FDR-CONTROL IN MULTISCALE CHANGE-POINT SEGMENTATION

HOUSEN LI, AXEL MUNK
INSTITUTE FOR MATHEMATICAL STOCHASTICS, UNIVERSITY OF GÖTTINGEN
AND MAX PLANCK INSTITUTE FOR BIOPHYSICAL CHEMISTRY
GÖTTINGEN, GERMANY
AND
HANNES SIELING
INSTITUTE FOR MATHEMATICAL STOCHASTICS, UNIVERSITY OF GÖTTINGEN
GÖTTINGEN, GERMANY

Abstract. Fast multiple change-point segmentation methods, which additionally provide faithful statistical statements on the number and size of the segments, have recently received great attention. For example, SMUCE, as introduced in (Frick, Munk, and Sieling, Multiscale change-point inference. J. R. Statist. Soc. B, 76:495-580, 2014), allows to control simultaneously over a large number of scales the error of overestimating the true number $K$ of change-points, $\mathbb{P}\{\hat{K} > K\} \leq \alpha_S$, for a preassigned significance level $\alpha_S$, independent of the underlying change-point function. The control of this family-wise error rate (FWER), however, makes this method generally conservative. In this paper, we propose a multiscale segmentation method, which controls the false discovery rate (FDR) instead. It can be efficiently computed by a pruned dynamic program. We show a non-asymptotic upper bound for its FDR in a Gaussian setting, which allows to calibrate the new segmentation method properly. By switching from FWER to FDR, the detection power of the proposed method significantly outperforms SMUCE. The favorable performance of the proposed method is examined by comparisons with some state of the art methods on both simulated and real datasets.

Keywords: Multiscale inference; Change-point regression; False discovery rate; Dynamic programming; Array CGH data; Ion channel recordings.

1. Introduction

To keep the presentation simple, we assume that observations are given by the regression model

$$Y_i = \mu \left( \frac{i}{n} \right) + \sigma \varepsilon_i, \quad i = 0, 1, \ldots, n - 1,$$

(1)

where $\varepsilon_0, \ldots, \varepsilon_{n-1}$ are independent standard normally distributed, and $\sigma > 0$. The mean-value function $\mu$ is assumed to be right-continuous and piecewise constant, i.e.

$$\mu = \sum_{k=0}^{K} 1_{[\tau_k, \tau_{k+1})} c_k.$$

(2)

Here $c_k \neq c_{k+1}$ for $k = 0, 1, \ldots, K - 1$, and $0 < \tau_1 < \ldots < \tau_K < 1$ denote the change-points of $\mu$, with the convention that $\tau_0 := 0$ and $\tau_{K+1} := 1$. For simplicity we will also use the notation $I_k = [\tau_k, \tau_{k+1})$ for the $k$-th segment. The value of $\mu$ on the $k$-th segment $I_k$ is denoted by $c_k$. We stress, however, that much of our
subsequent methodology and analysis can be extended to other models, e.g. when the observations come from an exponential family or more generally, errors obey certain moment conditions.

Estimation of $\mu$ and its change-points in this seemingly simple model (1) (and variations thereof) has a long history in statistical research (see e.g. Csörgő and Horváth [1997; Siegmund 2013; Frick et al. 2014] for a survey) and has recently gained renewed interest from two perspectives, in particular. Firstly, large scale applications such as from finance (see e.g. Inclán and Tiao 1994; Bai and Perron 1998; Lavielle and Teyssière 2007; Spokoiny 2009), signal processing (see e.g. Harchaoui and Lévy-Leduc 2008; Blythe et al. 2012; Hotz et al. 2013) or genetic engineering (see e.g. Braun et al. 2000; Olshen et al. 2004; Zhang and Siegmund 2007; Jeng et al. 2010) call for change-point segmentation methods which are computationally fast, say almost linear in the number of observations. Secondly, besides of a mere segmentation of the data into pieces of constancy certain evidence on the accuracy of the number and locations of change-points which comes with this segmentation is demanded.

Many of such methods are based on minimizing a penalized cost functional among different number of change-points $K$ and locations of change-points $\tau_k$. For a cost function $C$, which serves as goodness-of-fit measure of a constant function on an interval, and a penalty against over-fitting $f(K)$ these approaches search for a solution of the global optimization problem

$$\min_{K} \sum_{k=1}^{K+1} C(Y_{n\tau_{k-1}}, \ldots, Y_{n\tau_k-1}) + \gamma f(K). \quad (3)$$

Fast and exact algorithms for this kind of methods are often based on dynamic programming such as the optimal partitioning method (Jackson et al. 2005) and the Potts estimate (Boysen et al. 2009; Storath et al. 2014), who advocate the sparsest subset selection penalty

$$f(K) = l_0(\mu) = K. \quad (4)$$

For more general $f$, see e.g. the segment neighbor method (Auger and Lawrence 1989) or (Friedrich et al. 2008). More recently, Killick et al. (2012) introduced a pruned dynamic program (PELT) with expected linear complexity mainly for $f(K) = K$. From a computational point of view, approaches of type (3) seem therefore beneficial. Nevertheless, the choice of the balancing parameter $\gamma := \gamma_n(Y)$ in (3) is subtle. Birgé and Massart (2006) offer examples and discussion of this and other penalty choices, and Boysen et al. (2009) provide optimal choices of $\gamma_n$ as $n \to \infty$. Zhang and Siegmund (2007) proposed a penalty depending on $K$ and additionally on distances between consecutive change-points. However, given the data at hand, significance conclusions on the number, location and size of the change-point function are not an easy task for the above mentioned methods, although in many cases there is a good asymptotic understanding nowadays. A similar comment applies to other global segmentation methods which rely on an $l_1$ approximation of the $l_0$ penalty in (4) including lasso-type techniques possibly together with post filtering to further
enhance sparseness, see e.g. (Tibshirani et al. 2005; Friedman et al. 2007; Harchaoui and Lévy-Leduc 2010). To overcome the difficulty of choosing $\gamma$ properly, and for conclusions on the obtained segmentation with statistical evidence, Bayesian methods offer an attractive alternative as well, see (Barry and Hartigan 1993; Green 1995; Rigaill et al. 2012) and the references therein.

In contrast to solving the global optimization problem in (3) another prominent class of methods is based on the idea to iteratively apply a local segmentation method to detect a single change-point. If such a change-point is detected on a segment, it is split into two parts and the same routine is applied to both new segments. The method stops if no further change-points are found. This approach, referred to as binary segmentation (BS), is certainly among the most popular ones for change-point segmentation, in particular in the context of the analysis of copy number variation data and related biostatistical issues. It has already been suggested in Scott and Knott (1974) and more recently related methods have been proposed, such as circular binary segmentation (CBS) (Olshen et al. 2004; Venkatraman and Olshen 2007) and wild binary segmentation (WBS) (Fryzlewicz 2014). For these approaches, the to be specified parameter is the probability of including a false change-point in one iteration. Therefore, local error control can be provided, but the overall control on the error to include or exclude wrong segments appears to be elusive for these methods, as well.

Frick et al. (2014) suggest a hybrid method, simultaneous multiscale change-point estimator (SMUCE) (see also Boysen et al. 2009; Davies et al. 2012 in the context of variance estimation), which tries to address both tasks by minimizing the number of change-points under a local multiscale side-constraint. The side-constraint is based on a simultaneous multiple testing procedure on all scales (length of subsequent observations) which employs a scale calibrating penalty, borrowed from (Dümbgen and Spokoiny 2001). It can be shown that for the resulting segmentation $\hat{\mu}$ the number of change-points is not overestimated at a pre-defined probability, $1 - \alpha_S$ (i.e. family-wise error rate, FWER). This provides a direct statistical interpretation. In fact, the error of including $j$ false positives provided by SMUCE has exponential decay,

$$P\{\hat{K} \geq K + j\} \leq \alpha_S^{\left\lceil j/2 \right\rceil}, \quad j = 1, 2, \ldots$$

(5)

(see Sieling 2013; Frick et al. 2014), which in particular controls the overestimation of the number of jumps ($j = 1$ in (5))

$$\mathbb{P}\{\hat{K} > K\} \leq \alpha_S.$$  

(6)

Moreover, it can be shown that the method is able to detect the true number of change-points over a large range of scales with minimax detection power (see Theorem 5 in Frick et al. 2014). However, according to (6), in particular in situations with low signal to noise ratio (SNR) or with many change-points compared to the number of observations, this error control necessarily leads to a conservative estimate $\hat{\mu}$ of $\mu$ in (2), i.e. with fewer change-points than the true number $K$. Therefore, in
this paper we offer a strategy to overcome this drawback which might be beneficial
also for other related methods. This is based on the control of the \textit{false discovery rate} (FDR) \cite{Benjamini and Hochberg, 1995} instead of the FWER. Despite of the
huge literature about change-point detection, there is only a small number of papers
addressing the FDR issue in this context, \cite{Tibshirani and Wang, 2008, Siegmund et al., 2011, Hao et al., 2013}. These are multiple stage procedures and they only
control FDR in certain steps, not for the whole approach. In this work, we will
present \textit{FDR-SMUCE}, similar in spirit to SMUCE, which however, controls the FDR
of the whole segmentation, rather than the FWER. The significance statement given
by the method is quite intuitive and also holds for a finite number of observations.
This reveals that the contribution of this work is twofold: First, the new method
overcomes the conservative nature of SMUCE while maintaining a solid statistical
interpretation. In doing this, we provide a general framework how to combine FDR-
control with global segmentation methods. We are not aware of any other change-
point segmentation method which shares this property. Second, all results are non-
asymptotic and hold uniformly over all piecewise constant functions \( \mu \) in model \( \mu \).

Before going into details, we illustrate this by the example in Figure 1. We
employed the \textit{blocks signal} \cite{Donoho and Johnstone, 1994} with Gaussian obser-

observations of standard deviation \( \sigma = 10 \) (with average SNR \( \int |\mu(x)|dx/\sigma \approx 0.65 \)).

Very naturally we declare such discoveries (estimated change-points) true if they are
“close” (to be specified later) to true change-points. In this example FDR-SMUCE
(FDR \( \leq \beta = 0.1 \)) detects all the change-points correctly, while SMUCE (\( \alpha_S = 0.1 \))
finds only 6 out of 11, due to its requirement to control the FWER in \( \alpha \). On this
data, the smallest \( \beta \) for FDR-SMUCE which overestimates the number of change-
points is 0.5. With such choice of \( \beta \), FDR-SMUCE finds one additional false change-
point (at 0.17, marked by a vertical line and an associated interval defined in \( \beta \),
in the fourth panel) besides all the true ones. The proportion between false and all
discoveries plus one (number of segments) is hence \( 1/(12 + 1) \approx 0.08 \). Later we will
show that the FDR-SMUCE is indeed able to control this proportion in expectation
at the predefined level \( \beta \). For the other direction, the largest \( \beta \) for FDR-SMUCE
which underestimates the number of change-points is 0.07, which is shown in the
bottom. That is, FDR-SMUCE estimates the correct number of change-points for
every \( \beta \in (0.07, 0.5) \).

For our purpose it is helpful to interpret the “detection part” of the multiple change-
point regression problem as a multiple testing problem. In the literature methods
with this flavor often consider multiscale local likelihood tests. Whereas local tests
for the presence of a change-point on small systems of sets (e.g. the dyadics) of the
sampling points \( \{0, 1/n, \ldots, (n - 1)/n\} \) can be efficiently computed they may have
low detection power and highly redundant systems such as the system of all intervals
have been suggested instead \cite{Siegmund and Yakir, 2000, Dümbgen and Spokoiny, 2001, Davies et al., 2012, Frick et al., 2014}, see, however, \cite{Walther, 2010, Rivera and Walther, 2013} for sparser but still asymptotically efficient systems. It was pointed
Figure 1. Illustration of FDR-SMUCE. The noisy data together with the true signal is shown in the second panel. Below, FDR-SMUCE ($\beta = 0.1$), FDR-SMUCE ($\beta = 0.5$), and FDR-SMUCE ($\beta = 0.07$) are shown. As a comparison, SMUCE ($\alpha_s = 0.10$) is shown on the top. Each true discovery is indicated by a vertical dashed line and each false one by a vertical dotted line (red) and an associated interval defined in [7].

out by Siegmund et al. (2011) that classical FDR for redundant systems might be misleading, because such local tests are highly correlated and consequently tests on nearby intervals likely reject/accept the null-hypothesis together, see also Benjamini and Yekutieli (2001), Guo and Sarkar (2013) for a general discussion of this issue. Siegmund et al. (2011) therefore suggest to test for constancy on subintervals and to
group the nearby false (or true) rejections, and count them as a single discovery, which allows to control the FDR group-wise. In our approach, we circumvent this difficulty, but still are able to work with redundant systems, because instead we perform a multiple test for the change-points directly, i.e. we treat the multiple testing problem

\[ H_i : \frac{i}{n} \text{ is not a change-point} \quad \text{v.s.} \quad A_i : \frac{i}{n} \text{ is a change-point} \quad i = 0, \ldots, n - 1. \]

It remains to define a true discovery. This is done by identifying a rejection as a true discovery if it is “close” to a true change-point. To be specific, let \( \{\hat{\tau}_1, \ldots, \hat{\tau}_K\} \) be rejections (i.e. estimated change-points), and \( \hat{K} \) the estimated number of change-points. For each \( i \in \{1, \ldots, \hat{K}\} \), we classify \( \hat{\tau}_i \) as a true discovery if there is a true change-point lying in

\[
\left[ \frac{n(\hat{\tau}_{i-1} + \hat{\tau}_i)/2}{n}, \frac{n(\hat{\tau}_i + \hat{\tau}_{i+1})/2}{n} \right]
\]

(7)

where \( \hat{\tau}_0 := 0 \) and \( \hat{\tau}_{K+1} := 1 \); otherwise, it is a false discovery, see again panel 4 in Figure 1.

Similar to Benjamini and Hochberg (1995), we then define the false discovery rate (FDR) by

\[
\text{FDR} := E \left[ \frac{\text{FD}}{\hat{K} + 1} \right],
\]

(8)

where FD is the number of false discoveries in the above sense.

The rest of the paper is organized as follows. In Section 2, we introduce the new segmentation method (FDR-SMUCE) and show its control of FDR. In Section 3, we will develop a pruned dynamic program for its computation. The accuracy and efficiency of FDR-SMUCE is examined in Section 4 on both simulated and real datasets. The paper ends with a conclusion in Section 5. An implementation of FDR-SMUCE is provided in R package “FDRS”, available from http://www.stochastik.math.uni-goettingen.de/fdrs.

2. Method and Main Result

Now we will give a formal definition of the FDR-SMUCE. To simplify, we assume that the noise level \( \sigma \) is known. For methods to estimate \( \sigma^2 \), see e.g. (Rice, 1984), (Dette et al., 1998), or (Davies and Kovac, 2001) among many others. Assume that \( Y = (Y_0, \ldots, Y_{n-1}) \) is given by model (1). For an interval \( I \subset [0, 1) \) we consider the multiscale statistic with scale calibration

\[
T_I(Y, c) = \max_{|i/n, j/n| \subset I} \left| \frac{\sum_{i=1}^j (Y_i - c)}{\sigma \sqrt{j - i + 1}} \right| - \text{pen} \left( \frac{j - i + 1}{|I|} \right),
\]

(9)

where \( c \) is a real number, \( \text{pen}(x) = \sqrt{2 \ln(e/x)} \) the penalty term for the scale and \( |I| \) the number of observations in \( I \) (scale) with slight abuse of notation. For \( \alpha \in (0, 1) \), let us introduce \( q_\alpha(m) \) by

\[
q_\alpha(m) := \min \{ q : \mathbb{P} \{ T_I(\varepsilon, \bar{\varepsilon}_I) > q \} \leq \alpha \},
\]

(10)
where \( \epsilon = (\epsilon_0, \ldots, \epsilon_{n-1}) \) is standard normally distributed, \( \bar{\epsilon}_I = \sum_{i/n \in I} \epsilon_i/|I| \), and \( I \) a fixed interval with \( |I| = m \). It can be easily shown that \( q_\alpha(m) \) does not depend on the choice of \( I \) if \( |I| = m \), which justifies the definition (10).

**Remark 2.1.** As a direct consequence of (Dümbgen and Spokoiny, 2001) (see also Dümbgen and Walther, 2008; Frick et al., 2014) the limit distribution of \( T_I(\epsilon, \bar{\epsilon}_I) \) is finite almost surely and is continuous (Dümbgen et al., 2006), as \( |I| \to \infty \). The values \( q_\alpha(m) \) are therefore uniformly bounded for all \( m \).

For our purpose we have to introduce the set of step functions restricted to the multiscale side-constraint induced by (9) and (10)

\[
C_k = \left\{ \mu = \sum_{i=0}^k c_i 1_{I_i} : T_{I_i}(Y, c_i) - q_\alpha(|I_i|) \leq 0 \quad \text{for } i = 0, 1, \ldots, k \right\}. \tag{11}
\]

The estimated number of change-points \( \hat{K} \) will be given by

\[
\hat{K} := \min \{k : C_k \neq \emptyset\}. \tag{12}
\]

Then the FDR-SMUCE estimate \( \hat{\mu} \) is given by

\[
\hat{\mu} := \arg \min_{\mu \in C_{\hat{K}}} \sum_{i=0}^{n-1} \left( Y_i - \mu \left( \frac{i}{n} \right) \right)^2, \tag{13}
\]

that is, the constrained maximum likelihood estimator within \( C_{\hat{K}} \).

![Figure 2. Relation between tuning parameter \( \alpha \) and the bound of FDR \( \beta \).](image)

The main theorem of this paper is the FDR control, defined in (8), of the estimate \( \hat{\mu} \) in (13). More precisely, the FDR can be controlled explicitly by choosing the local level \( \alpha \) in (10) properly.

**Theorem 2.2.** Let \( Y \) be observations from model (1), and \( 0 < \alpha < 1/3 \). Then for the FDR-SMUCE \( \hat{\mu} \) in (13) with \( q_\alpha \) in (10) we have for the FDR defined in (8)

\[
\text{FDR}_{\hat{\mu}}(\alpha) \leq \frac{2\alpha}{1-\alpha} =: \beta. \tag{14}
\]
Proof. See Appendix A. □

Remark 2.3. To calibrate the method for given $\beta$, we simply rewrite (14) into

$$\alpha = \frac{\beta}{2 + \beta},$$

which is roughly, $\alpha = \beta/2$ for small $\beta$, see Figure 2.

![Figure 3](image)

**Figure 3.** Difference between SMUCE and FDR-SMUCE. The upper plots show the two estimates (solid line), respectively, together with the truth (dotted line) and the data (points). The lower left (right) shows all the intervals on which there is a constant function satisfying the multiscale side-constraint of SMUCE (FDR-SMUCE), with red ones chosen by the estimator, separately.

Remark 2.4 (Comparison of SMUCE and FDR-SMUCE). Let us stress some notable differences to SMUCE (Frick et al., 2014), which is based on restricting possible estimators to

$$C_k^0 = \left\{ \mu = \sum_{i=0}^{k} c_i 1_{I_i} : \max_{i=0,\ldots,k} T_{I_i}^0(Y,c_i) \le q_{\alpha_S} \right\},$$

where $T_{I_i}^0(Y,c)$ is as in (9), with penalty $\text{pen}( (j-i+1)/n )$ instead. Firstly, this penalty term underlying SMUCE on the interval $[i/n, j/n]$ only relates the ratio between the number of observations in $[i/n, j/n]$ and all the observations, while that of FDR-SMUCE relies on the ratio between the number of observations in $[i/n, j/n]$ and the corresponding segment length of $I$. This modification has a flavor similar to Zhang and Siegmund (2007)'s refined Bayes information criterion type of penalty. Secondly, the parameter $\alpha_S$ of SMUCE ensures that the true signal lies in the side-constraint $C_k^0$ with probability at least $1-\alpha_S$. In contrast, the FDR-SMUCE considers constant parts of the true signal individually, guaranteeing that the mean value of each segment $I_i$ lies in its associated side-constraint in $C_K$ with probability at least
$1 - \alpha$. This makes it much less conservative, and its error controllable in terms of FDR (see Theorem 2.2). This is a key idea underlying FDR-SMUCE. For an illustration of this effect see Figure 3. Thirdly, the quantile in SMUCE is universal and practically estimated by Monte-Carlo simulations using the worst case scenario, i.e. $\mu = 0$ on all the intervals, since the exact system of intervals on which the true signal is constant is unknown. In general, this leads to a larger quantile than what is required, making the method more conservative than necessary. In contrast, for FDR-SMUCE, the quantiles $q_\alpha$ in (10) are scale dependent, relying on all the intervals up to certain length, revealing the resulting method less conservative.

In situations with many change-points or low SNR, to overcome the conservative nature of SMUCE, the significance level $\alpha_S$ in (6), the overestimation error, has been suggested to be chosen close to one to produce an estimate with good screening properties (Frick et al., 2014). It follows from the arguments above that the parameter $\alpha$ of FDR-SMUCE relates to $\alpha_S$ roughly by

$$1 - (1 - \alpha)^{K+1} \approx \alpha_S$$

(15)

because the probability of coverage of the true signal by $C_K$ is $(1 - \alpha)^{K+1}$, where $K$ is the true number of change-points. This is confirmed by simulations. For example, consider the recovery of a teeth signal (adopted from Fryzlewicz, 2014) with $K = 50$ from 900 observations contaminated by standard Gaussian noise, see Figure 4. In Figure 5, the histogram of estimated number of change-points by SMUCE ($\alpha_S = 0.1$) and FDR-SMUCE ($\alpha = 0.1$) are shown in white bars from 1,000 repetitions. It can be seen that SMUCE ($\alpha_S = 0.1$) seriously underestimates the number of change-points, while FDR-SMUCE estimates the right number of change-points with high probability. If we adjust $\alpha_S$ according to (15), i.e. $\alpha_S = 1 - (1 - 0.1)^{51} \approx 0.995$, this leads to a significant improvement of detection power of SMUCE, as is shown by the corresponding histogram of estimated number of change-points in grey bars (left panel in Figure 5), however, at the expense of any provable statistical error control, i.e. the control of overestimating the true $K$ for SMUCE becomes increasingly more difficult as $K$ gets larger. On the other hand, FDR-SMUCE adapts to $K$ automatically, and works well with a choice of small values of $\beta$ in (14). Finally, it becomes apparent from a comparison of the two lower panels in Figure 7 that the local thresholding in (10) and (11) makes an important difference to SMUCE.

**Remark 2.5** (Discussion of the bound in Theorem 2.2). Various simulation studies (not displayed) suggest even the bound $\text{FDR} \leq \alpha$, improving (14) by a factor of 2. Although we were not able to prove this, we stress that this might be useful for practical purpose to select and interpret $\alpha$. For example in Figure 6 we display results for the teeth signal (see Figure 4), where the FDR is estimated by the empirical mean of 1,000 repetitions with $n = 600$. It shows that the bound (14) (dashed line) is good when $\alpha$ is small, and gets worse as $\alpha$ increases.
Figure 4. Estimation of teeth signal \((K = 50, n = 900)\) by SMUCE \((\alpha_S = 0.1)\), SMUCE \((\alpha_S = 0.995)\) and FDR-SMUCE \((\alpha = 0.1)\). The true signal (dashed line), together with data (points), is shown in each panel.

Figure 5. Histogram of number of change-points for SMUCE \((\alpha_S = 0.1)\), SMUCE \((\alpha_S = 0.995)\), and FDR-SMUCE \((\alpha = 0.1)\). The shaded bars correspond to the true number of change-points 50. The number of simulations is 1,000.
3. Implementation

It will be shown that the FDR-SMUCE can be efficiently computed by a pruned dynamic programming algorithm. For convenience let us introduce

\[
\mathcal{I}\left(\left[\frac{i}{n}, \frac{j}{n}\right]\right) = \begin{cases} 
1 & \text{if } T_{[i/n,j/n]}(Y, c) \leq q_\alpha(j - i) \text{ for some constant } c, \\
0 & \text{otherwise.}
\end{cases}
\]

We first consider the computation of \( \hat{K} \), which is defined in (12). Let \( \hat{K}[i] \) be the number of change-points of the FDR-SMUCE estimate when applying to \((Y_0, \ldots, Y_{i-1})\), that is,

\[
\hat{K}[i] := \min \left\{ k : \max_{0 \leq j \leq k} T_{I_j}(Y, c_j) - q_\alpha(|I_j|) \leq 0 \text{ for some } \mu = \sum_{j=0}^{k} c_j 1_{I_j} \text{ with } \biguplus_{j=0}^{k} I_j = [0, \frac{i}{n}] \right\}
\]

for \( i = 1, \ldots, n \), where \( \biguplus \) denotes disjoint union. Then the estimated number of change-points \( \hat{K} \) in (12) is given by \( \hat{K}[n] \). It can be shown that the following recursive relation

\[
\hat{K}[0] := -1,
\]

\[
\hat{K}[i] = \min \left\{ \hat{K}[j] + 1 : \mathcal{I}\left(\left[\frac{j}{n}, \frac{i}{n}\right]\right) = 1, \ j = 0, \ldots, i - 1 \right\}
\]

(16)

holds for \( i = 1, \ldots, n \). Eq. (16) is often referred to as Bellman equation (Bellman, 1957), also known as optimal substructure property in computer science community (Cormen et al., 2009). It justifies the use of dynamic programming (Bellman, 1957; Bellman and Dreyfus, 1962) for computing the FDR-SMUCE estimate. In this way, the computation of \( \hat{K} \) is decomposed into smaller subproblems of determining \( \hat{K}[i] \)'s. For each subproblem, it boils down to checking the existence of constant functions which satisfy the multiscale side-constraint on \([j/n, i/n]\) i.e. \( \mathcal{I}(j/n, i/n) = 1 \).

The \( \hat{K}[i] \) is computed, via the recursive relation (16), as \( i \) increases from 1 to \( n \).
For each $i$, this involves the search space of $\{0, \ldots, i - 1\}$, which increases as $i$ approaches $n$. However, some of such searches are, actually, not necessary and can be pruned. This can be seen by rewriting the recursive relation in terms of the number of change-points. Let $A_0 := \{0\}$ and $B_0 := \{1, 2, \ldots, n\}$. For $k = 1, 2, \ldots$, let

$$r_k := \max \left\{ j : T^0_{[i/n,j/n]}(Y, c) \leq \max_m q_\alpha(m) \text{ for some } i \in A_{k-1}, c \in \mathbb{R} \right\},$$

$$A_k := \{ i \in B_{k-1} \cap [1, r_k] : I \left( \frac{[j/n,i/n]}{1} \right) = 1 \text{ for some } j \in A_{k-1} \},$$

$$B_k := B_{k-1} \setminus A_k.$$ 

Then $\hat{K} = k^* - 1$ with $A_{k^*} \supset n$. The reason for introducing $r_k$ is that there is no need to consider larger intervals if the multiscale side-constraint on an interval does not allow a constant signal even with the maximal penalty and the maximal quantile. Now for each $i$ we only need to search in a subset $B_k \cap [1, r_k]$ of $\{0, \ldots, i - 1\}$, where $k := k(i)$. The complexity for computing $\hat{K}$ is bounded from above by

$$O \left( \sum_{k=0}^{\hat{K}} |A_k| (r_{k+1} - \min A_k - |A_k|/2)^2 \right) \leq O \left( n \max_{0 \leq k \leq \hat{K}} (r_{k+1} - \min A_k - |A_k|/2)^2 \right).$$

The value $\max_{0 \leq k \leq \hat{K}} (r_{k+1} - \min A_k - |A_k|/2)^2$ depends on the signal and the noise. If the signal has many change-points and segments have similar lengths, it is a constant independent of $n$. And the higher noise level, the larger it might be. In such situation, the computation complexity is linear, although in the worst case it can be cubic in $n$.

Indeed, the searches of $\hat{K}$ and the maximum likelihood estimate can be done simultaneously, if we record the likelihood for each point $i$. The complexity is again bounded above by (17) but with a possibly larger constant. The memory complexity of the whole algorithm is linear, i.e. $O(n)$. We omit technical details. The pruned algorithm is implemented in the statistical software R in the package “FDRS” (http://www.stochastik.math.uni-goettingen.de/fdrs).

4. Simulations and Applications

4.1. Simulation study. We now investigate the performance of FDR-SMUCE under situations with various SNRs or different number of change-points, and compare it with SMUCE (Frick et al., 2014), PELT (Killick et al., 2012), BS (Scott and Knott, 1974), CBS (Olshen et al., 2004; Venkatraman and Olshen, 2007) and WBS (Fryzlewicz, 2014). As mentioned in Section 1, these methods represent powerful state of the art procedures from two different view points: one is exact and fast optimization based on dynamic programming, including PELT, and SMUCE; the other is greedy methods based on single change-point detection, including BS, CBS and WBS. Concerning implementation, we use R packages “PSCBS” for CBS, “wbs” for BS and WBS, “changepoint” for PELT, and an efficient implementation in “FDRS” for SMUCE. All the packages, except for “FDRS”, are available on CRAN. For both
SMUCE and FDR-SMUCE, we estimate the α-quantile thresholds by 5,000 Monte-Carlo simulations. The penalty $2 \log(K)$ is chosen for PELT, which is dubbed by “SIC1” in the codes provided by its authors, and works much better than the default choice. If we identify a change-point with two parameters (location and jump-size), this is the same as the Schwarz information criterion (SIC). We use the automatic rule, strengthened SIC, recommended by the author for WBS. The default parameter setting provided in the packages was used for BS and CBS. In all simulated scenarios, we assume that the noise level $\sigma$ is known beforehand. For quantitative evaluation, we will use mean integrated square error (MISE), mean integrated absolute error (MIAE), the FDR defined in (8) and V-measure (Rosenberg and Hirschberg 2007). The V-measure, a segmentation evaluation measure, takes values in $[0, 1]$, with a larger value indicating higher accuracy. It is based upon two criteria for clustering usefulness, homogeneity and completeness, which capture a clustering solution’s success in including all and only data points from a given class in a given cluster. In particular, a V-measure of 1 shows a perfect segmentation. All the experiments are repeated 1,000 times.

4.1.1. Varying noise level. Let us consider the impact of different noise levels. To this end, we use the mix signal (adopted from Fryzlewicz 2014, see Figure 8) with additive Gaussian noise, which is a mix of prominent change-points between short intervals and less prominent change-points between longer intervals. The noise level $\sigma$ varies from 1 to 8, and the number of observations $n = 560$. For SMUCE and FDR-SMUCE, we choose the same parameter $\alpha_S = \alpha = 0.15$. As in Figure 7, FDR-SMUCE outperforms others in all noise levels, in terms of V-measure, MISE, MIAE, and detection power measured by the average number of detected change-points. As indicated by number of detected change-points, MISE and MIAE, the PELT ranks second followed by WBS, then CBS, SMUCE and lastly BS. The same order of performance is also seen from V-measure up to $\sigma = 5$, but SMUCE deteriorates slower as noise level $\sigma$ increases and achieves a better V-measure than CBS when $\sigma \geq 6$ and than WBS at $\sigma = 8$. It is worth noting that the empirical FDR of FDR-SMUCE is around 0.1, far away from the theoretical bound $\approx 0.35$ (indicated by the dashed horizontal line in the lower-left panel). The CBS has the second largest empirical FDR, while that of PELT, SMUCE, BS and WBS are almost zero. Once the quantiles for SMUCE and FDR-SMUCE are simulated, they can be stored and used for later computations, which are therefore excluded from the recorded computation time. The computation time of FDR-SMUCE is similar to the fastest ones, namely PELT, BS and SMUCE, at $\sigma = 1$ and increases with the noise level $\sigma$. The FDR-SMUCE is faster than WBS and CBS in all scenarios. To have a closer examination, we also illustrate histograms of the locations of change-points, for $\sigma = 8$ in Figure 8. In this situation, the FDR-SMUCE has uniformly the largest detection power over all change-points.
4.1.2. Varying frequency of change-points. In order to evaluate the detection power as \( K \) increases, we employed the teeth signal (see Figure 4) with \( n = 3,000 \), and \( K = n^\theta \), \( \theta = 0.1, 0.2, \ldots, 0.9 \), as its SNR remains the same for different number of change-points. The same parameter \( \alpha_S = \alpha = 0.1 \) is chosen for SMUCE and FDR-SMUCE. The results are summarized in Figure 9. The FDR-SMUCE and PELT perform comparably good in all situations in terms of number of detected change-points, V-measure, MISE and MIAE. As shown by V-measure, CBS and WBS fail when \( \theta \geq 0.7 \), BS fails when \( \theta \geq 0.8 \) and SMUCE deteriorates at \( \theta = 0.9 \). Similar trend can also be seen from number of estimated change-points, MISE, and MIAE. It is interesting that the empirical FDR of FDR-SMUCE is closer to the theoretical bound \( \approx 0.22 \) as \( \theta \to 1 \). The empirical FDR of CBS is large when the change-points are sparse, and decreases as \( K \) increases, while PELT, SMUCE, BS and WBS have a relatively small FDR close to zero in all cases. The computation time of FDR-SMUCE decreases as \( K \) increases, and is similar to the fastest ones, i.e. SMUCE, PELT and BS, when \( \theta \geq 0.4 \).

4.2. Array CGH data. Identifying the chromosomal aberration locations in genomic DNA samples is crucial in understanding the pathogenesis of many diseases,
Figure 8. The histogram of the locations of change-points for mix signal with noise level $\sigma = 8$. As a benchmark, the true signal is plotted.

in particular, various cancers. Array comparative genomic hybridization (CGH) provides the means to quantitatively measure such changes in terms of DNA copy number [Pinkel et al., 1998]. The statistical task is to determine accurately the regions of changed copy number. The model (1) has been well-justified and studied in this problem (Olshen et al., 2004; Zhang and Siegmund, 2007; Tibshirani and Wang, 2008; Jeng et al., 2010). We compared the FDR-SMUCE with SMUCE, and CBS, which is designed for the analysis of array CGH data, on the Coriel data set from (Snijders et al., 2001). An outlier smoothing procedure introduced in (Olshen et al., 2004) was applied before segmentation. The CBS estimate was computed using default parameters provided in the package “PSCBS”. The estimated copy number variations by each method are plotted with the data (points) for cell line GM01524 in Figure 10. CBS finds the largest number of change-points, which is 17 in total. The FDR-SMUCE with $\beta = 0.01$ (i.e. $\alpha \approx 0.005$) finds 2 more change-point in chromosome 11 than the SMUCE with $\alpha_S = 0.1$, which are in accordance with CBS. With a larger $\beta = 0.05$, the FDR-SMUCE detects 2 additional bumps: one at chromosomes 1, 2 and the other at chromosome 18, which are also found by CBS. If we continue to increase $\beta$ until 0.37, the FDR-SMUCE will detect the whole set of change-points by CBS, together with some additional change-points. Recall from Theorem 2.2 that the parameter $\beta$ controls the FDR. Hence, by applying FDR-SMUCE with a range
of $\beta$, we could provide a hierarchy of significant statements to the change-points by CBS (and any other method). This suggests that the FDR-SMUCE can also be used as a tool to interpret results by other methods which do not automatically come with statistical guarantees for its segmentation.

4.3. Ion channel idealization. Being prominent components of the nervous system, ion channels play major roles in cellular transporting (Hille, 2001), which are helpful in diagonalizing many human diseases such as epilepsy, cardiac arrhythmias, etc. (Kass et al., 2005). The data analysis is to obtain information about channel characteristics and the effect of external stimuli by monitoring their behavior with respect to conductance and/or kinetics (Chung et al., 2007). The measuring process involves an analog low-pass filter prior to digitization. As analyzed by Hotz et al. (2013), a realistic model for observations is

$$Y_i = (\rho * \mu)\left(\frac{i}{\Delta}\right) + \tilde{\varepsilon}_i,$$

where $\Delta$ is the sampling rate, and the convolution kernel $\rho$ of the low-pass filter has compact support in an interval of length $L$, such that $\int \rho(t)dt = 1$. Being the independent and identically distributed (i.i.d.) Gaussian noise after the low-pass filter $\rho$, the $\tilde{\varepsilon}_i$ is still Gaussian with mean zero, but it is correlated. We observe that
\( \bar{\varepsilon}_i \) is independent of \( \bar{\varepsilon}_j \) if \(|i - j| > L\Delta\), and that \( \rho \cdot \mu \) will equal to \( \mu \) on an interval of length \( T - L \) if \( \mu \) is constant for some time \( T \geq L \). Thus, if we undersample the observations at rate \( L\Delta \), the reduced data
\[
Y_{iL\Delta} = (\rho \cdot \mu)(iL) + \bar{\varepsilon}_{iL\Delta}
\] (19)
will satisfy our model (1).

We compare SMUCE (\( \alpha_S = 0.05 \)) and FDR-SMUCE (\( \beta = 0.05 \)) estimates on the subsampled data (19). As a benchmark, the *jump segmentation by multiresolution filter* (J-SMURF) estimate, introduced by Hotz et al. (2013), was computed on the full data (18) with \( \alpha = 0.05 \), which takes into account the dependence structure of the noise. The implementation of J-SMURF is provided in R package “stepR”, available from http://www.stochastik.math.uni-goettingen.de/smuce. Figure 11 shows a characteristic conductance trace of gramicidin A with a typical SNR, \( L = 30 \) and \( \Delta = 0.1 \) ms. Compared with SMUCE, the additional change-points (indicated by vertical lines) detected by FDR-SMUCE are clearly to be correct, which is reassured by J-SMURF. By the theoretical bound of FDR (see Theorem 2.2), it is known that out of 10 change-points detected by FDR-SMUCE there are at most 0.55 false ones.
on average. This shows that the ideas underlying FDR-SMUCE are very helpful to
detect change-points on various scales as it is to be expected for the investigated
gramicidin channel (for an explanation see [Lotz et al., 2013]). Therefore, it is promis-
ing to modify FDR-SMUCE such that local dependencies are taken fully into account,
which is postponed to future research.

![Graph showing conductance over time](image)

**Figure 11.** The time trace of conductance for gramicidin A. The J-
SMURF estimate (dashed line) together with data (points) is shown in
each panel.

5. Conclusion

In this work we proposed a multiple change-point estimate FDR-SMUCE, bearing
similar flavor to SMUCE, however with notable differences. Key is the relaxation
of the family-wise error to the false discovery rate. The new FDR multiscale side-
constraint succeeded in overcoming the conservativeness of SMUCE. By experiments
on both simulation and real data, the FDR-SMUCE shows a significant increase
in the detection power and meanwhile with controlled accuracy. A theoretical
bound is provided for its FDR, which provides a meaningful interpretation of the
only user-specified parameter $\alpha$.

Our method is not confined to i.i.d. Gaussian observations, although we restricted
our presentation to this in order to highlight the main ideas more concisely. Obviously,
it can be extended to more general additive errors, because the proof of Lemma A.1
only relies on Gaussianity for the independence of the residuals and the mean. In
the case of different models, e.g. exponential family regression, we believe that one
can argue in similar lines as in the proof of Theorem 2.2 but results will only hold
asymptotically. This, however, is above the scope of the paper, and postponed to
further research. Also as we have applied the CBS outlier smoothing procedure to the array CGH data, it might be of interest to have more robust versions of FDR-SMUCE. To this end, e.g. local median, instead of local mean, might provide useful results. Alternatively, one may transform this into a Bernoulli regression problem (see Dümbgen and Kovac, 2009; Frick et al., 2014), which might be interesting for future research.

**Appendix A. Appendix**

The proof of Theorem 2.2 relies on two lemmata. As a convention, all the results are concerning the FDR-SMUCE \( \hat{\mu} \) in (13) without explicit statement. The first one gives a bound for the expected number of false discoveries (FD) given no true discoveries (TD = 0), see Section 1 for the definitions.

**Lemma A.1.** Let \( Y \) be observations from model (1). Then for \( 0 < \alpha < 1/3 \)

\[
E[FD(\alpha)|TD(\alpha) = 0] \leq \frac{2\alpha}{1-3\alpha} =: G(\alpha).
\]

**Proof.** Note that it suffices to prove the result for a constant signal, which we assume w.l.o.g. to be constant zero. The proof is then based on the following observation. Assume there exists an estimate \( \tilde{\mu} = \sum_{k=0}^{\tilde{K}} c_k I_k \) with \( (\tilde{K}+1) \) segments \( \bigcup_{k=0}^{\tilde{K}} \tilde{I}_k = [0,1) \), which fulfills the multiscale side-constraint \( \mathcal{C}_{\tilde{K}} \) in (11). Then, the FD of FDR-SMUCE is bounded by \( \tilde{K} \), since it minimizes the number of change-points \( k \) among all nonempty \( \mathcal{C}_k \)'s. We will prove the result by constructing such an estimate \( \tilde{\mu} \) and show that \( E[\tilde{K}] \leq 2\alpha/(1-3\alpha) \). The estimate \( \tilde{\mu} \) is given by an iterative rule to include change-points until the multiscale side-constraint \( \mathcal{C}_{\tilde{K}} \) is fulfilled.

We first check the whole interval \([0,1)\) whether its mean value \( \bar{Y} \) satisfies the multiscale side-constraint. If \( T_{[0,1)}(Y, \bar{Y}) \leq q_\alpha(n) \), then \( \tilde{\mu} := \bar{Y} 1_{[0,1)} \). Otherwise, we randomly choose \( i^* \) and \( j^* \) from

\[
\left\{ (i,j) : t_{[0,1)} \left( \left[ \frac{i}{n}, \frac{j}{n} \right] \right) := \frac{\sum_{l=i}^{j} (Y_l - \bar{Y})}{\sigma \sqrt{j-i+1}} - \text{pen} \left( \frac{j-i+1}{n} \right) - q_\alpha(n) > 0 \right\},
\]

according to any distribution which is independent of the values of \( t_{[0,1)}(\left[ i/n, j/n \right]) \)'s. Then we check intervals \([0, i^*/n)\), \([i^*/n, j^*/n)\) and \((j^*/n, 1)\) individually, and split them further in the same manner if necessary. This procedure is repeated until on each resulting interval \( I \) its mean value \( \bar{Y}_I \) satisfies the multiscale side-constraint, i.e. \( T_I(Y, \bar{Y}_I) \leq q_\alpha(|I|) \). Finally, \( \tilde{\mu} := \sum_I \bar{Y}_I 1_I \).

Let \( D_k \) denote the number of change-points (discoveries) and \( S_k \) the number of segments introduced in the \( k \)-th step. We make the convention that \( D_k = S_k = 0 \) if the procedure stops before the \( k \)-th step. It follows from \( \mathbb{P}\left\{ T_{[0,1)}(Y, \bar{Y}) > q_\alpha(n) \right\} \leq \alpha \), cf. (10), (recall \( Y_i = \varepsilon_i \) here) that

\[
E[D_1] \leq 2\alpha \text{ and } E[S_1] \leq 3\alpha.
\]
Now we consider the three intervals \( I_1 = [0, i^*/n], I_2 = [i^*/n, j^*/n] \) and \( I_3 = (j^*/n, 1) \) and bound the probability of further splitting them into smaller intervals. It will be shown that
\[
\mathbb{P}\left\{ T_{i_k}(Y, \hat{Y}_{i_k}) > q_\alpha(|I_k|) \mid T_{(0,1)}(Y, \bar{Y}) > q_\alpha(n) \right\} \leq \alpha \quad \text{for } k = 1, 2, 3.
\]

Given \( I_2 = [i/n, j/n] \), the random variable \( T_{i_k}(Y, \hat{Y}_{i_k}) \) depends only on \( \{Y_i - \bar{Y}_{i_k}, i/n \in I_k\} \), which is independent of \( Y \) and \( \bar{Y}_{i_2} \). It follows from (20) that \( t_{(0,1)}(I_2) \) depends only on \( Y \) and \( \bar{Y}_{i_2} \). Thus \( T_{i_k}(Y, \hat{Y}_{i_k}) \) is independent of \( t_{(0,1)}(I_2) \) conditioned on \( I_2 \).

\[
\begin{align*}
\mathbb{P}\left\{ T_{i_k}(Y, \hat{Y}_{i_k}) > q_\alpha(|I_k|) \mid T_{(0,1)}(Y, \bar{Y}) > q_\alpha(n) \right\} &= \sum_{0 \leq i < j < n} \mathbb{P}\left\{ T_{i_k}(Y, \hat{Y}_{i_k}) > q_\alpha(|I_k|) \mid t_{(0,1)}(I_2) > 0, I_2 = \left[ \frac{i}{n}, \frac{j}{n} \right] \right\} \\
&\times \mathbb{P}\left\{ I_2 = \left[ \frac{i}{n}, \frac{j}{n} \right] \mid t_{(0,1)}(Y, \bar{Y}) > q_\alpha(n) \right\} \\
&= \sum_{0 \leq i < j < n} \mathbb{P}\left\{ T_{i_k}(Y, \hat{Y}_{i_k}) > q_\alpha(|I_k|) \mid I_2 = \left[ \frac{i}{n}, \frac{j}{n} \right] \right\} \cdot \mathbb{P}\left\{ I_2 = \left[ \frac{i}{n}, \frac{j}{n} \right] \mid t_{(0,1)}(Y, \bar{Y}) > q_\alpha(n) \right\} \\
&\leq \sum_{0 \leq i < j < n} \alpha \mathbb{P}\left\{ I_2 = \left[ \frac{i}{n}, \frac{j}{n} \right] \mid t_{(0,1)}(Y, \bar{Y}) > q_\alpha(n) \right\} \leq \alpha.
\end{align*}
\]

It follows that
\[
\mathbb{E}[D_2|S_1] \leq 2\alpha S_1 \quad \text{and} \quad \mathbb{E}[S_2|S_1] \leq 3\alpha S_1.
\]

Using the same line of argumentation we find in general that
\[
\mathbb{E}[D_k|S_{k-1}] \leq 2\alpha S_{k-1} \quad \text{and} \quad \mathbb{E}[S_k|S_{k-1}] \leq 3\alpha S_{k-1}.
\]

It implies
\[
\mathbb{E}[D_k] = \mathbb{E}[\mathbb{E}[D_k|S_{k-1}]] \leq 2\alpha \mathbb{E}[S_{k-1}] = 2\alpha \mathbb{E}[S_{k-1}|S_{k-2}] \leq 2\alpha \cdot 3\alpha \mathbb{E}[S_{k-2}] \leq 2\alpha (3\alpha)^{k-1}.
\]

Hence,
\[
\mathbb{E}[FD] \leq \mathbb{E}[\bar{K}] = \mathbb{E}\left[ \sum_{k=1}^{\infty} D_k \right] = \sum_{k=1}^{\infty} \mathbb{E}[D_k] \leq \sum_{k=1}^{\infty} 2\alpha (3\alpha)^{k-1} = \frac{2\alpha}{1 - 3\alpha}.
\]

The next lemma shows the expected FD conditioned on TD.

**Lemma A.2.** \( \mathbb{E}[FD(\alpha)|TD(\alpha) = \kappa] \leq (\kappa + 1)\mathbb{E}[FD(\alpha)] \leq (\kappa + 1)G(\alpha).\)

**Proof.**

\[
\mathbb{E}[FD | TD = \kappa] = \sum_{i_1 < \cdots < i_\kappa} \mathbb{E}[FD | \tau_{i_1}, \ldots, \tau_{i_\kappa} \text{ are true, } TD = \kappa] \mathbb{P}\{ \tau_{i_1}, \ldots, \tau_{i_\kappa} \text{ are true } | TD = \kappa\}
\]

\[
= \sum_{i_1 < \cdots < i_\kappa} \sum_{j=0}^{\kappa} \mathbb{E}[FD(\tau_{i_j}, \hat{\tau}_{i_{j+1}}) | \tau_{i_1}, \ldots, \tau_{i_\kappa} \text{ are true, } TD = \kappa] \mathbb{P}\{ \tau_{i_1}, \ldots, \tau_{i_\kappa} \text{ are true } | TD = \kappa\},
\]
where we set $\tau_{i_0} := 0$ and $\tau_{i_{\kappa+1}} := 1$. Note that there is no true discovery on $(\hat{\tau}_{ij}, \hat{\tau}_{ij+1})$, $j = 0, \ldots, \kappa$. By applying Lemma A.1 to each segment on $(\hat{\tau}_{ij}, \hat{\tau}_{ij+1})$, we have

\[
E[FD | TD = \kappa] \leq \sum_{i_1 < \cdots < i_\kappa} \sum_{j=0}^\kappa G(\alpha) \mathbb{P}\{\hat{\tau}_{i_1}, \ldots, \hat{\tau}_{i_\kappa} \text{ are true} | TD = \kappa\} \leq (\kappa + 1)G(\alpha).
\]

\[\square\]

Now we are ready to prove Theorem 2.2.

\textbf{Proof of Theorem 2.2.} For random variables $X, Y$ and $Z = X + Y$ we find by Jensen’s inequality that

\[
E\left[ E\left[ \frac{X}{Z} | Y \right] \right] \leq E\left[ \frac{E[XY]}{Y + E[X|Y]} \right].
\]

We set $X = FD$, $Y = TD + 1$. Together with Lemma A.2 this yields that

\[
FDR = E\left[ \frac{X}{Z} \right] = E\left[ E\left[ \frac{X}{Z} | Y \right] \right] \leq E\left[ \frac{E[XY]}{Y + E[X|Y]} \right] \leq \frac{G(\alpha)}{1 + G(\alpha)} = \frac{2\alpha}{1 - \alpha}.
\]

\[\square\]

\textbf{Acknowledgement}

H.L. acknowledges financial support by the China Scholarship Council, A.M. and H.S. DFG CRC 755, 803 and DFG FOR 916.

\textbf{References}

Auger, I. E. and Lawrence, C. E. (1989). Algorithms for the optimal identification of segment neighborhoods. \textit{Bull. Math. Biol.}, 51(1):39–54.

Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. \textit{Econometrica}, 66(1):47–78.

Barry, D. and Hartigan, J. A. (1993). A Bayesian analysis for change point problems. \textit{J. Am. Stat. Assoc.}, 88(421):309–319.

Bellman, R. (1957). \textit{Dynamic Programming}. Princeton University Press, Princeton, NJ, USA.

Bellman, R. E. and Dreyfus, S. E. (1962). \textit{Applied Dynamic Programming}. Princeton, NJ: Princeton University Press.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. \textit{J. R. Statist. Soc. Ser. B}, 57(1):289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. \textit{Ann. Statist.}, 29(4):1165–88.

Birgé, L. and Massart, P. (2006). Minimal penalties for Gaussian model selection. \textit{Probab. Theory Related Fields}, 138:33–73.

Blythe, D., von Bumau, P., Meinecke, F., and Muller, K. (2012). Feature extraction for change-point detection using stationary subspace analysis. \textit{IEEE Trans. Neur. Netwrks Learn. Syst.}, 23(4):631–643.
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., R.K., M., and Mueller, H.-G. (2000). Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation. *Biometrika*, 87(2):301–314.

Chung, S.-H., Andersen, O. S., and Krishnamurthy, V. (2007). *Biological Membrane Ion Channels*. Springer.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2009). *Introduction to Algorithms*. MIT press, third edition.

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

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

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

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

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

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.

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

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. *J. Computnl Graph. Statist.*, 17(1):201–224.

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

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

Guo, W. and Sarkar, S. (2013). Adaptive controls of fwer and fdr under block dependence. *Unpublished manuscript*. web.njit.edu/~wguo/Guo
Hao, N., Niu, Y., and Zhang, H. (2013). Multiple change-point detection via a screening and ranking algorithm. *Statist. Sinica, 23*(4):1553–1572.

Harchaoui, Z. and Lévy-Leduc, C. (2008). Catching change-points with lasso. *Adv. in Neur. Inform. Processing Syst., 20*:161–168.

Harchaoui, Z. and Lévy-Leduc, C. (2010). Multiple change-point estimation with a total variation penalty. *J. Am. Statist. Ass., 105*(492):1480–1493.

Hille, B. (2001). *Ion Channels of Excitable Membranes*. Sinauer Sunderland, MA.

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

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.

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 Process. Lett., 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.

Kass, R. S. et al. (2005). The channelopathies: novel insights into molecular and genetic mechanisms of human disease. *J. Clin. Invest., 115*(8):1986–1989.

Killick, R., Fearnhead, P., and Eckley, I. A. (2012). Optimal Detection of Change-points With a Linear Computational Cost. *J. Am. Statist. Ass., 107*(500):1590–1598.

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.

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.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.-L., Chen, C., Zhai, Y., et al. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet., 20*(2):207–211.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist., 12*(4):1215–1230.

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. (2013). Optimal detection of a jump in the intensity of a Poisson process or in a density with likelihood ratio statistics. *Scand. J. Stat., 40*:752–769.
Rosenberg, A. and Hirschberg, J. (2007). V-measure: a conditional entropy-based external cluster evaluation measures. *Proc. Conf. Empirical Methods Natural Lang. Process.*, (June):410–420.

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

Siegmund, D. (2013). Change-points: from sequential detection to biology and back. *Sequent. Anal.*, 32(1):2–14.

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

Siegmund, D. O., Zhang, N. R., and Yakir, B. (2011). False discovery rate for scanning statistics. *Biometrika*, 98(4):979–985.

Sieling, H. (2013). *Statistical Multiscale Segmentation: Inference, Algorithms and Applications*. PhD thesis, University of Göttingen, Germany.

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. *Nat. Genet.*, 29(3):263–264.

Spokoiny, V. (2009). Multiscale local change point detection with applications to value-at-risk. *Ann. Statist.*, 37(3):1405–1436.

Storath, M., Weinmann, A., and Demaret, L. (2014). Jump-sparse and sparse recovery using potts functionals. *IEEE Trans. Signal Process.*, 62(14):3654–3666.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Statist. Soc. B*, 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.

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

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

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.