Published in final edited form as:

Commun Stat Simul Comput. 2017 ; 46(1): 127–144. doi:10.1080/03610918.2014.957842.

FDR control of detected regions by multiscale matched filtering

Nezamoddin N. Kachouie,

Department of Mathematical Sciences, Florida Institute of Technology, Melbourne, FL, USA.

Xihong Lin,

Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA.

Armin Schwartzman

Department of Statistics, North Carolina State University, Raleigh, NC, USA.

Abstract

Feature extraction from observed noisy samples is a common important problem in statistics and engineering. This paper presents a novel general statistical approach to the region detection problem in long data sequences. The proposed technique is a multi-scale kernel regression in conjunction with statistical multiple testing for region detection while controlling the false discovery rate (FDR) and maximizing the signal to noise ratio (SNR) via matched filtering. This is achieved by considering a one-dimensional (1D) region detection problem as its equivalent 0D (zero dimensional) peak detection problem. The detection method does not require *a priori* knowledge of the shape of the non-zero regions. However, if the shape of the non-zero regions is known *a priori*, e.g. rectangular pulse, the signal regions can also be reconstructed from the detected peaks, seen as their topological point representatives. Simulations show that the method can effectively perform signal detection and reconstruction in the simulated data under high noise conditions, while controlling the FDR of detected regions and their reconstructed length.

Index Terms

False discovery rate; Feature extraction; Local polynomial regression; Kernel regression; Matched filtering; Multiple testing; Region detection; Signal reconstruction

1. INTRODUCTION

Extracting the key features of the original signal from observed noisy samples is a challenging problem in statistical signal processing. Features of interest can be a set of peaks, a set of change points, or a set of non-zero regions. Peak identification [1–6], change point localization [7–11], and region detection [12–16] are common problems in statistics, engineering, medicine, and biology, requiring the development of robust and efficient techniques. Among the features of interest, peaks and change points are zero-dimensional (0D) features while regions are one-dimensional (1D) and therefore harder to detect.

In previous work [17, 18], the peak detection problem was addressed, highlighting the importance of multiple testing considerations in order to achieve proper error control in this type of problems. In this paper, a novel statistical signal processing approach is proposed to address the more challenging problem of region detection and reconstruction. The contributions of the current paper are: 1) Addressing region detection by peak detection; 2) Multiscale design to deal with regions of different sizes; 3) Automatic matched filtering to choose the best match representing the hidden region size; 4) Signal reconstruction.

Rather than estimating the underlying signal, the problem of identifying non-zero signal regions is solved as a detection problem. The proposed approach is topological in the sense that it addresses the one-dimensional (1D) problem of detecting non-zero signal regions by a zero dimensional (0D) peak detection problem. The solution to the 0D problem is then achieved through a multiple testing approach [17–20], where each candidate peak (representing a candidate region) is tested for significance and the significance threshold is adjusted automatically in order to control the global false discovery rate (FDR).

Representation of regions by peaks for detection purposes is achieved by matched filtering via multi-scale kernel regression. This is based on the idea that convolution of a constant or unimodal non-zero signal region with a symmetric unimodal kernel produces a unimodal peak at the center of the region. Furthermore, the SNR of the peak is maximized when the kernel bandwidth matches the size of the region (up to a proportionality constant that depends on the kernel shape). Therefore, the peak with the strongest SNR among all bandwidths within a spatial neighborhood becomes the best candidate for detection of an underlying non-zero region.

If in addition, the shape of the non-zero regions is known *a priori*, the signal regions can be reconstructed from the detected peaks, seen as their topological point representatives. In particular, constant non-zero regions correspond to a rectangular signal shape. For such case, the maximal SNR of a peak is proportional to the height of the underlying constant signal, and hence the height and size of a constant non-zero signal region can be recovered from the height of the detected peak and its best matching bandwidth. The best matching bandwidth is the bandwidth that gives the highest estimated signal to noise ratio for the detected peak.

The performance of the proposed method is evaluated using simulated data inspired by genomic data analysis. Simulations show that the method can effectively perform signal detection and reconstruction under high noise conditions. It is shown too that the use of a multiple testing correction in the algorithm achieves the desired effect of controlling the false discovery rate (FDR) of detected regions as well as that of their reconstructed length.

2. METHOD

This approach proposes a multi-scale kernel regression in conjunction with statistical multiple testing to detect non-zero signal regions. The key idea is to use kernel regression to transform the region detection problem to a peak detection problem, where multiple testing is applied to identify the significant peaks while controlling type-I error rate [17–20]. Each

peak is a 0D representative of a 1D region. The main steps of the algorithm are summarized here (Fig. 1(A)).

1. Multi-scale kernel regression (Sec. 3.1):

To smooth observed noisy samples, a set of kernels with the same shape but different bandwidths is applied. Although the region sizes are unknown, in practice one often has some rough prior information about the range of region sizes that s/he is interested in detecting if they exist in the signal. The selected range of region sizes dictates the number of bandwidths in the kernel bank. Smaller range for the region size may suggest smaller number of bandwidths in the kernel bank, making the process more efficient.

2. Locating candidate peaks (Sec. 3.2):

After kernel regression, a high local maximum indicates a close match between a specific bandwidth and the underlying signal at that location. The set of peaks obtained with all bandwidths constitutes a set (hereafter candidate set) where each located peak, together with its location, height, and associated bandwidth, is a potential candidate to represent an underlying signal region. Note that, in this paper, kernel regression does not estimate the underlying function directly, but its purpose is to identify peaks as potential candidates for detection.

3. Bandwidth matching and pruning (Sec. 3.3):

For any true underlying non-zero region, bandwidths smaller than the best matching bandwidth may produce too many peaks while larger bandwidths may yield no peak representatives at all. This redundancy is eliminated by selecting the peak (and its associated bandwidth) that maximizes the SNR in the local neighborhood. A subset (hereafter selected set) is chosen by pruning the redundant peaks of the candidate set. In this way, each selected peak is the best single representative of a candidate region. Its location estimates the center of the region and the associated bandwidth is used to estimate its length. Pruning avoids overlap between detected regions.

4. Region identification by statistical multiple testing (Sec. 4):

To differentiate where the signal is zero and where it is not, a p-value is computed and assigned to each selected peak. Consequently significant p-values are identified as a subset of the selected set (hereafter significant set) by automatically setting a threshold to control the FDR using the Benjamini and Hochberg (BH) algorithm [19]. The identified significant peaks and their associated bandwidths locate the significant regions.

5. Signal reconstruction (Sec. 5):

If the shape of the non-zero regions is assumed to be known a priori, the signal is reconstructed as a combination of signal segments of that shape, with lengths and heights set by the corresponding best matching kernel bandwidths.

3. PEAK DETECTION

In this section we describe how a pool of candidate peaks will be obtained. The detailed steps of the peak detection process follow.

3.1 Multi-Scale Kernel regression

Assume that the underlying signal is a pulse train with variable height and length that can be represented by a set of non-zero non-overlapping regions as

$$\mu(x) = \sum_{t=1}^T \mu_t = \sum_{t=1}^T \beta_t \cdot \zeta(\tau_t - \delta_t, \tau_t + \delta_t), \quad (1)$$

where the unknown constants T , β_t , τ_t and $2\delta_t$ are respectively the number of non-zero regions, and the height, center, and length of the region indexed by t . The pulse shape ζ is assumed to be constant or unimodal and it does not need to be known for detection purposes [17]. In general, the coefficients β_t can be positive or negative. The pulse train (1) is corrupted with additive zero-mean Gaussian noise. The n pairs of observations

$$(x_i, Y_i), \quad i = 1, 2, \dots, n \quad (2)$$

consist of a response variable Y and a location x and are related by the signal-plus-noise model

$$Y_i = \mu(x_i) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad i \in [1, n] \quad (3)$$

where Y_i , $i \in [1, n]$ are observed noisy samples, the locations x_i , $i \in [1, n]$ need not be equally spaced, and ε_i , $i \in [1, n]$ is independent and identically distributed Gaussian noise $N(0, \sigma^2)$.

The goal is to recover the underlying structure (1) as a set of regions where the signal is non-zero while controlling the probability of falsely detected regions. We use kernel regression to transform the underlying non-zero signal regions into smooth peaks, which can be more easily detected. We apply the Nadaraya-Watson [21] linear operator:

$$\widehat{\mu}_\gamma(x) = \sum_{i=1}^n l_{\gamma,i}(x) Y_i = \sum_{i=1}^n l_{\gamma,i}(x) \mu(x_i) + \sum_{i=1}^n l_{\gamma,i}(x) \varepsilon_i = \mu_\gamma(x) + \varepsilon_\gamma(x) \quad (4)$$

where for each $x \in \mathbb{R}$, the i^{th} weight $l_{\gamma,i}(x)$ is the scalar

$$l_{\gamma,i}(x) = \frac{K_\gamma(x - x_i)}{\sum_{j=1}^n K_\gamma(x - x_j)}, \quad i = 1, 2, \dots, n \quad (5)$$

and $K_\gamma(x)$ is a kernel with bandwidth γ

$$K_\gamma(x) = \frac{1}{\gamma} K\left(\frac{x}{\gamma}\right), \quad \int K(x) dx = 1. \quad (6)$$

The smooth function $\widehat{\mu}_\gamma(x)$ estimates the smoothed signal $\mu_\gamma(x)$ defined in (4). The mean and variance of this estimator are

$$\begin{aligned} E(\widehat{\mu}_\gamma(x)) &= \sum_{i=1}^n l_{\gamma,i}(x) \mu(x_i) \\ \text{var}(\widehat{\mu}_\gamma(x)) &= \sigma^2 \sum_{i=1}^n l_{\gamma,i}(x)^2 = \sigma^2 \|l_\gamma(x)\|^2 \end{aligned} \quad (7)$$

Where $l_{\gamma,i}(x)$ given in (5) are elements of vector $\underline{l}_\gamma(x)$ defined by:

$$\underline{l}_\gamma(x) = (l_{\gamma,1}(x), l_{\gamma,2}(x), \dots, l_{\gamma,n}(x))^T \quad (8)$$

The kernel shape is chosen to satisfy the following constraints [21]:

- a. To have a finite support,
- b. To have continuous first and second derivatives within its support and at the boundaries.

Condition (a) limits the spread of the signal regions into their neighborhood as a result of kernel regression. Condition (b) is necessary for locating local maxima, as explained in the next section. As an example, in the simulations and data analysis we use a bank of tricube kernels (Fig. 1(B)) based on the tricube function

$$K_{\gamma=1}(x) = (1 - |x|^3)^3 \cdot \mathbf{1}_{\{|x| \leq 1\}} \quad (9)$$

which satisfies the aforementioned conditions. The choice of kernel K is not a major concern. Different kernels usually provide estimates which are numerically very similar. Theoretical calculations show that the risk is very insensitive to the choice of kernel [21, 22]. The choice of bandwidth γ which controls the amount of smoothing is much more important than the kernel shape [21]. Therefore we use tricube kernel which has a finite support and satisfies the required conditions for region detection. At the same time to choose the bandwidth, a multiscale approach in conjunction with matched filtering is proposed. In this way, each underlying region will be detected by the bandwidth γ that best matches the region size.

Let Γ be a discrete and limited set of bandwidths γ where $b1 < \gamma < b2$ for two pre-specified constants $b1$ and $b2$. A bank of smoothed signals is created by convolving the kernel K_γ with the noisy observations for the different values of γ in Γ . The range of bandwidths is chosen based on prior information from the application at hand. In principle, it should cover the range of region lengths to be detected, i.e. $b1 < \delta_t < b2$ for all t .

3.2 Locating Candidate Peaks

Because the underlying signal regions can be positive or negative, we are interested in locating both local maxima and local minima (together, local optima) of the estimated smooth signal $\widehat{\mu}_\gamma(x)$. The positive peaks of the estimated smooth signal $\widehat{\mu}_\gamma(x)$ are defined by the set of locations of local maxima

$$O_{\gamma, \max} = \left\{ \tau_{\gamma, m} \left| \begin{array}{l} \widehat{\mu}_\gamma'(\tau_{\gamma, m}) = \frac{d\widehat{\mu}_\gamma}{dx} \Big|_{t=\tau_{\gamma, m}} = 0 \\ \widehat{\mu}_\gamma''(\tau_{\gamma, m}) = \frac{d^2\widehat{\mu}_\gamma}{dx^2} \Big|_{t=\tau_{\gamma, m}} < 0 \end{array} \right. \right\}, m \in \{1, \dots, N_\gamma\} \quad (10)$$

Where N_γ is the number of located maxima along bandwidth γ . Note that each peak has an associated bandwidth and height. A pool of all local maxima located along different bandwidths where $b1 < \gamma < b2$ is then constructed as:

$$O_{Max} = \bigcup_{\gamma} O_{\gamma, \max} \quad (11)$$

Similarly, a pool of all local minima $O_{Min} = \bigcup_{\gamma} O_{\gamma, \min}$ located along different bandwidths is constructed, where $O_{\gamma, \min}$ contains local minima of the estimated smooth signal $\widehat{\mu}_\gamma(x)$ for bandwidth γ . Minima of $\widehat{\mu}_\gamma(x)$ are identified by locating the maxima of $-\widehat{\mu}_\gamma(x)$ and then negating their signs.

3.3 Pruning

A non-zero region of the underlying signal $\mu_\gamma(x)$ will typically produce several peaks of the estimated smooth signal $\widehat{\mu}_\gamma(x)$ at different bandwidths γ . These peaks along different bandwidths are related to the same underlying region and only one such representative is needed. The idea of pruning is to select the peak with the highest SNR within the neighborhood, as defined by the bandwidth, as a single candidate for that potential region. The SNR $\Omega_{\gamma, m}$ associated with a local maximum at location $\tau_{\gamma, m}$ in O_{Max} is defined as

$$\Omega_{\gamma, m}(\tau_{\gamma, m}) = \frac{\widehat{\mu}_\gamma(\tau_{\gamma, m})}{\sigma_\gamma(\tau_{\gamma, m})} \quad (12)$$

where $\widehat{\mu}_\gamma$ is the estimated smooth signal by kernel K_γ , and σ_γ is the noise standard deviation after kernel regression, computed using (4) as

$$\sigma_\gamma^2(x) = \text{var}[\varepsilon_\gamma(x)] = \text{var}\left[\sum_{i=1}^n l_{\gamma,i}(x) \varepsilon_i\right] = \sigma^2 \|l_\gamma(x)\|^2 \quad (13)$$

where the noise variance σ^2 is estimated using regression residuals and we have $\widehat{\sigma}_\gamma^2 = \widehat{\sigma}^2 \|l_\gamma(x)\|^2$. For the set of positive candidate peaks O_{Max} , an iterative pruning procedure is applied whose objective is to select a subset $O_{Max,Prune}$ (hereafter called selected set), representing non-overlapping regions. Let Ω_{Max} be the set of computed SNRs corresponding to the local maxima O_{Max} . The pruning procedure is as follows:

Repeat while Ω_{Max} is not empty:

1. Locate the largest SNR $\Omega_{\gamma,i}$ in Ω_{Max} and add its associated peak location $\tau_{\gamma,i}$ to the selected set $O_{Max,Prune}$.
2. Remove $\Omega_{\gamma,i}$ and any other SNR with location in the interval $[\tau_{\gamma,i} - \gamma_i, \tau_{\gamma,i} + \gamma_i]$.

The pruning procedure ensures that the selected peaks satisfy

$$d(\tau_{\gamma,i}, \tau_{\gamma,j}) \geq \gamma_i + \gamma_j \quad (14)$$

where d is the distance between two neighboring peak locations $\tau_{\gamma,i}$ and $\tau_{\gamma,j}$ in $O_{Max,Prune}$ with associated bandwidths γ_i and γ_j (and associated SNRs $\Omega_{\gamma,i}$ and $\Omega_{\gamma,j}$) respectively. As a consequence, the detected signal is guaranteed to have no overlapping non-zero regions. Similarly, a reduced set of peaks $O_{Min,Prune}$ is constructed containing the selected minima. Next we explain how the selected peaks are tested to identify the significant peaks.

4. MULTIPLE TESTING

Statistical multiple testing is used to detect significant peaks. After pruning, first, p-values are computed for the selected peaks. Then, to control the FDR, p-values of the selected peaks are adjusted using the distribution of height of local peaks.

4.1 Computing p-values

Statistical multiple testing is used to detect significant peaks. After pruning, p-values are computed for the selected maxima, each with an associated location and bandwidth, creating a pool of p-values P^+ regarding $O_{Max,Prune}$. Similarly, a pool of p-values P^- is created regarding the set of selected minima $O_{Min,Prune}$. Each p-value is calculated under the complete null hypothesis of no signal everywhere

$$\mu(x) = 0, \quad \forall x \quad (15)$$

so that $\widehat{\mu}_\gamma(x) = \varepsilon_\gamma(x)$. The p-value of a selected local maximum at a given location is defined as the probability, under the null hypothesis, that a local maximum at that location is as high or higher than the one observed. Let $F_\gamma(u | \tau_{\gamma,i})$ be the complementary cumulative distribution function (ccdf) or survival function of the height u of a local maximum of the process $\varepsilon_\gamma(x)$ at the local maximum $\tau_{\gamma,i} \in O_{\gamma,max}$:

$$F_\gamma(u, \tau_{\gamma,i}) = P\left\{\varepsilon_\gamma(\tau_{\gamma,i}) > u \mid \tau_{\gamma,i} \in O_{\gamma,max}\right\}. \quad (16)$$

Then the p-values of the selected local maxima $O_{Max,Prune}$ are calculated via:

$$p(\tau_{\gamma,m} \in O_{Max,Prune}) = F_\gamma(\widehat{\mu}_\gamma(\tau_{\gamma,m}), \tau_{\gamma,m}) \quad (17)$$

Similarly, p-values of selected local minima ($O_{Min,Prune}$) are calculated as the p-values of the local maxima of the process $-\widehat{\mu}_\gamma(x)$. The required ccdf $F_\gamma(u, x)$ is evaluated as follows.

4.2 Distribution of the height of local maxima

For a stationary ergodic process, the distribution of the heights of the local maxima has a closed-form expression that depends on the first three spectral moments of the process [20]. The random process $\varepsilon_\gamma(x)$ satisfies these conditions if the locations x are equally spaced. If not, simulations (results not shown) indicate that the formula continues to hold approximately when applied using the local spectral moments, as expected from the local nature of local maxima. Specifically, the distribution $F_\gamma(u, x)$ of the height of a local maximum at any point x is computed as

$$F_\gamma(u, x) = 1 - \Phi\left(u\sqrt{\frac{\lambda_{4,\gamma}}{\Delta}}\right) + \left(\sqrt{\frac{2\pi\lambda_{2,\gamma}^2}{\lambda_{4,\gamma}\sigma_\gamma^2}}\right)\phi\left(\frac{u}{\sigma_\gamma}\right)\Phi\left(u\sqrt{\frac{\lambda_{2,\gamma}^2}{\Delta\sigma_\gamma^2}}\right) \quad (18)$$

where the parameters σ_γ , $\lambda_{2,\gamma}$, $\lambda_{4,\gamma}$, and Δ depend on x and are given by

$$\begin{aligned} \Delta(x) &= \sigma_\gamma^2(x)\lambda_{4,\gamma}(x) - \lambda_{2,\gamma}^2(x), \\ \sigma_\gamma^2(x) &= \text{var}[\varepsilon_\gamma(x)], \\ \lambda_{2,\gamma}(x) &= \text{var}[\varepsilon'_\gamma(x)], \quad \lambda_{4,\gamma}(x) = \text{var}[\varepsilon''_\gamma(x)]. \end{aligned} \quad (19)$$

Reminding (7), the smoothed noise variance $\sigma_\gamma^2(x)$ is computed as in (13), while the spectral moments $\lambda_{2,\gamma}(x)$ and $\lambda_{4,\gamma}(x)$ are computed as

$$\lambda_{2,\gamma}(x) = \text{var} \left[\frac{d\varepsilon_\gamma}{dx} \right] = \sigma^2 \|l'_\gamma(x)\|^2 \quad (20)$$

$$\lambda_{4,\gamma}(x) = \text{var} \left[\frac{d^2\varepsilon_\gamma}{dx^2} \right] = \sigma^2 \|l''_\gamma(x)\|^2$$

where $\underline{l}_\gamma(x)$ is the vector (8) with derivatives

$$\underline{l}'_\gamma(x) = \frac{d}{dx} \underline{l}_\gamma(x), \quad (21)$$

$$\underline{l}''_\gamma(x) = \frac{d^2}{dx^2} \underline{l}_\gamma(x)$$

In practice, the derivatives are computed numerically at the data points x_i by

$$\begin{aligned} \underline{l}'_\gamma(x_i) &\approx \nabla [\underline{l}_\gamma](x_i), \quad i = 1, 2, \dots, n \\ \underline{l}''_\gamma(x_i) &\approx \nabla^2 [\underline{l}_\gamma](x_i), \quad i = 1, 2, \dots, n \end{aligned} \quad (22)$$

Where ∇ is backward difference operator.

4.3 Multiple Testing of p-Values to Identify Significant Peaks

Computed p-values associated with the selected peaks are adjusted to control the FDR of detected peaks. Let the number of true detected peaks (by rejecting the null hypothesis) be

$$T_R = \left| \left\{ \tau_{\gamma, i} \in \mathbf{1}(\mu(x) > 0) \right\} \right| \quad (23)$$

and the number of false detected peaks be

$$F_R = \left| \left\{ \tau_{\gamma, i} \in \mathbf{1}(\mu(x) = 0) \right\} \right| \quad (24)$$

Let also the total number of rejections at level α be

$$A_R = T_R + F_R = |O_{Max}| \quad (25)$$

The FDR is defined as the expected proportion of falsely rejected hypothesis:

$$FDR = E \left[\frac{F_R}{\max(A_R, 1)} \right] \quad (26)$$

FDR control for the positive peaks is achieved by applying the BH algorithm [19] on the pool of p-values P^+ . Similarly, FDR control for the negative peaks is achieved by applying the BH algorithm on the pool of p-values P^- .

The significant peaks indicate the location of detected non-zero regions. Each significant peak at location τ indicates that the underlying structure is significantly different from zero at τ . In turn, without prior knowledge of the underlying peak shape, the associated bandwidth γ provides an estimate of the support of the non-zero region as the interval $[\tau - \gamma, \tau + \gamma]$. Because each detected peak corresponds to a region, the FDR for detected peaks can be interpreted as the FDR of detected regions.

5. MULTI-SCALE SIGNAL RECONSTRUCTION

Once non-zero regions have been detected as significant peaks, the underlying structure (1) can be reconstructed as a set of regions where the signal is non-zero by estimating T, β, τ_t and γ_t . Precise estimation of these parameters is possible with prior knowledge of the underlying pulse shape ζ . Motivated by CN data, in this section we consider the case where the underlying non-zero signal regions are constant within their domain.

5.1 Estimation of Signal Regions for Rectangular Pulse Regions

Assuming the underlying function $\zeta = \mathbf{1}(\cdot)$ is a rectangular pulse in model (1), a non-zero region indexed by t takes the form $\mu_t = \beta_t \mathbf{1}(\tau_t - \delta_t, \tau_t + \delta_t)$, i.e. a constant non-zero region that is equal to β_t on the interval $[\tau_t - \delta_t, \tau_t + \delta_t]$. A non-zero region $\hat{\mu}_t$ is reconstructed for each significant p-value $p_{\gamma,t}$ identified by BH algorithm. The center ($\hat{\tau}_t$), half-width ($\hat{\delta}_t$) and height ($\hat{\beta}_t$) of the reconstructed non-zero region $\hat{\mu}_t$ are determined by the spatial location ($\tau_{\gamma,t}$), bandwidth (γ_t), and SNR ($\Omega_{\gamma,t}$) associated with $p_{\gamma,t}$ via

$$\hat{\tau}_t = \tau_{\gamma,t}, \quad \hat{\delta}_t = \gamma_t \times W_{factor}, \quad \hat{\beta}_t = \mu_{\gamma,t} \times H_{factor} \quad (27)$$

Where W_{factor} and H_{factor} are width and height factors respectively. As shown next, for finely spaced x , these factors are essentially constant and only depend on the kernel shape, so it is enough to compute them once at the beginning of the analysis.

5.2 The Width and Height Factors for Rectangular Pulse Regions

The factors W_{factor} and H_{factor} in (27) enable recovery of the width and height of a pulse after convolution. They are derived as follows. For a finely spaced grid, in the neighborhood of a single pulse center $\tau_{\gamma,t}$ from (4),

$$\mu_{\gamma, t}(\tau_{\gamma, t}) = \sum_{i=1}^n l_{\gamma, i}(\tau_{\gamma, t}) \mu_t(x_i) = \frac{\sum_{i=1}^n K_{\gamma}(\tau_{\gamma, t} - x_i) \mu_t(x_i)}{\sum_{j=1}^n K_{\gamma}(\tau_{\gamma, t} - x_j)} \approx \frac{\beta_t \int_{-\delta_t}^{+\delta_t} K_{\gamma}(x) dx}{1} \quad (28)$$

which does not depend on the location $\tau_{\gamma, t}$. In addition, from (5),

$$\sigma_{\gamma}^2(\tau_{\gamma, t}) = \sigma^2 \|l_{\gamma}(\tau_{\gamma, t})\|^2 = \sigma^2 \frac{\sum_{i=1}^n K_{\gamma}^2(\tau_{\gamma, t} - x_i)}{[\sum_{j=1}^n K_{\gamma}(\tau_{\gamma, t} - x_j)]^2} \approx \sigma^2 \frac{\int K_{\gamma}^2(x) dx}{1} \quad (29)$$

which again does not depend on the location $\tau_{\gamma, t}$. Hence, by (12), the SNR $\Omega_{\gamma, t}(\tau_{\gamma, t})$ is

$$\Omega_{\gamma, t}(\tau_{\gamma, t}) = \frac{\mu_{\gamma, t}(\tau_{\gamma, t})}{\sigma_{\gamma}(\tau_{\gamma, t})} \approx \frac{\beta_t \int_{-\delta_t}^{+\delta_t} K_{\gamma}(x) dx}{[\sigma^2 \int K_{\gamma}^2(x) dx]^{1/2}} \quad (30)$$

Mimicking the pruning process in which the bandwidth corresponding to the largest SNR is picked, here the kernel that best matches μ_t has the bandwidth γ_{match} that maximizes $\Omega_{\gamma}(\tau_{\gamma})$:

$$\gamma_{match} = \operatorname{argmax}_{\gamma} \Omega_{\gamma}(\tau_{\gamma}) \quad (31)$$

From (30), it is easy to see that this matching bandwidth is proportional to δ_t and does not depend on β_t or the original noise level σ^2 . The proportionality factor W_{factor} is obtained numerically by maximizing (30) for arbitrary δ_b , β_t and σ^2 . As depicted in Fig. 1(B) for the tricube function, a kernel with support of $[-1.6\delta_b + 1.6\delta_t]$ best matches the non-zero region μ_t with $\beta_t = 1$ on the interval $[-\delta_b + \delta_t]$. Therefore the width factor for a tricube kernel is:

$$W_{factor} = \frac{\delta_t}{\gamma_{match}} = \frac{\delta_t}{1.6\delta_t} = 0.625 \quad (32)$$

To obtain the height factor, notice that the height of the smoothed signal (28) is proportional to the height β_t of the pulse before convolution. Hence we define

$$H_{factor} = \left[\int_{-\delta_t}^{+\delta_t} K_{\gamma}(x) dx \right]^{-1} \text{ evaluated at } \gamma = \gamma_{match} \quad (33)$$

which is constant and only depends on the kernel shape. For a tricube kernel the factor is $H_{factor} = 1.1$.

6. RESULTS

The proposed method is first applied to a simulated pulse train with different lengths but the same heights. The second example is constructed based on some preliminary analysis of real genomic data and is composed of several non-zero regions with different lengths and different heights.

6.1 Simulated Sparse Pulse Train

A sparse pulse train is simulated by three regions, μ_1, μ_2, μ_3 , centered at $\tau_1 = 250, \tau_2 = 615, \tau_3 = 895$, having the same unit height $\beta_t = 1$, and different lengths $2 \times \delta_1 = 1, 2 \times \delta_2 = 0.3, 2 \times \delta_3 = 0.7$. The pulse train is then corrupted by zero-mean additive white Gaussian noise with variance $\sigma^2 = 0.5^2$. Original and corrupted signal are depicted in Fig. 2. A bank of tricube kernels is constructed with a set of 10 bandwidths uniformly spaced in the range $\gamma \in [0.1, 1]$. Multi-scale kernel regression is applied to the corrupted signal to locate candidate peaks. Fig. 2 (A-F) shows the smoothed signal for three different bandwidths.

The local maxima of the smoothed signals associated with different γ 's are then located as candidate representatives of the underlying pulse train (blue disks). Each non-zero region of the simulated pulse train is discovered by several located peaks of the smoothed signals at different bandwidths (Fig. 2 (G)). These peaks are associated to the same underlying non-zero region and only one is needed to represent that region. To avoid having overlapping detected non-zero regions, the located peaks are pruned by bandwidth matching based on their associated bandwidth (γ) and SNR. The remaining peaks after pruning are marked with black disks in Fig. 2 (A-F).

In the next step, statistical multiple testing is used to identify the significant peaks among the selected peaks after pruning. Among associated kernel bandwidths which capture a non-zero region, the best matching bandwidth enables a nearly perfect reconstruction. Therefore each region can be represented by a single local maximum associated with the best matching bandwidth (Fig. 2(H)). To identify the representative peaks of non-zero regions, p-values are calculated for the height of each candidate peak. The significant ones (marked by green disks in Fig. 2 (A-F)) are identified by the BH algorithm. Each significant peak at a spatial location τ and associated bandwidth γ identifies the presence of a non-zero region roughly on the interval $[\tau - \gamma, \tau + \gamma]$. Eventually, the set of detected peaks are used for reconstruction of the underlying signal (Fig. 2 (H)).

The representation of regions by local maxima at different bandwidths can be appreciated in the multi-scale local maxima map (Fig. 2(G)). In this map, each row shows the located peaks of the smoothed signal with bandwidth γ increasing from top ($\gamma_1 = 0.1$) to bottom ($\gamma_{10} = 1$). Notice that the number of located peaks decreases and converges to one by increasing the kernel bandwidth. The first region μ_1 with length of $2 \times \delta_1 = 1$ is represented by seven peaks of smoothed signal with $\gamma_1 = 0.1$, two peaks of $\gamma_2 = 0.2$, and one peak of $\gamma_3 = 0.3$ to $\gamma_{10} = 1$. The map also shows the best matching bandwidth for each true non-zero region (green).

6.2 Simulated Tightly Packed Pulse Train

The second example is another simulated signal which is composed of six significant non-zero regions $\{\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \mu_6\}$ tightly centered at $\tau_1 = 105$, $\tau_2 = 241$, $\tau_3 = 423$, $\tau_4 = 651$, $\tau_5 = 789$, and $\tau_6 = 915$ with different heights equal to $\beta_1 = 0.55$, $\beta_2 = 0.92$, $\beta_3 = 0.81$, $\beta_4 = 1.23$, $\beta_5 = 0.72$, and $\beta_6 = 0.59$, and different lengths equal to $2 \times \delta_1 = 1.22$, $2 \times \delta_2 = 0.97$, $2 \times \delta_3 = 0.38$, $2 \times \delta_4 = 1$, $2 \times \delta_5 = 0.41$, and $2 \times \delta_6 = 1.25$ (Fig. 3). The simulated signal was then corrupted by zero-mean additive Gaussian noise with variance σ^2 with seven different noise levels $\sigma \in \{0.1, [0.5:0.5:3]\}$. Original and corrupted simulated signal are depicted in Fig. 3 along with the detected regions applying the proposed method using tricube kernel for $\sigma = 0.5$ and ten bandwidths equal to $\gamma \in [0.1:0.1:1]$. Signal reconstruction was carried out assuming a rectangular pulse shape.

The multi-scale local maxima map in Fig. 3(A) shows the located peaks along smaller (top row) to larger (bottom row) bandwidths used in the signal reconstruction of Fig. 3(C). Significant peaks (marked with green in Fig. 3(A)) are associated with reconstructed regions (Fig. 3(C)). The length of each constructed non-zero region is directly related to the associated bandwidth of its representative peak. For example μ_6 , the widest non-zero region ($2 \times \delta_6 = 1.25$) in the simulated signal (Fig. 3(C)), is represented by the significant peak associated with the 10th bandwidth ($\gamma_{10} = 1$), so the estimated length of the reconstructed region $\hat{\mu}_6$ is $2 \times W_{factor} \times \gamma_{10} = 1.25$. Note that the proposed algorithm is able to reconstruct the underlying signal quite accurately under a very high level of noise (Fig. 3(D)).

6.3 Performance

The performance of the proposed method was quantified for the simulated signals corrupted with zero-mean additive Gaussian noise at seven different noise levels $\sigma \in \{0.1, 0.5, 1, 1.5, 2, 2.5, 3\}$ (Fig. 3). The proposed method was applied to each of 100 instances of noisy signal at each noise level. The performance in Fig. 3(E-F) was evaluated in terms of detection power, FDR, true positive coverage and false positive coverage for each signal instance. These quantities are defined below.

The mean detection power and mean FDR for each noise level are shown in Fig. 3(E).

Detection power is defined as $\pi = \frac{n_D}{n_T}$, where n_D is the number of detected regions and n_T is

the number of regions in the true signal (6 for this simulated signal). A region is considered detected if at least one peak among the significant peaks is located in the region's support. Thus, the detection power is equal to one if all the regions are detected and is equal to zero if no region is detected. The proposed method performs robustly and the detection power reaches about 90% for high noise $\sigma_{noise} = 1$ while for very high noise ($\sigma_{noise} = 2$), still 35% of the true regions are detected.

Reminding (26), false discovery rate is defined as $FDR = E\left[\frac{F_R}{\max(A_R, 1)}\right]$, where F_R is the

number of peaks among the detected peaks which are not located in any region's support and A_R is the total number of detected peaks. Hence, the FDR is equal to zero if all detected peaks are located in true regions' supports and is equal to one if no detected peak is located

in a true region's support. Application of the BH algorithm ensures FDR control of the number of falsely detected regions at the nominal level $q = 0.01$ regardless of the noise level, as shown in Fig. 3(E).

We have also investigated the performance of the signal reconstruction in terms of length coverage (Fig. 3(F)). In contrast with detection power, computed based on the number of detected regions, true positive coverage is defined as $T_C = \frac{L_D}{L_T}$, where L_D is the true detected length, i.e. length of the true non-zero signal that is covered by the reconstructed signal and L_T is the total length of the true non-zero signal. Similarly, false positive coverage is defined as $F_C = \frac{L_F}{L_Z}$, where L_F is the false detected length, i.e. length of the true zero-signal that is covered by the reconstructed non-zero signal and L_Z is the total length of the true zero signal. Depicted in Fig. 3(F), we can see that the true positive length coverage of the proposed method follows the same trend as its detection power. It achieves a coverage about 90% of the true signal regions for high noise ($\sigma_{noise} = 1$) whereas still 35% of the true regions are covered for very high noise ($\sigma_{noise} = 2$). False positive coverage is below 5% for all noise levels.

The performance of the proposed method for the same 100 instances of noisy signal at each noise level was also evaluated in terms of mean integrated squared error (MISE) as well as ratio of MISE to the integrated true signal (MISE Ratio) depicted in Fig. 4(A-B). As it can be observed in Fig. 4(A-B) the MISE (and MISE Ratio) increases by increasing the noise level which agrees with the evaluated performance depicted in Fig. 3(E-F). MISE Ratio is close to zero for low noise ($\sigma_{noise} = 0.1$) where the region detection power and true positive coverage (Fig. 3(E-F)) are about 100%. MISE Ratio reaches to about 20% (Fig. 4(B)) for high noise ($\sigma_{noise} = 1$) where the true positive coverage decreases to 80% (Fig. 3(F)). MISE Ratio increases to 40% (Fig. 4(B)) for very high noise ($\sigma_{noise} = 2$) where the true positive coverage drops to below 40% (Fig. 3(F)).

Moreover we evaluated the performance of the proposed method for the simulated constant zero signal (with no change point) corrupted with zero-mean additive Gaussian noise at seven different noise levels $\sigma \in \{0.1, 0.5, 1, 1.5, 2, 2.5, 3\}$ (Fig. 4(C-E)). The proposed method was applied to each of 100 instances of noisy zero signal at each noise level. The performance in Fig. 4(E) was evaluated in terms of type I error (number of false positives). The mean number of optima is depicted in Fig. 4(E) as follows: I) primary located optima (marked in blue); II) selected optima after pruning using SNR (marked in black); and III) identified optima after controlling the FDR (type I error) by applying the BH algorithm (marked in red). As we can see in Fig. 4(E), because of the added noise, there are primarily several located optima (representing false regions) across multiple scales with different bandwidths; however after pruning and controlling the FDR, the false positive rate (identified significant optima) is literally zero. This result agrees with the performance evaluated based on FDR depicted in Fig. 3(E-F) for tightly packed pulse train where the FDR was close to zero.

6.4 Comparison with Circular Binary Segmentation (CBS)

The performance of the proposed method is compared with CBS [14], a widely used method for segmentation of genomic data by detecting significant change points. CBS was applied to the same simulated signal introduced in Sec. 6.2 (Fig. 5). The same instances of the noisy signal corrupted with zero-mean additive Gaussian noise with seven different noise levels $\sigma_{noise} \in \{0.1, [0.5: 0.5: 3]\}$ were used. The segmented regions (green) obtained by CBS for medium $\sigma_{noise} = 0.5$ and high noise $\sigma_{noise} = 1.5$ are depicted in Fig. 5(A, B) and correspond to the same noisy instances in the application of our method in Fig. 3(C, D). To compare the results obtained by these two methods in terms of discrimination between zero and non-zero regions in this particular instance, CBS segmented regions whose height is lower than a threshold 0.2 are set to zero. At medium noise, the proposed method detects all 6 non-zero regions, but CBS often splits them into 8 regions (Fig. 5(A)). At high noise, CBS produces a single long constant region (Fig. 5(B)), covering all true non-zero regions but being unable to discriminate between the zero and non-zero regions. In contrast, the proposed method is able to detect and reconstruct 5 out of the 6 true non-zero regions for the same noisy instance of the signal (Fig. 3(D)).

To compare the performance of CBS with the proposed method using similar criteria, the point location of each non-zero region produced by CBS (after thresholding to discriminate between the zero and non-zero regions) is defined as the center of that region. A detected region is a true positive if its center falls inside the domain of a true non-zero region; otherwise it is a false positive. Given these definitions, detection power, FDR, true positive and false positive coverage were quantified for CBS in the same way as described for our method in Sec. 6.3. Fig. 5(C) shows that, in comparison with the performance of the proposed method (Fig. 3(E)), the CBS detection power is slightly lower, falling below 65% for $\sigma_{noise} = 1$ and below 15% for very high noise ($\sigma_{noise} = 2$). Moreover, FDR is not controlled, being below 2% for medium noise level $\sigma_{noise} = 0.5$ but reaching above 15% for $\sigma_{noise} = 1$ and above 60% for $\sigma_{noise} = 2$.

Fig 5(D) shows that, in comparison with the performance of the proposed method (Fig. 3(F)), CBS achieves above 90% true positive length coverage regardless of the noise level. However, the false positive length coverage is not controlled, increasing to above 30% for high noise $\sigma_{noise} = 1$ and above 80% for very high noise $\sigma_{noise} = 2$. The false positive length coverage is the same as the true positive length coverage for $\sigma_{noise} = 3$. These results show that despite a high true detection rate, CBS suffers from high false detection rate. This is because under high noise condition CBS does not discriminate between zero and non-zero regions (Fig. 5(D)) but segments the whole observed signal to a single constant region.

7. CONCLUSION

A new generic approach for detecting non-zero signal regions and signal reconstruction has been proposed. This approach solves the 1D spatial region detection problem by an equivalent 0D peak detection problem and reconstructs the signal regions from their topological point representatives. We combine multi-scale kernel regression for locating peaks, statistical multiple testing for controlling the FDR of the detected peaks, and matched

filtering for maximizing the SNR of the detected non-zero regions and estimating their height and length.

Each candidate peak of a smoothed signal aggregates neighborhood information, whose extent is determined by the associated kernel bandwidth. Therefore, the heights of the located peaks can be considered as test statistics of the integrated information along the neighborhood. The number of candidate peaks are much smaller than the original number of observed data points (scale: tens to hundreds instead of thousands). In summary, the proposed method improves the detection power by both taking advantage of the spatial structure in the data and reducing the number of tests in the multiple comparisons problem.

Bandwidth resolution, i.e. number of bandwidths in the chosen range, is only a computational issue, as redundancies in local maxima among bandwidths in the signal region will be eliminated by pruning, and the total number of local maxima will be used for FDR control, no matter what bandwidths they are associated with. As we choose a limited number of bandwidths to cover the range, the detected peaks are not exactly optimal but approximately so.

In our case, Eq. (18) was applied to obtain p-values because t-statistics with large number of degrees of freedom are approximately normally distributed. In the absence of normality, permutation methods or Monte Carlo simulations could be used instead [18]. When valid, using formula (18) saves computational cost. In addition, using random permutations may result in minor differences in repeated analyses of the same sample, as we have observed in CBS (results not shown).

In comparison with two widely used methods, CBS by Olshen et. al [14] and the wavelet based method by Hsu et. al [15], our proposed method controls for multiple testing while the others do not. The CBS authors state explicitly that CBS does not correct for multiple testing and therefore the probability of finding spurious change-points could be larger than the significance level α set by the user [14]. In Sec. 6.4, we showed that CBS does not control FDR at the detection step, which also results in lack of control of the false positive coverage. In contrast, our proposed method ensures a controlled false discovery rate (here 1%), together with apparent control of false positive length coverage (below 5%). Setting the proper threshold in CBS is subjective, causing under-smoothing (small threshold) or over-smoothing (large threshold). Since thresholding is applied before segmentation, CBS does not control for false detections at the segmentation step.

In comparison with wavelet-based methods, the proposed method is also multi-scale; however in contrast, it generates a bank of smoothed signals instead of a single smoothed signal which is, the purpose of kernel regression in our method is not denoising but locating candidate peaks. In other words, we smooth the signal to locate the peaks whereas the wavelet-based methods locate the large coefficients (by thresholding) to smooth the signal.

As opposed to CBS and the wavelet-based methods, our proposed method is by design concerned with multiple testing and thus capable of controlling FDR for detection of non-zero regions. The proposed method shows promise for signal detection and reconstruction in different areas of engineering and science.

ACKNOWLEDGMENT

We would like to thank Dr. Yen-Tsun Huang from HSPH and Prof. Giovanni Parmigiani from DFCI for discussions about DNA copy number data of lung cancer. We also thank NIH for supporting this research through Grant P01 CA134294-03.

REFERENCES

- [1]. Du P, Kibbe WA, and Lin SM, Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, 2006, 22(17): p. 2059–65. [PubMed: 16820428]
- [2]. Gregoire JM, Dale D, and van Dover RB, A wavelet transform algorithm for peak detection and application to powder x-ray diffraction data. *The Review of scientific instruments*, 2011, 82(1): p. 015105. [PubMed: 21280856]
- [3]. Kim J, et al., A robust peak detection method for RNA structure inference by high-throughput contact mapping. *Bioinformatics*, 2009, 25(9): p. 1137–44. [PubMed: 19246511]
- [4]. Wee A, et al., A continuous wavelet transform algorithm for peak detection. *Electrophoresis*, 2008, 29(20): p. 4215–25. [PubMed: 18924102]
- [5]. Zhang J, Zhou X, Wang H, Suffredini A, Zhang L, Huang Y, Wong S, Bayesian Peptide Peak Detection for High Resolution TOF Mass Spectrometry, *Signal Processing, IEEE Transactions on*, 2010, 58(11): p. 5883–5894.
- [6]. Scholkman F, Boss J, and WolfAn M, Efficient Algorithm for Automatic Peak Detection in Noisy Periodic and Quasi-Periodic Signals, *Algorithms*, 2012, 5: p. 588–603.
- [7]. Minin VN, et al., Dual multiple change-point model leads to more accurate recombination detection. *Bioinformatics*, 2005, 21(13): p. 3034–42. [PubMed: 15914546]
- [8]. Muggeo VM and Adelfio G, Efficient change point detection for genomic sequences of continuous measurements. *Bioinformatics*, 2011, 27(2): p. 161–6. [PubMed: 21088029]
- [9]. Wang Y, et al., Non-parametric change-point method for differential gene expression detection. *PloS one*, 2011, 6(5): p. e20060. [PubMed: 21655325]
- [10]. Kawahara Y, and Sugiyama M, Change-Point Detection in Time-Series Data by Direct Density-Ratio Estimation, *SDM*, 2009, 9.
- [11]. Leisch F, Hornik K, and Kuan CM, Monitoring structural changes with the generalized fluctuation test, *Econometric Theory*, 2000, 166: 835–854.
- [12]. Lai Y and Zhao H, A statistical method to detect chromosomal regions with DNA copy number alterations using SNP-array-based CGH data. *Computational biology and chemistry*, 2005, 29(1): p. 47–54. [PubMed: 15680585]
- [13]. Li W, Lee A, and Gregersen PK, Copy-number-variation and copy-number-alteration region detection by cumulative plots. *BMC bioinformatics*, 2009, 10 Suppl 1: p. S67. [PubMed: 19208171]
- [14]. Olshen AB, et al., Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 2004, 5(4): p. 557–72. [PubMed: 15475419]
- [15]. Hsu L and et al., Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics*, 2005, 6(2): p. 211–226. [PubMed: 15772101]
- [16]. Dobigeon N, Tourneret JY, and Davy M, Joint segmentation of piecewise constant autoregressive processes by using a hierarchical model and a Bayesian sampling approach, *IEEE Tran Signal Processing*, 2007, 55(4): p. 1251–1263.
- [17]. Schwartzman A, Gavrilov Y, and Adler RJ, Multiple Testing of Local Maxima for Detection of Peaks in 1D. *Annals of Statistics*, 2011, 39(6): p. 3290–3319. [PubMed: 23576826]
- [18]. Schwartzman A, Jaffe A, Gavrilov Y, and Meyer CA, Multiple Testing of Local Maxima for Detection of Peaks in ChIP-Seq Data. *Annals of Applied Statistics, Ann. Appl. Stat*, 2013 7(1): p. 471–494.
- [19]. Benjamini Y and Hochberg Y, Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society B*, 1995 57: p. 12.

- [20]. Cramér H and Leadbetter MR, Stationary and related stochastic processes; sample function properties and their applications. Wiley series in probability and mathematical statistics 1967, New York, Wiley xii, 348 p.
- [21]. Wasserman L, All of nonparametric statistics. Springer texts in statistics 2006, New York; London: Springer xii, 268 p.
- [22]. Scott DW, Multivariate Density Estimation: Theory, Practice, and Visualization. Wiley 1992, New York.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Fig. 1.

The proposed algorithm. (A) Schematics. (B) Matched filtering: (Left) A bank of four tricube kernels with different bandwidths $\gamma \in \{0.1, 0.5, 0.8, 1.5\}$. (Middle) Smoothed signals obtained by convolution of tricube kernels with underlying structure (unit pulse). (Right) Locating the bandwidth that optimizes the SNR (SNR vs. kernel bandwidth) for a region with unit height and unit length ($\gamma = 0.8$).

Fig. 2.

Underlying signal (black), noisy observations (purple) and smoothed observations (blue) shown for three different bandwidths. Blue disks show local maxima of the smoothed signals. Black disks mark the selected peaks (after pruning) and green disks show the significant peaks (using BH). The associated bandwidth γ with each detected region is written on top of the region. (A, B) $\gamma = 0.7$ (region 1). (C, D) $\gamma = 0.3$ (region 2). (E, F) $\gamma = 0.5$ (region 3). (B), (D), (F): zoomed in. (G) Multi-scale local maxima map where local maxima of the smoothed signals are marked in blue, selected peaks in black, and significant peaks in green. (H) Reconstructed signal (green), true signal (black), and noisy observations (purple) where estimated region lengths ($2 \times \hat{\delta} = 2 \times \gamma \times W_{factor}$) are written on top of them.

Fig. 3.

Simulated example based on analysis of real genomic data. (A) Multi-scale local maxima map showing located peaks (blue), selected peaks after pruning (black), and significant peaks after applying BH algorithm (green) along different bandwidths for $\sigma = 0.5$. (B-D) Simulated signal (black), noisy observations (purple), and reconstructed signal (green) for different noise levels σ . (B) 0.1. (C) 0.5. (D) 1.5. (E, F) Performance of the proposed method based on 100 simulation instances for each noise level: (E) Region detection power (blue) and FDR (red) for $q = 0.01$ (black) vs. noise level (σ_{noise}). Reference black line is $q = 0.01$. (F) True positive coverage (blue) and false positive coverage (red) vs. noise level.

Fig. 4.

Performance of the proposed method. (A-B) Performance based on the same 100 simulation instances as Fig. 3(E-F) for each noise level: MISE (A) and MISE Ratio (B). (C-E) Performance based on 100 simulation instances of a constant zero signal without change point for each noise level: (C) Simulated constant zero signal (black) and noisy observations corrupted with low noise level $\sigma = 0.1$ (purple). (D) Simulated constant zero signal (black) and noisy observations corrupted with extreme noise level $\sigma = 3$ (purple). (E) The mean number of optima over multiple scales with different bandwidths: primarily located peaks (blue), selected peaks after pruning (black), and significant peaks after applying BH algorithm (red) for each noise level for $q = 0.01$.

Fig.5.

Performance of CBS obtained by applying CBS to simulated signal. (A, B) Simulated signal (black), noisy observations (purple), CBS segmentation (green), and thresholded CBS segmentation (red) at threshold height 0.2 (blue) for different noise levels σ . (A) 0.5. (B) 1.5 (same noise level as Fig. 3). (C, D) Performance of CBS based on the same 100 simulation instances for each noise level as Fig. 3. (C) Region detection power (blue) and FDR (red) for $\alpha = 0.01$ (black) vs. noise level (σ_{noise}). Reference black line is $\alpha = 0.01$. (D) True positive coverage (blue) and false positive coverage (red) vs. noise level.