat{l}_{k+1}} \left(\frac{\left| \sum_{l=i}^{j} Y_l - \hat{\mu}_k \right|}{\sigma \sqrt{j-i+1}} - \sqrt{2 \log \frac{en}{j-i+1}} \right).$$

After selecting $\hat{K}(q)$ as the minimal value of (1.4), SMUCE becomes

$$\hat{\mu}(q) = \operatorname*{argmin}_{\hat{\mu} \in \mathcal{S}_n[\hat{K}(q)]} \sum_{k=0}^{\hat{K}(q)} (\hat{l}_{k+1} - \hat{l}_k)(\overline{Y}_{\hat{l}_k}^{\hat{l}_{k+1}} - \hat{\mu}_k)^2 \quad \text{s.t.} \quad T_n(Y, \hat{\mu}) \le q.$$

In our simulation study we compare our approach to the following change-point methods. A large group follows the common paradigm of maximizing a penalized likelihood criterion of the form

$$\mu \mapsto l(Y, \mu) - \mathrm{pen}(\mu) \tag{6.5}$$

over $\mu \in \mathcal{S}_n[k]$ for $k = 1, \ldots, n$, where the function $\mathrm{pen}(\mu)$ penalizes the complexity of the model. This includes the *Bayes Information Criterion (BIC)* introduced in Schwarz (1978). As it was for instance stressed in Zhang and Siegmund (2007), the formal requirements to apply the BIC are not satisfied for the change-point problem. Instead the authors propose the following penalty function, denoted as modified BIC:

$$\mathrm{pen}(\mu) = -\frac{1}{2} \left(3 \# J(\mu) \log n + \sum_{k=1}^{\#J(\mu)+1} \log(\tau_k - \tau_{k-1}) \right).$$

They compare their mBIC method with the traditional BIC as well as with *circular binary segmentation* (Olshen et al., 2004) and the method in Fridlyand et al. (2004) by means of a comprehensive simulation study and demonstrated the superiority of their method with respect to the number of correctly estimated change-points. We only consider the method of

Zhang and Siegmund (2007) and CBS in our simulations.

In addition, we will include the *penalized likelihood oracle (PLoracle)* as a benchmark, which is defined as follows: Recall that K denotes the true number of change-points. For given data Y, define ω_l and ω_u as the minimal and maximal element of the set

$$\left\{\omega \in \mathbb{R} \;:\; \operatorname*{argmax}_{\hat{\mu}\in\mathcal{S}_n} \left(l(Y,\hat{\mu}) - \omega \# J(\hat{\mu})\right) \text{ has } K \text{ change-points}\right\},$$

respectively. In particular, for $\omega_m := 2(\omega_l + \omega_u)$ the penalized maximum likelihood estimator, i.e. a maximizer of (6.5) obtained with penalty $\mathrm{pen}(\mu) = \omega_m \# J(\mu)$, has exactly K change-points. For our assessment, we simulate 10^4 instances of data Y and compute the median ω^* of the corresponding ω_m's. We then define the PLoracle to be a maximizer of (6.5) with $\mathrm{pen}(\mu) = \omega^* \# J(\mu)$. Of course, PLoracles are not accessible in practice (since K and μ are unknown). However, they represent benchmark instances within the class of estimators given by (6.5) and penalties of the form $\mathrm{pen}(\mu) = \omega \# J(\mu)$. We stress again that even if SMUCE and the PLoracle have the same number of change-points they are in general not equal, since the likelihood in (2.18) is maximized only over the set $\mathcal{H}(q)$.

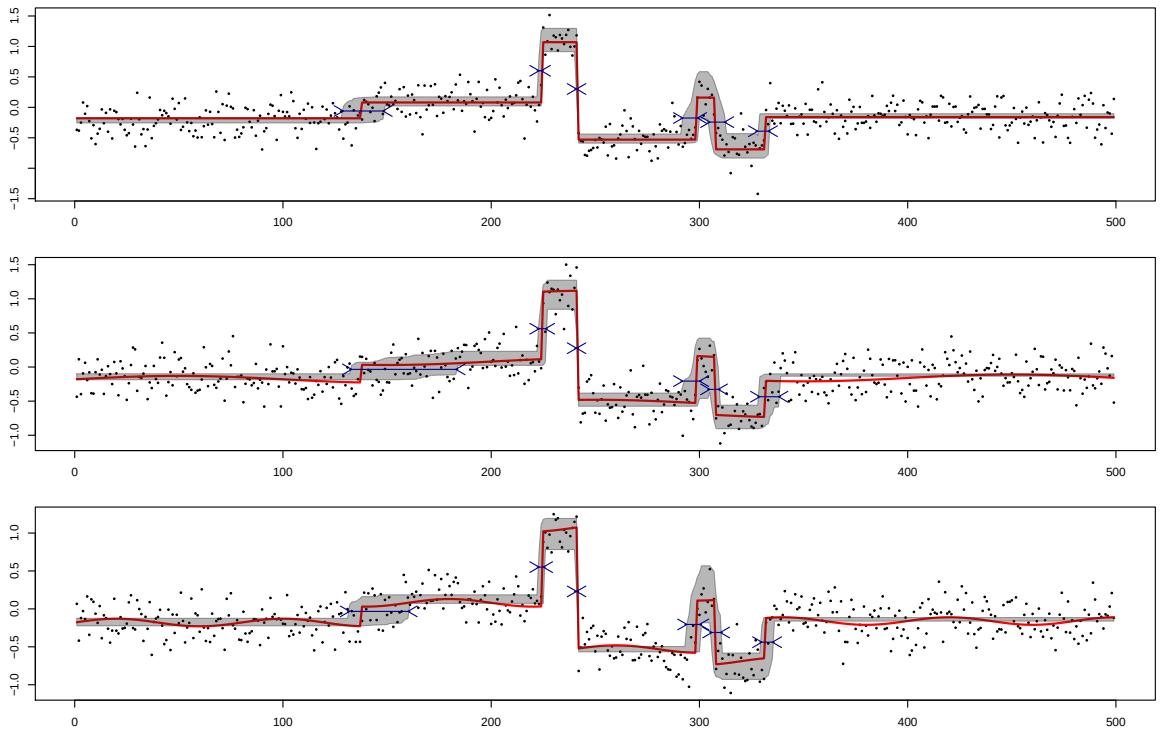


Figure 16: True signal (solid line), simulated data (dots) and confidence bands (grey hatched) and confidence intervals for the change-points (inwards pointing arrows) for $a = 0$ (left), $a = 0.01$ (middle) and $a = 0.025$ (right) and $\sigma = 0.2$.

Moreover, we consider the *fused lasso* algorithm which is based on computing solutions of

$$\min_{\hat{\mu}\in S} \sum_{i=1}^{n} (Y_i - \hat{\mu}(i/n))^2 + \lambda_1 \|\hat{\mu}\|_1 + \lambda_2 \|\hat{\mu}\|_{\mathrm{TV}}, \tag{6.6}$$

where $\|\cdot\|_1$ denotes the l_1-norm and $\|\cdot\|_{\mathrm{TV}}$ the total variation semi-norm (see also Harchaoui and Lévy-Leduc (2010)). The fused lasso is not specifically designed for the change-point problem. However, due to its prominent role and its application to change-point problems (Tibshirani and Wang, 2008), we include it into our simulations. An optimal choice of the parameters (λ_1, λ_2) is crucial and in our simulations we consider two *fused lasso oracles* $\mathrm{FL}^{\mathrm{IMSE}}$ and $\mathrm{FL}^{\text{c-p}}$. In 500 Monte Carlo simulations (using the true signal) we compute λ_1 and λ_2 such that the mean integrated squared error (MISE) is minimized for the $\mathrm{FL}^{\mathrm{IMSE}}$ and such that the frequency of correctly estimated number of change-points is maximized for $\mathrm{FL}^{\text{c-p}}$.

In summary, we compare SMUCE with the modified BIC approach suggested in Zhang and Siegmund (2007), the CBS algorithm[1] proposed in Olshen et al. (2004), the fused lasso algorithm[2] suggested in Tibshirani et al. (2005) and the PLoracle as defined above. Since the CBS algorithm tends to overestimate the number of change-points, the authors included a pruning step which requires the choice of an additional parameter. The choice of the parameter is not explicitly described in Olshen et al. (2004) and here we only consider the un-pruned algorithm.

We follow the simulation setup considered in Zhang and Siegmund (2007). The application they bear in mind is the analysis of array-based comparative genomic hybridization (array-CGH) data. Array-CGH is a technique for recording the number of copies of genomic DNA (cf. Kallioniemi et al. (1992)). As pointed out in Olshen et al. (2004), piecewise constant regression is a natural model for array DNA copy number data. We will discuss the application of SMUCE to these data sets in Section 7.

Here, one has $n = 499$ observations with constant variance σ^2 and the true regression function has 6 change-points at locations $\tau_i = l_i/n$ and $(l_1, \ldots, l_6) = (138, 225, 242, 299, 308, 332)$ with intensities $(\mathfrak{m}_0, \ldots, \mathfrak{m}_6) = (-0.18, 0.08, 1.07, -0.53, 0.16, -0.69, -0.16)$. In order to investigate robustness against small deviations from the model a small deterministic sinusoidal local trend component is included in these simulations, i.e.

$$Y_i \sim \mathcal{N}(\mu(i/n) + 0.05 \sin(a\pi i), \sigma^2), \quad i = 1, \ldots, n. \tag{6.7}$$

Following Zhang and Siegmund (2007) we simulate data for $\sigma = 0.2$ and $a = 0$ (no trend), $a = 0.01$ (long trend) and $a = 0.025$ (short trend), see Figure 16 for an illustration. Moreover,

[1] R package available at `http://cran.r-project.org/web/packages/PSCBS`

[2] R package available at `http://cran.r-project.org/web/packages/flsa/`

	trend	σ	≤ 4	5	6	7	≥ 8	MISE	MIAE
SMUCE $(1-\alpha = 0.55)$	no	0.1	0.000	0.000	0.988	0.012	0.000	**0.00019**	0.00891
PLoracle	no	0.1	0.000	0.000	**1.000**	0.000	0.000	**0.00019**	**0.00874**
mBIC	no	0.1	0.000	0.000	0.964	0.031	0.005	0.00020	0.00888
CBS	no	0.1	0.000	0.000	0.922	0.044	0.034	0.00023	0.00903
$FL^{c\text{-}p}$	no	0.1	0.124	0.122	0.419	0.134	0.201	0.00928	0.15821
FL^{IMSE}	no	0.1	0.000	0.000	0.000	0.000	1.000	0.00042	0.00274
SMUCE $(1-\alpha = 0.55)$	no	0.2	0.000	0.000	**0.986**	0.014	0.000	**0.00117**	**0.01887**
PLoracle	no	0.2	0.024	0.001	0.975	0.000	0.000	0.00138	0.01915
mBIC	no	0.2	0.000	0.000	0.960	0.037	0.003	0.00120	0.01894
CBS	no	0.2	0.000	0.000	0.870	0.089	0.041	0.00146	0.01969
$FL^{c\text{-}p}$	no	0.2	0.184	0.162	0.219	0.174	0.261	0.08932	0.23644
FL^{IMSE}	no	0.2	0.000	0.000	0.000	0.000	1.000	0.00297	0.03692
SMUCE $(1-\alpha = 0.55)$	long	0.2	0.000	0.000	0.825	0.171	0.004	**0.00209**	**0.03314**
PLoracle	long	0.2	0.026	0.030	**0.944**	0.000	0.000	0.00245	0.03452
mBIC	long	0.2	0.000	0.000	0.753	0.215	0.032	0.00214	0.03347
CBS	long	0.2	0.000	0.000	0.708	0.130	0.162	0.00266	0.03501
$FL^{c\text{-}p}$	long	0.2	0.078	0.112	0.219	0.215	0.376	0.08389	0.22319
FL^{IMSE}	long	0.2	0.000	0.000	0.000	0.000	1.000	0.00302	0.03782
SMUCE $(1-\alpha = 0.55)$	short	0.2	0.000	0.002	0.903	0.088	0.007	**0.00235**	**0.03683**
PLoracle	short	0.2	0.121	0.002	0.877	0.000	0.000	0.00325	0.03846
mBIC	short	0.2	0.000	0.000	0.878	0.107	0.015	0.00238	0.03695
CBS	short	0.2	0.000	0.000	0.675	0.182	0.143	0.00267	0.03806
$FL^{c\text{-}p}$	short	0.2	0.175	0.126	0.192	0.210	0.297	0.08765	0.23105
FL^{IMSE}	short	0.2	0.000	0.000	0.000	0.000	1.000	0.00331	0.04111
SMUCE $(1-\alpha = 0.55)$	no	0.3	0.030	0.340	0.623	0.007	0.000	0.00660	0.03829
PLoracle	no	0.3	0.181	0.031	0.788	0.000	0.000	0.00505	0.03447
mBIC	no	0.3	0.015	0.006	**0.927**	0.050	0.002	**0.00364**	**0.03123**
CBS	no	0.3	0.006	0.019	0.764	0.157	0.054	0.00449	0.03404
$FL^{c\text{-}p}$	no	0.3	0.038	0.059	0.088	0.115	0.700	0.08792	0.23496
FL^{IMSE}	no	0.3	0.531	0.200	0.125	0.078	0.066	0.09670	0.24131
SMUCE $(1-\alpha = 0.4)$	no	0.3	0.000	0.099	0.798	0.089	0.000	0.00468	0.03499

Table 2: Frequencies of estimated number of change-points and MISE by model selection for SMUCE, PLoracle, mBIC (Zhang and Siegmund, 2007), CBS (Olshen et al., 2004), the fused lasso oracles $FL^{c\text{-}p}$ and FL^{IMSE}. The results are obtained from 500 simulations and the true signals, shown in Figure 16, have each six change-points.

we included a scenario with a smaller signal-to-noise ratio, i.e. $\sigma = 0.1$ and $a = 0$ and one with a higher signal-to-noise ratio, i.e. $\sigma = 0.3$ and $a = 0$. For $\sigma = 0.1$ and $\sigma = 0.3$ we do not display results with a local trend, since we found the effect to be quite similar to the results with $\sigma = 0.2$.

Table 2 shows the frequencies of the number of detected change-points for all methods mentioned and the corresponding mean integrated squared error (MISE) and mean integrated absolute error (MIAE). Moreover, in Figure 17 we displayed typical observation of model (6.7) with $a = 0.1$ and $b = 0.1$ and the aforementioned estimators. The results show that the SMUCE outperforms the mBIC (Zhang and Siegmund, 2007) slightly for $\sigma = 0.2$ and appears to be less vulnerable for trends, in particular. Notably, SMUCE often performs even better than the PLoracle. For $\sigma = 0.3$ SMUCE has a tendency to underestimate the number of change-points by one, while CBS and in particular mBIC estimates the true number $K = 6$ with high probability correctly. As it is illustrated in Figure 18, this is due to the fact that SMUCE can not detect all change-points at level $1 - \alpha \approx 0.55$ as we have chosen

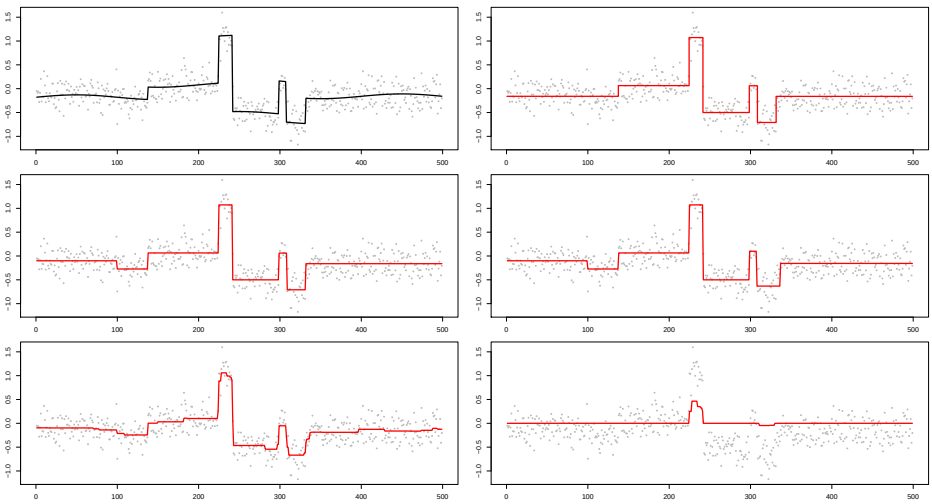


Figure 17: An example of model (6.7) for $a = 0.01$, $b = 0.1$ and $\sigma = 0.2$. From top left to bottom right: true signal, SMUCE, mBIC, CBS, FL^{IMSE} and FL^{cp}.

it following the simple rule (6.4) in Section 4. For further investigation, we lowered the level to $1 - \alpha = 0.4$ (see last row in Table 2). Even though this improves estimation, SMUCE performs comparably to CBS and the PLoracle now, it is still worse than mBIC.

For an evaluation of FL^{MSE} and $\text{FL}^{\text{c-p}}$ one should account for the quite different nature of the fused lasso: The weight λ_1 in (6.6) penalizes estimators with large absolute values, while λ_2 penalizes the cumulated jump height. However, none of them encourages directly sparsity with respect to the number of change-points. That is why these estimators often incorporate many small jumps (well known as the *staircase effect*). In comparison to SMUCE one finds that SMUCE outperforms the FL^{IMSE} with respect to the MISE and it outperforms $\text{FL}^{\text{c-p}}$ with respect to the frequency of correctly estimated the number of change-points. The example in

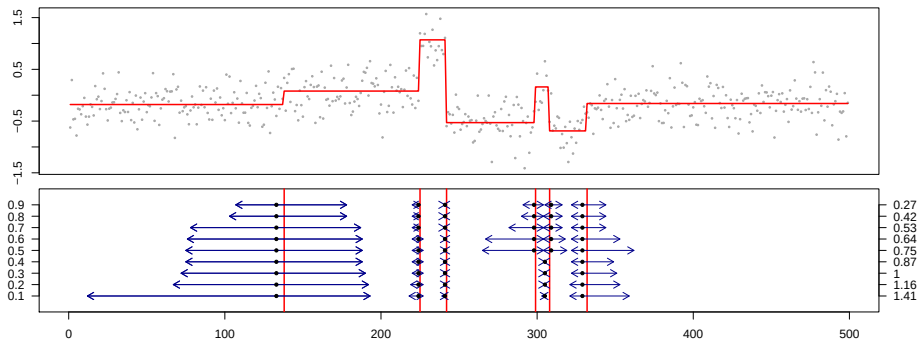


Figure 18: Top: typical example of model (6.7) for $b = 0$ and $\sigma = 0.3$; bottom: change-points and confidence intervals for SMUCE with $\alpha = 0.1, \ldots, 0.9$ (left y-axis) and the corresponding quantiles $q(\alpha)$ (right y-axis).

Figure 17 suggests that the major features of the true signal are recovered by $\mathrm{FL}^{\mathrm{IMSE}}$. But additionally, there are also some artificial features in the estimator which indicate that an additional filtering step has to be included (see Tibshirani and Wang (2008)). Again, we note that Table 2 can be complemented by the simulation study in Zhang and Siegmund (2007) which accounts for the classical BIC (Schwarz, 1978) and the method suggested in Fridlyand et al. (2004).

6.3 Gaussian variance regression

Again, we consider normal data Y_i, however, in contrast to the previous section we aim to estimate the variance $\sigma^2 \in \mathcal{S}$. For simplicity we set $\mu = 0$. This constitutes a natural exponential family with natural parameter $\theta = -(2\sigma^2)^{-1}$ and $\psi(\theta) = -\log(-2\theta)/2$ for the sufficient statistic $Z_i = Y_i^2$, $i = 1, \dots, n$. It is easily seen that the multiscale statistic in this case reads as

$$T_n(Z, \hat{\sigma}^2) = \max_{0 \le k \le \hat{K}} \max_{\hat{l}_k < i \le j \le \hat{l}_{k+1}} \left(\frac{\sqrt{j-i+1}}{\sqrt{2}} \sqrt{\frac{\overline{Z}_i^j}{\hat{\sigma}_k^2} - \log \frac{\overline{Z}_i^j}{\hat{\sigma}_k^2} - 1} - \sqrt{2 \log \frac{en}{j-i+1}} \right).$$

After selecting $\hat{K}(q)$ as the minimal value of (1.4), SMUCE is given by

$$\hat{\sigma}^2(q) = \operatorname*{argmax}_{\hat{\sigma}^2 \in \mathcal{S}_n[\hat{K}(q)]} \sum_{k=0}^{\hat{K}(q)} (\hat{l}_{k+1} - \hat{l}_k) \left(\log \frac{1}{\hat{\sigma}_k^2} - \frac{\overline{Z}_{\hat{l}_l}^{\hat{l}_{k+1}}}{\hat{\sigma}_k^2} \right), \quad \text{s.t.} \quad T_n(Z, \hat{\sigma}^2) \le q.$$

We compare our method to the method proposed in Höhenrieder (2008), see also Davies et al. (2012). Similar to SMUCE they propose to minimize the number of change-points under a multiscale constraint. They additionally restrict their final estimator to coincide with the local maximum likelihood estimator on constant segments. This may increase the number of detected change-points, as pointed out by the authors and confirmed in our simulations.

	k	-3	-2	-1	0	+1	+2	+3	MISE	MIAE
SMUCE	0	0.000	0.000	0.000	**0.945**	0.053	0.002	0.000	**0.00072**	**0.02040**
(Davies et al., 2012)	0	0.000	0.000	0.000	0.854	0.127	0.019	0.000	0.00093	0.02122
SMUCE	1	0.000	0.000	0.000	**0.975**	0.024	0.001	0.000	**0.00653**	**0.04295**
(Davies et al., 2012)	1	0.000	0.000	0.000	0.901	0.089	0.009	0.001	0.00935	0.04648
SMUCE	4	0.000	0.000	0.000	**0.997**	0.003	0.000	0.000	**0.02153**	**0.07967**
(Davies et al., 2012)	4	0.000	0.000	0.000	0.957	0.042	0.001	0.000	0.03378	0.09655
SMUCE	9	0.000	0.001	0.023	**0.973**	0.003	0.000	0.000	**0.06456**	**0.13206**
(Davies et al., 2012)	9	0.000	0.000	0.009	0.968	0.023	0.000	0.000	0.11669	0.18297
SMUCE	19	0.000	0.027	0.222	0.751	0.000	0.000	0.000	**0.26076**	**0.27468**
(Davies et al., 2012)	19	0.000	0.008	0.074	**0.912**	0.006	0.000	0.000	0.47105	0.40606

Table 3: Comparison of SMUCE and the method in Davies et al. (2012). The table shows the frequencies of the differences between the estimated and the true number of change-points for $k = 0, 1, 4, 19$ as well as MISE and MIAE for both estimators.

Following their simulation study we consider test signals σ_k with $k = 0, 1, 4, 9, 19$ equidistant change-points and constant values alternating from 1 to 2 ($k = 1$), from 1 to 2 ($k = 4$), from 1 to 2.5 ($k = 9$) and from 1 to 3.5 ($k = 19$). For this simulation the parameter of both procedures are chosen such that the number of changes should not be overestimated with probability 0.9. For any signal we computed both estimates in 1,000 simulations. The difference of true and estimated number of change-points as well as the MISE and MIAE are shown in Table 3. Considering the number of correctly estimated change-points, it shows that SMUCE performs better for few changes ($k = 1, 4, 9$) and worse for many changes ($k = 19$). This may be explained by the fact that the multiscale test in Davies et al. (2012) does not include a scale-calibration and is hence more sensible on small scales than on larger ones, see also Subsection 8.4. With respect to MISE and MIAE, SMUCE outperforms in every scenario, even for $k = 19$, where Davies et al. (2012) performs better with respect to the estimated number of change-points.

6.4 Poisson regression

We consider the Poisson-family of distributions with intensity $\mu > 0$. Then, $\theta = \log \mu$ and $\psi(\theta) = \exp \theta$. The multiscale statistic is computed as

$$T_n(Y, \hat{\mu}) = \max_{0 \le k \le \hat{K}} \max_{\hat{l}_k < i \le j \le \hat{l}_{k+1}} \left(\sqrt{2(j-i+1)} \sqrt{\overline{Y}_i^j \log \frac{\overline{Y}_i^j}{\mu_k} + \mu_k - \overline{Y}_i^j} - \sqrt{2 \log \frac{en}{j-i+1}} \right).$$

After selecting $\hat{K}(q)$ as the minimal value of (1.4), the SMUCE is given by

$$\hat{\mu}(q) = \operatorname*{argmax}_{\hat{\mu} \in \mathcal{S}_n[\hat{K}(q)]} \sum_{k=0}^{\hat{K}(q)} (\hat{l}_{k+1} - \hat{l}_k)(\overline{Y}_{\hat{l}_l}^{\hat{l}_{k+1}} \log \hat{\mu}_k - \hat{\mu}_k) \quad \text{s.t.} \quad T_n(Y, \hat{\mu}) \le q.$$

In applications one is often faced with the problem of *low count* Poisson data, i.e. when the intensity μ is small. It will turn out that in this case, data transformation towards Gaussian variables such as variance stabilizing transformations are not always sufficient and it pays off to take into account the Poisson likelihood into SMUCE.

	≤ 5	6	7	8	≥ 9	MISE	MIAE	Kullback-Leibler
SMUCE	0.000	0.067	**0.929**	0.004	0.004	**0.274**	**0.217**	**0.0187**
BIC	0.000	0.000	0.080	0.094	0.920	0.575	0.313	0.0417
SMUCE_{mm}	0.013	0.420	0.561	0.005	0.006	0.434	0.364	0.0418
PLoracle	0.045	0.014	0.942	0.000	0.000	0.275	0.217	0.0185
MLoracle	0.000	0.000	1.000	0.000	0.000	0.258	0.208	0.0143

Table 4: Frequencies of $\hat{K}$ and distance measures for SMUCE, the BIC (Schwarz, 1978), SMUCE for variance stabilized signals as well as the PLoracle and MLoracle.

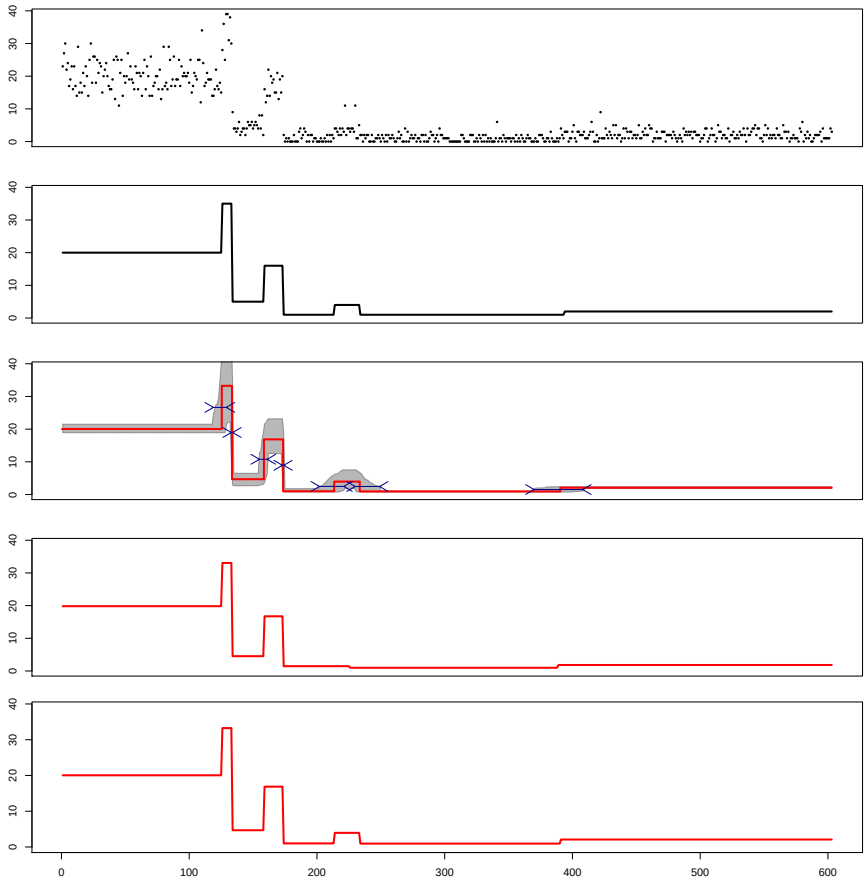


Figure 19: From top to bottom: simulated data, true signal, SMUCE with confidence bands for the signal intensities (gray area) and confidence intervals for the change-points (inward pointed arrows), SMUCE_{mm} and Ploracle.

In the following, we perform a simulation study where we use a signal with a low count and a spike part (see top panel of Figure 19). In order to evaluate the performance of the SMUCE, we compare it to the BIC estimator and the PLoracle as described before. Moreover, we included a version of SMUCE which is based on variance stabilizing transformations of the data. To this end, we applied the *mean-matching* transformation (Brown et al., 2010) to pre-process the data. We then compute SMUCE under a Gaussian model and re-transform the obtained estimator by the inverse mean-matching transform. The resulting estimator is referred to as SMUCE_{mm}. Moreover, as a benchmark, we compute the (parametric) maximum likelihood estimator with $K = 7$ change-points, which is referred to as MLoracle.

Table 4 summarizes the simulation results. As to be expected the standard BIC performs far from satisfactorily. We stress that SMUCE clearly outperforms the SMUCE_{mm}, which is based on Gaussian transformations. Note that SMUCE_{mm} systematically underestimates the number of change-points $K = 7$ which highlights the difficulty to capture those parts of

the signal correctly, where the intensity is low (see also the example in Figure 19). Again, SMUCE performs almost as good as the Poisson-oracle PLoracle. To get a visual impression along with the results of Table 4, we illustrated these estimators in Figure 19.

6.5 Quantile regression

In the following we compare the Q-SMUCE from Section 5.2 with a generalized taut string algorithm which was proposed in Dümbgen and Kovac (2009). Their estimate is constructed in such a way that it minimizes the number of local extreme values among a specified class of functions. Here, a local extreme value is either a local maximum or a local minimum.

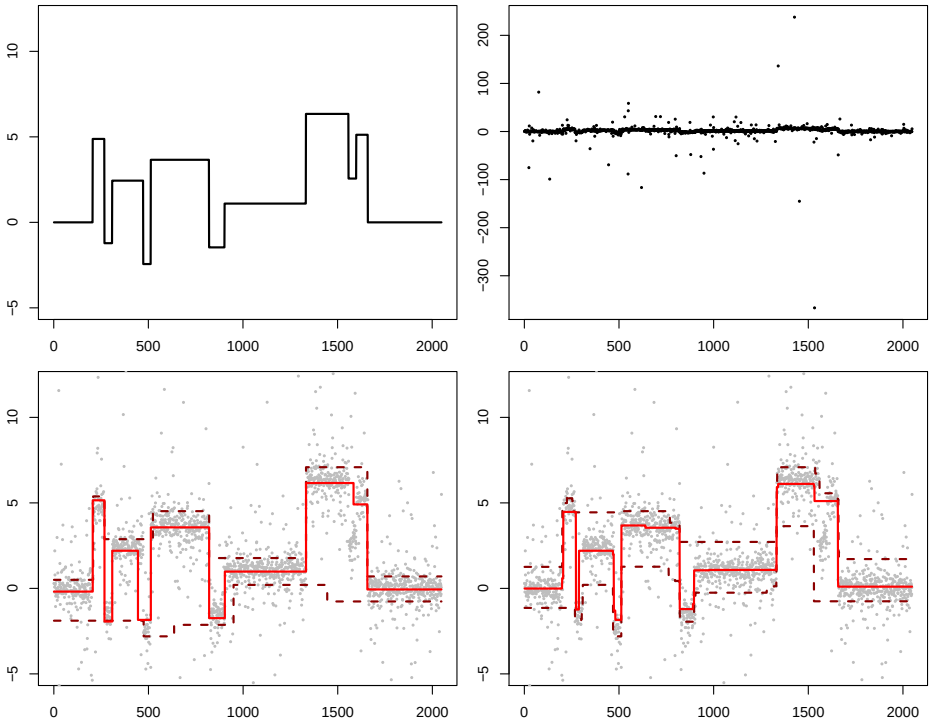


Figure 20: First row: signal *blocks* (left) and simulated Cauchy data (right) for sample size $n = 2048$. Second row: Estimator for median (solid), 0.1 and 0.9-quantiles (dashed) from SMUCE (left) and generalized taut string (right).

In contrast to SMUCE the number of change-points is not penalized. In a simulation study the authors showed that their method is particularly suitable to detect local extremes of a signal. We follow this idea and repeated their simulations for the signals *blocks* and *doppler* (Donoho et al., 1995), see Figure 20 and Figure 21 for an illustration. For the simulations we considered independent Cauchy observations given by

$$Y_i \sim C(\vartheta(i/n), 0.4), \quad i = 1, \ldots, n,$$

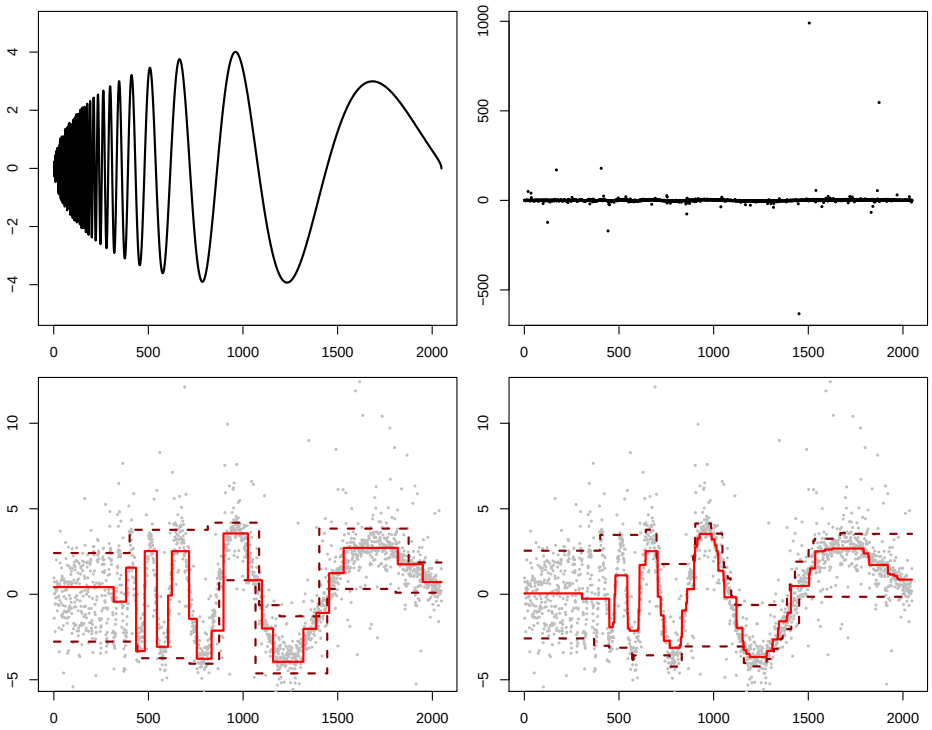


Figure 21: First row: signal *doppler* (left) and simulated Cauchy data (right) for sample size $n = 2048$. Second row: Estimator for median (solid), 0.1 and 0.9-quantiles (dashed) from SMUCE (left) and generalized taut string (right).

where $C(l, s)$ denotes the Cauchy distribution with location l and scale s. The results which also include the estimated number of change-points, are shown in Table 5. Even though detection of local extremes and change-points is almost the same for the blocks signal, it can be seen that the generalized taut string estimates the number of local extremes slightly better than SMUCE, while the number of change-points is overestimated for $n = 2048$ and $n = 4096$. Clearly, this may be explained by the fact that the generalized taut string is primarily designed to have few local extremes instead of change-points. The results are similar for the *doppler* signal, even though it is not piecewise constant. The generalized taut string approximated the signal by step functions which typically incorporate many more change-points than SMUCE. The performance with respect to the number of detected local extremes is better throughout all scenarios. However, the difference is rather small, indicating that SMUCE is able to provide a good approximation of the signal, even though the assumption of a piecewise step function is violated.

			local extreme values			change-points		
signal		n	$\beta = 0.5$	$\beta = 0.1$	$\beta = 0.9$	$\beta = 0.5$	$\beta = 0.1$	$\beta = 0.9$
blocks	SMUCE	512	3 (5.9)	1 (7.9)	2 (7.4)	5 (5.8)	2 (9.1)	3 (8.3)
	gen. taut string	512	3(6.0)	3 (6.6)	3 (6.6)	12 (2.0)	6 (4.9)	7 (4.0)
	SMUCE	2048	9 (0.4)	4 (5.4)	3 (5.8)	11 (0.1)	6 (5.2)	5 (5.9)
	gen. taut string	2048	9 (0.7)	5 (4.0)	3 (5.7)	26 (15.3)	18 (7.1)	16 (5.7)
	SMUCE	4096	9 (0.1)	4 (4.3)	5 (4.5)	11 (0.2)	8 (3.1)	6 (4.8)
	gen. taut string	4096	9 (0.0)	6 (3.1)	3 (5.3)	35 (24.1)	25 (13.8)	21 (9.9)
doppler	SMUCE	512	5	2	1	8	3	2
	gen. taut string	512	5	3	2	38	18	9
	SMUCE	2048	10	4	4	26	7	7
	gen. taut string	2048	11	5	6	132	38	43
	SMUCE	4096	15	6	7	43	13	14
	gen. taut string	4096	16	8	9	266	70	75

Table 5: Comparison of SMUCE and generalized taut string (Dümbgen and Kovac, 2009). Median of local extreme values/ change-points of the estimators and mean absolute difference (in brackets) to true number of local extremes/ change-points. For *blocks* the true number of local extremes equals 9 and the true number of change-points equals 11.

6.6 Uniform noise

In this section we consider uniform additive noise, i.e. for $\mu \in \mathcal{S}$ and some $u > 0$ we assume

$$Y_i = \mu(i/n) + \epsilon_i \quad \text{with} \quad \epsilon_i \overset{i.i.d.}{\sim} U[-u, u].$$

Here $U[-u, u]$ denotes the uniform distribution with support $[-u, u]$. For our simulations we let μ be the *blocks* signal from the previous section. A realization of Y together with the signal μ is shown in Figure 22.

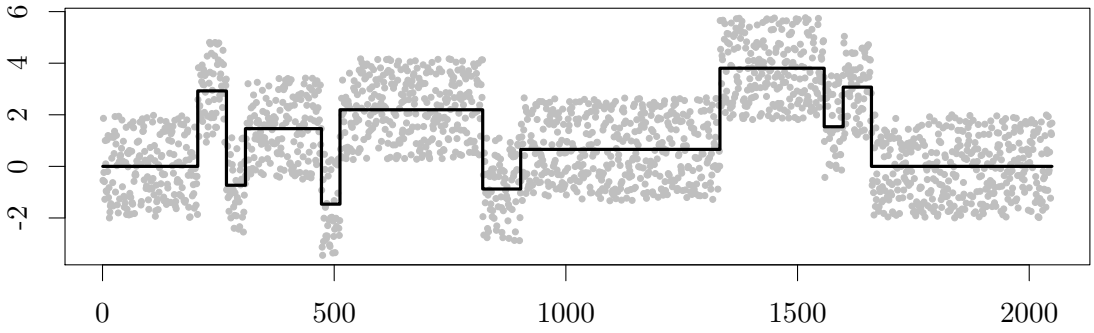


Figure 22: *Blocks* signal (black solid) and simulated uniform noise supported on $[-2, 2]$.

For our simulations we consider the SMUCE with Gaussian likelihood (see also Section 5.1 for a theoretical justification). Clearly, the Gaussian SMUCE is not an optimal estimator for uniform noise. The motivation for this simulation study is to investigate the robustness of SMUCE against violations of the assumed distribution.

		10	11	12	13	≥ 14
SMUCE	$u = 1.5$	0.026	0.974	0	0	0
CBS (Olshen et al., 2004)	$u = 1.5$	0.616	0.863	0.09	0.036	0.011
SMUCE	$u = 2$	0.616	0.384	0	0	0
CBS (Olshen et al., 2004)	$u = 2$	0.002	0.569	0 0.259	0.117	0.053

Table 6: Frequencies of estimated number of change-points for SMUCE and CBS (Olshen et al., 2004) for the blocks signal and independent uniform noise $\epsilon_i \sim U[-u, u]$.

In 1,000 simulations we compare the SMUCE for Gaussian likelihoods and the CBS (Olshen et al., 2004). In the CBS procedure permutations of the data are used for the choice of the required thresholds, which makes it applicable to any distributions. For two different signal-to-noise-ratios ($u = 1.5$ and $u = 2$) the frequencies of the estimated number of change-points are shown in Table 6.

Similar to the results in Section 6.2, we find that SMUCE performs very well for data with a large signal-to-noise ratio, while CBS is superior for signals with higher variance.

In order to assess the locations of the change-points, we depicted histograms of the locations of estimated change-points for both procedures and $u = 2$ in Figure 23. It can be seen that the estimated locations by SMUCE and CBS is quite similar in both scenarios, even though CBS tends to include more change-points.

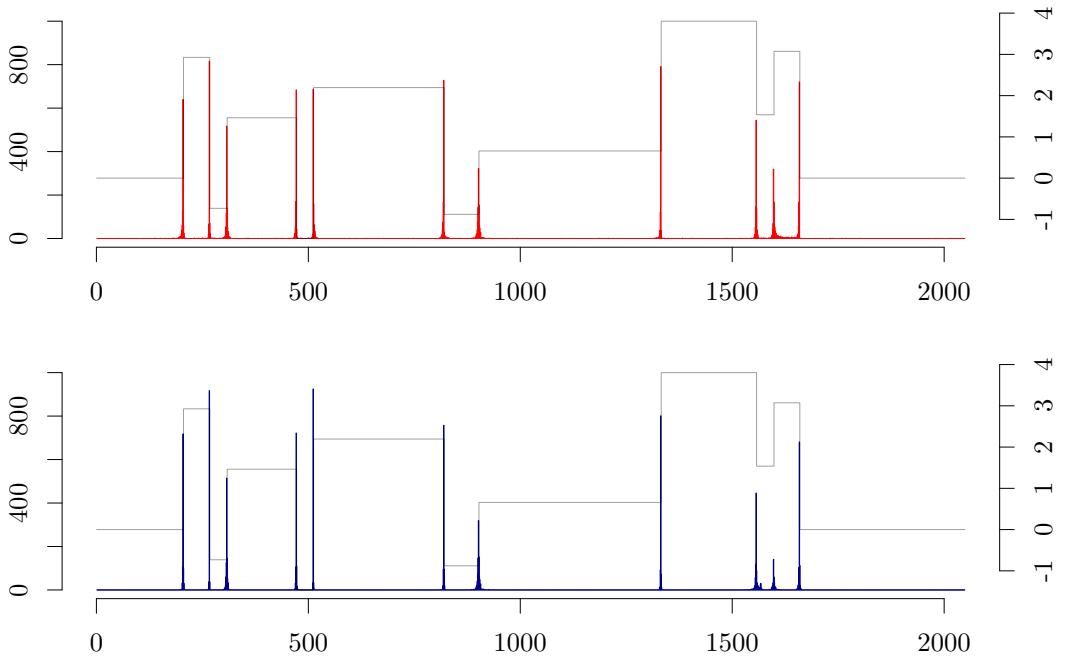


Figure 23: Histograms of the locations of estimated change-points for SMUCE (blue) and CBS (red) and the true signal. The results are based on 1,000 simulations with $u = 2$.

6.7 Application to DNA segmentation

As we have stressed in Section 1 the applications for change-points methods are vast. So far, the SMUCE methodology has been applied thoroughly for the idealization of ion channel recordings (Hotz et al., 2012) and for segmentation of DNA sequences (Futschik et al., 2013). The application to binary data in Futschik et al. (2013) does not need any modification of the approach in Section 2.3. Here, we briefly introduce the application and, as an extension of the empirical findings in Futschik et al. (2013), we illustrate the applicability of SMUCE by means of a data set from the literature.

Change-point methods for binary responses have frequently been used to identify regions of homogeneous GC-content in DNA sequences. Each base of a DNA sequence is one of the four-adenosine(A), guanine(G), cytosine(C) or thymine(T). The DNA sequence is often converted into a binary sequence where G and C are set to "1" and A and T to "0". The relative frequency of bases G and C is referred to as the *GC-content*.

The GC-content is typically not homogeneously composed and detection of these inhomogeneities is important as it correlates with many features of biological interest, see Futschik et al. (2013) for a detailed list. Various methods have been suggested for the segmentation of DNA sequences, including Bayesian methods as in Boys and Henderson (2004) and the quasi-likelihood approach in Braun et al. (2000) (see also Elhaik et al. (2010) for an overview and comparison of available methods).

We think that SMUCE is suitable for these data since it is designed to detect variations on small and large scales simultaneously and can moreover provide significant statements about the detected changes. We illustrate this by means of segmentation of the bacteriophage *lambda*. This virus has been recently analyzed, see e.g. again Braun et al. (2000), Boys and Henderson (2004), Churchill (1992) and the references therein. It became a common benchmark sequence for segmentation algorithms. The whole genome is available from the National Center for Biotechnology Information (NCBI) and can be obtained online[3]. The GC-content of the sequence is shown in the top panel of Figure 24.

The third panel shows the change-points estimated by SMUCE for α varying from 0.1 up to 0.9. For $0.3 \leq \alpha \leq 0.85$ it can be seen that SMUCE detects $\hat{K} = 8$ change-points, which is in accordance with the results from Braun et al. (2000).

We also included the CBS estimate as introduced in Section 6.2 into Figure 24 (second panel). It consists of 12 change-points, which includes the eight change-points, that were detected by SMUCE. Since CBS is not taking into account the multiple testing problem it has the tendency to overestimate change-points. The fact that SMUCE is constructed to control the error of overestimation can therefore be utilized in combination with the results from CBS: a comparison of the estimated change-points by SMUCE and CBS reveals that some of the

[3] `http://www.ncbi.nlm.nih.gov/`

detected changes from CBS can be confirmed with high significance by SMUCE.

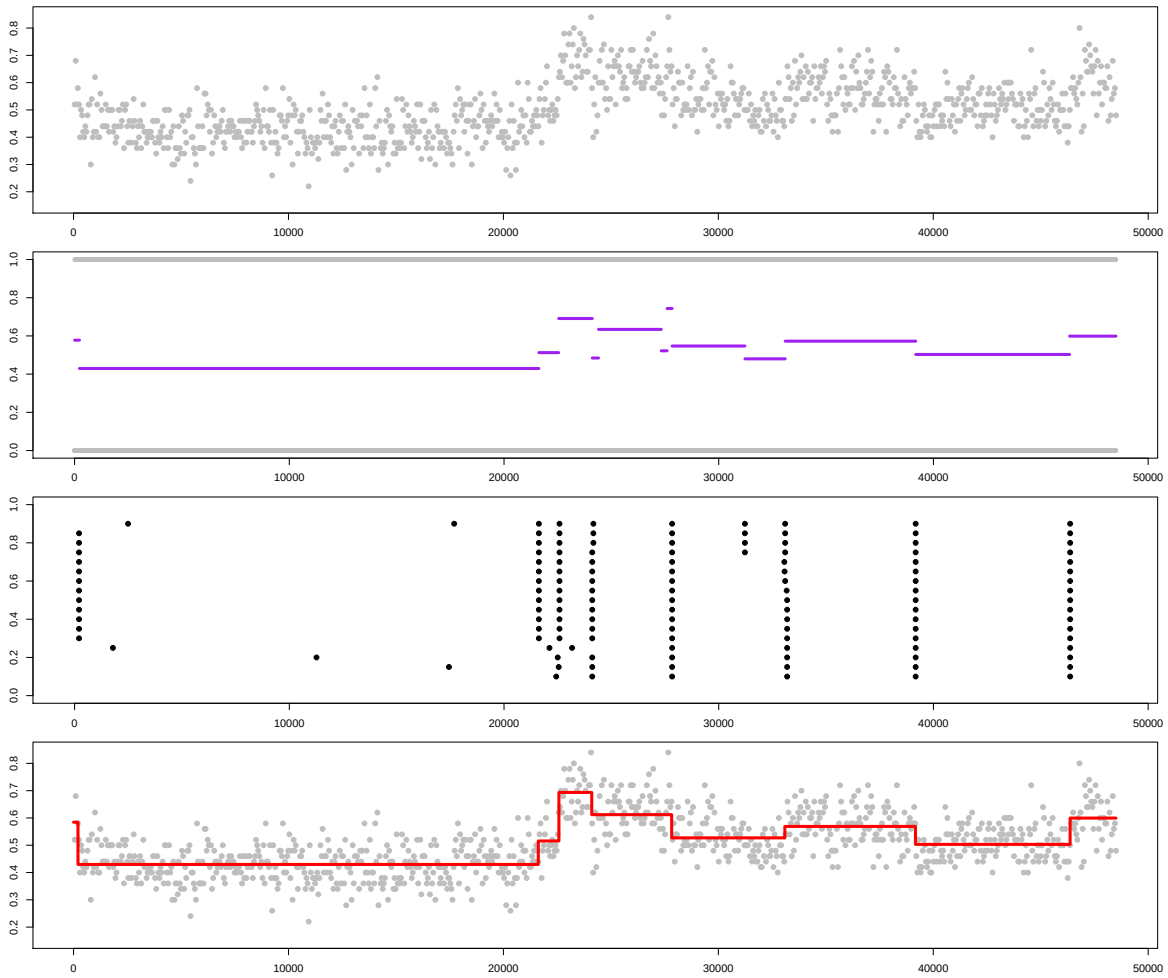


Figure 24: From top to bottom: binned GC-content of bacteriophage *lambda*; segmentation by CBS (Olshen et al., 2004) (purple); estimated change-points by SMUCE for $\alpha \in \{0.1, 0.15, \ldots, 0.9\}$ (y-axis); SMUCE for $\alpha = 0.3$ (red line).

SECTION 7

Multiscale segmentation with few levels

In this section we assume that the mean regression function $\mu \in \mathcal{S}$ has only a finite number of values, which is typically much smaller than the number of change-points. Throughout this section, we restrict ourselves to Gaussian observations. Nevertheless, an extension to additive sub-Gaussian noise as in Section 5.1 is straightforward. More precisely, we consider the following model:

Model 4. *Let $\epsilon_1, \dots, \epsilon_n$ be independent and identically distributed Gaussian observations with $\mathbf{E}[\epsilon_i] = 0$ and $\mathbf{Var}[\epsilon_i] = \sigma^2$. We assume that $\mu \in \mathcal{S}$ is a piecewise constant, right-continuous function with L different values, denoted by $\mathcal{L}(\mu) = \{l_1, \dots, l_L\}$. Further, let the observations $Y_1 \dots, Y_n$ be given by*

$$Y_i = \mu(i/n) + \epsilon_i, \quad i = 1, \dots, n.$$

Henceforth, we will refer to $\mathcal{L}(\mu)$ as the *levels* of μ. SMUCE, as it was defined in (2.18), is entitled to sparsity with respect to the number change-points. However, in general the estimated function values will be different on different segments. In other words, the estimate $\hat{\mu}$ is not enforced to sparsity with respect to its level $\mathcal{L}(\hat{\mu})$. In many applications, however, it is known that the mean function μ takes values from a finite but unknown number of states, which is typically much smaller than the number of change-points. One prominent example is the analysis of array CGH data, which we will discuss in Section 7.3. It may seriously weaken inference in applications if this discreteness is not taken into account. In a first step we will illustrate this and show how SMUCE can be modified, assuming the true levels $\mathcal{L}(\mu)$ are known (Section 7.1). Subsequently, we incorporate an estimation step for the values $\mathcal{L}(\mu)$ into the SMUCE procedure. To this end, we will take advantage of the fact that the multiscale constraint $T_n(Y, \mu) \leq q$ carries two kinds of information about μ simultaneously: first about the location of change-point and second about the values of μ.

7.1 A modification for known levels

Let us assume that the levels $\mathcal{L}(\mu)$ are known. We define the function

$$\Gamma(x,y) := \left(y + \sqrt{2\log\frac{e}{x}}\right) x^{-1/2}.$$

Further, for any $1 \le i \le j \le n$ let $\mu_{i,j} = (j-i+1)^{-1}\sum_{l=i}^{j}\mu_l$ denote the mean value of μ on the interval $[i/n, j/n]$. Then, Theorem 3 yields that with probability greater than $1-\alpha$

$$\mu_{i,j} \in [\underline{b}_{i,j}(q(\alpha)), \overline{b}_{i,j}(q(\alpha))], \tag{7.1}$$

for all $1 \le i \le j \le n$ such that μ is constant on $[i/n, j/n]$. Here $\underline{b}_{i,j}(q) = \overline{Y}_i^j - \Gamma((j-i+1)/n, q)$, $\overline{b}_{i,j}(q) = \overline{Y}_i^j + \Gamma((j-i+1)/n, q)$ and $q(\alpha)$ is the $(1-\alpha)$-quantile of M. Assuming that $\mathcal{L}(\mu)$ is known, (7.1) can easily be refined to

$$\mu_{i,j} \in \left\{[\underline{b}_{i,j}(q), \overline{b}_{i,j}(q)] \cap \mathcal{L}(\mu)\right\},$$

for all $1 \le i \le j \le n$ such that μ is constant on $[i/n, j/n]$. Following the estimation methodology in Section 2.3 we now consider the optimization problem

$$\inf_{\hat{\mu}\in\mathcal{S}} \#J(\hat{\mu}) \quad \text{s.t. } \hat{\mu}_{i,j} \in \left\{[\underline{b}_{i,j}(q), \overline{b}_{i,j}(q)] \cap \mathcal{L}(\mu)\right\}, \tag{7.2}$$

for all $1 \le i \le j \le n$ such that $\hat{\mu}$ is constant on $[i/n, j/n]$. Note that for $q = q(\alpha)$ the side-constraint is fulfilled for the true signal μ with probability greater than $1-\alpha$. As the side-constraint in (7.2) is more restrictive than in (2.15) the estimate will in general incorporate more change-points than the regular SMUCE.

We illustrate the advantage of this approach in the two upper panels of Figure 25, where we compare the regular SMUCE with the modification from (7.2) that incorporates the true levels $\mathcal{L}(\mu) = \{0, 2\}$. Clearly, the additional information result in a better estimation.

7.2 A modification for unknown levels

In real world applications the assumption that $\mathcal{L}(\mu)$ is known is often not realistic. Therefore, we will incorporate estimation of $\mathcal{L}(\mu)$ into the methodology in this section. We begin with a simplification of Model 4, in which we assume that the number of levels $L \le 2$. As shown in the following this leads to a natural way to include inference on $\mathcal{L}(\mu)$ into the procedure. In Section 7.2.2, these ideas will be extended to an arbitrary number of levels.

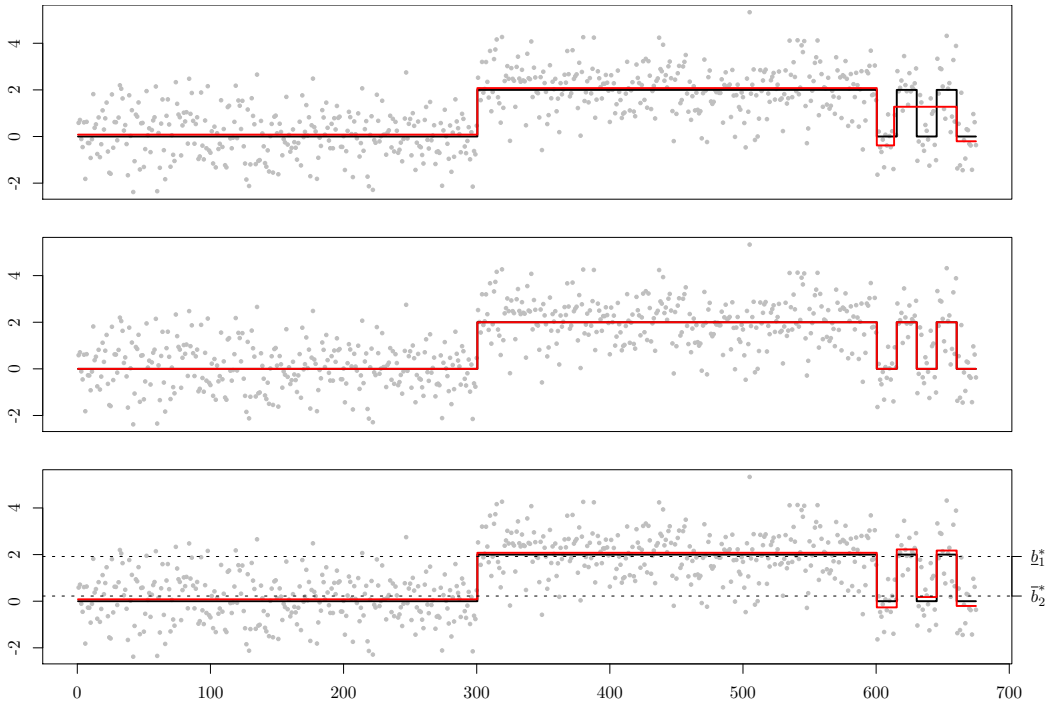


Figure 25: Upper panel: SMUCE (red) with data and true signal (black); Middle panel: modified SMUCE for known levels $\mathcal{L} = \{0, 2\}$ (red) with data and true signal (black); modified SMUCE (as in (7.5)) for estimated levels $\hat{\mathcal{L}}(q)$ as in (7.4) (red) with data and true signal (black). For all estimates q was chosen to be the 0.9-quantile of the null-distribution.

7.2.1 Estimation of signals with two levels

We assume that μ has not more than two different values, i.e. $\mathcal{L}(\mu) = \{l_1, l_2\}$, where we allow for $l_1 = l_2$. By choosing $q = q(\alpha)$, it holds as a consequence of Corollary 25 with probability greater than $1 - \alpha$

$$\mu_{i,j} \in \left\{[\underline{b}_{i,j}(q), \overline{b}_{i,j}(q)]\right\} \quad \text{for all} \quad 1 \le i \le j \le n. \tag{7.3}$$

In particular, this implies

$$\underline{l}_2^*(q) := \max_{1\le i\le j\le n} \underline{b}_{i,j}(q) \le l_2 \quad \text{and} \quad \overline{l}_1^*(q) := \min_{1\le i\le j\le n} \overline{b}_{i,j}(q) \ge l_1.$$

Hence, $\mathbb{R}\setminus\left(\overline{l}_1^*(q), \underline{l}_2^*(q)\right)$ is a confidence set for the true levels $\mathcal{L}(\mu)$.

We set

$$\mathcal{M}(q) := \left\{\mathbb{R} \setminus \left(\underline{l}_1^*(q), \overline{l}_2^*(q)\right)\right\} \tag{7.4}$$

and consider the optimization problem

$$\inf_{\hat{\mu}\in\mathcal{S}} \#J(\hat{\mu}) \quad \text{s.t. } \hat{\mu}_{i,j} \in \left\{[\underline{b}_{i,j}(q), \overline{b}_{i,j}(q)] \cap \mathcal{M}(q)\right\}, \tag{7.5}$$

for all $1 \leq i \leq j \leq n$ so that $\hat{\mu}$ is constant on $[i/n, j/n]$. We then proceed analogously to the method in Section 1.1. The minimal value of (7.5) gives an estimate for the number of change-points and the final estimate for μ is chosen as to be the solution of (7.5) with maximal likelihood.

This approach is illustrated in the lower panel of Figure 25. The signal contains large and small segments of its two levels. The large segments allow us to obtain sharp bounds $\underline{l}_1^*(q)$ and $\overline{l}_2^*(q)$ which in turn strengthen inference on small scales. For this reason, we obtain a considerably better reconstruction than from the regular SMUCE. For an empirical assessment, we ran 100 simulations for the signal in Figure 25 with standard Gaussian noise. Table 7 shows the frequencies of estimated number of change-points. In general, the results show that both modified estimates outperform the regular SMUCE by far. In particular, the estimate based on (7.2) is superior to SMUCE, which does not account for the information that the number of levels is bounded by two.

Remark 27. Note that whenever $T_n(Y, \mu) \leq q$ we find that $l_1 \leq \underline{l}_1^*$ and $l_2 \geq \overline{l}_2^*$. Hence, for all intervals $[i/n, j/n]$, which the true regression function μ is constant on, it holds

$$\mu_{i,j} \in \left\{[\underline{b}_{i,j}(q), \overline{b}_{i,j}(q)] \cap \hat{\mathcal{L}}(q)\right\}.$$

Consequently (given $T_n(Y, \mu) \leq q$), the number of change-points is not overestimated. Which in turn implies that

$$\mathbf{P}\left(\hat{K}(q(\alpha)) > K\right) \leq \alpha$$

remains true for the modified approach given by (7.5).

	2	3	4	5	6	7
SMUCE for known levels	0	0	0	0.19	0.79	0.02
modified SMUCE for $\hat{\mathcal{L}}(q)$ as in (7.4)	0	0	0.8	0.26	0.66	0
SMUCE	0.29	0.59	0.11	0.01	0	0

Table 7: Frequencies of estimated change-points for the modified SMUCE with known levels as in (7.2), the modified SMUCE as in (7.5) and the regular SMUCE. For all estimates q is chosen as the 0.9-quantile of the null-distribution. The results are obtained from 100 simulations with standard Gaussian noise and the signal from Figure 25.

7.2.2 Signals with an unknown number of levels

Finally, we assume the number of levels M to be finite but unknown. We propose a more general modification of SMUCE which is designed to incorporate few different levels.

Assume an interval I contains two subintervals $[i_1/n, j_1/n] \subset I$ and $[i_2/n, j_2/n] \subset I$ such that $\underline{b}_{i_1,j_1}(q) > \overline{b}_{i_2,j_2}(q)$ or vice versa. Clearly, this implies that any function fulfilling the multiscale constraint $T_n(Y, \hat{\mu}) \leq q$ (i.e. $\hat{\mu} \in \mathcal{C}(q)$) has at least one change-point in I. Let $\Psi(q)$ denote the set of all such intervals, i.e. define

$$\Psi(q) := \left\{ [k/n, l/n] : \underline{b}_{i_1,j_1}(q) > \overline{b}_{i_2,j_2}(q) \text{ or } \underline{b}_{i_2,j_2}(q) > \overline{b}_{i_1,j_1}(q) \text{ and } k \leq i_1 \leq j_1 \leq i_2 \leq j_2 \leq l \right\}.$$

We then find that any $\hat{\mu} \in C(q)$ has a change-point in every interval $I \in \Psi(q)$. More precisely,

$$J(\hat{\mu}) = \left(\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{K}}\right) \cap I \neq \emptyset \quad \text{for all } I \in \Psi(q).$$

Hence, $\Psi(q)$ constitutes a confidence region for the location of change-points. The estimated number of change-points $\hat{K}(q)$ as in (2.15) is the minimal value of

$$\inf_{\mu \in \mathcal{S}} \#J(\mu) \quad \text{s.t.} \quad J(\mu) \cap I \neq \emptyset \qquad \text{for all } I \in \Psi(q).$$

Moreover, we can employ $\Psi(q)$ to construct confidence regions for $\mathcal{L}(\mu)$. To this end, consider the complement of $\Psi(q)$ defined as

$$\Psi^C(q) := \left\{ [i/n, j/n] : i < j \text{ and } [i/n, j/n] \notin \Psi(q) \right\}.$$

Then $\Psi^C(q)$ is a confidence set of all intervals the true regression function is constant on. Consequently, a confidence set for the levels $\mathcal{L}(\mu)$ is given by

$$\Upsilon(q) := \bigcup_{[i/n, j/n] \in \Psi^C(q)} [\underline{b}_{i,j}(q), \overline{b}_{i,j}(q)].$$

Recall that in Section 1.1 we proposed first to minimize the number of change-points and estimate the function values $\mathfrak{m}_1, \ldots, \mathfrak{m}_{\hat{K}(q)}$ by the maximum likelihood step in (2.18).

In comparison to SMUCE, we will now interchange the role of change-points and levels. In a first step we concern estimating the levels $\mathcal{L}(\mu)$. For this purpose, we consider the optimization problem

$$\inf_{\mu \in \mathcal{S}} \#\mathcal{L}(\mu) \quad \text{s.t.} \quad \mathcal{L}(\mu) \cap J \neq \emptyset, \quad \forall J \in \Upsilon(q). \tag{7.6}$$

Let $\mathcal{M}(q)$ denote the set of all solutions of (7.6) and $\hat{L}(q)$ its minimal value. For estimation

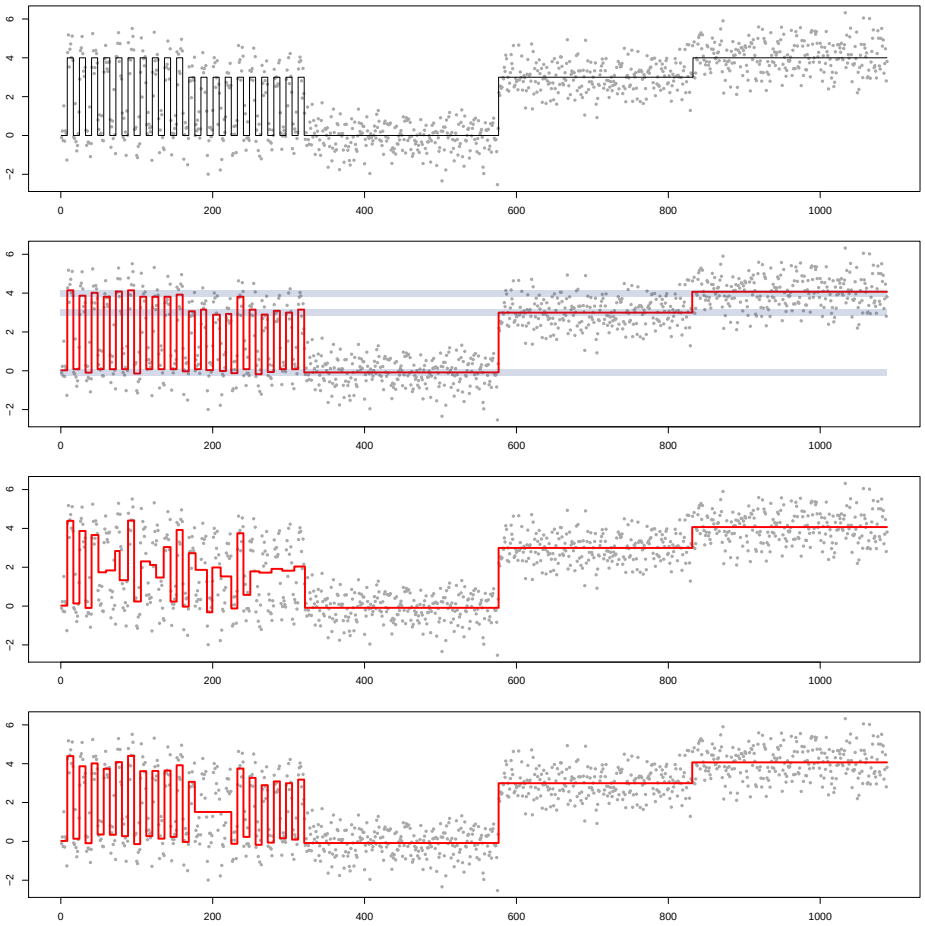


Figure 26: From top to bottom: true signal (black line) and simulated data for Gaussian noise with variance $\sigma = 0.8$; modified SMUCE for $\alpha = 0.1$ together with $\overline{\mathcal{M}}$ (blue hatched area); SMUCE (as in (2.18)) for $\alpha = 0.1$; mBIC estimate (Zhang and Siegmund, 2007).

of μ we then propose to consider the optimization problem

$$\inf_{\mu \in \mathcal{S}} \#J(\mu) \quad \text{s.t.} \quad T_n(Y, \mu) \leq q \quad \text{and} \quad \mathcal{L}(\mu) \in \mathcal{M}(q). \tag{7.7}$$

In other words, we minimize the number of change-points under the additional side-constraint that it has a minimal number of levels (in the sense of (7.6)). In order to make fast computation possible, we consider a relaxation of (7.7) by replacing $\mathcal{M}(q)$ by a superset $\overline{\mathcal{M}}(q) \supseteq \mathcal{M}(q)$. In general, this relaxation decreases the number of detected change-points. This superset $\overline{\mathcal{M}}(q)$ can be constructed very similar in spirit as the confidence intervals in Section 4.5.1. It consists of $\hat{L}(q)$ disjoint intervals $[\underline{l}_1, \overline{l}_1], \ldots, [\underline{l}_{\hat{L}(q)}, \overline{l}_{\hat{L}(q)}]$. A formal proof for $\overline{\mathcal{M}}(q) \supset \mathcal{M}(q)$ is analog to Section 4.5.1 and is omitted.

A pseudo-code for the computation is given in Algorithm 2. The relaxation of (7.7) is then

Data: $\Psi^C(q)$ and $\{\underline{b}_{i,j}(q), \overline{b}_{i,j}(q)\}_{i,j\in\Psi^C(q)}$
Result: $\underline{l}, \overline{l}$
$I \leftarrow \Psi^C(q)$;
for $m = 1, 2, \ldots$ **do**
  $\overline{l}_m \leftarrow \min_{[i/n,j/n]\in I} \overline{b}_{i,j}(q)$;
  $I \leftarrow I \backslash \{[i/n, j/n] : \underline{b}_{i,j}(q) < \overline{l}_m\}$ `// remove already considered intervals;`
  if $I = \emptyset$ **then**
    break
  end
end
$I \leftarrow \Psi^C(q)$;
for $m = 1, 2, \ldots$ **do**
  $\underline{l}_m \leftarrow \max_{[i/n,j/n]\in I} \underline{b}_{i,j}(q)$;
  $I \leftarrow I \backslash \{[i/n, j/n] : \overline{b}_{i,j}(q) > \underline{l}_m\}$;
  if $I = \emptyset$ **then**
    break
  end
end
return $\underline{l}, \overline{l}$

Algorithm 2: Algorithm for the computation of $\overline{\mathcal{M}}(q)$

explicitly given by

$$\inf_{\mu\in\mathcal{S}} \#J(\mu) \quad \text{s.t. } \mu_{i,j} \in \left\{[\underline{b}_{i,j}(q), \overline{b}_{i,j}(q)] \cap \overline{\mathcal{M}}(q)\right\}, \tag{7.8}$$

for all $1 \le i \le j \le n$ so that $\hat{\mu}$ is constant on $[i/n, j/n]$. This relaxed optimization problem can be solved by dynamic programing as in Section 4. Analogously to Section 2.3, let $\hat{K}^*(q)$ denote the minimal value of (7.8) and $\hat{\mu}^*(q)$ be the maximum likelihood estimate among all solutions of (7.8).

Figure 26 shows a simulated data set, together with the modification of SMUCE defined by (7.8). The set $\overline{\mathcal{M}}(q)$ is shown in the second panel (blue hatched area). For comparison, we also depicted the regular SMUCE and the mBIC estimate (Zhang and Siegmund, 2007) as introduced in Section 6.2. Clearly, the regular SMUCE does not give a precise reconstruction of the signal. The mBIC performs slightly better. However, with none of both approaches, the different levels with values three and four can be separated in the first part of the signal. In contrast, the modified approach distinguishes between them almost perfectly.

In 100 simulations for the signal in Figure 26 we found that modified SMUCE $\hat{\mu}^*$ detected the 43 change-points in all 100 simulations at level $\alpha = 0.1$. In contrast, the regular SMUCE at $\alpha = 0.1$ detected 28.7 change-point in average and 35 maximum. This confirms that inference is considerably strengthened by including the information that only few levels are present in the true signal. In the upcoming section we apply the modified approach to a real data set.

7.3 Application to array CGH data

The statistical analysis of array CGH data has drawn a lot of attention recently, see e.g. Fridlyand et al. (2004), Venkatraman and Olshen (2007), Lai et al. (2008) and Jeng et al. (2010).

The analysis of CGH data concerns detection of aberrations in DNA copy number. In normal diploid cells, the autosomal chromosomes each have two copies. The earliest observed aberration is a trisomy of chromosome 21 in Down's Syndrome. Nowadays it is known that changes in copy number occur on parts of chromosomes of different lengths. For example, in cancer cells parts of chromosomes can be present in zero copies (loss) as well as in two or more copies (gain). Detection of gains or losses is e.g. important in order to identify certain cancer genes.

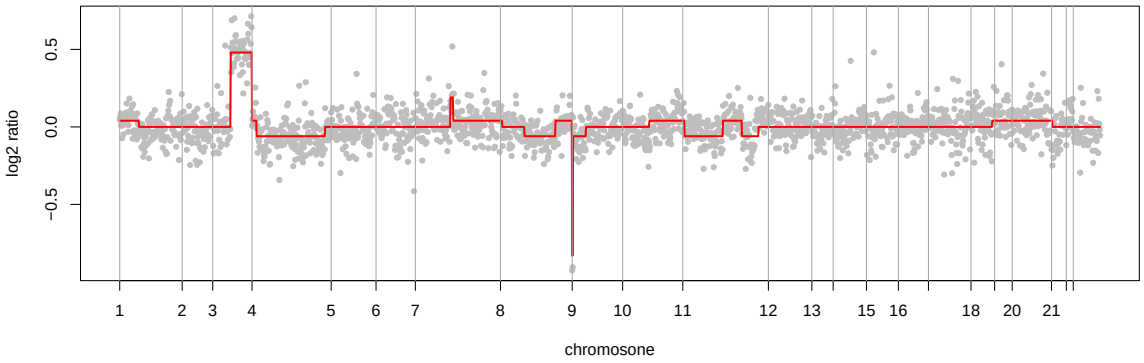


Figure 27: Log ratios and the modified SMUCE for the cell line GM03563.

In array CGH analysis genomic DNA is isolated from a test and a reference sample and labeled differently. In a second step the DNA is hybridized to a DNA micro-array. Ideally, the hybridization intensity for a segment gives the proportion of the copy number of the test and reference sample. For the (statistical) analysis of array CGH data the $\log_2$ ratios of the intensities are considered. Since the ratios at each position are half-integer-valued, the log ratios are discrete. However, due to normal tissue contamination and other effects the log2 ratios differ from the expected values $\log_2(1/2), \log_2(1), \log_2(3/2), \ldots$ and it cannot assumed that these values are known.

We apply the modified approach from the previous section (see (7.8)) to a data set from Snijders et al. (2001), which was also considered in Olshen et al. (2004) and is available online [1]. In total the data set consists of 15 cell lines with 2,276 observations each. In each cell line there are one or two aberrations present as was shown by spectral karyotyping. For comparison we consider the same nine cell lines as in Olshen et al. (2004). They showed that

[1] `http://nature.com/ng/journal/v29/n3/suppinfo/ng754_S1.html`

12 from 15 aberrations are detected by their approach. With the modified SMUCE we could identify the same 12 aberrations at level $\alpha = 0.25$. In addition, in chromosome 9 on GM03563 we could detect an aberration on a segment consisting of only two observations, which is not detected by CBS. This observation is in accordance with the results from the simulations in the previous section. There, it was shown that in particular inference on small segment can be strengthened by the modified approach. Figure 27 shows the data and estimate $\hat{\vartheta}^*(q)$ for the entire cell line GM03563. Further, we depicted magnifications of chromosomes 3 and chromosome 9 in Figure 28 as examples.

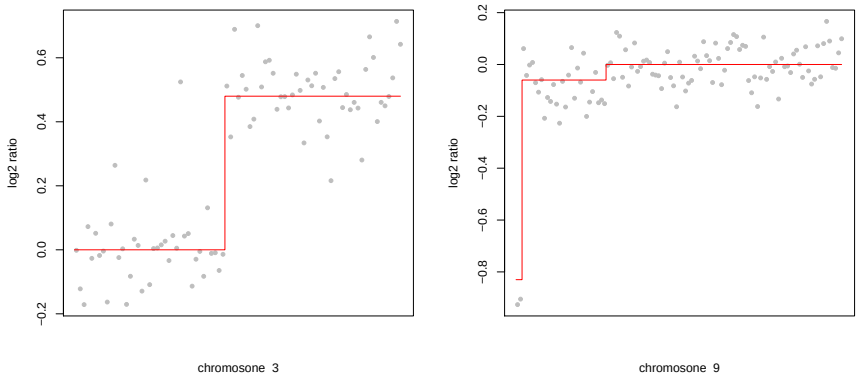


Figure 28: Log ratios and the modified SMUCE for chromosome 3 and chromosome 9.

SECTION 8

Outlook and discussion

In this section we discuss extensions and modifications of the methodology in this thesis. Recall that SMUCE is primarily based on the idea of controlling the probability of overestimation of the number of change-points. We first discuss a less conservative approach by controlling the false discovery rate. Moreover, we give a brief motivation how the method may be extended to dependent data and mention alternative penalizations of the multiscale statistic. Finally, we show how the computation time may be reduced considerably by restricting the multiscale statistic to fewer intervals.

8.1 False discovery rate

In this section we will relate the findings of Chapter 3 to the frequently considered quantities *true and false discoveries*. To this end, we consider Gaussian random variables and stress that asymptotic versions can be obtained for exponential families in general. The following results are a direct consequence of the proofs of Section 3. The false discovery rate as it was introduced in the celebrated work of Benjamini and Hochberg (1995) is a criterion for controlling the Type 1 error in multiple testing which has drawn lots of attention since. Following their notation, let R denote the number of discoveries of a statistical procedure and let V denote the number of *false discoveries*. The *false discovery rate (FDR)* is then given as $\mathbf{E}\left[\mathbf{1}_{\{R>0\}}V/R\right]$.

We consider the multiple testing problem underlying SMUCE in (2.5). As it was pointed out in Siegmund et al. (2011) such local tests are highly correlated and consequently tests on nearby intervals will likely reject the (true) null-hypotheses together. These rejections, however, typically lead to detection of only one single (false) change-point. Instead of considering the number of false rejected null-hypotheses $H_{i,j}$ it is more intuitively to balance false and true discoveries in terms of change-points. The specification of true and false change-points is ambiguous, here we agree upon the following definitions, which are tailor-suited to the findings in Section 3.

Definition 28. For the true signal $\vartheta \in \mathcal{S}$ and an estimate $\hat{\vartheta}$ with change-points $(\tau_1, \dots, \tau_{\hat{K}})$

- an estimated change-point $\hat{\tau}_i$ is a *false positive* (*false discovery*), if $(\hat{\tau}_{i-1}, \hat{\tau}_{i+1}] \subseteq (\tau_j, \tau_{j+1}]$ for some $j = 1, \dots, K$.
- a true change-point τ_i is a *false negative*, if the estimate $\hat{\vartheta}$ is constant on $\left(\frac{\tau_{i-1}+\tau_i}{2}, \frac{\tau_i+\tau_{i+1}}{2}\right]$.

Figure 29 illustrates these definitions by means of an example.
For the SMUCE at level α, the number of discoveries is given by $R(\alpha) = \hat{K}(q(\alpha))$ and we denote the number of false positives for SMUCE at level α by $V(\alpha)$. For $K > 0$ the *sensitivity* rate is then defined as $\mathbf{E}\left[(R(\alpha) - V(\alpha))/K\right]$ and the *false discovery rate* is defined as $\mathbf{E}\left[\mathbf{1}_{\{R(\alpha)>0\}} V(\alpha)/R(\alpha) > 0\right]$.
As a straightforward consequence of Definition 28, the expression $\hat{K}(q(\alpha)) - K$ can be replaced by the number of false positives $V(\alpha)$ in Corollary 6 and Theorem 5. Hence,

$$\mathbf{P}\left(V(\alpha) > 0\right) \leq \alpha \quad \text{and} \quad \mathbf{E}\left[V(\alpha)\right] \leq \frac{2\alpha}{1-\alpha}. \tag{8.1}$$

Similarly, (3.19) implies for $K > 0$ that

$$\mathbf{E}\left[\frac{R(\alpha) - V(\alpha)}{K}\right] \geq 1 - \beta_n(q), \tag{8.2}$$

where β_n is as in (3.17). The bound in (8.1) reveals the nature of SMUCE with respect to false discoveries. The *absolute number* of false discoveries is controlled uniformly over all $\vartheta \in \mathcal{S}$. In contrast to that SMUCE uniformly controls the *sensitivity rate*: the bound in (8.2) does not depend on K but on $\beta_n(q)$ only.
These findings give motivation to a different parameter choice which depends on the number of discoveries and which is related to ideas in Siegmund et al. (2011). This approach will be designed in such a way that not the probability of overestimation is bounded but rather the

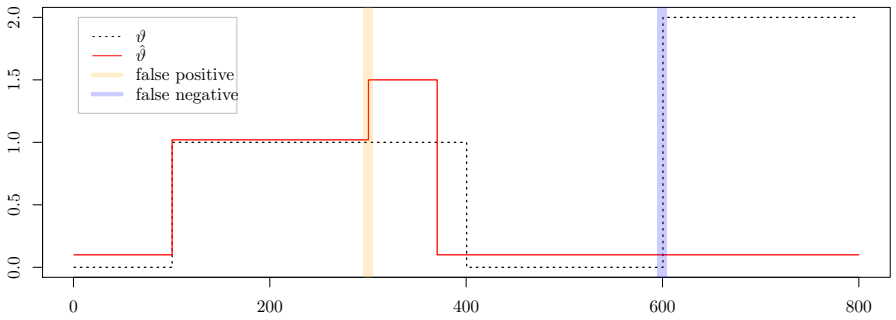


Figure 29: Illustration of false positives and false negatives as in Definition 28

false discovery rate, which leads to a data-driven choice of the threshold q. We chose the level α in such a way that the false discovery rate is bounded by some $\gamma \in (0,1)$. To this end, let

$$\alpha^*(\gamma) := \max\left\{\alpha \in (0,1) : \mathbf{1}_{\{R(\alpha)=1\}}\alpha + \mathbf{1}_{\{R(\alpha)>1\}}\frac{2\alpha}{(1-\alpha)R(\alpha)} \leq \gamma\right\}. \tag{8.3}$$

If the threshold parameter $q = q(\alpha^*)$ is chosen to be the $(1-\alpha^*)$-quantile of the null-distribution of T_n the false discovery rate can be controlled. This is due to the following bound which is based on (8.1) and the definition of α^*. We find that

$$\mathbf{E}\left[\mathbf{1}_{\{R(\alpha^*)>0\}}\frac{V(\alpha^*)}{R(\alpha^*)}\right] \leq \max\left\{\mathbf{E}\left[\frac{V(\alpha^*)}{R(\alpha^*)}\middle| R(\alpha^*)=1\right], \mathbf{E}\left[\frac{V(\alpha^*)}{R(\alpha^*)}\middle| R(\alpha^*)>1\right]\right\}. \tag{8.4}$$

First, we find from the definition of α^* and the r.h.s. of (8.1) that

$$\begin{aligned}\mathbf{E}\left[\frac{V(\alpha^*)}{R(\alpha^*)}\middle| R(\alpha^*)>1\right] &= \mathbf{E}\left[\left(\frac{2\alpha^*}{1-\alpha^*}\frac{1}{R(\alpha^*)}\right)\left(\frac{1-\alpha^*}{2\alpha^*}V(\alpha^*)\right)\middle| R(\alpha^*)>1\right] \\ &\leq \gamma\mathbf{E}\left[\frac{1-\alpha^*}{2\alpha^*}V(\alpha^*)\middle| R(\alpha^*)>1\right] \\ &\leq \gamma.\end{aligned} \tag{8.5}$$

Second, the r.h.s. of (8.1) together with the definition of α^* yield

$$\mathbf{E}\left[\frac{V(\alpha^*)}{R(\alpha^*)}\middle| R(\alpha^*)=1\right] = \mathbf{P}\left(V(\alpha^*)>0\middle| R(\alpha^*)=1\right) \leq \alpha^* \leq \gamma. \tag{8.6}$$

Plugging (8.5) and (8.6) into (8.4) finally gives

$$\mathbf{E}\left[\mathbf{1}_{\{R(\alpha^*)>0\}}\frac{V(\alpha^*)}{R(\alpha^*)}\right] \leq \gamma.$$

This proves that the false discovery rate for the SMUCE at level $\alpha^*(\gamma)$ is bounded from above by γ. Overall, this provides a method in order to control the false discovery rate by choosing $q = q(\alpha^*)$. In order to solve the optimization problem underlying (8.3), one has to compute the path of solutions $R(\alpha) = \hat{K}(q(\alpha))$ for all $\alpha \in (0,1)$. We use an approximation by computing $\hat{K}(q(\alpha))$ for the discretization $\alpha = 5i/100, i = 1,\ldots,20$. Clearly, this will give an approximation for α^*, however, the false discovery rate is controlled for this approximation. We illustrate this approach for two different signals (see Figure 30). For the two data sets in the first row of Figure 30 we computed the selection criterion in (8.3) and chose $\alpha^*(0.15)$ accordingly (vertical gray line), i.e. we bound the false discovery by $\gamma = 0.15$. For the signal with many change-points (left), this leads to $\alpha^*(0.15) = 0.55$ and for the signal with one change-point(right) to $\alpha^*(0.15) = 0.15$. The resulting estimates for this choices are shown in the bottom row. In both scenarios the number of change-points is estimated correctly.

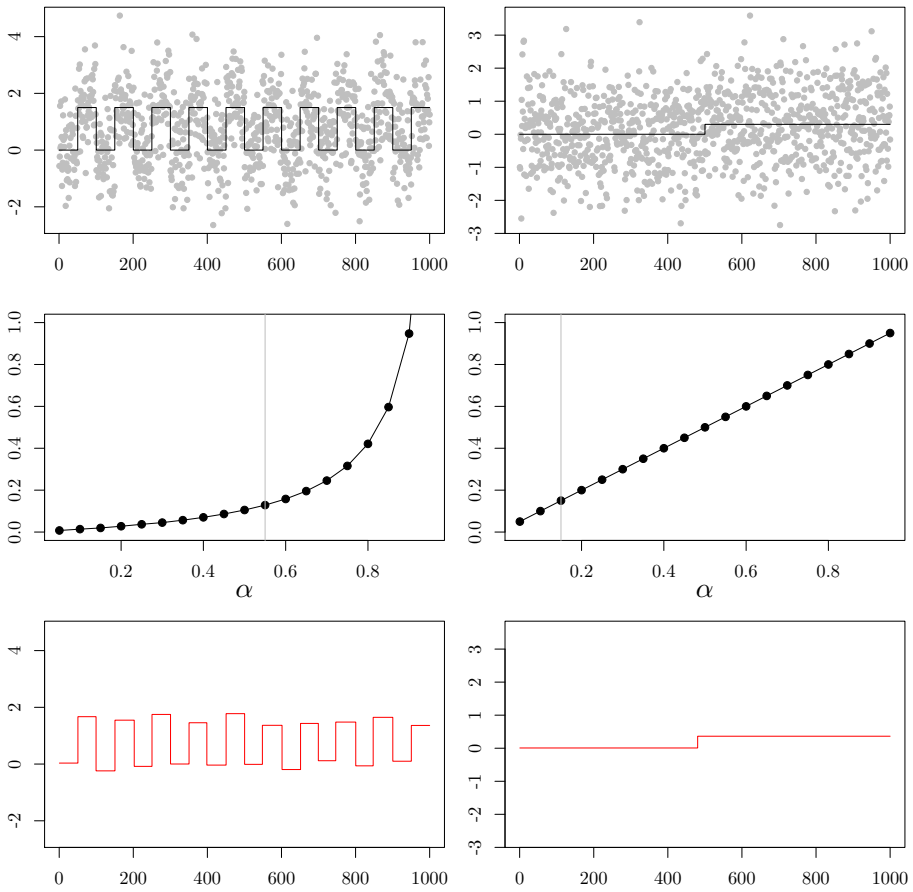


Figure 30: First row: simulated standard Gaussian data and true regression functions (solid line); second row: selection criterion as in (8.3) and optimal choice $\alpha^*(0.15)$ (vertical line); third row: SMUCE for $\alpha^*(0.15)$.

In order to assess the performance empirically we simulated data for both signals in 100 runs with standard Gaussian noise. For each simulation we computed the regular SMUCE at level $\alpha = 0.15$ as well as SMUCE for $\alpha^*(\gamma)$ with $\gamma = 0.15$. The frequency of estimated change-points for both procedures is shown in Table 8. For the signal with one change-point both methods perform equally well. For the signal with 19 change-points, we find that controlling the FDR leads to considerably better results. This is due to the less conservative approach of controlling the false discovery rate instead of the probability of overestimation.

	$K = 19$					$K = 1$		
	≤ 15	16	17	18	19	0	1	2
SMUCE with $\alpha = 0.15$	0.37	0.21	0.31	0.17	0.04	0.07	0.92	0.01
SMUCE with $\alpha = \alpha^*(0.15)$	0	0.01	0.02	0.19	0.78	0.07	0.92	0.01

Table 8: Frequencies of estimated change-points for the signals in Figure 30 by the SMUCE for $\alpha = 0.15$ and the SMUCE for $\alpha^*(\gamma)$ with $\gamma = 0.15$. The results are obtained from 100 simulations.

8.2 Reducing computation time

In order to apply SMUCE to large data sets, one has to reduce the number of considered intervals in the multiscale constraint in order to make fast computation possible. For the applications in Hotz et al. (2012) and Futschik et al. (2013) this was achieved by considering only intervals of dyadic length. However, an interesting strategy to reduce the computational costs even further can be adapted from Walther (2010), see also Rivera and Walther (2012). There it was suggested to restrict the multiscale constraint to a specific system of intervals, which is of size $\mathcal{O}(n)$. The authors could prove that this still guarantees optimal detection. We used the system of intervals as in Rivera and Walther (2012) but also included intervals with size smaller than $\log n$, which were not considered in Rivera and Walther (2012) in the context of density estimation. For the same signal as in Section 6.2 (without a deterministic trend) and for the same level of significance as before we compute the SMUCE for the reduced set of intervals. The results from 1,000 simulations are shown in Table 9. It turns out that the performance is only slightly worse than for the regular SMUCE for $\sigma = 0.1$ and $\sigma = 0.2$ and in fact better for $\sigma = 0.3$.

More striking, however, is the decrease of computation, that comes along with the reduction to fewer intervals. The data consists of $n = 499$ observations and hence potentially $n(n-1)/2 = 124{,}750$ subintervals have to be considered for the computation of the regular SMUCE, whereas the reduced system consists of only 2,207 intervals. This reduction leads to a considerable speed up of the computation times: in average the computation of SMUCE took 0.1384 seconds, whereas the modified version was in average computed in 0.0078 seconds (on a single-core system with 2.67 GHz and 8 GB RAM in a 64-bit OS).

	4	5	6	7
$\sigma = 0.1$	0	0.004	0.979	0.020
$\sigma = 0.2$	0	0.003	0.978	0.022
$\sigma = 0.3$	0.022	0.242	0.702	0.040

Table 9: Frequencies of estimated number of change-points for the modified SMUCE (with $\alpha = 0.1$) for different noise levels σ. The true signal (in Figure 16) has 6 change-points.

8.3 Dependent data

So far the theoretical justification for SMUCE relies on the independence of the data in Model 1. The methodology underlying SMUCE can be extended to piecewise constant regression problems with *serially dependent* data, if the dependence structure is known. We will outline this in the following for a simple example.

Example 29. For a piecewise constant function $\mu \in \mathcal{S}$ we consider the MA(1) model

$$Y_i = \mu(i/n) + \varepsilon_i + \beta\varepsilon_{i-1} \quad \text{for} \quad i = 1, \dots, n,$$

where $\beta < 1$ and $\varepsilon_0, \varepsilon_1, \dots, \varepsilon_n \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$. We aim to adapt the SMUCE to this situation. Following (1.4), one might simply replace the local statistic $\sqrt{2T_i^j(Y, \mu_0)}$ for $\mu_0 \in \mathbb{R}$ in (2.6) by the (modified) local statistics

$$\sqrt{2\tilde{T}_i^j(Y, \mu_0)} = \frac{\left|\sum_{l=i}^j Y_l - \mu_0\right|}{\sqrt{\sigma^2\left[(j-i+1)(1+\beta^2) + (j-i)\beta\right]}}. \tag{8.7}$$

This is motivated by the fact that $\operatorname{Var}(\sum_{l=i}^j Y_l) = \sigma^2\left[(j-i+1)(1+\beta^2) + (j-i)\beta\right]$. Under the null-hypothesis the local statistics $\tilde{T}_i^j$ then marginally have χ_1^2 distributions, as T_i^j in (2.6) for independent Gaussian observations.

In order to control the overestimation error as in Section 3.2, one now has to compute the null-distribution of

$$\tilde{T}_n(Y, \mu) = \max_{\substack{1 \le i < j \le n \\ \mu(t) = \mu_0 \text{ for } t \in [i/n, j/n]}} \left(\sqrt{2\tilde{T}_i^j(Y, \mu_0)} - \sqrt{2 \log \frac{en}{j-i+1}} \right).$$

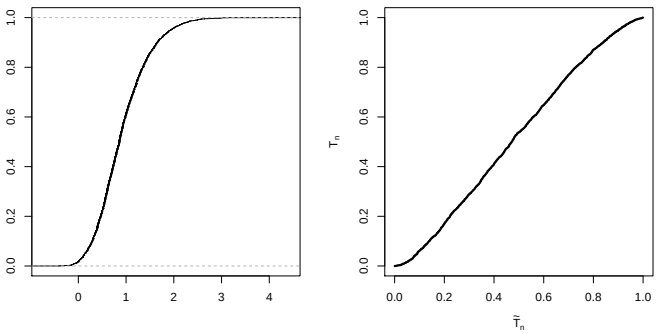


Figure 31: Empirical distribution functions of the null-distribution for dependent observations with $\beta = 0.3$ and probability-probability plot against the null-distribution for independent observations.

To this end, we used Monte-Carlo simulations for a sample size of $n = 500$. We reconsider the test signal from Section 6.2 with $\sigma = 0.2$ and $a = 0$. The empirical null-distribution of $\tilde{T}_n$ and a probability-probability plot of the null-distribution of T_n against $\tilde{T}_n$ are shown in Figure 31.

For $\beta = 0.1$ and $\beta = 0.3$, which corresponds to a correlation of $\rho = 0.1$ and $\rho = 0.27$, we run 1,000 simulations each. We compute the modified SMUCE, as in (8.7), and the SMUCE for independent Gaussian observations. For both procedures we choose q to be the 0.75-quantile of the corresponding null-distribution. The results are shown in Table 10. For $\beta = 0.1$ both procedures perform similarly, which indicates that SMUCE is robust to such weak dependences, while for $\beta = 0.3$ the modified version performs much better with respect to the estimated number of change-points.

	β	5	6	7	8	≥ 9	MISE	MIAE
modified SMUCE	0.1	0.02	**0.98**	0.00	0.00	0.00	0.00154	**0.02104**
SMUCE	0.1	0.00	0.95	0.04	0.00	0.00	**0.00142**	0.02117
modified SMUCE	0.3	0.27	**0.73**	0.00	0.00	0.00	0.00435	**0.03084**
SMUCE	0.3	0.00	0.29	0.34	0.24	0.13	**0.00277**	0.03229

Table 10: Frequencies of estimated number of change-points and MISE by model selection for the modified SMUCE and SMUCE.

The example illustrates that the ideas underlying SMUCE can be successfully applied to dependent data after an adjustment of the underlying multiscale statistic T_n to the dependence structure.

This strategy has been elaborated for m-dependent data in a more complex model in Hotz et al. (2012) in order to apply the SMUCE methodology to estimating the channels conductivity in ion channel recordings. In this application an analog filter is applied before the data analysis which yields dependent observations. The methods capacity is shown in simulations as well as in real data examples.

8.4 Penalizations

The penalization of different scales we use for SMUCE in (2.12) was established in Dümbgen and Spokoiny (2001) and calibrates the number of intervals on a given scale. This prevents small intervals from dominating the statistic. For this purpose one might also consider two alternative penalizations methods given by

$$T_n^1(Y,\vartheta) = \max_{\substack{1\leq i<j\leq n \\ \vartheta(t)=\theta \text{ for } t\in[i/n,j/n]}} \left(\sqrt{\log \frac{en}{j-i+1}} \frac{\sqrt{2T_i^j(Y,\theta)} - \sqrt{2\log \frac{n}{j-i+1}}}{\log(e\log(en/(j-i+1)))} \right)$$

$$T_n^2(Y,\vartheta) = \max_{\substack{1\leq i<j\leq n \\ \vartheta(t)=\theta \text{ for } t\in[i/n,j/n]}} \left(\frac{T_i^j(Y,\theta) - 2\log \frac{n}{j-i+1}}{\log\log \frac{e^e n}{j-i+1}} \right)$$

which are both finite a.s. as $n \to \infty$ (see again (Dümbgen and Spokoiny, 2001, Theorem 6.1) and Dümbgen and Walther (2008)). A multiscale statistic without scale calibration, i.e.

$$T_n^3(Y, \vartheta) = \max_{\substack{1 \le i < j \le n \\ \vartheta(t) = \theta \text{ for } t \in [i/n, j/n]}} T_i^j(Y, \theta)$$

was e.g. considered similarly in Davies et al. (2012) and Davies and Kovac (2001). We illustrate the calibration effect of the statistics T_n, as in (2.12), T_n^1, T_n^2 and T_n^3 in Figure 32. The graphic shows the frequencies at which the multiscale constraint is violated for T_n, T_n^1, T_n^2 and T_n^3 at a certain scale (scales are displayed on the x-axis). It can be seen that T_n^3 puts much emphasis on small scales, while the penalized statistics T_n, T_n^1 and T_n^3 treats the scales more uniformly. The difference between the various penalizations, however, is rather small. For our purposes calibration is beneficial in two ways: First it is required to obtain the optimal detection rates in Theorem 19 as it was shown in Chan and Walther (2013). Second, we find it to work well in practice. However, there is no uniformly superior type of penalization: in application were changes are known to occur mainly on the smallest scales the choice of T_n^3 is most appropriate.

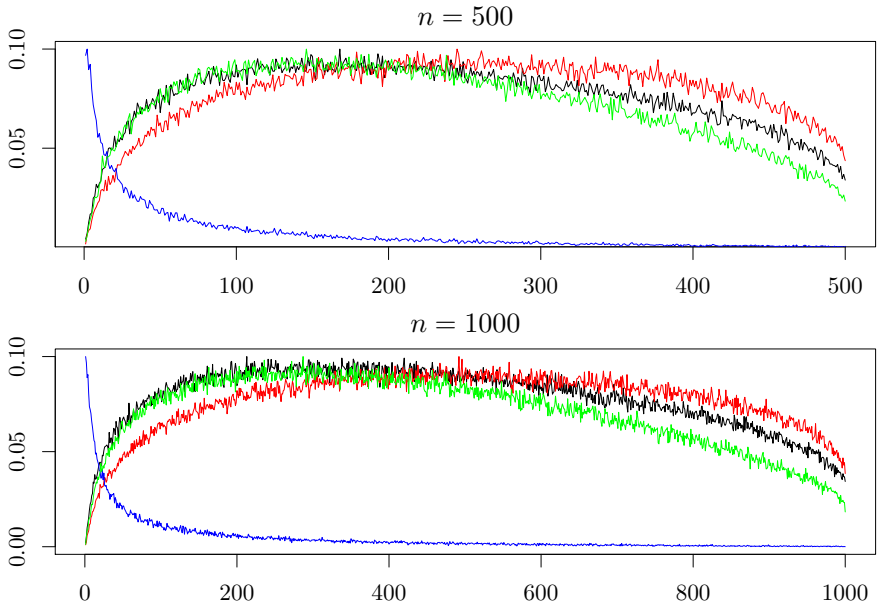


Figure 32: Frequencies at which the corresponding 0.75-quantiles of the statistics T_n, T_n^1, T_n^2 and T_n^3 is exceeded at a certain scale (x axis). Results are obtained from 10,000 simulations with Gaussian observations.

8.5 Piecewise parametric models

In this section, we discuss how SMUCE may be extended to more general right-continuous piecewise models. To this end, let $\Gamma \subset \mathbb{R}^k$ and $h_\gamma(x) : [0,1] \times \Gamma \mapsto \mathbb{R}$. We will assume that the mean regression function μ is in the class of piecewise parametric functions

$$\left\{ \mu(x) = \sum_{k=1}^{K} \mathbf{1}_{[\tau_k, \tau_{k+1})}(x) h_{\gamma_k}(x) : 0 = \tau_0 < \tau_1 < \ldots < \tau_K < \tau_{K+1} = 1, \gamma_1, \ldots, \gamma_k \in \Gamma, K < \infty \right\}.$$

Example 30. Setting $\gamma_k = (\gamma_k^1, \gamma_k^2) \in \mathbb{R}^2$ and $h_{\gamma_k}(x) = \gamma_k^1 + \gamma_k^2 x$ yields piecewise linear regression functions with jumps between its linear segments.

In this more general setting, the observations Y are not identically distributed within one segment. We leave the generality of exponential families and consider only such families that fulfill the following *reproducing property*.

Assumption 31. *Let $\mathcal{F} = \{F_\mu\}$ be a one-dimensional, standard exponential family, parameterized in the mean value $\mu \in m(\Theta)$. Further, let $Y_i \sim \mathcal{F}_{\mu_i}$ be an arbitrary sequence of independent random variables with $\mu_i \in m(\Theta)$ for all $i = 1, \ldots, n$. We assume that a sequence of positive numbers $(a_n)_{n\in\mathbb{N}}$ exists, so that $1/a_n \sum_{i=1}^n Y_i \sim \mathcal{F}_{\tilde{\mu}}$, with $\tilde{\mu} = 1/a_n \sum_{i=1}^n \mu_i$.*

Two examples of such reproducing exponential families are Gaussian distributions with a fixed variance and Poisson distributions. We suggest a modifications of the local tests in (2.5) for a fixed candidate function $\hat{\mu} \in \mathcal{S}$ and the observations $Y_i, \ldots, Y_j$. Recall that $\frac{1}{a_{j-i+1}} \sum_{l=i}^j Y_l \sim F_{\tilde{\mu}}$ for some $\tilde{\mu} \in m(\Theta)$. We consider the local tests

$$H_{i,j} : \tilde{\mu} = \frac{1}{a_{j-i+1}} \sum_{l=i}^{j} \hat{\mu}(l/n) \quad \text{vs.} \quad K_{i,j} : \tilde{\mu} \neq \frac{1}{a_{j-i+1}} \sum_{l=i}^{j} \hat{\mu}(l/n).$$

By choosing $\hat{\theta}_{i,j}$ such that $m(\hat{\theta}_{i,j}) = \frac{1}{a_{j-i+1}} \sum_{l=i}^j \hat{\mu}(l/n)$, the local likelihood-ratios are given by

$$T_i^j(Y, \hat{\mu}) = \sup_{\theta\in\Theta} \left(\theta \frac{1}{a_{j-i+1}} \sum_{l=i}^{j} Y_l - \psi(\theta) \right) - \left(\hat{\theta}_{i,j} \frac{1}{a_{j-i+1}} \sum_{l=i}^{j} Y_l - \psi(\hat{\theta}_{i,j}) \right).$$

Again, we consider the multiscale statistic which evaluates the maximum over all local statistics on intervals between two change-points

$$T_n(Y, \hat{\mu}) = \max_{0\le k\le \hat{K}} \max_{\hat{\tau}_k \le i/n \le j/n < \hat{\tau}_{k+1}} \left(\sqrt{2T_i^j\,(Y, \hat{\mu})} - \sqrt{2 \log \frac{en}{j-i+1}} \right).$$

The null-distribution of $T_n(Y, \mu)$ can be computed, following the ideas in Section 3.1. This is

due to the fact that the minimal sufficient statistic $1/a_{j-i+1}\sum_{l=i}^{j} Y_l$ is again in the exponential family.

A crucial task is to compute the estimate efficiently. In particular it is important to provide a fast computation of local optimal costs as in Section 4.3. More precisely, the local optimal costs $\theta^*_{r,p}$ are a solution for the multiscale constrained maximum likelihood problem

$$\max_{\gamma\in\Gamma}\sum_{l=r}^{p} l(Y_l, h_\gamma(l/n)) \quad \text{s.t.} \quad \underline{b}_{i,j} \le \sum_{l=i}^{j} h_\gamma(l/n) \le \overline{b}_{i,j} \text{ for all } r \le i \le j \le p,$$

where the constraints $\underline{b}_{i,j}$ and $\overline{b}_{i,j}$ are computed analogously to Section 4.3.

In particular, it is important for a fast implementation that these solutions can be updated efficiently, i.e. that $\theta^*_{r,p+1}$ can be computed fast if $\theta^*_{r,p}$ is given. However, even if the computation of the restricted maximum likelihood estimate is too expensive, the number of change-points may be estimated and confidence regions may be constructed, as we will illustrate by means of the piecewise linear regression model in Example 30.

Example 32 (Example 30 revisited)**.** We reconsider the example of piecewise linear regression, i.e. $\gamma_k = (\gamma_k^1, \gamma_k^2) \in \mathbb{R}^2$ and $h_{\gamma_k}(x) = \gamma_k^1 + \gamma_k^2 x$. For this problem, a dynamic programming has been used in Bellman (1961) for the computation of least square solutions for a given number of segments. The special case, in which no jumps between the linear segments are allowed is often referred to as the *broken-line problem* and has been studied extensively (see e.g. Feder (1975) and Siegmund and Zhang (1994)). We briefly discuss the computation of solutions, fulfilling the constraint $T_n(Y, \hat{\mu}) \le q$. From

$$\sum_{l=i}^{j} h_\gamma(l/n) = (j-i+1)h_\gamma(i+j/(2n))$$

we find that the multiscale constraint on an interval $[p/n, r/n]$ is fulfilled by γ if

$$\underline{b}_{i,j}(j-i+1)^{-1} \le h_\gamma\left(\frac{i+j}{2n}\right) \le \overline{b}_{i,j}(j-i+1)^{-1} \quad \text{for all} \quad r \le i \le j \le p. \tag{8.8}$$

In other words, there exists a parameter γ which fulfills the multiscale constraint on $[p/n, r/n]$, whenever a linear function exists, that lies below the points $\big((i+j)/2, \overline{b}_{i,j}(j-i+1)^{-1}\big)$ and above $\big((i+j)/2, \underline{b}_{i,j}(j-i+1)^{-1}\big)$ for all $p \le i \le j \le r$. These conditions can be restated with the notions of *greatest convex minorants* and *least concave majorants* (Barlow et al., 1972). The latter conditions are fulfilled, if the greatest convex minorant of

$$\big((i+j)/2, (j-i+1)^{-1}\overline{b}_{i,j}\big)_{p\le i\le j\le r}$$

and the least concave majorant of

$$\big((i+j)/2, (j-i+1)^{-1}\underline{b}_{i,j}\big)_{p\le i\le j\le r}$$

do not intersect. Hence, confidence intervals for the change-points $\tau_1, \tau_2, \ldots$ can be constructed similar to Section 4.5. Moreover, on the intervals between these confidence intervals we obtain simultaneous confidence bands for the graph of ϑ by the greatest convex minorant and the least concave majorant, as above. Without discussing any algorithmic details, we stress that these can be used to construct simultaneous confidence bands. Figure 33 illustrates this for an example with Poisson observations: any estimator fulfilling the multiscale constraint with minimal number of change-points $\hat{K}$ has exactly one change-point in any of the (blue hatched) confidence regions and its graph lies within the (red hatched) confidence bands.

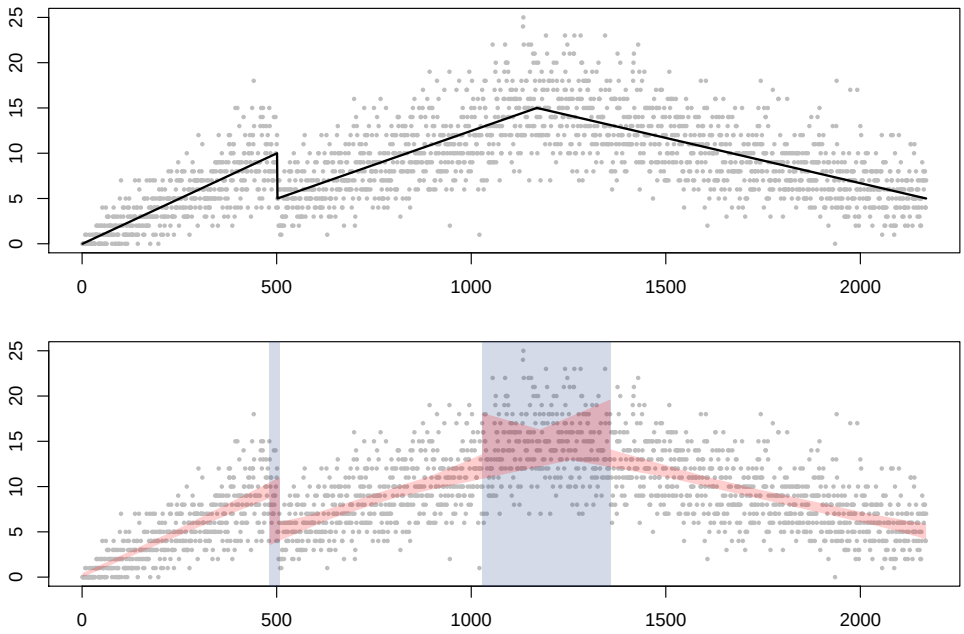


Figure 33: Top: Poisson data and a piecewise linear regression function (black, solid line); bottom: confidence bands for ϑ (red hatched) and confidence intervals for the change-point location (blue hatched). Here q is chosen as the 0.9-quantile of M.

APPENDIX A

Proofs

A.1 Auxiliary Results

A.1.1 Large deviation and power results

We begin by showing some large deviation results for exponential families. Recall that by $D(\theta||\tilde{\theta})$ we denote the *Kullback-Leibler divergence* of F_θ and $F_{\tilde{\theta}}$, i.e.

$$D(\theta||\tilde{\theta}) = \int_{\mathbb{R}} f_\theta(x) \log \frac{f_\theta(x)}{f_{\tilde{\theta}}(x)} \, \mathrm{d}\nu(x) = \psi(\tilde{\theta}) - \psi(\theta) - (\tilde{\theta} - \theta) m(\theta).$$

With the techniques used in Brown (1986) [Thm 7.1] it is readily seen that for a sequence of independent and F_θ-distributed r.v. $Y_1, \ldots, Y_n$ one has that

$$\mathbf{P}\left(\overline{Y} - m(\theta) \geq \eta\right) \leq e^{n(D(\theta||\theta+\varepsilon) - \eta\varepsilon)} \tag{A.1}$$

for all $\varepsilon > 0$ such that $\theta + \varepsilon \in \Theta$. The following restatement of inequality (A.1) turns out to be very useful.

Lemma 33. *Let $Y = (Y_1, \ldots, Y_n)$ be independent random variables such that $Y_i \sim F_\theta$ and assume that $\delta > 0$ is such that $\theta + \delta \in \Theta$. Then,*

$$\mathbf{P}(m^{-1}(\overline{Y}) \geq \theta + \delta) \leq e^{-nD(\theta+\delta||\theta)}.$$

Proof. First observe that according to (A.1)

$$\begin{aligned} \mathbf{P}(m^{-1}(\overline{Y}) \geq \theta + \delta) &= \mathbf{P}(\overline{Y} - m(\theta) \geq m(\theta + \delta) - m(\theta)) \\ &\leq \exp(n(D(\theta||\theta + \delta) - (m(\theta + \delta) - m(\theta))\delta)). \end{aligned}$$

Now it follows from (3.11) that

$$\begin{aligned} D(\theta||\theta+\delta) - (m(\theta+\delta) - m(\theta))\delta &= \psi(\theta+\delta) - \psi(\theta) - m(\theta+\delta)\delta \\ &= -(\psi(\theta) - \psi(\theta+\delta) - (\theta - (\theta+\delta))m(\theta+\delta)) \\ &= -D(\theta+\delta||\theta). \end{aligned}$$

□

From (A.1) we further derive a basic power estimate for the local likelihood-ratio statistic.

Lemma 34. *Let $Y = (Y_1, \ldots, Y_n)$ be independent random variables such that $Y_i \sim F_\theta$ and assume that $\delta \in \mathbb{R}$ is such that $\theta + \delta \in \Theta$. Then, for all $x > 0$*

$$\mathbf{P}\left(\sqrt{2T_1^n(Y, \theta+\delta)} \geq x\right) \geq 1 - \exp\left(n \inf_{\varepsilon \in [0,\delta]} \left[D(\theta||\theta+\varepsilon) - \frac{\varepsilon}{\delta} D(\theta||\theta+\delta) + \frac{\varepsilon x^2}{2n\delta}\right]\right).$$

Proof. For

$$J(\overline{Y}, \theta) = \phi(\overline{Y}) - \big(\overline{Y}\theta - \psi(\theta)\big)$$

we obtain

$$J(\overline{Y}, \theta+\delta) = J(\overline{Y}, \theta) - \delta\overline{Y} - \psi(\theta) + \psi(\theta+\delta). \tag{A.2}$$

Thus, we have for any $z > 0$

$$\begin{aligned} \Pi(z, n, \delta) &:= \mathbf{P}\big(T_1^n(Y, \theta+\delta) \geq z\big) \\ &= \mathbf{P}\left(J(\overline{Y}, \theta+\delta) \geq \frac{z}{n}\right) \\ &= \mathbf{P}\left(J(\overline{Y}, \theta) - \delta\overline{Y} \geq \frac{z}{n} - \psi(\theta+\delta) + \psi(\theta)\right) \\ &\geq \mathbf{P}\left(-\delta\overline{Y} \geq \frac{z}{n} - \psi(\theta+\delta) + \psi(\theta)\right), \end{aligned}$$

where in the last inequality holds since $J(z, \theta) \geq 0$ for all $z \in \mathbb{R}$ and $\theta \in \Theta$. Now, let us first assume that $\delta > 0$. Then by (3.11) we find

$$\mathbf{P}\left(-\delta\overline{Y} \geq \frac{z}{n} - \psi(\theta+\delta) + \psi(\theta)\right) = \mathbf{P}\left(\overline{Y} - m(\theta) \leq -\frac{z}{\delta n} + \frac{D(\theta||\theta+\delta)}{\delta}\right). \tag{A.3}$$

Combining this with the large deviation inequality (A.1) yields

$$\mathbf{P}\left(\sqrt{2T_1^n(Y, \theta+\delta)} \geq x\right) \geq 1 - \exp\left(n(D(\theta||\theta+\varepsilon) - \frac{\varepsilon}{\delta} D(\theta||\theta+\delta)) + \frac{\varepsilon\sqrt{2x}}{\delta}\right),$$

for all $0 \leq \varepsilon \leq \delta$. The case when $\delta < 0$ follows analogously. □

For Gaussian observations the result can be sharpened, as shown in the following lemma.

Lemma 35. *Let $Y_1, \dots, Y_n$ be i.i.d. random variables such that $Y_1 \sim \mathcal{N}(0,1)$ and let $x_+ = \max(0, x)$ for $x \in \mathbb{R}$. Then,*

$$\mathbf{P}\left(\sqrt{2T_1^n(Y,\delta)} \geq x\right) \geq 1 - \exp\left(-\frac{(\sqrt{n}\delta - x)_+^2}{2}\right). \tag{A.4}$$

Proof. Assume w.l.o.g. that $\delta > 0$ and observe that

$$\sqrt{2T_1^n(Y,\delta)} = \left|\sqrt{n}\,\overline{Y}_1^n - \sqrt{n}\delta\right| \geq \sqrt{n}\delta - \sqrt{n}\,\overline{Y}_1^n.$$

Since $\sqrt{n}\,\overline{Y}_1^n$ is standard normal distributed, we find for any $z > 0$

$$\mathbf{P}\left(\sqrt{n}\,\overline{Y}_1^n \geq z\right) \leq \exp(-z^2/2).$$

Therefore, we find

$$\mathbf{P}\left(\sqrt{n}\delta - \sqrt{n}\,\overline{Y}_1^n \geq x\right) \geq 1 - \mathbf{P}\left(\sqrt{n}\,\overline{Y}_1^n \geq \sqrt{n}\delta - x\right),$$

which proves the assertion. □

A.1.2 On the limit distribution M

In this section we collect some known facts and proof some new properties of the random variable M defined as in (3.3). These results will be employed frequently for the proofs of Section 3 but might also be of interest on its own. We will first fix some notations. Throughout this Section, let B denote the standard Brownian motion. For $0 \leq s < t \leq 1$ define the calibrated absolute increments of the Brownian motion as

$$\xi(s,t) = \frac{|B(t) - B(s)|}{\sqrt{t-s}} - \sqrt{2\log\frac{e}{t-s}}. \tag{A.5}$$

Thus, we find that

$$M = \sup_{0 \leq s < t \leq 1} \xi(s,t). \tag{A.6}$$

The first result gives a bound for the probability that these calibrated increments exceed a threshold value on multiple disjoint subintervals of the unit interval.

Theorem 36. *Let $k \in \mathbb{N}$ with $k \geq 1$ and $q(\alpha)$ be the $(1-\alpha)$-quantile of M. Then,*

$$\mathbf{P}\left(\min_{l=1,\dots,k} \xi(s_l, t_l) > q(\alpha) \text{ for some } 0 \leq s_1 < t_1 \leq s_2 < t_2 \leq \dots \leq s_k < t_k \leq 1\right) \leq \alpha^k. \tag{A.7}$$

Proof. For a fixed $q > 0$ we iteratively define the stopping times $\zeta_0(q), \zeta_1(q), \zeta_2(q), \ldots$ by

$$\begin{aligned}\zeta_0(q) &= 0,\\ \zeta_i(q) &= \inf\left\{t > 0 : \sup_{\zeta_{i-1}(q)\le s\le t} \xi(s,t) > q\right\} \quad \text{for} \quad i = 1, 2, \ldots\end{aligned}$$

From the strong Markov property of the Brownian motion, we obtain that the waiting times

$$\zeta_1(q), \zeta_2(q) - \zeta_1(q), \zeta_3(q) - \zeta_2(q), \ldots$$

are independent and identically distributed. Therefore, for any $x > 0$ and for any $k \ge 1$ it follows that

$$\begin{aligned}\mathbf{P}\left(\zeta_k(q) \le x\right) = \mathbf{P}\left(\sum_{l=1}^{k} \zeta_l(q) - \zeta_{l-1}(q) \le x\right) \le &\mathbf{P}\left(\zeta_l(q) - \zeta_{l-1}(q) \le x \ \forall l = 1, \ldots, k\right) \quad \text{(A.8)}\\ = &\mathbf{P}\left(\zeta_1(q) \le x\right)^k.\end{aligned}$$

Next, note that by definition $\zeta_1(q) \le 1$ implies that $M > q$. Therefore,

$$\mathbf{P}\left(\zeta_1(q(\alpha)) \le 1\right) \le \mathbf{P}(M > q(\alpha)) \le \alpha, \tag{A.9}$$

where $q(\alpha)$ denotes the $(1-\alpha)$-quantile of M. Combing the results from (A.8) and (A.9) leads to

$$\mathbf{P}\left(\zeta_k(q(\alpha)) \le 1\right) \le \alpha^k.$$

The proof is now completed by the observation that the l.h.s. in (A.7) already implies that $\zeta_k(q(\alpha)) \le 1$. □

It was shown in Dümbgen and Spokoiny (2001) that M is finite almost surely and in Dümbgen et al. (2006) that it has a continuous distribution which is supported on $[0, \infty)$. Moreover, we can prove the following lemma about the tails of M.

Theorem 37. *Let M be as in* (3.3)*. Then,*

(i) $\mathbf{E}[M] < \infty$,

(ii) $\mathbf{med}[M] < \infty$,

(iii) $\mathbf{P}(M \ge t) \le 2\exp(-(t - \mathbf{E}[M])^2/2)$ *for all* $t > \mathbf{E}[M]$,

(iv) $\mathbf{P}(M \ge t) \le \exp(-(t - \mathbf{med}[M])^2/2)$ *for all* $t > \mathbf{med}[M]$,

(v) $\mathbf{P}(M \ge t) \le 2\exp\left(-t^2/8\right)$ *for all* $t > 2\mathbf{E}[M]$.

The proof needs some preparation. It is essentially build on a corollary of *Borell's inequality* (Borell, 1975), see also van der Vaart and Wellner (1996).

Corollary 38. *Let $(X_t)_{t\in T}$ be a separable Gaussian process on T, such that*

$$\Psi := \sup_{t\in T} |X_t| < \infty \text{ a.s.} \quad \text{and} \quad \sup_{t\in T} \mathbf{Var}\left[X_t\right] \leq 1.$$

Then, $\mathbf{E}\left[\Psi\right] < \infty, \mathbf{med}\left[\Psi\right] < \infty$ and for all $t > 0$

$$\begin{aligned}\mathbf{P}\left(\Psi - \mathbf{E}\left[\Psi\right] > t\right) &\leq \exp(-t^2/2),\\ \mathbf{P}\left(\Psi - \mathbf{med}\left[\Psi\right] > t\right) &\leq 1/2\exp(-t^2/2).\end{aligned}$$

The main idea the proof follows van der Vaart and Wellner (1996) [Proposition A.2.1.] and at some points Adler and Taylor (2007). However, since some modifications are needed, we give the complete proof for the sake of clarity. It strongly relies on the following lemma (see van der Vaart and Wellner (1996) [Lemma A.2.2.]).

Lemma 39. *Let $Z \sim \mathcal{N}(0, I_d)$, where $I_d \in \mathbb{R}^{d\times d}$ denotes the d-dimensional unit matrix. Then, for every function $f : \mathbb{R}^d \to \mathbb{R}$ that is Lipschitz-continuous with constant 1, it holds for all $t > 0$ that*

$$\begin{aligned}\mathbf{P}\left(f(Z) - \mathbf{E}\left[Z\right] > t\right) &\leq \exp\left(-\frac{t^2}{2}\right),\\ \mathbf{P}\left(f(Z) - \mathbf{med}\left[Z\right] > t\right) &\leq \frac{1}{2}\exp\left(-\frac{t^2}{2}\right).\end{aligned}$$

Proof of Corollary 38. The proof is done in two steps: first the result is shown for a finite T and extended to separable spaces in the second step. We prove the result only for the version with the expected value, as the argumentation is the same for the median. Let us assume for the moment that T is finite with $|T| = n$ and covariance matrix

$$\Sigma = \{\sigma_{ij}\}_{1\leq i,j\leq n} \in \mathbb{R}^{n\times n}.$$

Then, there exists a matrix $Q \in \mathbb{R}^{n\times n}$, such that $Q^TQ = \Sigma$. Further, for $Z \sim \mathcal{N}(0, I_n)$ and $\mu = \mathbf{E}\left[(X_t)_{t\in T}\right] \in \mathbb{R}^n$ one finds that

$$(X_t)_{t\in T} \stackrel{\mathcal{D}}{=} QZ + \mu.$$

We define $f(x) = \|Qx\|_\infty$ and will show that f is Lipschitz-continuous with constant one, which is a consequence of the following observation. Let $e_i \in \mathbb{R}^n$ denote the i-th unit vector.

Then,

$$\|Qx\|_\infty \leq \max_{1\leq i\leq n} \|Qe_i\|_\infty \|x\|_\infty \leq \max_{1\leq i\leq n} \sqrt{e_i^T Q^T Q e_i}\, \|x\|_\infty = \max_{1\leq i\leq n} \sigma_{ii} \|x\|_\infty \leq \|x\|_\infty .$$

We have used the Cauchy-Schwarz inequality as well as $Q^TQ = \Sigma$ and $\sigma_{ii} \leq 1$ (by assumption). Therefore, we can now apply Lemma 39, which completes the proof for finite T.

We extend the result to general T by taking into account separability. To this end, let T_n be a sequence of finite subsets of T, such that $T_{n-1} \subset T_n$ and T_n increases to a subset which is dense in T. Then,

$$\Psi_n := \max_{t\in T_n} |X_t| \to \Psi \text{ a.s.}$$

From the monotonicity of the convergence, we also deduce

$$\lim_{n\to\infty} \mathbf{P}\left(\Psi_n > x\right) = \mathbf{P}\left(\Psi > x\right) \quad \text{and} \quad \lim_{n\to\infty} \mathbf{E}\left[\Psi_n\right] = \mathbf{E}\left[\Psi\right].$$

It remains to ensure that $\mathbf{E}[\Psi] < \infty$ to prove the assertion. To this end, let us assume that $\mathbf{E}[\Psi] = \infty$. We choose some $x_0 \in \mathbb{R}$ that satisfies

$$\exp\left(-\frac{x_0^2}{2}\right) \leq \frac{1}{2} \quad \text{and} \quad \mathbf{P}\left(\Psi < x_0\right) \geq \frac{3}{4}.$$

Since $\lim_{n\to\infty} \mathbf{E}[\Psi_n] = \mathbf{E}[\Psi] = \infty$ we can select $n_0 \in \mathbb{N}$ large enough that $\mathbf{E}[\Psi_n] > 2x_0$ for all $n \geq n_0$. Using the argumentation as for finite sets T, we obtain on the one hand

$$\mathbf{P}\left(\mathbf{E}\left[\Psi_{n_0}\right] - \Psi_{n_0} > x_0\right) \leq \exp\left(-\frac{x_0^2}{2}\right) \leq \frac{1}{2}. \tag{A.10}$$

On the other hand, since Ψ_n is monotonically increasing in n and $\mathbf{E}[\Psi_{n_0}] > 2x_0$, we find

$$\mathbf{P}\left(\mathbf{E}\left[\Psi_{n_0}\right] - \Psi_{n_0} > x_0\right) \geq \mathbf{P}\left(2x_0 - \Psi_{n_0} > x_0\right) = \mathbf{P}\left(\Psi_{n_0} < x_0\right) \geq \mathbf{P}\left(\Psi < x_0\right) \geq \frac{3}{4}.$$

Since this contradicts (A.10) to the latter inequality, we have shown that $\mathbf{E}[\Psi] < \infty$. The proof for $\Psi - \mathbf{med}[\Psi]$ is obtained analogously, for uniqueness of the median see e.g. van der Vaart and Wellner (1996). □

Proof of Theorem 37. After having established Corollary 38 the proof of (i)-(iv) is straightforward. Let $T = \{(t_0, t_1) : 0 \leq t_0 < t_1 \leq 1\}$ and define the Gaussian-processes

$$X^1_{(t_0,t_1)} = \frac{B_{t_1} - B_{t_0}}{\sqrt{t_1 - t_0}} - \sqrt{2\log\frac{e}{t_1 - t_0}} \quad \text{and} \quad X^2_{(t_0,t_1)} = -\frac{B_{t_1} - B_{t_0}}{\sqrt{t_1 - t_0}} - \sqrt{2\log\frac{e}{t_1 - t_0}}.$$

We then observe that

$$M = \sup_{(t_0,t_1)\in T} \frac{|B_{t_1} - B_{t_0}|}{\sqrt{t_1 - t_0}} - \sqrt{2\log\frac{e}{t_1 - t_0}} \leq \max\left(\sup_{(t_0,t_1)\in T} X^1_{(t_0,t_1)}, \sup_{(t_0,t_1)\in T} X^2_{(t_0,t_1)}\right).$$

Then, the assertions (i)-(iv) follow by applying Corollary 38 to X^1 and X^2 together with the observation

$$\begin{aligned}\mathbf{E}[M] &> \mathbf{E}\left[\sup_{(t_0,t_1)\in T} X^1_{(t_0,t_1)}\right] = \mathbf{E}\left[\sup_{(t_0,t_1)\in T} X^2_{(t_0,t_1)}\right]\\ \mathbf{med}[M] &> \mathbf{med}\left[\sup_{(t_0,t_1)\in T} X^1_t\right] = \mathbf{med}\left[\sup_{(t_0,t_1)\in T} X^2_{(t_0,t_1)}\right].\end{aligned}$$

For (v) note that from $t > 2\mathbf{E}[M]$ we find that $t/2 + \mathbf{E}[M] < t$ and hence (v) follows from (iii). □

A.2 Proofs of Section 3

In this section we collect the proofs of Section 3. We begin with results on the asymptotic null-distribution.

A.2.1 Proof of Section 3.1

We will assume for now that $Y = (Y_1, \dots, Y_n)$ are independent and identically distributed random variables with $Y_1 \sim F_\theta$ and $\theta \in \Theta$, i.e. we consider the situation of no change-point. Without loss of generality we will assume that $m(\theta) = \dot{\psi}(\theta) = 0$ and $v(\theta) = \ddot{\psi}(\theta) = 1$. Moreover, assume that $(c_n)_{n\in\mathbb{N}}$ satisfies (3.2) and introduce the notation

$$\mathcal{I}(c_n) = \{(i,j) : j - i + 1 \geq c_n n\}.$$

We will show that in this scenario $T_n(Y, \vartheta, c_n) \xrightarrow{\mathcal{D}} M$. The proof is divided into several steps. First, we use Taylor expansions and strong approximation results to approximate the local likelihood-ratios uniformly by a function of Gaussian partial sums (Proposition 42). This function is then shown to converge to the random variable M weakly, which completes the proof for signals without change-point. The actual assertion is then derived at the end of this section.

Lemma 40.

$$\max_{(i,j)\in\mathcal{I}(c_n)} \left| \sqrt{2T_i^j(Y,\theta)} - \sqrt{j-i+1}\left|\overline{Y}_i^j\right| \right| = o_{\mathbf{P}}(1)$$

Proof. Set $\xi = m^{-1}$ and note that ξ is strictly increasing. Since Θ is open, there exists for each given $\delta' > 0$ a $\delta > 0$ such that $\xi(B_\delta(0)) \subset B_{\delta'}(\theta) \subset \Theta$. Next define the random variable

$$L_n = \max_{1\le i<j\le n} \left|\overline{Y}_i^j\right| \sqrt{j-i+1}.$$

Then, it follows from Shao's Theorem (Shao, 1995) that $L_n/\sqrt{\log n}$ converges a.s. to some finite constant and we hence find that

$$\max_{(i,j)\in\mathcal{I}(c_n)} \left|\overline{Y}_i^j\right| \le \sqrt{\frac{\log n}{nc_n}} \frac{L_n}{\sqrt{\log n}} \to 0 \quad \text{a.s.}$$

Thus, for each $\varepsilon > 0$ there exists an index $n_0 = n_0(\varepsilon) \in \mathbb{N}$ such that for all $n \ge n_0$

$$\mathbf{P}\left(\max_{(i,j)\in\mathcal{I}(c_n)} \left|\overline{Y}_i^j\right| \ge \delta \right) \le \varepsilon.$$

In other words, $\xi(\overline{Y}_i^j) \in B_\delta(\theta)$ uniformly over $\mathcal{I}(c_n)$ with probability not less than $1-\varepsilon$. Note that $\phi(\overline{Y}_i^j) = \max_{\theta\in\Theta} \theta\overline{Y}_i^j - \psi(\theta) = \xi(\overline{Y}_i^j)\overline{Y}_i^j - \psi(\xi(\overline{Y}_i^j))$ which in turn implies that

$$J(\overline{Y}_i^j,\theta) = \phi(\overline{Y}_i^j) - \theta\overline{Y}_i^j + \psi(\theta) = (\xi(\overline{Y}_i^j) - \theta)\overline{Y}_i^j - (\psi(\xi(\overline{Y}_i^j)) - \psi(\theta)).$$

Taylor expansion of ψ around θ gives (recall that $m(\theta) = \dot{\psi}(\theta) = 0$ and $v(\theta) = \ddot{\psi}(\theta) = 1$)

$$\psi(\xi(\overline{Y}_i^j)) - \psi(\theta) = \frac{1}{2}(\xi(\overline{Y}_i^j) - \theta)^2 + \frac{1}{6}\dddot{\psi}(\tilde{\theta})(\xi(\overline{Y}_i^j) - \theta)^3$$

for some $\tilde{\theta} \in B_\varepsilon(\theta)$. This in turn implies

$$J(\overline{Y}_i^j,\theta) = (\xi(\overline{Y}_i^j) - \theta)(\overline{Y}_i^j) - \frac{1}{2}(\xi(\overline{Y}_i^j) - \theta)^2 - \frac{1}{6}\dddot{\psi}(\tilde{\theta})(\xi(\overline{Y}_i^j) - \theta)^3.$$

Again, Taylor expansion of $\xi = m^{-1}$ around 0 shows

$$\xi(\overline{Y}_i^j) - \theta = \overline{Y}_i^j - \frac{\dddot{\psi}(\tilde{\theta})}{2(v(\tilde{\theta}))^2}(\overline{Y}_i^j)^2$$

for some $\tilde{\theta} \in B_{\delta'}(\theta)$. This finally yields

$$2T_i^j(Y,\theta) = (j-i+1)J(\overline{Y}_i^j,\theta) = (j-i+1)(\overline{Y}_i^j)^2 + (j-i+1)r_n(\overline{Y}_i^j)$$

where r_n is such that $\left|r_n(\overline{Y}_i^j)\right| \le C^2(\overline{Y}_i^j)^3$ for a constant $C = C(\delta') > 0$ (independent of ε, i

and j) and for all $n \geq n_0$. It thus holds with probability not less than $1-\varepsilon$ that

$$\begin{aligned}\max_{(i,j)\in\mathcal{I}(c_n)} \left|\sqrt{2T_i^j(Y,\theta^*)} - \sqrt{j-i+1}\left|\overline{Y}_i^j\right|\right| \leq & C \max_{(i,j)\in\mathcal{I}(c_n)} \left|(j-i+1)\left(\overline{Y}_i^j\right)^3\right|^{1/2} \\ = & C \max_{(i,j)\in\mathcal{I}(c_n)} \left|\frac{\sum_{l=i}^j Y_l}{\sqrt{j-i+1}}(j-i+1)^{-1/6}\right|^{3/2} \\ \leq & C\left(\frac{L_n}{\sqrt{\log n}}\right)^{3/2} \sqrt[4]{\frac{\log^3 n}{nc_n}}.\end{aligned}$$

From Shao's Theorem it follows that the r.h.s. vanishes almost surely as $n \to \infty$. □

We proceed with some strong approximation results for $\overline{Y}_i^j$, which is due to Komlós et al. (1976).

Lemma 41. *There exist i.i.d standard normally distributed r.v. $Z_1, \ldots, Z_n$ such that*

$$\lim_{n\to\infty} \sqrt{\log n} \max_{(i,j)\in\mathcal{I}(c_n)} \left(\sqrt{j-i+1}\left|\left|\overline{Y}_i^j\right| - \left|\overline{Z}_i^j\right|\right|\right) = 0 \quad a.s.$$

Proof. We define the partial sums $S_0^Y = 0$ and $S_l^Y = Y_1 + \ldots + Y_l$ and find that $(j-i+1)\left|\overline{Y}_i^j\right| = \left|S_j^Y - S_{i-1}^Y\right|$. Analogously we define S_l^Z. Now let $(i,j) \in \mathcal{I}(c_n)$ and observe that

$$\left|\frac{\left|S_j^Y - S_{i-1}^Y\right|}{\sqrt{j-i+1}} - \frac{\left|S_j^Z - S_{i-1}^Z\right|}{\sqrt{j-i+1}}\right| \leq \frac{\left|S_j^Y - S_j^Z\right|}{\sqrt{nc_n}} + \frac{\left|S_i^Y - S_i^Z\right|}{\sqrt{nc_n}} \leq 2 \max_{0\leq l\leq n} \frac{\left|S_l^Y - S_l^Z\right|}{\sqrt{nc_n}}.$$

It follows from the *KMT inequality* in Komlós et al. (1976)[Thm. 1] and (3.2) that

$$\sqrt{\log n} \max_{(i,j)\in\mathcal{I}(c_n)} \left(\sqrt{j-i+1}\left|\left|\overline{Y}_i^j\right| - \left|\overline{Z}_i^j\right|\right|\right) = \sqrt{\log n} \max_{0\leq l\leq n} \frac{\left|S_l^Y - S_l^Z\right|}{\sqrt{nc_n}} \leq \mathrm{o}(1) \quad \text{a.s.}$$

□

By combining Lemma 41 and 40 we obtain

Proposition 42. *There exist i.i.d standard normally distributed r.v. $Z_1, \ldots, Z_n$ such that*

$$\max_{(i,j)\in\mathcal{I}(c_n)} \left|\sqrt{2T_i^j(Y,\theta)} - \sqrt{j-i+1}\left|\overline{Z}_i^j\right|\right| = \mathrm{o}_{\mathbf{P}}(1).$$

Lemma 43. *For $n \in \mathbb{N}$, define the continuous functionals $h, h_n : \mathcal{C}([0,1]) \to \mathbb{R}$ by*

$$h(x,c) = \sup_{\substack{0\le s<t\le 1\\ t-s\ge c}} \left(\frac{|x(t)-x(s)|}{\sqrt{t-s}} - \sqrt{2\log\frac{e}{t-s}} \right) \quad \text{and}$$

$$h_n(x,c) = \max_{\substack{1\le i<j\le n\\ (j-i+1)/n\ge c}} \left(\frac{|x(j/n)-x(i/n)|}{\sqrt{(j-i+1)/n}} - \sqrt{2\log\frac{en}{j-i+1}} \right),$$

respectively. Moreover assume that $\{x_n\}_{n\in\mathbb{N}} \subset \mathcal{C}([0,1])$ is such that $x_n \to x$ for some $x \in \mathcal{C}([0,1])$. Then $h_n(x_n,c) \to h(x,c)$.

Proof. Let $\delta > 0$. Then there exists an index $n_0 \in \mathbb{N}$ such that $|x_n(t) - x(t)| \le \delta$ for all $n \ge n_0$ and $t \in [0,1]$. Thus, it follows directly from the definition that $h_n(x) = h_n(x_n) + \mathcal{O}(\delta)$ for $n \ge n_0$. Since $u \mapsto \sqrt{2\log e/u}$ is uniformly continuous on $[c,1]$ we consequently have that $h_n(x) \to h(x)$ as $n \to \infty$ and the assertion follows. □

Before we proceed, recall the definition of M in (3.3). Moreover, we introduce for $0 < c \le 1$ the statistic

$$M(c) := \sup_{\substack{0\le s<t\le 1\\ t-s>c}} \left(\frac{|B(t)-B(s)|}{\sqrt{t-s}} - \sqrt{2\log\frac{e}{t-s}} \right). \tag{A.11}$$

From (Dümbgen and Spokoiny, 2001, Thm. 6.1) (and the subsequent Remark 1) it can be seen that $M(c)$ converges weakly to M as $c \to 0^+$.

We continue by showing the convergence of the multiscale statistic with a constant lower bound on the interval length. More precisely we consider the statistic

$$T_n^c(Y,\theta) = \max_{(i,j)\in\mathcal{I}(c)} \left(\sqrt{2T_i^j(Y,\theta)} - \sqrt{2\log\frac{en}{j-i+1}} \right). \tag{A.12}$$

Proposition 44. *For $c > 0$ and the statistic T_n^c, as in* (A.12)*, we find that*

$$\lim_{c\to 0^+} \lim_{n\to\infty} T_n^c(Y,\theta) = M,$$

weakly.

Proof. Let S^Z be the partial sums of Z as in Lemma (41) and let $\{X_n(t)\}_{t\ge 0}$ be the process that is linear on the intervals $[i/n,(i+1)/n]$ with values $X_n(i/n) = S_i^Z/\sqrt{n}$. We obtain from Donsker's Theorem that $X_n \xrightarrow{\mathcal{D}} B$. Now, recall the definition of h and h_n in Lemma 43 and observe that

$$h_n(X_n,c) = \max_{(i,j)\in\mathcal{I}(c)} \left(\sqrt{j-i+1}|\overline{Z}_i^j| - \sqrt{2\log\frac{en}{j-i+1}} \right).$$

It hence follows from Proposition 42 that

$$|T_n^c(Y,\theta) - h_n(X_n,c)| \le \max_{(i,j)\in\mathcal{I}(c)} \left| \sqrt{2T_i^j(Y,\theta)} - \sqrt{j-i+1}\,|\overline{Z}_i^j| \right| = \mathrm{o}_{\mathbf{P}}(1). \tag{A.13}$$

Since $X_n \xrightarrow{\mathcal{D}} B$, Lemma 43 and (Billingsley, 1968, Thm. 5.5) imply that

$$h_n(X_n,c) \xrightarrow{\mathcal{D}} h(B,c) \stackrel{\mathcal{D}}{=} M(c).$$

Together with (A.13) one hence finds that for all $c > 0$

$$T_n^c(Y,\theta) \xrightarrow{\mathcal{D}} h(B,c) = M(c) \quad \text{as} \quad n \to \infty.$$

Thus, the assertion finally follows, since $M(c) \to M$ weakly as $c \to 0^+$. □

Theorem 45. *Let $\vartheta \equiv \theta$ and recall from the definition of T_n that*

$$T_n(Y,\vartheta,c_n) = \max_{(i,j)\in\mathcal{I}(c_n)} \left(\sqrt{2T_i^j(Y,\theta)} - \sqrt{2\log\frac{en}{j-i+1}} \right).$$

Then, $T_n(Y,\vartheta,c_n) \to M$ weakly as $n \to \infty$.

Proof. First observe that according to Proposition 42 we have for all $t > 0$ that

$$\mathbf{P}\left(T_n(Y,\vartheta;c_n) \le t\right) = \mathbf{P}\left(\max_{(i,j)\in\mathcal{I}(c_n)} \left(\sqrt{j-i+1}\,|\overline{Z}_i^j| - \sqrt{2\log\frac{en}{j-i+1}} \right) \le t \right) + \mathrm{o}(1).$$

Since furthermore

$$\begin{aligned} &\mathbf{P}\left(\max_{(i,j)\in\mathcal{I}(c_n)} \left(\sqrt{j-i+1}\,|\overline{Z}_i^j| - \sqrt{2\log\frac{en}{j-i+1}} \right) \le t \right) \\ \ge &\mathbf{P}\left(\sup_{0\le s<t\le 1} \left(\frac{|B(t)-B(s)|}{\sqrt{t-s}} - \sqrt{2\log\frac{e}{t-s}} \right) \le t \right), \end{aligned}$$

this shows that for all $t > 0$

$$\liminf_{n\to\infty} \mathbf{P}(T_n(Y,\vartheta,c_n) \le t) \ge \mathbf{P}(M \le t).$$

Now let $c > 0$ be fixed and assume w.l.o.g. $c_n < c$ for all $n \in \mathbb{N}$. With T_n^c as defined in Proposition 44 we conversely find

$$\limsup_{n\to\infty} \mathbf{P}(T_n(Y,\vartheta,c_n) \le t) \le \limsup_{n\to\infty} \mathbf{P}(T_n^c(Y,\vartheta,c_n) \le t) = \mathbf{P}(M(c) \le t).$$

Hence, the assertion follows from Proposition 44 with $c \to 0^+$ and the fact that $M > 0$ a.s. □

Proof of Theorem 3. Let $T_n(Y, \vartheta; c_n)$ be defined as in (2.12). From Theorem 45 it then follows that

$$T_n(Y, \vartheta; c_n) \xrightarrow{\mathcal{D}} \max_{0 \le k \le K} \sup_{\tau_k \le s < t \le \tau_{k+1}} \left(\frac{|B(t) - B(s)|}{\sqrt{t-s}} - \sqrt{2 \log \frac{e}{t-s}} \right).$$

The limiting statistic on the right hand side is stochastically bounded from above by M, since the maximum is taken over a smaller set. Conversely, we observe by the scaling property of the Brownian motion, its stationarity and by choosing $\tilde{s} = s/(\tau_{k+1} - \tau_k)$ and $\tilde{t} = t/(\tau_{k+1} - \tau_k)$

$$\begin{aligned} &\sup_{\tau_k \le s < t \le \tau_{k+1}} \left(\frac{|B(t) - B(s)|}{\sqrt{t-s}} - \sqrt{2 \log \frac{e}{t-s}} \right) \\ &\overset{\mathcal{D}}{=} \sup_{0 \le \tilde{s} < \tilde{t} \le 1} \left(\frac{\left|B(\tilde{t}) - B(\tilde{s})\right|}{\sqrt{\tilde{t} - \tilde{s}}} - \sqrt{2 \log \frac{e}{\tilde{t} - \tilde{s}} + 2 \log \frac{1}{\tau_{k+1} - \tau_k}} \right) \overset{\mathcal{D}}{\ge} M - \sqrt{2 \log \frac{1}{\tau_{k+1} - \tau_k}}. \end{aligned}$$

□

A.2.2 Proofs of Section 3.2

We now give the proofs for the bounds for the probability of overestimation. These essentially rely on the results in Section 3.1 and Section A.1.2.

Proof of Theorem 5. We first note that it suffices to give the proof for constant $\vartheta \equiv \theta_0$, i.e. $K = 0$, which will ease notation. Second, observe that $\hat{K}(q(\alpha)) > K + 2k$ implies that the multiscale constraint for true regression function ϑ is violated on at least k disjoint intervals. This amounts to say that for k disjoint intervals $[i_1/n, j_1/n], \ldots, [i_k/n, j_k/n] \subset [0, 1]$ it holds that

$$\sqrt{2T_{i_s}^{j_s}(Y, \theta_0)} - \sqrt{2 \log \frac{en}{j_s - i_s + 1}} \ge q(\alpha) \quad \text{for all } 1 \le s \le k.$$

As a consequence of Proposition 42 we find that there exist i.i.d. standard normally distributed random variables $Z_1, \ldots, Z_n$ so that

$$\max_{(i,j) \in \mathcal{I}(c_n)} \left| \sqrt{2T_i^j(Y, \theta_0)} - \sqrt{j - i + 1} \left| \overline{Z}_i^j \right| \right| = o_{\mathbf{P}}(1).$$

As before, we set

$$\mathcal{I}(c_n) = \{(i, j) : 1 \le i \le j \le n \quad \text{and} \quad j - i + 1 \ge c_n n\}$$

and moreover define

$$\mathcal{D}_k := \left\{ ((i_1, j_1), \ldots, (i_k, j_k)) \in (\mathcal{I}(c_n))^k : 1 \le i_1 < j_1 < \ldots < i_k < j_k \le n \right\}.$$

Next we observe that

$$\begin{aligned}
&\lim_{n\to\infty} \mathbf{P}\left(\exists((i_1,j_1),\dots,(i_k,j_k)) \in \mathcal{D}_k : \min_{1\le s\le k} \sqrt{2T_{i_s}^{j_s}(Y,\theta_0)} - \sqrt{2\log\frac{en}{j_s-i_s+1}} \ge q(\alpha)\right)\\
&= \lim_{n\to\infty} \mathbf{P}\left(\exists((i_1,j_1),\dots,(i_k,j_k)) \in \mathcal{D}_k : \min_{1\le s\le k} \sqrt{j_s-i_s+1}\left|\overline{Z}_{i_s}^{j_s}\right| - \sqrt{2\log\frac{en}{j_s-i_s+1}} \ge q(\alpha)\right)\\
&= \lim_{n\to\infty} \mathbf{P}\left(\exists((i_1,j_1),\dots,(i_k,j_k)) \in \mathcal{D}_k : \min_{1\le s\le k} \frac{\left|B(\frac{i_s}{n}) - B(\frac{j_s}{n})\right|}{\sqrt{j_s-i_s+1}} - \sqrt{2\log\frac{en}{j_s-i_s+1}} \ge q(\alpha)\right)\\
&\le \alpha^{k+1}.
\end{aligned}$$

Here the last inequality follows from Theorem 36. □

With Theorem 5 we can proof Corollary 6.

Proof of Corollary 6. For the proof we will use that for a random variable supported on $\mathbb{N}_0$ it holds that

$$\mathbf{E}[X] = \sum_{i=0}^{\infty} \mathbf{P}(X > i).$$

Together with Theorem 5 this shows that

$$\begin{aligned}
\lim_{n\to\infty} \mathbf{E}\left[(\hat{K}(q(\alpha)) - K)_+\right] &= \lim_{n\to\infty} \sum_{s=0}^{\infty} \mathbf{P}(\hat{K}(q(\alpha)) - K > s)\\
\le \lim_{n\to\infty} 2\sum_{s=0}^{\infty} \mathbf{P}(\hat{K}(q(\alpha)) - K > 2s) &\le 2\sum_{s=0}^{\infty} \alpha^{s+1} = \frac{2\alpha}{1-\alpha},
\end{aligned}$$

which completes the proof. □

A.2.3 Proofs of Section 3.3

In this section we prove the bounds for the probability of underestimation. We begin with the result for Gaussian observations (Theorem 14) and then turn to the general case (Theorem 7). This eases presentation, since the idea of both proofs is the same, but the Gaussian case requires less technicalities.

Proof of Theorem 14. For the proof we define for $k = 1,\dots,K$ the pairwise disjoint intervals

$$I_k = \left(\frac{\tau_{k-1}+\tau_k}{2}, \frac{\tau_k+\tau_{k+1}}{2}\right]. \tag{A.14}$$

Recall that the value of μ on the segment I_k is denoted by $\mathfrak{m}_k$. Let $\mathfrak{m}_k^+ = \max\{\mathfrak{m}_k, \mathfrak{m}_{k+1}\}$,

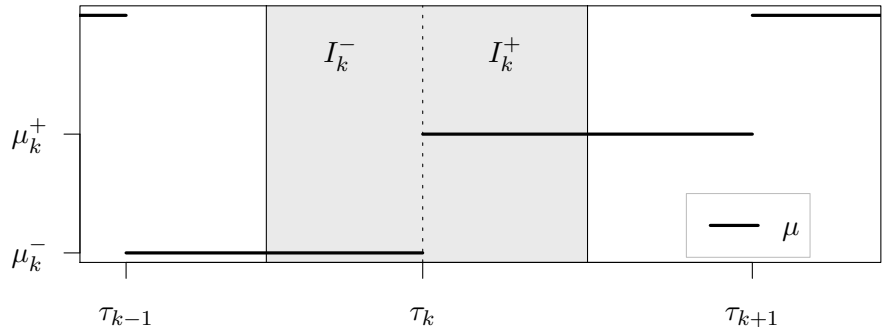


Figure 34: Illustration of I_k^-, I_k^+, $\mathfrak{m}_k^-$ and $\mathfrak{m}_k^+$ as in (A.14) and (A.15).

$\mathfrak{m}_k^- = \min\{\mathfrak{m}_k, \mathfrak{m}_{k+1}\}$ and split each interval I_k accordingly, i.e.

$$I_k^+ = \{t \in I_k : \mu(t) = \mathfrak{m}_k^+\} \quad \text{and} \quad I_k^- = \{t \in I_k : \mu(t) = \mathfrak{m}_k^-\}. \tag{A.15}$$

Clearly, it holds that $I_k = I_k^+ \cup I_k^-$. We define the event that a function exists, which is constant on I_k and fulfills the multiscale constraint on I_k^+ and I_k^-, i.e.

$$\Omega_k = \left\{ \exists \hat{\mathfrak{m}} \in \mathbb{R} : \sqrt{2T_{I_k^+}(Y, \hat{\mathfrak{m}})} - \sqrt{2\log\frac{en}{\#I_k^+}} \le q \text{ and } \sqrt{2T_{I_k^-}(Y, \hat{\mathfrak{m}})} - \sqrt{2\log\frac{en}{\#I_k^-}} \le q \right\}.$$

Here $\#I_k$ denotes the number of observations in the interval I_k. We proceed by computing an upper bounds for $\mathbf{P}(\Omega_k)$. To this end, observe that either $\hat{\mathfrak{m}} \le \mathfrak{m}_k^+ - \delta_k/2$ or $\hat{\mathfrak{m}} \ge \mathfrak{m}_k^- + \delta_k/2$. Following this idea we define

$$\begin{aligned} \Omega_k^+ &= \left\{ \exists \hat{\mathfrak{m}} \le \mathfrak{m}_k^+ - \delta_k/2 : \sqrt{2T_{I_k^+}(Y, \hat{\mathfrak{m}}_0)} - \sqrt{2\log\frac{en}{\#I_k^+}} \le q \right\} \quad \text{and} \\ \Omega_k^- &= \left\{ \exists \hat{\mathfrak{m}} \ge \mathfrak{m}_k^- + \delta_k/2 : \sqrt{2T_{I_k^-}(Y, \hat{\mathfrak{m}}_0)} - \sqrt{2\log\frac{en}{\#I_k^-}} \le q \right\}. \end{aligned} \tag{A.16}$$

Next, observe that $\mathbf{P}(\Omega_k) \le 1 - (1 - \mathbf{P}(\Omega_k^+))(1 - \mathbf{P}(\Omega_k^-))$, due to independence of Ω_k^- and Ω_k^+ and the fact that $\Omega_k \subset \{\Omega_k^- \cup \Omega_k^+\}$. In other words, the event Ω_k implies either Ω_k^- or Ω_k^+. We proof an upper bound for $\mathbf{P}(\Omega_k^-)$ only, the same bound can be obtained for $\mathbf{P}(\Omega_k^+)$ by symmetry arguments.

Recall that $x \mapsto T_{I_k^-}(Y, x)$ is convex with global minimum at $\overline{Y}_{I_k^-}$. Thus, for all $\hat{\mathfrak{m}} \ge \mathfrak{m}_k^- + \delta_k/2$ one obtains

$$T_{I_k^-}(Y, \hat{\mathfrak{m}}) \ge T_{I_k^-}(Y, \mathfrak{m}_k^- + \delta_k/2)$$

whenever $\overline{Y}_{I_k^-} \leq \mathfrak{m}_k^- + \delta_k/2$. This yields

$$\begin{aligned}
\mathbf{P}\left(\Omega_k^-\right) &\leq \mathbf{P}\left(\Omega_k^- \cap \left\{\overline{Y}_{I_k^-} \leq \mathfrak{m}_k^- + \frac{\delta_k}{2}\right\}\right) + \mathbf{P}\left(\overline{Y}_{I_k^-} > \mathfrak{m}_k^- + \frac{\delta_k}{2}\right) \\
&\leq \mathbf{P}\left(\sqrt{2T_{I_k^-}\left(Y, \mathfrak{m}_k^- + \frac{\delta_k}{2}\right)} \leq \left(q + \sqrt{2\log(e/\lambda_k)}\right)\right) + \mathbf{P}\left(\overline{Y}_{I_k^-} > \mathfrak{m}_k^- + \frac{\delta_k}{2}\right) \\
&\leq \exp\left(-\frac{\left(\sqrt{n\lambda_k}\delta_k - 2q - \sqrt{8\log\frac{e}{\lambda_k}}\right)^2}{8}\right) + \exp\left(-\frac{n\lambda_k\delta_k^2}{8}\right),
\end{aligned}$$

where the last inequality stems from Lemma 33 and Lemma 35. Hence,

$$\begin{aligned}
\mathbf{P}(\Omega_k) \leq& 1 - (1 - \mathbf{P}(\Omega_k^+))(1 - \mathbf{P}(\Omega_k^-)) \\
\leq& 1 - \left(1 - \exp\left(-\frac{\left(\sqrt{n\lambda_k}\delta_k - 2q - \sqrt{8\log\frac{e}{\lambda_k}}\right)^2}{8}\right) - \exp\left(-\frac{n\lambda_k\delta_k^2}{8}\right)\right)^2 \\
=& 1 - \beta_{nk}(q). \qquad \text{(A.17)}
\end{aligned}$$

Next, for $k = 1, \ldots, K$ we define the random variables

$$Z_k(\omega) = \begin{cases} 0 & \text{if} \quad \omega \in \Omega_k \quad \text{and} \\ 1 & \text{otherwise.} \end{cases}$$

Observe that $Z_k = 1$ implies that any function $\hat{\mu} \in \mathcal{S}$ with $T_n(Y, \hat{\mu}) \leq q$ has a least one change-point on the interval I_k. Since the intervals $I_1, \ldots, I_K$ are pairwise disjoint, this yields $\hat{K}(q) \geq \sum_{k=1}^K Z_k$. Therefore, we find

$$\mathbf{P}\left(\hat{K}(q) \geq K\right) \geq \mathbf{P}\left(\sum_{k=1}^K Z_k \geq K\right) = \prod_{k=1}^K (1 - \mathbf{P}(\Omega_k)) = \prod_{k=1}^K \beta_{nk}(q),$$

which completes the proof of the first part. As a result of (A.17), Z_k can be bounded in probability by a Bernoulli random variable with success probability β_{nk}. Therefore,

$$\mathbf{E}\left[\hat{K}(q)\right] \geq \mathbf{E}\left[\sum_{k=1}^K Z_k\right] \geq \sum_{k=1}^K \beta_{nk}(q)$$

and hence

$$\mathbf{E}\left[\left(K-\hat{K}(q)\right)_+\right] \leq K - \sum_{k=1}^{K} \beta_{nk}(q) = \sum_{k=1}^{K}(1-\beta_{nk}(q)).$$

□

We now turn to the proof of Theorem 7, which follows the proof Theorem 14. But here we will employ a general large deviation results for exponential families (Lemma 34), instead of Lemma 35.

Proof of Theorem 7. First, let I_k, I_k^- and I_k^+ as in (A.14) and (A.15) and define θ_k^- and θ_k^+ accordingly. We again consider the events

$$\Omega_k = \left\{\exists\hat{\theta}\in\Theta : \sqrt{2T_{I_k^+}(Y,\hat{\theta})} - \sqrt{2\log\frac{en}{\#I_k^+}} \leq q \text{ and } \sqrt{2T_{I_k^-}(Y,\hat{\theta})} - \sqrt{2\log\frac{en}{\#I_k^-}} \leq q\right\},$$

Ω_k^- and Ω_k^+ analog to (A.16). We provide an upper bound for $\mathbf{P}\left(\Omega_k^-\right)$ and $\mathbf{P}\left(\Omega_k^+\right)$. Again, we only show the proof for $\mathbf{P}\left(\Omega_k^-\right)$, since the bound for $\mathbf{P}\left(\Omega_k^+\right)$ follows by symmetry. To this end, we find from Lemma 34 and (A.1)

$$\begin{aligned}
&\mathbf{P}\left(\Omega_k^-\right)\\
\leq&\mathbf{P}\left(\Omega_k^- \cap \left\{\overline{Y}_{I_k^-} \leq m^{-1}(\theta_k^-) + \frac{\delta_k}{2}\right\}\right) + \mathbf{P}\left(\overline{Y}_{I_k^-} > m^{-1}(\theta_k^-) + \frac{\delta_k}{2}\right)\\
\leq&\mathbf{P}\left(\sqrt{2T_{I_k^-}\left(Y,\theta_k^- + \frac{\delta_k}{2}\right)} \leq \left(q + \sqrt{2\log(e/\lambda_k)}\right)\right) + \mathbf{P}\left(\overline{Y}_{I_k^-} > m^{-1}(\theta_k^-) + \frac{\delta_k}{2}\right)\\
\leq&\exp\left(\lambda_k n \inf_{\varepsilon\in[0,\delta_k/2]}\left(D(\theta_k^-||\theta_k^-+\varepsilon) - \frac{2\varepsilon}{\delta_k}D(\theta_k^-||\theta_k^-+\delta_k/2) + \frac{\varepsilon\left(q+\sqrt{2\log(e/\lambda_k)}\right)^2}{\delta_k\lambda_k n}\right)\right)\\
&+\exp\left(-\lambda_k n D(\theta_k^- + \delta_k/2||\theta_k^-)\right).
\end{aligned}$$

From $\mathbf{P}(\Omega_k) \leq 1-(1-\mathbf{P}(\Omega_k^+))(1-\mathbf{P}(\Omega_k^-))$ and the definitions of κ_k^1 and κ_k^2 in (3.14) we then find

$$\mathbf{P}(\Omega_k) \leq 1-(1-e^{n\lambda_k\kappa_k^1}-e^{n\lambda_k\kappa_k^1})^2 = 1-\beta_{nk}(q). \tag{A.18}$$

With this inequality, the rest of the proof is identical to the proof of Theorem 14. □

Proof of Lemma 8. First observe from (3.11) that for any $\theta\in\Theta$ and $\varepsilon>0$ such that $\theta+\varepsilon\in\Theta$

one has $D(\theta||\theta+\varepsilon) = \int_\theta^{\theta+\varepsilon}(\theta+\varepsilon-t)v(t)\,\mathrm{d}t$. Thus, it follows that for all $0 \le \varepsilon \le x$

$$\begin{aligned}\frac{\varepsilon}{x}D(\theta||\theta+x) - D(\theta||\theta+\varepsilon) &= \frac{\varepsilon}{x}\int_\theta^{\theta+x}(\theta+x-t)v(t)\,\mathrm{d}t - \int_\theta^{\theta+\varepsilon}(\theta+\varepsilon-t)v(t)\,\mathrm{d}t \\ &\ge \frac{\varepsilon x}{2}\inf_{t\in[\theta,\theta+x]}v(t) - \frac{\varepsilon^2}{2}\sup_{t\in[\theta,\theta+x]}v(t).\end{aligned}$$

Maximizing over $0 \le \varepsilon \le x$ then yields

$$\sup_{\varepsilon\in[0,x]}\frac{\varepsilon}{x}D(\theta||\theta+x) - D(\theta||\theta+\varepsilon) \ge \frac{x^2}{8}\frac{\inf_{t\in[\theta,\theta+x]}v(t)^2}{\sup_{t\in[\theta,\theta+x]}v(t)}.$$

This proves that

$$\kappa_1^+(v,w,x,y) \ge \frac{x^2}{8}\frac{\inf_{v\le t\le w}v(t)^2}{\sup_{v\le t\le w}v(t)} - y.$$

Likewise, one finds

$$\kappa_2^+(v,w,x) \ge \frac{x^2}{2}\inf_{v\le t\le w}v(t).$$

The estimates for κ_1^- and κ_2^- are derived analogously. □

A.2.4 Proofs of Section 3.4

Proof of Corollary 9 and Corollary 15. First recall, that $\vartheta \in \mathcal{S}$ is fixed and therefore K, Λ and Δ are constant. From (3.20) we find that there exists a constant $C < \infty$, so that

$$\mathbf{P}\left(\hat{K}(q) < K\right) \le 2Ke^{-Cn\Lambda\Delta^2}\left[e^{\left(q+\sqrt{2\log(2e/\Lambda)}\right)^2}+1\right]. \tag{A.19}$$

On the other hand, Corollary 13 combined with Corollary 4 yields for sufficiently large values of q_n that

$$\lim_{n\to\infty}\mathbf{P}\left(\hat{K}(q_n) > K\right) \le 2e^{-q_n^2/8}. \tag{A.20}$$

Therefore, a sufficient condition for $\mathbf{P}\left(\hat{K}(q_n) = K\right) \to 1$, is that the r.h.s. in (A.19) and (A.20) are converging to zero. It is clear that this is true, whenever $q_n/\sqrt{n} \to 0$ and $q_n \to \infty$, which proves Corollary 9.

We now proof the almost sure statement in Corollary 15, i.e. we consider Gaussian observations. Note that in this case, inequality (A.20) holds for finite n. We employ the Borel-Cantelli-Lemma. Let $0 < \zeta < 0.5$ assume that $q_n/\sqrt{\log n} \to \infty$ and $q_n n^{-\zeta} \to 0$ holds. First,

we show that $q_n n^{-\zeta} \to 0$ implies that the r.h.s. in (A.19) is summable. To this end, observe

$$\begin{aligned}&\exp\left[-Cn\Lambda\Delta^2 + \left(q_n + \sqrt{2\log(2e/\Lambda)}\right)^2\right]\\ =&\exp\left[-n^{2\zeta}\left(Cn^{1-2\zeta}\Lambda\Delta^2 - \left(\frac{q_n}{n^\zeta} + \frac{\sqrt{2\log(2e/\Lambda)}}{n^\zeta}\right)^2\right)\right].\end{aligned}$$

Since $1 - 2\zeta > 0$ and $qn^{-\zeta} \to 0$ as $n \to \infty$, the latter expression is summable (and therefore the r.h.s. in (A.19)). Summability of the r.h.s. of (A.20) follows directly from from $q_n/\sqrt{\log n} \to \infty$. This shows that

$$\sum_{n=1}^{\infty} \mathbf{P}\left(\hat{K}(q_n) \neq K\right) < \infty$$

and almost sure convergence from $\hat{K}(q_n)$ to K follows from the Borel-Cantelli-Lemma. □

We close this section with the proof of Theorem 11 which is in the spirit of the proof of Theorem 7 above.

Proof of Theorem 11. Let again Δ be the smallest jump of the true signal ϑ and recall that $\vartheta(t) \in [\underline{\theta}, \overline{\theta}]$ for all $t \in [0,1]$. Further, as in the proof of Theorem 7, define the K disjoint intervals $I_k := (\tau_k - \epsilon_n, \tau_k + \epsilon) \subset [0,1]$ and I_k^-, I_k^+ and θ_k^-, θ_k^+ accordingly.
Now assume that $\hat{K} \in \mathbb{N}_0$ and that $\hat{\vartheta} \in \mathcal{S}_n[\hat{K}]$ is an estimator of ϑ such that $T_n(Y, \hat{\vartheta}) \leq q$ and

$$\max_{0 \leq k \leq K} \min_{0 \leq l \leq \hat{K}} |\hat{\tau}_l - \tau_k| > \epsilon_n.$$

Put differently, there exists an index $k \in \{1, \ldots, K\}$ such that $|\hat{\tau}_l - \tau_k| > \epsilon_n$ for all $0 \leq l \leq \hat{K}$ or, in other words, $\hat{\vartheta}$ contains no change-point in the interval I_k. With the very same reasoning as in the proof of Theorem 7 we find that

$$\begin{aligned}&\mathbf{P}\left(\exists \hat{K} \in \mathbb{N}, \hat{\vartheta} \in \mathcal{S}_n[\hat{K}] : T_n(Y, \hat{\vartheta}) \leq q \text{ and } \max_{0 \leq k \leq K} \min_{0 \leq l \leq \hat{K}} |\hat{\tau}_l - \tau_k| > \epsilon_n\right)\\ \leq&\mathbf{P}\left(\exists \hat{\theta} \in \Theta \text{ and } k \ : T_{I_k^+}(Y, \hat{\theta}) \leq \frac{1}{2}\left(q + \sqrt{\log\frac{e}{\epsilon_n}}\right)^2 \text{ and } T_{I_k^-}(Y, \hat{\theta}) \leq \frac{1}{2}\left(q + \sqrt{\log\frac{e}{\epsilon_n}}\right)^2\right).\end{aligned}$$

By replacing λ_k in the proof of Theorem 7 by ϵ_n and the assertion follows from (3.20). □

A.2.5 Proofs of Section 3.5

Proof of Theorem 16. W.l.o.g. we shall assume that $\Delta_n \geq 0$. The main idea of the proof is as follows: Let $J_n = \operatorname{argmax}\{|J| : J \subset [0,1],\ J \cap I_n = \emptyset\}$. In order to show that (3.29) holds,

we prove

$$\sup_{\mu_0 \equiv \mathfrak{m} \in \Theta} \mathbf{P}_{\mu_n} \left(T_n(Y, \mu_0) \leq q_n \right) \to 0. \tag{A.21}$$

For this purpose we construct a sequence $\mathfrak{m}_n^* \in \mathbb{R}$ such that

$$\sup_{\mathfrak{m} \geq \mathfrak{m}_n^*} \mathbf{P}\left(\sqrt{2T_{J_n}(Y, \mathfrak{m})} \leq q_n + \sqrt{2\log\left(e/\left|J_n\right|\right)} \right) \to 0 \text{ and} \tag{A.22}$$

$$\sup_{\mathfrak{m} \leq \mathfrak{m}_n^*} \mathbf{P}\left(\sqrt{2T_{I_n}(Y, \mathfrak{m})} \leq q_n + \sqrt{2\log(e/\left|I_n\right|)} \right) \to 0. \tag{A.23}$$

Recall that the true signal μ_n takes the value $\mathfrak{m}_0 + \Delta_n$ on I_n and $\mathfrak{m}_0$ on J_n. Without loss of generality we assume that $\inf_{n \in \mathbb{N}} |J_n| > 0$. We will construct a sequence of functions

$$\mathfrak{m}_n^* = \mathfrak{m}_0 + \sqrt{\beta_n / n}$$

for a sequence $(\beta_n)_{n \in \mathbb{N}}$ that satisfies $\sqrt{\beta_n}/q_n \to \infty$, (A.22) and (A.23), where we consider (A.22) first. Observe that for all $t \in J_n$ we have $|\mathfrak{m}_n^* - \mu_n(t)| \sqrt{|J_n|\, n} = \sqrt{\beta_n |J_n|}$. We further find that

$$\Gamma_{J_n} := \sqrt{\beta_n |J_n|} - q_n - \sqrt{2\log(e/|J_n|)} = q_n \left(\frac{\sqrt{\beta_n |J_n|}}{q_n} - 1 - \frac{\sqrt{2\log(e/|J_n|)}}{q_n} \right) \to \infty.$$

With this preparations, we can apply (A.4) and find for all $\mathfrak{m} \geq \mathfrak{m}_n^*$

$$\mathbf{P}\left(\sqrt{2T_{J_n}(Y, \mu)} \leq q_n + \sqrt{2\log(e/|J_n|)} \right) \leq \exp\left(-\frac{\Gamma_{J_n}^2}{2} \right) \to 0.$$

Now observe that for $t \in I_n$ we have $|\mathfrak{m}_n^* - \mu_n(t)| \sqrt{|I_n|\, n} = \Delta_n \sqrt{|I_n|\, n} - \sqrt{\beta_n |I_n|}$. Thus, by again applying (A.4) we can show (A.23) by proving

$$\Gamma_{I_n} := \Delta_n \sqrt{|I_n|\, n} - \sqrt{\beta_n |I_n|} - q_n - \sqrt{2\log(e/|I_n|)} \to \infty.$$

It hence remains to construct sequences (β_n) for each case (1.) and (2.) in the assumptions, such that the previous condition holds while $\sqrt{\beta_n}/q_n \to \infty$. We assume $\liminf_{n \to \infty} |I_n| > 0$ and define β_n through the equation

$$\sqrt{\beta_n |I_n|} = c\left(\Delta_n \sqrt{|I_n|\, n} - q_n - \sqrt{2\log(e/|I_n|)} \right)$$

for some arbitrary $0 < c < 1$. Clearly, this implies that

$$\frac{\sqrt{\beta_n |I_n|}}{q_n} = c\left(\frac{\Delta_n \sqrt{|I_n|\, n}}{q_n} - 1 - \frac{\sqrt{2\log(e/|I_n|)}}{q_n} \right).$$

From the condition in case (1.) of the theorem, the fact that $|I_n|$ is bounded away from zero for large n and $\sqrt{\beta_n}/q_n \to \infty$ we find

$$\Gamma_{I_n} = (1-c)\sqrt{\beta_n |I_n|} \to \infty.$$

Finally, we consider the case when $|I_n| \to 0$ and define β_n through the equation

$$\sqrt{\beta_n |I_n|} = c\varepsilon_n \sqrt{-\log |I_n|}. \tag{A.24}$$

From the conditions in case (2.) of the theorem and the inequality $\sqrt{x+1} - \sqrt{x} \leq 1/(2\sqrt{x})$, which holds for any $x > 0$, one obtains

$$\begin{aligned}
\Gamma_{I_n} &\geq (\sqrt{2} + \varepsilon_n)\sqrt{-\log |I_n|} - \sqrt{\beta_n |I_n|} - q_n - \sqrt{2\log(e/|I_n|)} \\
&= (\sqrt{2} + (1-c)\varepsilon_n)\sqrt{-\log |I_n|} - q_n - \sqrt{2}\sqrt{1 + \log(1/|I_n|)} \\
&\geq (1-c)\varepsilon_n \sqrt{-\log |I_n|} - \frac{1}{\sqrt{-2\log |I_n|}} - q_n.
\end{aligned}$$

This shows that $\Gamma_{I_n} \to \infty$ for a suitable small c, such that

$$\sup_{n\in\mathbb{N}} q_n/(\epsilon_n \sqrt{\log(1/|I_n|)}) \leq 1 - 2c,$$

which is not restrictive since c was only assumed to be in $(0,1)$. □

Proof of Theorem 19. The proof will be essentially based on Theorem 14. First, we define β, $\delta_{n1}, \ldots, \delta_{nK}$ and $\lambda_{n1}, \ldots, \lambda_{nK}$ as in Theorem 14. From Theorem 14 and the subsequent remarks we find that $K_n(1 - \beta_n(q_n)) \to 1$ is a sufficient conditions for

$$\mathbf{P}\left(\hat{K}(q_n) \geq K_n\right) \to 1.$$

By definition we find $K_n \leq 1/\Lambda_n$, $2\lambda_{nk} \leq \Lambda_n$ and $\delta_{nk} \leq \Delta_n$ for all $1 \leq k \leq K$. Therefore,

$$\begin{aligned}
K(1 - \beta_n(q)) \leq &\exp\left(-\frac{\left(\sqrt{n\Lambda_n}\Delta_n - 2\sqrt{2}q - 4\sqrt{\log(2e/\Lambda_n)}\right)_+^2}{8\sqrt{2}} + \log(K_n)\right) \\
&+ \exp\left(-\frac{n\Lambda\Delta_n^2}{16} + \log(K_n)\right) \\
=: &\exp\left(-\Gamma_{1,n}\right) + \exp\left(-\Gamma_{2,n}\right).
\end{aligned}$$

Hence, the proof is completed by showing that $\Gamma_{1,n} \to \infty$ and $\Gamma_{2,n} \to \infty$. It is easy to see that any of the conditions (1.)-(3.) implies $\Gamma_{2,n} \to \infty$. Therefore, it only remains to ensure

that $\Gamma_{1,n} \to \infty$. Under condition (1.) we find that $1/\Lambda_n$ is bounded and observe that

$$\frac{\Gamma_{1,n}}{q_n^2} = \frac{1}{8\sqrt{2}} \left(\frac{\sqrt{n\Lambda_n}\Delta_n}{q_n} - \frac{2\sqrt{2}q_n + 4\sqrt{\log(2e/\Lambda_n)}}{q_n} \right)_+^2 - \frac{\log 1/\Lambda_n}{q_n^2} \to \infty.$$

Since q_n is bounded away from zero, the assertion follows. Next, we consider conditions (2.) and (3.). To this end, assume that $\sqrt{n\Lambda_n}\Delta_n \geq (C+\varepsilon_n)\sqrt{\log(1/\Lambda_n)}$ for some constant $C > 0$ and a sequence ε_n such that $\varepsilon_n\sqrt{\log(1/\Lambda_n)} \to \infty$. We find that

$$\begin{aligned}
\Gamma_{1,n} &\geq \frac{1}{8\sqrt{2}} \left((C+\varepsilon_n)\sqrt{\log \frac{1}{\Lambda_n}} - 2\sqrt{2}q_n - 4\sqrt{\log(2e/\Lambda_n)} \right)_+^2 - \log K_n \\
&\geq \frac{1}{8\sqrt{2}} \left(\varepsilon_n \sqrt{\log \frac{1}{\Lambda_n}} + (C-4)\sqrt{\log \frac{1}{\Lambda_n}} - 2\sqrt{2}q_n - 4\frac{1+\log 2}{2\sqrt{\log(1/\Lambda_n)}} \right)_+^2 - \log K_n,
\end{aligned}$$

where we have used the inequality $\sqrt{x+y} - \sqrt{x} \leq y/(2\sqrt{x})$. Under condition (2.), i.e. if $\sup_{n\in\mathbb{N}} K_n < \infty$, the choice $C = 4$ implies $\Gamma_{1,n} \to \infty$. Otherwise, we use the estimate $K_n \leq 1/\Lambda_n$ which results in $C = 8$ as a sufficient condition for $\Gamma_{1,n} \to \infty$. □

Proof of Theorem 18. The proof is build on a result on Gaussian likelihood-ratios which we state here, see Ingster (1993) or Dümbgen and Spokoiny (2001)[Lemma 6.2] for a proof.

Lemma 46. *Let $Z_1, Z_2 \ldots$ be independent standard Gaussian random variables. If $\omega_m = \sqrt{2\log m}(1-\epsilon_m)$ with $\lim_{m\to\infty} \epsilon_m = 0$ and $\lim_{m\to\infty} \epsilon_m\sqrt{\log m} = \infty$, then*

$$\mathbf{E} \left| \frac{1}{m} \sum_{j=1}^{m} \exp\left(\omega_m Z_j - \omega_m^2/2\right) - 1 \right| \to 0.$$

With this lemma we can now give the proof of Theorem 18 which follows ideas from Dümbgen and Spokoiny (2001). Let $l_n = \lfloor 1/\Lambda_n \rfloor$ and define the piecewise constants functions

$$\mu_0 \equiv 0, \quad \mu_{n,j} = \mathbf{1}_{[(j-1)\Lambda_n, j\Lambda_n)}\Delta_n,$$

for $j = 1, \ldots, l_n$. Clearly, $\{\mu_{n,j}\}_{1\leq j\leq l_n} \subset \tilde{S}_n$ (as in (3.30)) for any n. We will show that for any test $\phi_n(Y)$

$$\lim_{n\to\infty} \mathbf{E}_{\mu_0}\phi_n(Y) - \alpha = 0 \quad \Rightarrow \quad \lim_{n\to\infty} \inf_{1\leq j\leq l_n} \mathbf{E}_{\mu_{n,j}}\phi_n(Y) - \alpha = 0.$$

To this end, let ϕ_n be a test, so that $\mathbf{E}_{\mu_0}\phi_n(Y) \leq \alpha + \mathrm{o}(1)$. Let f_μ denote the Lebesgue density of a Gaussian random variable with mean $\mu \in \mathbb{R}$ and variance one and define the

likelihood-ratios

$$L_{n,j}(Y) = \prod_{i=1}^{n} \frac{f_{\mu_{n,j}(i/n)}(Y_i)}{f_0(Y_i)}.$$

We then find that,

$$\begin{aligned}
\inf_{1\le j\le l_n} \mathbf{E}_{\mu_{n,j}}\phi_n(Y) - \alpha \le& \frac{1}{l_n}\sum_{j=1}^{l_n}\left[\mathbf{E}_{\mu_{n,j}}\phi_n(Y) - \alpha\right] \\
\le& \frac{1}{l_n}\sum_{j=1}^{l_n}\mathbf{E}_{\mu_{n,j}}\left[\phi_n(Y) - \mathbf{E}_{\mu_0}\phi_n(Y)\right] + \mathrm{o}(1) \\
=& \mathbf{E}_{\mu_0}\left[\left(\frac{1}{l_n}\sum_{j=1}^{l_n} L_{n,j}(Y) - 1\right)\phi_n(Y)\right] + \mathrm{o}(1) \\
\le& \mathbf{E}_{\mu_0}\left|\frac{1}{l_n}\sum_{j=1}^{l_n} L_{n,j}(Y) - 1\right| + \mathrm{o}(1).
\end{aligned}$$

Next observe that for i.i.d. standard Gaussian observations $Z_1, Z_2, \ldots, Z_{l_n}$

$$\mathbf{E}_{\mu_0}\left|\frac{1}{l_n}\sum_{j=1}^{l_n} L_{n,j}(Y) - 1\right| = \mathbf{E}\left|\frac{1}{l_n}\sum_{j=1}^{l_n}\exp\left(\sqrt{|I_n|\,n}\Delta_n Z_j - |I_n|\,n\Delta_n^2/2\right) - 1\right|,$$

which is a straightforward computation. Since the r.h.s. converges to zero by Lemma 46, this completes the proof. □

A.3 Proof of Section 5

Proof of Theorem 24. To begin the proof we show that T_n is finite almost surely, which follows from Dümbgen and Walther (2008)[Theorem 7.1]. More precisely, their result states that

$$\max_{1\le i\le j\le n}\left\{D((j-i+1)/n)\left(\frac{\left|\sum_{l=i}^{j}\epsilon_l\right|}{\sqrt{j-i+1}\sigma} - \sqrt{2\log\frac{en}{j-i+1}}\right)\right\} < \infty \quad \text{a.s.} \tag{A.25}$$

where $D(x) = \log(e/x)^{1/2}\log(e\log(e/x))^{-1}$. Note that $D(x)\to\infty$ as $x\to 0$. For the proof we divide the set of intervals into sets of "large" intervals

$$\mathcal{I}_n = \{(i,j) : 1\le i\le j\le n \text{ and } j-i+1\ge c_n\}$$

and "small" intervals

$$\mathcal{J}_n = \{(i,j) : 1 \le i \le j \le n \text{ and } j - i + 1 < c_n\}$$

for some sequence c_n such that $c_n / \log n \to \infty$. We first consider the small intervals in $\mathcal{J}_n$. Note that $\min_{(i,j)\in\mathcal{J}_n} D((j-i+1)/n) \to \infty$ as $n \to \infty$. Therefore, we find from (A.25) that

$$\max_{(i,j)\in\mathcal{J}_n} \left\{ \frac{\left|\sum_{l=i}^{j} \epsilon_l\right|}{\sqrt{j-i+1}\sigma} - \sqrt{2 \log \frac{en}{j-i+1}} \right\} \le 0 \quad \text{a.s.} \tag{A.26}$$

In order to deal with the intervals in $\mathcal{I}_n$ we use strong Gaussian approximation results from Sakhanenko (1985) (see also Zaitsev (2002)[Theorem 1] and the subsequent remark) which provides a generalization and a refinement of the results in Komlós et al. (1976).

Corollary 47 (Sakhanenko (1985))**.** *Given the random variables $\epsilon_1, \ldots, \epsilon_n$, one can construct a sequence of independent Gaussian random variables $\zeta_1, \ldots, \zeta_n$, such that $\mathbf{E}[\zeta_i] = 0$, $\mathbf{Var}[\zeta_i] = \mathbf{Var}[\epsilon_i] = \sigma^2$ and for all $x > 0$*

$$\mathbf{P}\left(C_1 \Delta(\epsilon, \zeta) \ge x\right) \le \exp\left(\log\left(1 + C_2\sqrt{n}\sigma\right) - x\right),$$

for some constants $C_1 < \infty$ and $C_2 < \infty$ and $\Delta(\epsilon,\zeta) = \max_{i\le n} \left|\sum_{l=1}^{i}(\epsilon_l - \zeta_l)\right|$.

From Corollary 47 and $c_n / \log n \to \infty$ we deduce that there exists a sequence of independent Gaussian random variables $\zeta_1, \ldots, \zeta_n$ such that

$$\max_{(i,j)\in\mathcal{I}} \frac{\left|\left|\sum_{l=i}^{j} \epsilon_l\right| - \left|\sum_{l=i}^{j} \zeta_l\right|\right|}{\sqrt{j-i+1}} \le \frac{2\Delta(\epsilon,\zeta)}{\sqrt{j-i+1}} \le \frac{2\Delta(\epsilon,\zeta)}{c_n} \to 0 \quad \text{a.s.}$$

From this in turn we observe that

$$\lim_{n\to\infty} \left| \max_{(i,j)\in\mathcal{I}_n} \left\{ \frac{\left|\sum_{l=i}^{j} \zeta_l\right|}{\sqrt{j-i+1}\sigma} - \sqrt{2\log\frac{en}{j-i+1}} \right\} - \max_{(i,j)\in\mathcal{I}_n} \left\{ \frac{\left|\sum_{l=i}^{j} \epsilon_l\right|}{\sqrt{j-i+1}\sigma} - \sqrt{2\log\frac{en}{j-i+1}} \right\} \right| = 0 \text{ a.s.}$$

We have shown in Theorem 3 that the l.h.s. converges in distribution to the M. The assertion then follows together with (A.26) and the fact that M is concentrated on the positive reals, as it was shown in Dümbgen et al. (2006). □

Proof of Theorem 26. We first define

$$\Gamma(x,y) = \left(y + \sqrt{2\log\frac{en}{x}}\right) x^{-1/2}.$$

Further, let $\hat{\mu}(q_n)$ be the SMUCE estimate with threshold q_n. For any interval $[i/n, j/n]$ which $\hat{\mu}(q_n)$ is constant on, let $\hat{\mathfrak{m}}_i^j$ denote the value of $\hat{\mu}$ on $[i/n, j/n]$ and $\mathfrak{m}_i^j$ denote the mean value of μ_n on $[i/n, j/n]$, i.e. $\mathfrak{m}_i^j = (j-i+1)^{-1}\sum_{l=i}^{j}\mu(l/n)$. Then, for $\alpha(q_n) = \mathbf{P}(M > q_n)$ we find

$$\mathbf{P}\left(\max_{[i/n,j/n]\in\hat{I}} \left|\mathfrak{m}_i^j - \hat{\mathfrak{m}}_i^j\right| - 2\Gamma(j-i+1, q_n) > 0\right) \leq \alpha(q_n). \tag{A.27}$$

The inequality is based on the following observations: first note, that $T_n(W, \hat{\mu}(q_n)) \leq q_n$ implies that $\left|\hat{\mathfrak{m}}_i^j - \overline{W}_i^j\right| \leq \Gamma(j-i+1, q_n)$ for all intervals $[i/n, j/n]$, which $\hat{\mu}$ is constant on. Here, $\overline{W}_i^j = (j-i+1)^{-1}\sum_{l=i}^{j} W_l$. Second, Corollary 25 yields that with probability greater than $1-\alpha(q_n)$ it holds uniformly over all $1 \leq i \leq j \leq n$ that $\left|\overline{W}_i^j - \mathfrak{m}_i^j\right| \leq \Gamma(j-i+1, q_n)$. Combining both observations together with the triangle inequality yields (A.27).

Now, define I_k, I_k^-, I_k^+, $\mathfrak{m}_k^-$ and $\mathfrak{m}_k^+$ as in (A.15). Assume that $\hat{K}(q_n) < K$ which implies that $\hat{\mu}(q_n)$ is constant on I_k with value $\hat{\mathfrak{m}}_k$ for some k . Since $\left|\mathfrak{m}_k^+ - \mathfrak{m}_k^-\right| \geq \Delta_n$, this means that either

$$\left|\mathfrak{m}_k^- - \hat{\mathfrak{m}}_k\right| \geq \Delta_n/2 \quad \text{or} \quad \left|\mathfrak{m}_k^+ - \hat{\mathfrak{m}}_k\right| \geq \Delta_n/2.$$

Furthermore, by straightforward calculations we find

$$\frac{\Delta_n}{2} - 2\Gamma(\Lambda_n n/2, q_n) = \frac{\Delta_n\sqrt{\Lambda_n n} - \sqrt{32}q_n - 8\log(2e/\Lambda_n)}{2\sqrt{\Lambda_n n}}. \tag{A.28}$$

Under any of the two assumptions (1.) or (2.) the last term in (A.28) is greater than zero, if n is sufficiently large. Under assumptions (1.) this is clear. For assumption (2.) we set $\Delta_n\sqrt{\Lambda_n n} = (8+\epsilon)\sqrt{\log\frac{e}{\Lambda_n}}$. With the inequality $\sqrt{x+y} - \sqrt{x} \leq y/(2\sqrt{x})$ we find that the r.h.s. of (A.28) is greater than or equal to

$$\frac{\epsilon\sqrt{\log\Lambda_n} - \sqrt{32}q_n - \frac{\log(2e)}{2\sqrt{-\log\Lambda_n}}}{2\sqrt{\Lambda_n n}},$$

which is positive for large n, by assumption. Summarizing, $\hat{K}(q_n) < K$ implies that for large n there exists an interval $[i/n, j/n]$, which $\hat{\mu}(q_n)$ is constant on and

$$\left|\mathfrak{m}_i^j - \hat{\mathfrak{m}}_i^j\right| - 2\Gamma(j-i+1, q_n) > 0. \tag{A.29}$$

As shown in (A.27), the probability for (A.29) can be bounded by $\alpha(q_n)$ and consequently

$$\mathbf{P}\left(\hat{K}(q_n) < K\right) \leq \alpha(q_n) \to 0,$$

since $\alpha(q_n) \to 0$ as $q_n \to \infty$ (see e.g. Theorem 37). □

Bibliography

Adler, R. J. and Taylor, J. E. (2007). *Random fields and geometry*. Springer Monographs in Mathematics. Springer, New York.

Arrow, K. J., Blackwell, D., and Girshick, M. A. (1949). Bayes and minimax solutions of sequential decision problems. *Econometrica*, 17:213–244.

Auger, I. E. and Lawrence, C. E. (1989). Algorithms for the optimal identification of segment neighborhoods. *Bull. Math. Biol.*, 51(1):39–54.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972). *Statistical inference under order restrictions. The theory and application of isotonic regression*. John Wiley & Sons, London-New York-Sydney. Wiley Series in Probability and Mathematical Statistics.

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.

Bellman, R. (1961). On the approximation of curves by line segments using dynamic programming. *Commun. ACM*, 4(6):284–.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57(1):289–300.

Billingsley, P. (1968). *Convergence of probability measures*. Wiley New York.

Birgé, L. and Massart, P. (2001). Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268.

Borell, C. (1975). The Brunn-Minkowski inequality in Gauss space. *Invent. Math.*, 30(2):207–216.

Boys, R. J. and Henderson, D. A. (2004). A Bayesian approach to DNA sequence segmentation. *Biometrics*, 60(3):573–588.

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.

Braun, J., Mueller, R., and Mueller, H.-G. (2000). Multiple changepoint fitting via quasilikelihood, with application to dna sequence segmentation. *Biometrika*, 87(2):301–314.

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.

Brown, L. D. (1986). *Fundamentals of statistical exponential families with applications in statistical decision theory*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 9. Institute of Mathematical Statistics, Hayward, CA.

Brown, L. D., Cai, T., and Zhou, H. H. (2010). Nonparametric regression in exponential families. *Ann. Stat.*, 38(4):2005–2046.

Candès, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351.

Carlstein, E., Müller, H.-G., and Siegmund, D. (1994). *Change-point problems*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 23. Institute of Mathematical Statistics, Hayward, CA. Papers from the AMS-IMS-SIAM Summer Research Conference held at Mt. Holyoke College, South Hadley, MA, July 11–16, 1992.

Chan, H. P. and Walther, G. (2013). Detection with the scan and the average likelihood ratio. *arXiv:1107.4344v1*.

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

Chernoff, H. and Zacks, S. (1964). Estimating the current mean of a normal distribution which is subjected to changes in time. *Ann. Math. Statist.*, 35:999–1018.

Churchill, G. A. (1992). Hidden markov chains and the analysis of genome structure. *Computers & Chemistry*, 16(2):107 – 115.

Csörgö, M. and Horváth, L. (1997). *Limit theorems in change-point analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons Ltd., Chichester. With a foreword by David Kendall.

Davies, L., Höhenrieder, C., and Krämer, W. (2012). Recursive computation of piecewise constant volatilities. *Computational Statistics & Data Analysis*, 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.

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

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

Donoho, D. L. (1988). One-sided inference about functionals of a density. *Ann. Stat.*, 16(4):1390–1420.

Donoho, D. L. (1995). De-noising by soft-thresholding. *Information Theory, IEEE Transactions on*, 41(3):613 –627.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1995). Wavelet shrinkage: asymptopia? *J. Roy. Statist. Soc. Ser. B*, 57(2):301–369. With discussion and a reply by the authors.

Du, C. and Kou, S. (2012). Stepwise signal extraction via marginal likelihood. *Harvard preprint*.

Dümbgen, L. (1991). The asymptotic behavior of some nonparametric change-point estimators. *Ann. Statist.*, 19(3):1471–1495.

Dümbgen, L. and Kovac, A. (2009). Extensions of smoothing via taut strings. *Electron. J. Stat.*, 3:41–75.

Dümbgen, L., Piterbarg, V. I., and Zholud, D. (2006). On the limit distribution of multiscale test statistics for nonparametric curve estimation. *Math. Methods Statist.*, 15(1):20–25.

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

Dümbgen, L. and Walther, G. (2008). Multiscale inference about a density. *Ann. Statist.*, 36(4):1758–1785.

Elhaik, E., Graur, D., and Josić, K. (2010). Comparative testing of dna segmentation algorithms using benchmark simulations. *Molecular Biology and Evolution*, 27:1015–24.

Fearnhead, P. (2006). Exact and efficient Bayesian inference for multiple changepoint problems. *Stat. Comput.*, 16(2):203–213.

Feder, P. I. (1975). The log likelihood ratio in segmented regression. *Ann. Statist.*, 3:84–97.

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. (2013). Multiscale change-point inference. *to appear in Journal of the Royal Statist. Society, Ser. B, with discussion and rejoinder by the authors.* preprint avaiable at http://arxiv.org/abs/1301.7212.

Fridlyand, J., Snijders, A. M., Pinkel, D., Albertson, D. G., and Jain, A. N. (2004). Hidden Markov models approach to the analysis of array CGH data. *J. Multivariate Anal.*, 90(1):132–153.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Statist.*, 1(2):302–332.

Friedrich, F., Kempe, A., Liebscher, V., and Winkler, G. (2008). Complexity Penalized M-Estimation : Fast Computation. *Journal of Computational and Graphical Statistics*, 17(1):201–224.

Fryzlewicz, P. (2012). Wild binary segmentation for multiple change-point detection. Technical report.

Futschik, A., Hotz, T., Munk, A., and Sieling, H. (2013). Multiresolution DNA partitioning: statistical evidence for segments. *submitted*.

Harchaoui, Z. and Lévy-Leduc, C. (2010). Multiple change-point estimation with a total variation penalty. *J. Amer. Statist. Assoc.*, 105(492):1480–1493.

Hinkley, D. V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika*, 57:1–17.

Hinkley, D. V. and Hinkley, E. A. (1970). Inference about the change-point in a sequence of binomial variables. *Biometrika*, 57(3):pp. 477–488.

Höhenrieder, C. (2008). *Nichtparametrische Volatilitäts-und Trendapproximation von Finanzdaten*. PhD thesis.

Hotz, T., Schütte, O. M., Sieling, H., Polupanow, T., Diederichsen, U., Steinem, C., and Munk, A. (2012). Idealizing ion channel recordings by jump segmentation and statistical multiresolution analysis. *to appear in IEEE Transactions on NanoBioscience*. avaiable at http://www.stochastik.math.uni-goettingen.de/preprints/IonMRC.pdf.

Hotz, T. and Sieling, H. (2013). *stepR: Fit Step-Functions*. R package version 1.0.

Hušková, M. and Antoch, J. (2003). Detection of structural changes in regression. *Tatra Mt. Math. Publ.*, 26:201–215.

Inclán, C. and Tiao, G. C. (1994). Use of cumulative sums of squares for retrospective detection of changes of variance. *J. Amer. Statist. Assoc.*, 89(427):913–923.

Ingster, Y. I. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives. I-III. *Math. Methods Statist.*, 2(2):85–114, 171–189,249–268.

Jackson, B., Sargle, J., Barnes, D., Arabhi, S., Alt, A., Gioumousis, P., Gwin, E., Sangtrakulcharoen, P., Tan, L., and Tsai, T. (2005). An algorithm for optimal partitioning of data on an interval. *IEEE, Signal Processing Letters*, 12(2):105–108.

Jeng, X. J., Cai, T. T., and Li, H. (2010). Optimal sparse segment identification with application in copy number variation analysis. *J. Amer. Statist. Assoc.*, 105(491):1156–1166.

Kallioniemi, A., Kallioniemi, O., Sudar, D., Rutovitz, D., Gray, J., Waldman, F., and Pinkel, D. (1992). Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, 258(5083):818–821.

Kander, Z. and Zacks, S. (1966). Test procedures for possible changes in parameters of statistical distributions occurring at unknown time points. *Ann. Math. Statist.*, 37:1196–1210.

Khodadadi, A. and Asgharian, M. (2008). Change-point problems and regression: An annotated bibliography. *Collection of Biostatistics Research Archive (COBRA)*.

Killick, R., Fearnhead, P., and Eckley, I. A. (2011). Optimal detection of changepoints with a linear computational cost. *ArXiv e-prints*.

Komlós, J., Major, P., and Tusnády, G. (1976). An approximation of partial sums of independent RV's, and the sample DF. II. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 34(1):33–58.

Lai, T. L., Xing, H., and Zhang, N. (2008). Stochastic segmentation models for array-based comparative genomic hybridization data analysis. *Biostatistics*, 9(2):290–307.

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.

Lavielle, M. and Teyssière, G. (2007). Adaptive detection of multiple change-points in asset price volatility. In *Long memory in economics*, pages 129–156. Springer, Berlin.

Li, K.-C. (1989). Honest confidence regions for nonparametric regression. *Ann. Statist.*, 17(3):1001–1008.

Luong, T. M., Rozenholc, Y., and Nuel, G. (2012). Fast estimation of posterior probabilities in change-point models through a constrained hidden Markov model. *ArXiv e-prints*.

Nemirovski, A. (1985). Nonparametric estimation of smooth regression functions. *Tekhnicheskaya Kibernetika*, 3:50–60.

Nielsen, B. O. (1973). *Exponential Families and Conditioning*. Wiley.

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.

Page, E. S. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42:523–527.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rigaill, G., Lebarbier, E., and Robin, S. (2012). Exact posterior distributions and model selection criteria for multiple change-point detection problems. *Stat. Comput.*, 22(4):917–929.

Rivera, C. and Walther, G. (2012). Optimal detection of a jump in the intensity of a Poisson process or in a density with likelihood ratio statistics. *ArXiv e-prints*.

Sakhanenko, A. (1985). Convergence rate in invariance principle for non-identically distributed variables with exponential moments. In Borovkov, A. and Balakrishnan, A., editors, *Limit theorems for sums of random variables*, Advances in probablity theory. Springer-Verlag.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464.

Sen, A. and Srivastava, M. S. (1975). On tests for detecting change in mean. *Ann. Statist.*, 3:98–108.

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

Siegmund, D. (1986). Boundary crossing probabilities and statistical applications. *Ann. Statist.*, 14(2):361–404.

Siegmund, D. (1988). Confidence sets in change-point problems. *Internat. Statist. Rev.*, 56(1):31–48.

Siegmund, D. and Venkatraman, E. S. (1995). Using the generalized likelihood ratio statistic for sequential detection of a change-point. *Ann. Statist.*, 23(1):255–271.

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

Siegmund, D. O. and Zhang, H. (1994). Confidence regions in broken line regression. In *Change-point problems (South Hadley, MA, 1992)*, volume 23 of *IMS Lecture Notes Monogr. Ser.*, pages 292–316. Inst. Math. Statist., Hayward, CA.

Siegmund, D. O., Zhang, N. R., and Yakir, B. (2011). False discovery rate for scanning statistics. *Biometrika*, 98(4):979–985.

Snijders, A. M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A. K., Huey, B., Kimura, K., et al. (2001). Assembly of microarrays for genome-wide measurement of dna copy number. *Nature genetics*, 29(3):263–264.

Spokoiny, V. (2009). Multiscale local change point detection with applications to value-at-risk. *Ann. Statist.*, 37(3):1405–1436.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108.

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

van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York. With applications to statistics.

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.

Vostrikova, L. J. (1981). Discovery of "discord" in multidimensional random processes. *Dokl. Akad. Nauk SSSR*, 259(2):270–274.

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

Winkler, G. and Liebscher, V. (2002). Smoothers for discontinuous signals. *J. Nonparametr. Stat.*, 14(1-2):203–222. Statistical models and methods for discontinuous phenomena (Oslo, 1998).

Wittich, O., Kempe, A., Winkler, G., and Liebscher, V. (2008). Complexity penalized least squares estimators: analytical results. *Math. Nachr.*, 281(4):582–595.

Worsley, K. J. (1983). The power of likelihood ratio and cumulative sum tests for a change in a binomial probability. *Biometrika*, 70(2):455–464.

Worsley, K. J. (1986). Confidence regions and test for a change-point in a sequence of exponential family random variables. *Biometrika*, 73(1):91–104.

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.

Yakir, B. and Pollak, M. (1998). A new representation for a renewal-theoretic constant appearing in asymptotic approximations of large deviations. *Ann. Appl. Probab.*, 8(3):749–774.

Yao, Y.-C. (1988). Estimating the number of change-points via Schwarz' criterion. *Statist. Probab. Lett.*, 6(3):181–189.

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

Zaitsev, A. Y. (2002). Estimates for the strong approximation in multidimensional central limit theorem. In *Proceedings of the International Congress of Mathematicians, Vol. III (Beijing, 2002)*, pages 107–116, Beijing. Higher Ed. Press.

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

Curriculum Vitae

Hannes Sieling
born 29th April 1985 in Oldenburg (Oldb), Germany

September 1991 – June 2004
Schooling
General Qualification for University Entrance
Herbartgymnasium Oldenburg

Oktober 2004 – September 2005
Voluntary Year of Social Service
Oldenburger Turnerbund

October 2005 – November 2010
Study of Business Mathematics
Faculty of Mathematics, University of Göttingen
Diploma thesis: *Statistical Multiscale Methods in Piecewise Constant Poisson Regression*
supervised by *Prof. Dr. Axel Munk*

since December 2010
Ph.D. Studies in Mathematics
Faculty of Mathematics, University of Göttingen
supervised by *Prof. Dr. Axel Munk*

since April 2011
Member of the *DFG-SNF research group 916*
"Statistical Regularization and Qualitive Constraints"