Multiscale Change Point Analysis in Poisson count data

Maarten Jansen
TU Eindhoven, dept. of Math. and Comp. Sci.
K.U.Leuven, dept. of Comp. Sci.

May 2006

Abstract

Motivated by the application of single molecule detection in a highly dilute solution, we discuss the problem of multiple change point detection in the intensity curve of low-intensive Poisson observations. It is explained that the *multiple* change point detection problem is inherently a multiscale problem. We analyze the data using an extension of the Continuous Wavelet Transform (CWT), the so-called Unbalanced Wavelet Transform. The presence of change points in the underlying intensity curve is revealed by a multiscale chain of local maxima in the CWT analysis. We present a new algorithm for the reconstruction of the chains by linking local maxima across scales. The new algorithm outperforms the existing ones in case of low-intensive signals, where noisy fluctuations are relatively dominant. Low intensities also motivate the extension of the CWT towards the Unbalanced Wavelet Transform. This extension is crucial in detecting small changes against intensive noise.

1 Introduction

The algorithm discussed in this paper is motivated by the following problem in biochemistry: fluorescent molecules, typically DNA molecules, in a highly dilute solution are detected using a confocal microscope. This microscope emits laser light which is then focused onto a spot in the solution. A fluorescent molecule in this focus re-emits photons that can be detected in a detector. If the concentration of molecules is such that on average only one molecule or less is in the focus of the microscope, this is called single molecule detection. The setting allows to detect rare events, to sequence single-molecule DNA and so on. Analysis of the detected photons allows even to identify the single molecule through its fluorescence properties. On the other hand, in such low concentrations, the signal has low intensity and background noise plays an important role. The objective of this paper is to detect the presence of one or more molecules in the focus. The intensity of the detected photons, the duration and the starting point of the molecule's presence in the focus are three parameters that characterize (the size, velocity, *etc* of) the molecule.

In order to develop a mathematical procedure, we generalize the problem formulation: count data such as photons in a confocal microscope are typical examples of Poisson data. We now consider general instances from such Poisson count processes, where the intensity, i.e., the average number of photons per microsecond, is piecewise constant. In other words, we consider signals with several levels of intensity. The assumption that the observations have piecewise constant intensity could easily be relaxed towards piecewise linear or even piecewise smooth intensities. Figure 1 shows a simulated test example. The lowest level of intensity is probably background noise, while the others contain significant data. The objective is to identify these levels of intensity. This identification includes the precise starting point of the significant interval, its exact duration and the (average) level of intensity. All three properties characterize the observed data.

It should be emphasized that the beginning and end points of the subsequent levels of intensity are unknown. The issue of this paper is *not* in the first place finding a smooth curve that fits the observations between the beginning and end points, for instance, using splines with (multiple) knots on the beginning and end points. That problem is rather a subsequent step, after the

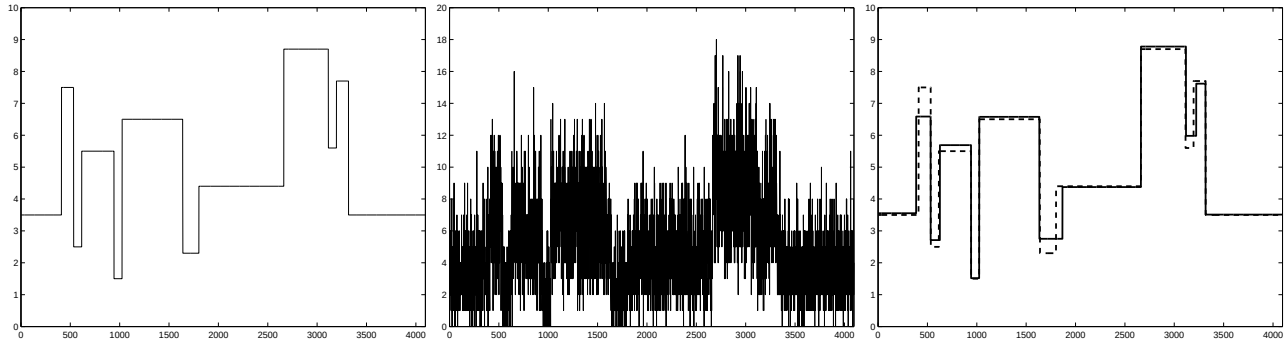

Figure 1: A simulated example of Poisson data with time varying intensities. On the left the plot of intensity curve. This is a scaled and vertically translated version of the well known 'Blocks' test example [**?**]. (The translation was achieved by adding 3.5 to all values of the Block test signal.) In the middle a random realization. On the right the estimation from that realization, using the procedure proposed in this paper.

one discussed in this paper. The problem discussed in this paper is the location of the beginning and end points themselves. A statistical description of this problem is change point detection. As we explain in Section 2, the case of multiple change points is intrinsically a multiscale problem. Change point analysis using multiscale techniques, such as wavelet decompositions, has been investigated in several papers [**?**, **?**, **?**, **?**, **?**, **?**, **?**]. Although not relying on wavelets, a methodology quite similar to multiscale analysis proceeds by kernel methods [**?**]. The kernel can be seen as a scale parameter, and it can be chosen in a data-adaptive way, leading to a procedure somehow related to a first step in the procedure presented in this paper. The work summarized in this paper concentrates on finding change points in Poisson data. Similar problem descriptions appear in earlier papers [**?**, **?**, **?**], but none of these applied a multiscale analysis to the problem.

The contribution of this paper consists of the combination of broad assumptions and an original approach applying new techniques in wavelet analysis. More specifically, on the assumption side, we allow more than one change point and we do not specify how many of them are present in the data. There is no upper bound on the possible number of change points. The background noise level is assumed unknown. The method is not restricted to high intensity signals. Indeed, the single molecule detection example has an extremely low intensive signal.

As for the new techniques, the majority of the existing wavelet based change point analyzes proceed in three steps: first perform a multiscale decomposition, i.e., a wavelet transform. This step transforms the observations into a sequence of numbers, called wavelet coefficients. In a second step, these coefficients are modified. A typical example is thresholding, where small coefficients are considered as insignificant contributions and hence replaced by zero. The third step is the reconstruction of the data from the modified coefficients. Thresholds operate on *individual* wavelet coefficients. Individual coefficients represent information at specific scales. Our method uses the information of a wavelet decomposition as a tool in an active search for the optimal scales to describe each of the change points. This search proceeds by scanning and connecting *local* maxima of coefficients within each scale of a (discretized) *continuous* wavelet analysis. In this respect, our method is different from an existing analysis, based on results in extreme value theory [**?**]. That method is based on *global* maxima within a single (well chosen) scale. It is also different from earlier methods that connect local maxima at successive scales [**?**, **?**]. Those methods look at the finest scale to determine the location of the change point, whereas the presented method finds an optimal scale to work in. Starting from the optimal scale found adaptively by connecting local maxima, we extend the analysis to *unbalanced* transforms. As explained later, this increases the (statistical) power of the detection procedure.

Although the techniques are presented in the framework of Poisson counting data, the method can easily be extended towards

other types of random data.

This paper is organized as follows. Section 2 deduces the idea and principles of a multiscale analysis. As a conclusion to this section, we find that the detection of *multiple* change points has an essential multiscale character. Section 3 refines the multiscale analysis, leading to the so-called Continuous Wavelet Transform (CWT). The benefits from this continuous version of the multiscale analysis in the context of change point detection is explained. Section 4 constructs lines of wavelet maxima out of the CWT. These lines are the basic tools to search the observations for significant features at all possible locations and with any range (i.e., scale, or duration). The relationship between these maxima lines and change points is further explored in Section 5. Next, Section 6 introduces unbalanced Haar transforms in order to increase the statistical power of the change point detection. After the CWT analysis, the construction of wavelet maxima lines out of it, and the extension towards unbalanced analyzes, we are ready to identify candidate change points, each with a level of statistical significance. As this results in a problem of multiple testing, Section 7 discusses a method to select change points among the candidates. Finally, Section 8 summarizes the algorithm, runs it on a simulation example and on real data, and formulates a conclusion.

2 Multiscale binning: Haar wavelet analysis

Suppose we are given observations $x_k, k = 0, \ldots, n-1$ of random variables X_k that are Poisson distributed, i.e.,

$$P(X_i = x) = \frac{e^{-\mu_k}\mu_k^x}{x!}, \tag{1}$$

where $\mu_k = EX_k$ is the expected value of the kth observation, also called the *intensity* of the counting process at location k.

The intensity μ_k is not constant, but depends on the location (or time point) k. More specifically, we assume that the intensity is piecewise constant, i.e., consecutive observations have the same intensity, except at some transition points:

$$\mu_k = \mu_{\tau_r}, \text{ for } k = \tau_r, \ldots, \tau_{r+1} - 1, \tag{2}$$

where $0 < \tau_0 < \ldots < \tau_r < \tau_{r+1} < \ldots < \tau_R \leq n-1$ is a sequence of R change points, and $0 \leq R \leq n-1$ is unknown. The change points are specified by the (integer) index τ_r of the first observation from the segment with a certain intensity. In other words, we are not interested in an estimate of the exact moment between $\tau_r - 1$ and τ_r where the change takes place. Obviously, the exact values of μ_k are unknown. We want to estimate those values from the observations. A central question in this estimation is to find good estimates for the locations τ_r of jumps. Not only are these locations crucial as such for several applications, they also allow good estimates of the intermediate intensities.

The problem of detecting a jump can be formulated in terms of statistical testing. Given two observations x_k and x_{k+1}, we want to test the null hypothesis $H_0 : \mu_k = \mu_{k+1}$ against the alternative $H_a : \mu_k \neq \mu_{k+1}$. Since we have only two observations, it is unlikely that those values are found significantly different.

In order to increase the power of this statistical test, we can involve the neighbors of these adjacent observations and compute the sums $x_{k-1} + x_k$ and $x_{k+1} + x_{k+2}$. It is well known that those sums are again Poisson distributed with intensities that are the sums of the individual intensities. As the relative standard deviation (i.e., the standard deviation divided by the intensity) decreases for increasing intensities, it is easier to detect changes in intensity after binning together adjacent pairs of observations. If the difference between two adjacent bins of two observations is still not significant, then binned observations can be binned again and so on. This leads to a multiscale processing.

Since we do not know the number of change points in the data, it is possible that at a given scale j, the sums of bins $\sum_{i=1}^{2^j} x_{k+i}$ and $\sum_{i=0}^{2^j-1} x_{k-i}$ contain one or more change points. The presence of neighboring changes within the range of scale j, obviously disturbs the test $\mu_k = \mu_{k+1}$ in location k at that scale j. Since we cannot know in advance which scale is appropriate, it is interesting to keep all levels of binning in a single multiscale (or multiresolution) analysis of the data, so that we can pick an appropriate level of binning based on a full, multiscale analysis. In the first instance, we consider the following decomposition:

1. Let $S_{J,k} = X_k$ be the finest level scaling coefficients. These coefficients are nothing but the observations at locations k, with $k = 0, \ldots, n-1$. The integer J is an arbitrary number assigned to the finest *scale*. Subsequent coarser scales get a lower index j.

2. Let $S_{j,k} = S_{j+1,2k+1} + S_{j+1,2k}$ be the summed values at level j of binned pairs at finer level $j+1$.

3. Let $W_{j,k} = S_{j+1,2k+1} - S_{j+1,2k}$ be the differences between scaling coefficients at level j.

If such a difference is significantly different from zero, then we know that somewhere in the interval covered by the associated two bins, a change point must have occured.

The decomposition is the Haar-wavelet-transform. To check whether a wavelet coefficient $W_{j,k}$ is significant, it is required to normalize it:

$$Z_{j,k} = W_{j,k}/\sqrt{S_{j,k}}. \tag{3}$$

The values $Z_{j,k}$ have an asymptotically normal distribution [**?**, **?**] and their variance is constant if one leaves out the cases where $S_{j,k} = 0$:

$$V(Z_{j,k}|S_{j,k} \neq 0) = 1.$$

As a conclusion to this section, we found that the *multiple* change point detection problem is essentially a multiscale problem: subsequent change points occur after an unknown and arbitrary number of observations, i.e., at an unknown scale. The dyadic haar-wavelet transform is a computationally fast multiscale analysis of the observations, but — as argued in the following section — the analysis may miss change points. We therefore proceed to the (well-known) Continuous Wavelet Transform (CWT) in Section 3. In Section 6 we further extend the continuous wavelet analysis to the *Unbalanced* Wavelet Transform, for a still more powerful detection of change points.

3 Continuous wavelet analysis: wavelet maxima

The Haar decomposition presented in the previous section is a *dyadic* transform. This means that the computation of $W_{j,k}$ involves a dyadic number of observations, i.e., the number of observations involved is an integer power of two. Also, the subsequent coefficients at a given level are based on mutually disjoint sets of observations, starting at a dyadic location $k2^{j^*+1}$:

$$W_{j,k} = \sum_{i=0}^{2^{j^*}-1} X_{k2^{j^*+1}+2^{j^*}+i} - \sum_{i=0}^{2^{j^*}-1} X_{k2^{j^*+1}+i},$$

where $j^* = J - j - 1$. These two properties obviously limit the power of the statistical test, see Figure 2. If the coefficients were not limited to dyadic locations and dyadic bins, we would certainly find a coefficient where at least one of the adjacent bins perfectly coincides with a complete interval of constant intensity.

The analysis with arbitrary locations and arbitrary length of bins is a maximal discretization of the continuous Haar wavelet transform. The continuous wavelet transform (CWT) of $x(t)$ in general is a two-dimensional function $W(a,b)$, defined by

$$W(a,b) = \int_{-\infty}^{\infty} \psi\left(\frac{t-a}{b}\right) f(t)dt,$$

where $\psi(t)$ is a wavelet function. In practice, the conditions for a function to be a wavelet function, reduce to

$$\int_{-\infty}^{\infty} \psi(t)dt = 0.$$

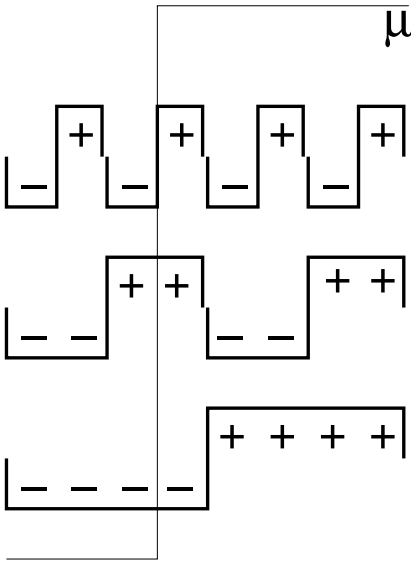

Figure 2: The Haar-wavelet analysis is a dyadic analysis: it does not include all possible bins. As a consequence, the differences between bins at coarse scales may not coincide with the real locations of transitions (change points).

The Haar wavelet is defined by $\psi(x) = \chi_{[0,1)} - \chi_{[1,2)}$, where χ_A is the characteristic function on set A, i.e., $\chi_A(x) = 1$ if $x \in A$ and $\chi_A(x) = 0$ otherwise. If X_i are the observations from the function $f(t)$, then maximal discretization of the continuous wavelet transform leads to the following wavelet coefficients:

$$W_{j,k} = \sum_{i=0}^{j^*-1} X_{k+j^*+i} - \sum_{i=0}^{j^*-1} X_{k+i},$$

where $j^* = n - j$ and $j = n-1, \ldots, 1$.

Such a complete analysis allows to find at each scale j the locations k where the normalized wavelet coefficient $Z_{j,k} = W_{j,k}/\sqrt{S_{j,k}}$, with $S_{j,k} = \sum_{i=0}^{2j^*-1} X_{k+i}$, reaches a local maximum. The image in Figure 3 displays the absolute values of the normalized continuous Haar transform, where white pixels correspond to large values and dark pixels are close to zero. Figure 4 plots the coefficients at the coarsest scale. It illustrates that for a low intensity signal with a small jump in intensities, even at the coarsest scales, the maximum values are not very prominent. If a local maximum at a certain scale is sufficiently large (say, if its absolute value is larger than 3), the corresponding location is marked as a candidate change point. Let $\mathcal{M}_j$ denote the set of indices corresponding to these selected maxima at scale j, i.e.,

$$\mathcal{M}_j = \{k = 0, \ldots, n-1 : |Z_{j,k}| \geq |Z_{j,k\pm1}| \text{ and } |Z_{j,k}| \geq 3\}$$

In the presence of noise, the fine scales of the wavelet transform have a lot of local maxima. In order to save computations, we first smooth the wavelet transform within the scale and compute the local maxima of that smoothed version. The obtained values serve as initial estimates of the local maxima. In a second step, we compute the global maxima of the original, non-smoothed

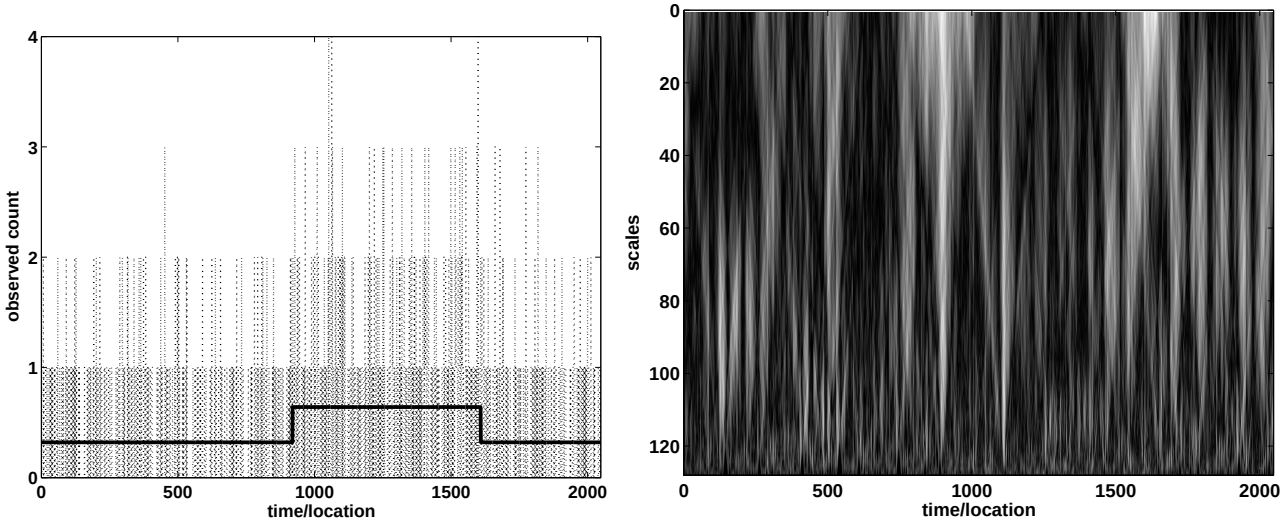

Figure 3: A low intensity test signal with two change points. The plot on the left hand side depicts the intensity curve along with a realization in 2048 data points. All observations are integers with values between 0 and 4. The figure on the right is a grey value pixel representation of the continuous wavelet transform, where white pixels correspond to coefficients with large absolute values.

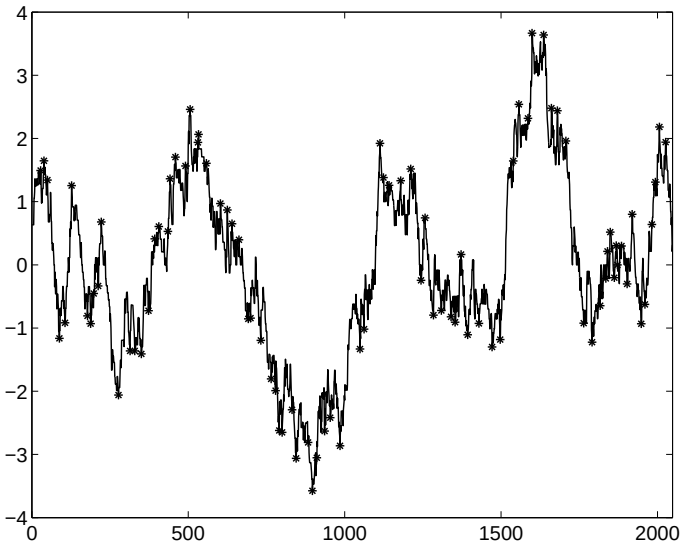

Figure 4: A plot of the coefficients at the coarsest scale in the analysis of Figure 3. Local maxima at this scale are marked with a asterisk.

transform, in the neighborhoods of each of the provisional maxima and we replace the provisional values by their corresponding new values.

4 Lines of wavelet maxima

Before the final selection of significant change points takes place, we first want to select the optimal scale for each candidate. To this end, the locations of local maxima are linked into lines of local maxima across scales [**?**, **?**], as in Figure 5.

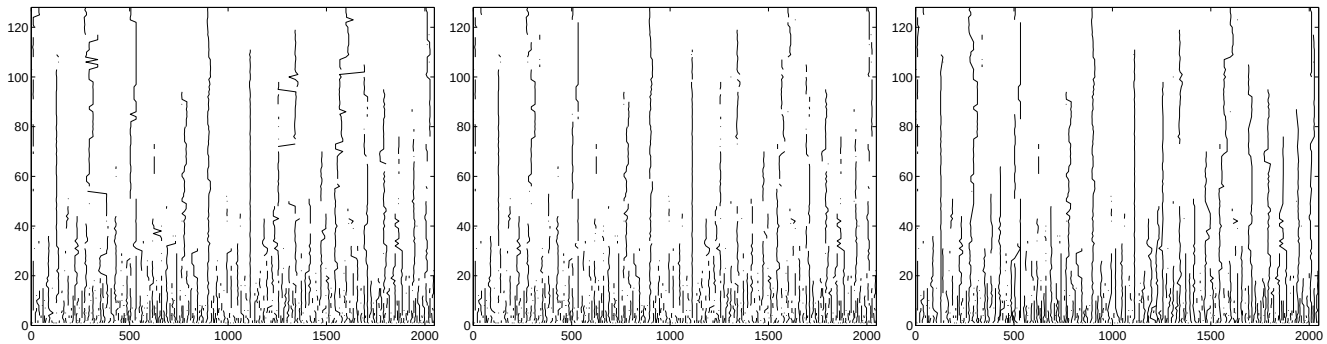

Figure 5: Consecutive steps in the construction of lines of local maxima across scales. The horizontal axis contains the locations, the vertical axis depicts the successive scales. The steps depicted in these figures operate on the geometry of the lines only, i.e., they do not take the values of the observed maxima into account.

As the location of local maxima may shift across scales and local maxima may even disappear at some scales, the construction of these maxima lines is non-trivial. On a non-discretized continuous wavelet transform, these lines can be found exactly. Because of the intensive noise, the heuristic algorithm in [**?**, page 632] is not sufficiently accurate in our application. That algorithm finds the correct lines of maxima if the resolution of the discretization is sufficiently fine, but in the case of observations with intense noise, the required resolution is generally finer than the resolution of the input observations. We therefore need a more sophisticated construction of maxima lines, so that it can find the correct lines even at coarser resolution levels. The proposed approach proceeds in five steps. Some of these steps are characterized by tuning parameters, which are filled in on a heuristics basis. Although a motivated choice may be a subject of further investigation, the heuristically choices work well in nearly all cases.

1. Eliminate: Eliminate maxima that are close to another, higher maximum. More precisely, remove a location k from $\mathcal{M}_j$ is there exists another maximum $m \in \mathcal{M}_j$ such that $|k - m| \leq j$ and $|Z_{j,k}| \leq |Z_{j,m}|$. The idea is that most of these secondary maxima are due to noise and even if they reflect the existence of a change point, their contribution at finer scales is relatively more significant. Indeed, at scale j, a wavelet coefficient centered at location k stretches beyond the neighboring change point near location m, thereby accumulating the effects of two change points.

2. Connect: Link local maxima at successive scales. Two maxima at successive scales are linked if both are the closest maximum to the other one. More precisely, a maximum at scale j, location k_j is connected to a maximum at scale $j + 1$,

location k_{j+1} if and only if

$$\begin{aligned} k_j &= \arg\min_{l \in \mathcal{M}_j} |l - k_{j+1}|, \text{ and} \\ k_{j+1} &= \arg\min_{l \in \mathcal{M}_{j+1}} |l - k_j|. \end{aligned}$$

Figure 5(Left) shows the output of this process for the test example in Figure 3. The result is a set $\mathcal{L}$ of maxima lines:

$$\begin{aligned} \mathcal{L} = \{\boldsymbol{\ell} = [(L_\ell, k_{L_\ell}), \ldots, (j, k_j), \ldots, (J_\ell, k_{J_\ell})] : k_j \in \mathcal{M}_j \\ \text{and } k_j \text{ is connected to } k_{j+1}, \forall j = L_\ell, \ldots, J_\ell - 1\}. \end{aligned}$$

3. Disconnect: The construction in Figure 5 has a few lines that jump from one to another, clearly distinctive position. Therefore, the next step disconnects local maxima if the connection is an outlier among the lengths of connections in the line of maxima. Figure 5(Middle) shows the result of this step. The exact definition of an outlier is a matter of fine tuning.

4. Merge: Merge lines with (nearly) overlapping locations into a single line: some different lines show up at the same location, but different scales, for instance if there is a gap between scales of local maxima.

Definition 1 *Let $\boldsymbol{\ell_1}, \boldsymbol{\ell_2} \in \mathcal{L}$ be two maxima lines. Suppose $k_{i,1} = \min\{k | \exists j \in \{L_{\ell_i}, \ldots, J_{\ell_i}\} : \boldsymbol{\ell_i}_j = (j, k)\}$ is the left most position on $\boldsymbol{\ell_i}, i = 1, 2$. Similarly, $k_{i,2}$ is the right most position on maxima line i. Two lines $\boldsymbol{\ell_1}$ and $\boldsymbol{\ell_2}$ are said to have nearly overlapping locations if*

(a) for $\Delta k_i = k_{2,i} - k_{1,i}$, it holds that

$$\left[k_{1,1} - \Delta k_1, k_{2,1} + \Delta k_1\right] \cap \left[k_{1,2} - \Delta k_2, k_{2,2} + \Delta k_2\right] \neq \emptyset.$$

(b) $\min\{J_{\ell_1}, J_{\ell_2}\} - \max\{L_{\ell_1}, L_{\ell_2}\} < r \cdot \min\{J_{\ell_1} - L_{\ell_1}, J_{\ell_2} - L_{\ell_2}\}$, with $0 \leq r \leq 1$ a tuning parameter, which was taken $1/4$ in our implementation.

The first condition of this definition is the actual (near-) location overlap, while the second condition prevents parallel lines, close to each other to be merge into a single line if they co-exist at the majority of their common scales. The algorithm starts from the longest existing lines. If such a line does not continue all the scales down, we check if a bridge can be constructed from its end point to another line with overlapping locations. If there is more than one candidate, take the shortest bridge, where the length of the bridge between is defined as:

$$d(\boldsymbol{\ell_0}, \boldsymbol{\ell_1}) = \min_{j_0 \in \{J_{\ell_0}, L_{\ell_0}\}} \min_{j_1 \in \{J_{\ell_1}, \ldots, L_{\ell_1}\}} \frac{|j_0 - j_1|}{J_{\ell_0} - L_{\ell_0}} + \#\mathcal{M}_{j_0} \cdot |k_{j_0} - k_{j_1}|.$$

Note that the definition is non-symmetric: the jump across scales is weighted by the length of the initiating line ($\boldsymbol{\ell_0}$): short lines are discouraged to catch up longer lines if there is a large gap in scales between them. On a scale with a high density of maxima lines ($\#\mathcal{M}_{j_0}$ large), bridges to neighboring lines are discouraged and bridges across scales are favored. Also, a bridge can only start from the end points of the initiating line (the minimum is taken over a set of two end points). The goal of the bridge can be any point on the other line. If two candidate lines can be reached by bridges of equal lengths, we select the line whose average location over all scales is closest to the average of the original line that we want to extend. As soon as a candidate line of maxima is selected, the original line is completed by filling in the locations of the secondary line at scales where the original line had no maxima. The secondary line is then removed from the set of maxima lines. Experiments confirm, not surprisingly, that the output of this step depends on which lines are taken first in consideration for extension. As a general rule of thumb, priority is given to lines that connect a lot of maxima already, lines that start at coarse scales and lines that have little variation in locations at successive scales. The output of such a merging procedure appears in Figure 5(Right).

5. Select: Remove lines where no maximum reaches a significant value, say 3. We do not remove individual maxima based on magnitudes only, because a small maximum may be a necessary link in the construction of an important line of maxima.

5 Maxima lines and change point locations

Once we have linked the maxima at successive scales, we can select the scale and location with the highest absolute normalized coefficient value. This is a new approach. Existing algorithms based on wavelet maxima [**?**, **?**] follow the lines up to the finest resolution level. The argument for following the lines up to the finest level is based on a theory of pointwise regularity characterization. That means that, under some conditions, the continuous wavelet transform, is able to characterize the regularity of a function in each point. Once we can quantify the regularity of a function in each point, we define a change point, also called a singularity, as a point where the regularity of the function is discontinuous. The regularity of a function in a point is defined by its Lipschitz coefficient γ.

Definition 2 *A function f is (pointwise) Lipschitz α at x if there exists $K > 0$ and a polynomial $p_x(t)$ of degree $\lfloor \alpha \rfloor$ such that $\forall t \in [0,1] : |f(t) - p_x(t)| \leq K|t - x|^\alpha$. The Lipschitz coefficient is defined as $\gamma = \sup\{\alpha | f \text{ is Lipschitz} \alpha\}$.*

The following theorem [**?**] states that all change points (i.e., all points with a discontinuity in the Lipschitz smoothness) must have local wavelet maxima at sufficiently fine scales.

Theorem 1 *Suppose (wavelet) $\psi \in C^p$ and $\psi = \theta^{(p)}$ with $\int_{-\infty}^{\infty} \theta(t)dt \neq 0$. Let $f \in L_1([0,1])$. If there exists a scale s_0 such that*

$$F(u,s) = \int_{-\infty}^{\infty} f(t)\psi\left(\frac{t-u}{s}\right) dt$$

has no local maxima for $u \in [0,1]$ and $s < s_0$, then f is uniformly Lipschitz p on $[\epsilon, 1-\epsilon]$ for any $\epsilon > 0$.

The theorem does not give any guarantee that wavelet maxima at successive scales, corresponding to the same singularity point, can be connected into a maxima line. Lines can be interrupted at some scales, or mixed up between nearby singularities. A guarantee for the existence of proper maxima lines exists if the wavelet function ψ is a derivative of a Gaussian function [**?**, Proposition 6.1], i.e., if, up to a constant and a scaling,

$$\psi(x) = \frac{d^k}{dx^k} e^{-x^2}.$$

No such guarantee has been proven for other wavelet functions ψ. In order to distinguish between data singularities and noise, it is possible to look at the evolution of the maxima along a line [**?**, **?**, page 171]. Indeed, discontinuities due to noise have short maxima lines with main contribution at the finest scale. Wide scale singularities generally have a long maxima line, where the maxima decrease across scale when moving to finer scales. The evolution across scales is used to distinguish between noise and change points, while the fine scales are used to locate the exact position of the change points.

This approach is problematic in our setting for a few reasons.

1. The characterization of singularities according to the evolution of wavelet maxima across scales has been derived in theoretical, noise-free situations. We are dealing with intense noise.

2. We do not use derivatives of Gaussian as wavelet functions. The Haar function that we use, is even not continuous, as required by Theorem 1. As a consequence, we cannot be sure that the location of a singularity can be found by following the maxima line up to fine scales. Moreover, due to the heavy noise, a jump between two adjacent maxima lines at fine scales is very likely to happen.

Because of these reasons, the procedure in this paper considers every maxima line, whether or not it continues to exist at fine scales. The exact location of the change point is not estimated by following a line up to fine scales. The estimation of that location takes place at the scale with the maximum value on the given maximum line, as explained in the next section. Also, since maxima lines can be interrupted, several candidate locations of change points may eventually refer to the same change point. Such duplicates will be automatically removed during the selection of significant change points, explained in Section 7.

6 Unbalanced wavelet analysis

Even the continuous wavelet transform may not be able to detect all change points. Figure 6(a) shows an example of a wavelet analysis where one interval of constant intensity is fully covered but the other (left) interval is not completely observed by this wavelet coefficient. If the difference between the intensities in the two intervals is small, it may be crucial to estimate both intensities with the highest possible accuracy, i.e., the lowest possible variance. This can be done with a so-called unbalanced Haar transform [**?**], defined by:

$$W_{j_l,j_r,k} = \frac{1}{j_r} \sum_{i=0}^{j_r-1} X_{k+j_l+i} - \frac{1}{j_l} \sum_{i=0}^{j_l-1} X_{k+i},$$

for location k, left scale j_l and right scale j_r.

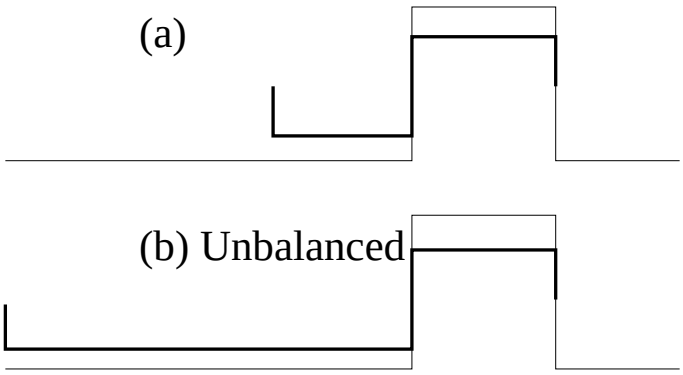

Figure 6: A symmetric wavelet analysis (bold line) is not optimal for detecting change points. If one allows unbalanced analyses, the resulting wavelet coefficients may be more significant if the analysis covers two complete, adjacent intervals of constant intensity. This may be crucial in change points with small jumps.

We now try to find all significant values of $W_{j_l,j_r,k}$. A straightforward calculation of these values would require $\mathcal{O}(n^4)$ computations. The subsequent optimization would be computationally impossible for large values of n. If we limit our search by starting of from the previously selected wavelet maxima, then this becomes an easy optimization. In a first step, we optimize over the two scale values, j_l and j_r, leading to a left and right end point. In a second step, we take the opportunity to further optimize the location k, within the optimal end points. In principle, this two step procedure could be iterated, but in practice, one iteration is sufficient. The reason for optimizing over the location k has to do with the difference between our procedure and the maxima propagation methods described in Section 5. As explained, in theory and in low noise level cases, the exact location of the change point is best estimated by following the maxima lines up to the finest scale. At coarse scales, the maximum balanced wavelet coefficient depends on a wide range of noisy observations and therefore may be shifted with respect to the actual change point location. As explained in Section 5, following the lines up to fine scales suffers from practical problems in our case. Instead

of following the maxima line to finer scales, we optimize within the given scale by letting the analysis to be unbalanced. If $\boldsymbol{\ell}$ is the current maxima line, then the position k is optimized over a set

$$K_\ell = \{k \in \mathbb{Z} | \exists j \in \{L_\ell, \ldots, J_\ell\} : \boldsymbol{\ell}_j = (j, k)\}. \tag{4}$$

In other words the search for an optimal k given j_l and j_r, is limited to those values of k that are reached by the line of maxima we started from. This search restriction avoids that we end up in a point on a neighboring line of maxima, thereby possibly missing a change point.

The normalized coefficients in an unbalanced analysis follow from a general expression in [**?**]. We now have two scales, left and right, denoted as j_l and j_r, and the unnormalized wavelet coefficients are:

$$W_{j_l,j_r,k} = \frac{1}{j_r} \sum_{i=0}^{j_r^*-1} X_{k+j_l^*+i} - \frac{1}{j_l} \sum_{i=0}^{j_l^*-1} X_{k+i},$$

where $j_l^* = n - j_l$ and $j_r^* = n - j_r$. Note that we use averages instead of sums as before. Applied to a symmetric analysis, a difference of averages is just proportional to a difference of sums. This is no longer the case for unbalanced analyses, where taking differences of sums would not lead to proper wavelet coefficients: indeed, the coefficient would not have a zero expectation if the intensities are the same on both sides. The normalization is now:

$$Z_{j_l,j_r,k} = \frac{W_{j_l,j_r,k}}{\sqrt{S_{j_l,j_r,k}}} \cdot \sqrt{j_l\, j_r}.$$

The output of this step is a set of candidate change points $T_\ell \in \{1, \ldots, n\}$, each with a value Z_{j_l,j_r,T_ℓ} — which is a measure of significance, and each corresponding to a line of local maxima $\ell \in \mathcal{L}$.

7 Selection of change points

The last phase of the algorithm is the selection of change points. Candidate change points are given by locations of local wavelet maxima and the unbalanced extension indicates the range of a change point. As Figure 7 indicates, two successive change points may reinforce each other's significance, by sharing observations. This occurs if two successive change points are both jumps up or both jumps down, thereby forming a staircase. Among all candidates, the most significant one, i.e., the one with the highest absolute $Z_{j_l,j_r,k}$-value, is selected as primary covariate, and its location is now considered as an inpenetrable boundary: the significance of the adjacent candidate change points are recomputed within this new situation. This prevents insignificant candidates from being reinforced by a significant neighbor. Indeed, as the more significant neighbor is also reinforced by this insignificant candidate, its $Z_{j_l,j_r,k}$-value remains the larger of the two, so that it is selected first. At that moment, the insignificant candidate loses its reinforcement. The recomputation of the $Z_{j_l,j_r,k}$-values after the selection of a candidate includes a recomputation of the most significant position for each remaining candidate.

A natural question arising from this procedure is how many candidate change points should be included. One way to look at this, is to view the procedure as an example of a multiple, sequential testing procedure. The selection of change points then continues as long as the corresponding $|Z_{j_l,j_r,k}|$-value is larger than a threshold of significance. The design of a proper threshold, with appropriate statistical limit properties, is far from trivial in the given settings. Moreover, this threshold would not be adaptive to the actual data. The selection of candidate change points should depend on the actual data. More precisely, let $\mathcal{T} \subset \{T_\ell | \ell \in \mathcal{L}\}$ denote the current set of selected candidates and let $\hat{R} = \#\mathcal{T}$. Based on this set, an estimate $\hat{\boldsymbol{\mu}}$ of the unknown, piecewise constant Poisson intensity vector $\boldsymbol{\mu}$, defined in Equation 2, is given by

$$\hat{\mu}_k = \hat{\mu}_{T_{(r)}}, \text{ for } k = T_{(r)}, \ldots, T_{(r+1)} - 1, \tag{5}$$

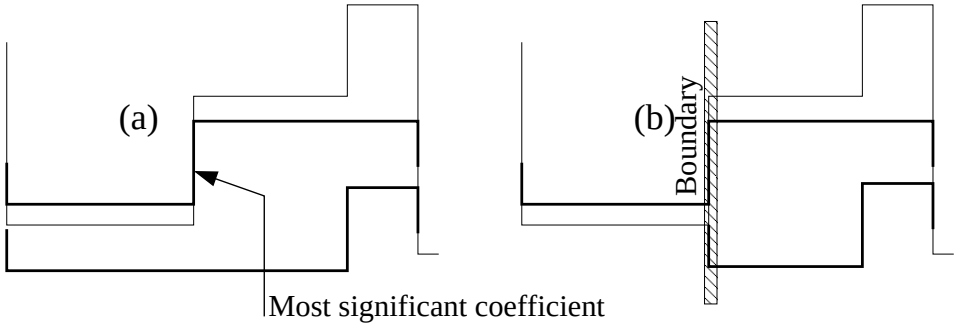

Figure 7: Selection of most significant change point and update of the range and significance of the remaining candidates.

where $T_{(r)}$ are the ordered elements of the current selection $\mathcal{T}$ and

$$\hat{\mu}_{T_{(r)}} = \frac{1}{T_{(r+1)} - T_{(r)}} \sum_{k=T_{(r)}}^{T_{(r+1)}-1} X_k. \tag{6}$$

The quality of the selection $\mathcal{T}$ is measured by the quality of the fit $\hat{\boldsymbol{\mu}}$ compared to $\boldsymbol{\mu}$. In theory, an optimal selection minimizes an error norm $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|$ (typically the ℓ_2-norm). In practice, the vector $\boldsymbol{\mu}$ is unknown, so the exact error norm cannot be computed and minimized. The optimal selection is then estimated as the best compromise (according to a given criterion, explained below) between closeness of fit to the input data and sparsity. Closeness of fit is measured by the log-likelihood function, which is, in the case of Poisson observations,

$$\mathrm{LL}(\hat{\boldsymbol{\mu}}) = \sum_{k=1}^{n} \left(\hat{\mu}_k + X_k \log \hat{\mu}_k - \log X_k!\right). \tag{7}$$

This likelihood on its own is maximized if the selection equals the set of all points of observations, i.e., if $\mathcal{T} = \{1, \ldots, n\}$, which is obviously too large. Therefore, the likelihood is penalized by the number of selected candidates, $\hat{R}$. The expression used in this paper for the compromise between likelihood and penalty is the AIC (Akaike's Information Criterion) [**?**].

$$\mathrm{AIC}(\hat{\boldsymbol{\mu}}) = -2\mathrm{LL}(\hat{\boldsymbol{\mu}}) + 2\hat{R}. \tag{8}$$

In our problem, the AIC values can be plotted as a function of the number of selected change points. An example of such a plot, corresponding to the data in Figure 1 is depicted in Figure 8. In principle, we should pick the number of selected change points with the absolute minimum on the AIC curve. As the example in Figure 8 illustrates, however, after an initial clear descend, the curve is quite flat in the neighborhood of the minimum, with generally more than one local minimum. This situation makes the eventual decision dependent on noisy fluctuations. Moreover, it is known that the AIC procedure has a tendency to overfit, i.e., to select too many change points [**?**]. For these reasons, we do not select the global minimum of the AIC curve, but rather the first local minimum. In the simulation study, described below, this procedure gave nearly identical results as obtained by (global) minimization of the BIC (Bayesian Information Criterion) [**?**].

$$\mathrm{BIC}(\hat{\boldsymbol{\mu}}) = -2\mathrm{LL}(\hat{\boldsymbol{\mu}}) + \log(n)\hat{R}. \tag{9}$$

In some simulation runs, the AIC based procedure in this paper finds a change point overlooked by the BIC.

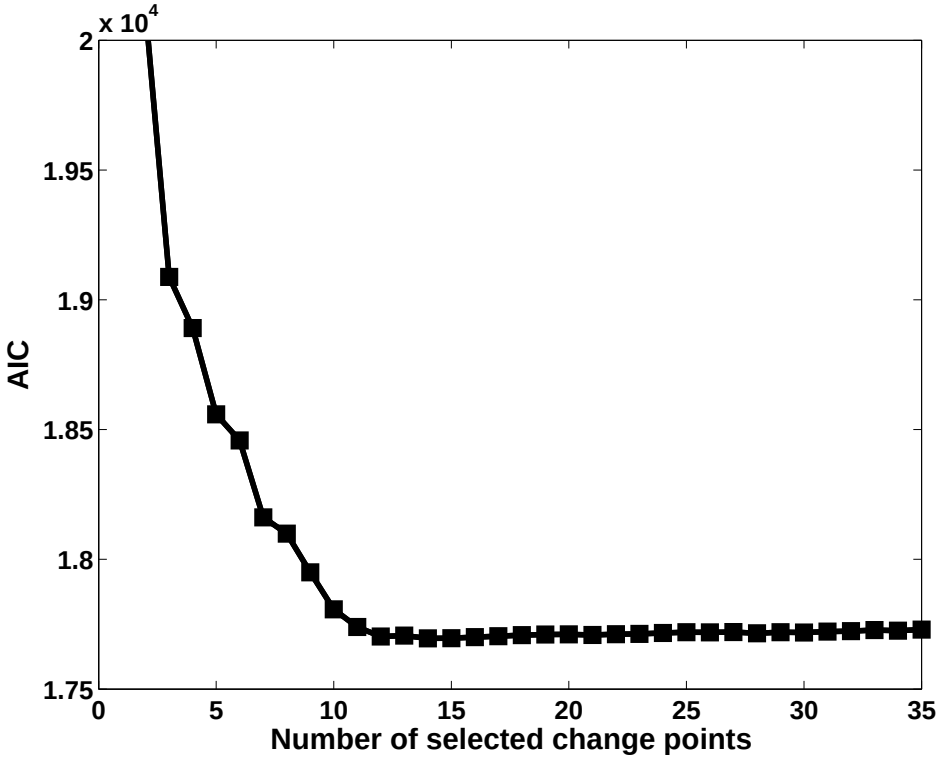

Figure 8: Plot of the AIC values as a function of the number of selected change points. The procedure picks the first local minimum of this curve as an estimate of the true number of change points.

8 Discussion and simulations

Before discussing the performance of our proposed algorithm on a simulated test example, we summarize the successive steps of the algorithm.

8.1 Overview of the algorithm

Given a vector $\boldsymbol{X}$ of n independent Poisson counts with intensities $\boldsymbol{\mu}$, at equidistant locations $\{1, \ldots, n\}$, the algorithm estimates locations and size of changes in $\boldsymbol{\mu}$. The algorithm proceeds in the following steps.

1. Compute a discretized version of the continuous Haar transform. The discretization along the location axis equals the discretization of the input. The discretization along the time axis is a tuning parameter of the algorithm. The finer the discretization, the larger the probability that all maxima lines correctly link maxima at successive scales. Let $W_{j,k}$, with $k = 1, \ldots, n$ and $j \in \mathcal{J} \subset \{1, \ldots, n\}$ the Haar coefficients at scales j and locations k. (See Section 3) Normalize the coefficients according to Equation (3). Let $Z_{j,k}$ be the normalized coefficients.

2. Pre-smooth the data $Z_{j,k}$ within the scale j and define $\mathcal{M}_j$ the set of locations corresponding to a local maximum of the smoothed coefficients $Z_{j,k}$, at scale j. (See Section 3)

3. Connect the maxima at successive scales, using the steps enumerated in Section 4, leading to the set $\mathcal{L}$ of maxima lines.

4. (a) On each maxima line $\boldsymbol{\ell} \in \mathcal{L}$, find the maximum normalized coefficient $\max_{(j,k)\in\boldsymbol{\ell}} |Z_{j,k}|$.
   (b) Starting from $Z_{j,k}$ as initial value, find the maximum normalized unbalanced coefficient $Z_{j_l,j_r,k}$ in the same neighborhood, as described in Section 6.

5. Let $\mathcal{T} = \emptyset$ be the current set of selected change points.
   Repeat
   (a) Find $\boldsymbol{\ell} \in \mathcal{L}$ with the largest $Z_{j_l,j_r,k}$-value and define $T_\ell = k$.
   (b) Compute $\hat{\boldsymbol{\mu}}$ based on $\mathcal{T} \cup \{T_\ell\}$, using Expression (5).
   (c) If $\mathrm{AIC}(\hat{\boldsymbol{\mu}})$ decreases compared to the previous value, then let $\mathcal{T} = \mathcal{T} \cup \{T_\ell\}$, $\mathcal{L} = \mathcal{L} - \{\boldsymbol{\ell}\}$.
   (d) If for any $\boldsymbol{\ell'} \in \mathcal{L}$, $T_\ell \in K_{\ell'}$, with $K_{\ell'}$ defined in (4), then recompute $T_{\ell'}$ and $Z_{j_l,j_r,T_{\ell'}}$ as explained in Section 7 and Figure 7.

   until $\mathrm{AIC}(\hat{\boldsymbol{\mu}})$ increases.

8.2 Simulation study

Table 1 and Figure 9 summarize the results of 200 runs on $N = 4096$ Poisson observations from an intensity function equal to the Block test function [**?**, page 430] *plus* a constant value of 3.5. Except for the 10th change point (see numbering in the figure), all change points were detected in every run of the experiment. We conjecture that this is essentially the best possible result in the given setting of piecewise constant intensities with small signal-to-noise ratios. Other change point techniques that we tried did not cope well with the heavy noise. The 10th change point was not recovered in all runs. The algorithm always found a maxima line corresponding to this change point, but the corresponding candidate change point was not always selected, since it was not always sufficiently significant: the candidate was ranked among noisy candidates behind the first local minimum of the AIC-curve. Nearly 20 % of the reconstructions also showed one or a few false positives. Those false positives appear if the AIC curve does not have a local minimum after the 10th or 11th (depending on the number of detected true positives) inserted change point.

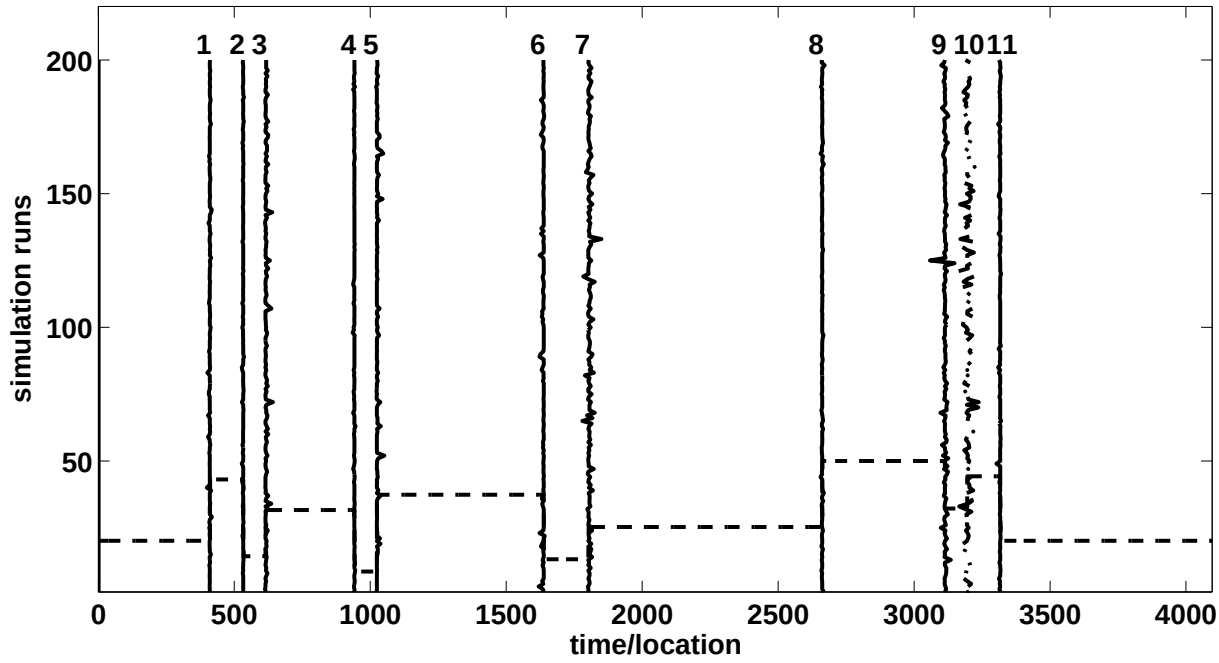

Figure 9: Summary of 200 simulations. The vertical lines plot the positions of the estimates of the 11 change points in the signal of Figure 1. Not surprisingly, the variance of the estimate depends on the altitude of the jump, and the width of the interval bounded by the two adjacent change points. Change point number 10 is not always detected, see Table 1 for numeric details. The dashed line on the figure is (a scaled version of) the underlying intensity function.

	Number of missing change points		
Number of false positives	0	1	-
0	60	20.5	80.5
1	12	3.5	15.5
2	3.5	0	3.5
3	0	0	0
4	0.5	0	0.5
-	76	24	100

Table 1: Percentages of reconstructions subdivided according to number of missing change points and number of false positives.

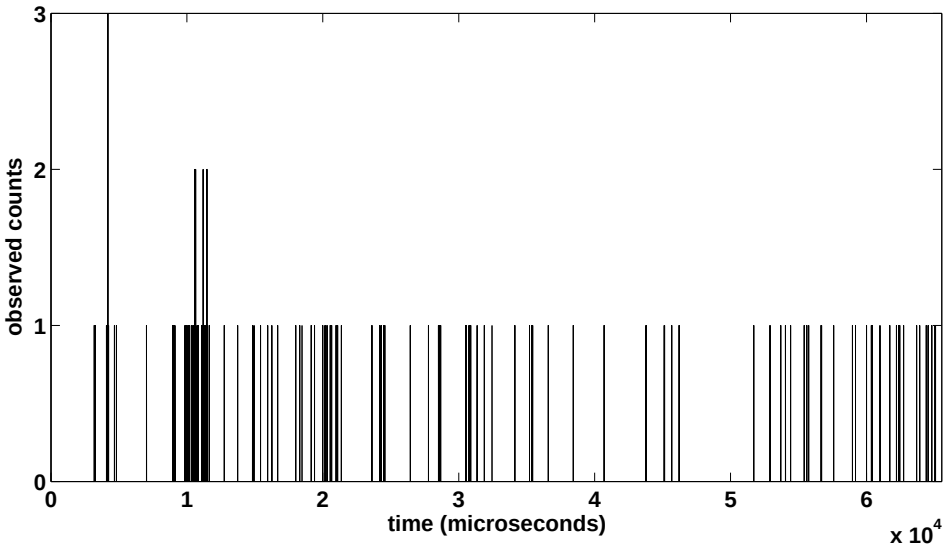

Figure 10: A real data example. The observations are counts of photons within 65 536 time spans of one microseconds. These photons were registered by a confocal microscope checking a solution for the presence of single fluorescent molecules.

8.3 Detection of single molecules in a solution

The count data in Figure 10 are the number photons detected by a confocal microscope within 65536 time spans of one microsecond. Within the total time of 65.536 milliseconds, a total number of 164 photons were registered, so the count process has a very low intensity. Yet, we suspect the presence of a few molecules, which are characterized by change points in the intensity curve (see the Introduction).

The change point detection procedure described in the previous sections reveals the simultaneous presence of three molecules in the focus, see Figure 11. A fourth passage of a molecule is detected at a different time point. The background noise intensity before and after the passages of the molecules is remarkably constant.

8.4 Directions for further research

In order to detect changes in piecewise smooth, not necessarily piecewise constant functions, or in order to detect more subtle changes (i.e., changes in a derivative, rather than plain discontinuities), it is necessary to use smoother wavelets. The construction of unbalanced versions of such wavelets requires the use of projection methods to 'stretch' existing balanced wavelet basis functions [**?**, **?**, **?**]. An alternative for this projection, is the use of the lifting scheme [**?**, **?**, **?**], which constructs new basis functions adapted to irregular (or unbalanced) data.

A second direction of further research is to incorporate a full analysis of the evolution of coefficients along a maxima line. Indeed, it has been proven [**?**, **?**, page 171] that this evolution across scale offers additional information about the character of the singularity, which can be used to distinguish between noise and signal jumps.

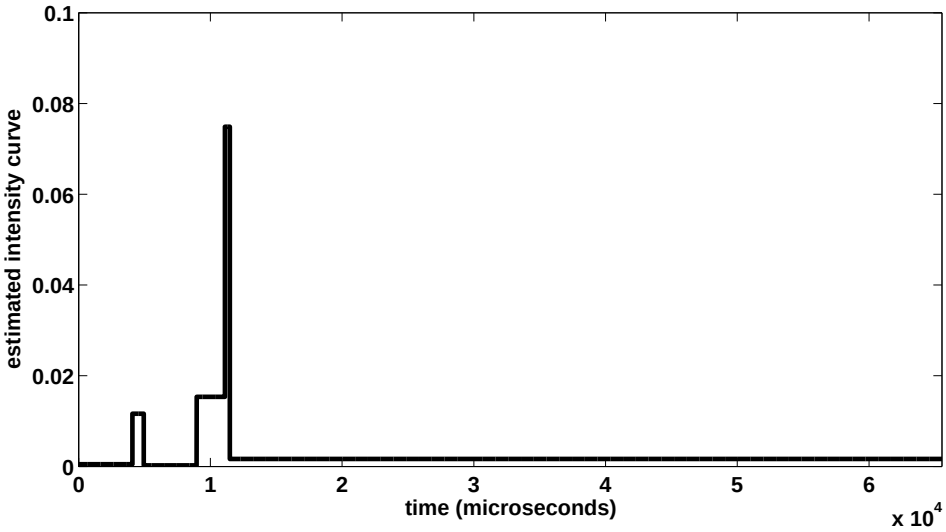

Figure 11: Change point detection and intensity estimation for the data in Figure 10.

8.5 Conclusion

This paper has elaborated a fast, multiscale based search for local maxima of normalized differences of the form

$$W_{j_l,j_r,k} = \frac{1}{j_r} \sum_{i=0}^{j_r-1} X_{k+j_l+i} - \frac{1}{j_l} \sum_{i=0}^{j_l-1} X_{k+i}.$$

The search for local maxima proceeds through a Continuous Wavelet Transform, extended towards an Unbalanced Wavelet Transform. The method does not involve wavelet thresholding, but concentrates on the local maxima of the coefficients within each scale. Through the construction along lines of maxima, the resulting values are maxima with respect to adjacent locations *and* adjacent scales. These local maxima point at a possible location of a sudden change in the underlying intensity of the given data. Candidate change points are selected in order of their significance, until the AIC-value of the resulting fit reaches a local minimum. Simulations illustrate that the method is very effective in cases of a low signal-to-noise ratio.

Acknowledgment

The author thanks Koen Dierckxs and the laboratory of biochemistry at the K.U.Leuven for kindly providing the measurement data, which were at the origin of this research.

References