Multi-Channel Change-Point Detection Statistic with Applications in DNA Copy-Number Variation and Sequential Monitoring

Benjamin Yakir

Department of Statistics, The Hebrew University,
Mount Scopus,
Jerusalem, 91905, Israel
msby@mscc.huji.ac.il

Abstract. We propose a generalized likelihood statistic for monitoring a multi-channel system that is composed of a collection of independent and identically distributed processes. The statistic may be used to monitor the system in order to detect a potential local divergence from the null distribution in a sub-collection of the processes.

A motivating application is the detection of inheritable DNA Copy Number Variation in samples that are genotyped using modern microchip technologies. The statistic can be used in this application in order to carry out off-line scanning. Sequential on-line monitoring schemes for a change-point in a sub-collection of processes are proposed based on the same statistic.

The null distribution of the resulting off-line scanning statistic is analyzed using a likelihood ratio identity and a localization argument. The approximation of the distribution of the off-line procedure is used in order to characterize the distribution of the sequential stopping rules.

Keywords. Change-point detection, likelihood ratio identity, scanning statistic, sequential analysis.

1 Introduction

The classical theory of change-point detection is usually introduced in the context of a sequence of observations $X_1, X_2, \ldots, X_n, \ldots$. These observations are taken to be univariate and their distribution characterized by a small number of parameters. The main split in the literature of change-point detection is between sequential and non-sequential, or retrospective, inference. In the former it is assumed that the system is monitored on-line and the main goal is to detect a change in the regime as soon as possible after it took place. The statistical issue is the construction of an appropriate stopping rule, typically of the form of some monitoring sequence exceeding an appropriate threshold, and the investigation of characteristics of the distribution of the stopping rule. In the latter one assumes an off-line monitoring. The entire sequence of observations is given and the goal is to estimate locations of change-points and the distribution of the system between these points.

On the technical level, with respect to the analysis of the probabilistic characteristics of change-point detection tools, the split between on-line and off-line settings is less pronounced. One of the goals of this paper is to demonstare that the computation of the null characteristics can be carried out for both off- and on-line procedures using the same tool. This phenomena may come as a surprise if we recall that the null characteristics are measured in the context of off-line problems by the consideration of probability of false alarms. On the other hand, early stopping in on-line monitoring is usually controlled via the expectation of the stopping rule. Nonetheless, a carefull examination will reveal that in both cases the characteristics may be formulated as functionals of random fields of likelihood ratios. If the random fields are properly defined then one may obtain similar forms for the functionals involved. Consequently, computations that are conducted in one context are relevant to the other context, and vice versa.

The key scenario we would like to pursue in this paper is one where a system is examined by several, possible a large number of, independent and identically distributed observational processes. At the time of change some of these processes may be affected while others are not. Yet, like in the classical setting, the goal in the case of sequential detection is to raise an alarm as soon as possible after the change or, in the case of prospective change-point inference, to test and estimate the change.

A naïve approach may consider the application of classical procedure to each of the observational processes separately. Hence, for example, one may raise an alarm if a change is detected in any of the observational processes. However, such an approach does not take into account the cumulative effect of the change on several processes at once and is necessarily less efficient. A more satisfactory solution will combine the information from all sequences to produce a single monitoring process.

A natural method for the construction of a monitoring statistic is to apply the same principles that are used in the classical setting in order to produce tools for dealing with this more complex setting. Hence, one may consider the subset of processes being affected by the change and/or the structure of the change as yet another unknown parameter that enters into the definition of the random field. Plugging in maximum-likelihood estimates for these unknown will produce generalized likelihood ratio statistics. If the location of change-points is treated in the same way then one obtains the Cusum statistics. The Shiryaev-Roberts statistic is constructed by the summation of likelihood ratios over the locations of the change instead of selecting the maximizer.

The main application of multi-channel change-point detection that will be examined in this paper is related to the problem of detecting DNA Copy-Number Variation. One type of variation that may be found in the genome is the presence or absence of an entire segment of the DNA molecule. Modern microchip technology enables the measurement of the entire genome over hundreds of thousands locations in the attempt to identify such segments. The intensity of the measurement at each of the location depends on the number of copies of the segment the person has (2, if there is no variation, or a different number if there is) but it is subject to random noise. Yet, since absence or access of segments may be an inheritable trait it is expected to be found at a proportion of subjects. Consequently, it was proposed to combine the measurements of a sample in order to increase the chances of detection (Zhang et al., 2008). The discussion in this paper is in response to the statistical approach that was proposed in that paper for construction of the combined monitoring sequence.

A different approach for combining information from a multi-channel process to produce a monitoring sequence was proposed in Mei (2008) in the sequential context. The proposal we make in the off-line scenario of DNA copy-number variation may yield an alternative approach for producing a monitoring sequence in the on-line setting.

Poisson approximations and computations based on measure transformation may be used in order to assess the null characteristics of our procedure in both off- and on-line setting. We describe the outcomes of these assessments and outline the approach for obtaining them. The analysis is specified to the case where the observations are normally distributed. Generalizations to other exponential families can be carried out along the lines of the proofs presented here.

In the next section we present the off-line monitoring statistic that emerges from the consideration of generalized likelihood-ratios statistics and give the resulting sequential Cusum and Shiryaev-Roberts procedure. In the last section a Poisson approximation is applied in order to obtain the exponential limit for the distribution of the tail of the monitoring statistic in the off-line setting and for the distribution of the stopping time in the on-line scenario. In the next to last section measure-transformation techniques are applied in order to obtain the rate of the exponential limit.

2 Generalized likelihood ratio statistics

Let us formulate a simple model of multi-channel monitoring in the context of testing for a shift from the target number of DNA copies.

Consider an array of independent normal observations $\{X_{ij}\}$, for subjects $1 \le i \le n$ and locations $1 \le j \le m$. The target is to detect the possibility that for some of the subjects the mean level is shifted over some interval. Specifically, consider testing that the shift in average intensity, if occurred, is to a given mean level δ. Given an interval $t = [t_1, t_2]$ of loci and a sub-collection $K \subset \{1, \cdots, n\}$ of subjects the log likelihood ratio statistic for testing shift to δ for the given interval and sub-collection is

$$\ell(t, K) = \sum_{i \in K} \sum_{j=t_1}^{t_2} [\delta X_{ij} - \delta^2/2] = \sum_{i \in K} \ell_i^\delta(t) \ .$$

The sub-collection K and the interval t are a-priori unknown. Maximization with respect to K over all subsets produces

$$\max_K \sum_{i \in K} \ell_i(t) = \sum_{i=1}^{n} [\ell_i^\delta(t)]^+ = Y_t \ ,$$

where $[\ell_i^\delta(t)]^+$ is the positive part of $\ell_i^\delta(t)$. The generalized log-likelihood statistic is $\max_t Y_t$, which is the statistic we propose. Two-sided alternatives may be handled by the addition of a term associated with the positive part of a likelihood ratio for testing for a negative shift.

In Zhang et al. (2008) a similar statistic is proposed. Specifically, they considered the subject-specific standardized sums $Z_i(t) = \sum_{j=t_1}^{t_2} X_{ij}/\sqrt{t_2 - t_1}$ and use the sum of functions of these standardized sums: $U_t = \sum_{i=1}^n g(Z_i(t))$ as the basis for scanning. Again, a variation is detected whenever $\max_t U_t$ crosses a threshold. For the case of the chi-square statistic, which corresponds to $g(z) = z^2$, they provided analytical approximation of the associated p-value. However, their recommendation is to use other functions g that give less weight to smaller values of $Z_i(t)$, since such values are more likely to be associated with random noise.

From a statistical prospective their chi-square statistic may emerge as a score statistic. Accordingly, a derivative with respect to δ is taken for each subject separately and a subject-specific maximum likelihood estimate of that parameter is plugged in. The underlying methodological concept assumes subject-specific alternative intensities δ_i and these intensities are considered in the context of local alternatives. In our approach, on the other hand, a common alternative rate is set for all subjects.

The formulation of our proposal implicitly assume fixed alternatives. An equivalent formulation that is associated with local alternatives may set a target value of $\delta/\sqrt{n}$. Consequently, one obtains $Y_t = \sum_{i=1}^n [\delta Z_i(t) - \delta^2/2]^+$. Thus, our statistic can be considered as a member in the family of statistics proposed in Zhang et al. (2008) that sets $g(z) = [\delta z - \delta^2/2]^+$ for one-sided alternatives and $g(z) = [\delta z - \delta^2/2]^+ + [-\delta z - \delta^2/2]^+$ for two-sided ones. Observe that these functions give zero weight for small values of the standardized sums.

A sequential approach for dealing with a similar type of multi-channel monitoring is discussed in Mei (2008). Motivated by the Cusum approach Mei proposes to use as the monitoring statistic computed at time t_2 the statistic $M(t_2) = \max_K \sum_{i\in K} \max_{t_1} \ell_i^\delta(t)$. The associated stopping time declares a change once this process crosses the threshold $\log A$. In contrast, we may propose to use $Y(t_2) = \max_{t_1} \max_K \sum_{i\in K} \ell_i^\delta(t) = \max_{t_1} Y_t$ instead and stop once a threshold is crossed. Both statistics can be interpreted as generalized likelihood ratio statistics that are based on the information collected up to time t_2. The difference between these two statistics is that the latter assumed a common change-point for all subjects that are affected by the change whereas the former opens the door to the possibility of subject-specific locations of the change. Hence, the terms that enter the sum are produced by estimating for each subject its private time of change.

An alternative to the statistic $Y(t_2)$, which is motivated by the Cusum approach, is the statistic $[\sum_{t_1} \exp\{\theta Y_t\}]^{1/\theta} = R(t_2)$, which is motivated by the Shiryaev-Roberts approach. The constant θ that is used in the definition is described in the next section. Again, a change is declared once the statistic crosses a threshold.

In the next two sections we deal with the asymptotic distribution of the off-line scanning statistic and the sequential stopping times that emerges from the on-line formulation of our procedure. In the next section we consider the probabilities $\mathbb{P}(\max_t Y_t \geq \log A) = \mathbb{P}(\max_{t_2\leq m} Y(t_2) \geq \log A)$ and $\mathbb{P}(\max_{t_2\leq m} R(t_2) \geq A)$. In the following section we use the results on the probabilities and a Poisson approximation in order to assess the null distribution of the rescaled change-point detection stopping time.

3 The rate of false detection

The discussion in this section is based on Sections 5 and 6 of Siegmund, Yakir and Zhang (2008).

Let n, the number of observational processes, be proportional to $\log A$. We will consider scanning statistics with a restricted window width $w = c \log A$. Consequently, $t_2 - t_1 \leq w$. For any such interval t one may consider a likelihood ratio $\exp\{\theta Y_t - n\psi(\theta)\}$, where $\psi(\theta) = \log \mathbb{E} \exp\{\theta g(Z)\}$, $Z \sim N(0,1)$, and θ is selected by solving the equation $\dot{\psi}(\theta) = (\log A)/n$. A likelihood ratio identity, which is based on transforming the null distribution to the measure determined by the summation of likelihood ratios

for all $t \subset [1,m]$, $t_2 - t_1 \le w$, will produce the representation

$$\mathbb{P}\big(\max_t Y_t \ge \log A\big) = \frac{[\dot\psi(\theta)]^{1/2}}{[\log A]^{1/2}} \frac{1}{A^{\theta-\psi(\theta)/\dot\psi(\theta)}} \sum_t \sqrt{n}\,\mathbb{E}_t\left[(M_t/S_t)e^{-\tilde\ell_t - \log M_t}; \tilde\ell_t + \log M_t \ge 0\right] ,$$

where $\tilde\ell_t = \sum_{i=1}^n \theta[g(Z_i(t)) - \dot\psi(\theta)]$ converges to a Gaussian limit and the two local terms are given by $S_t = \sum_\tau \exp\{\sum_{i=1}^n \theta[g(Z_i(\tau)) - g(Z_i(t))]\}$ and $M_t = \max_\tau \exp\{\sum_{i=1}^n \theta[g(Z_i(\tau)) - g(Z_i(t))]\}$.

We would like to apply Theorem 5.1 of Siegmund, Yakir and Zhang (2008) and the analysis given in Section 5 of that paper. Unfortunately, that cannot be carried out directly, since it is assumed for the analysis of the local terms that the function $g(z)$ is twice differentiable, which is not the case in our procedure. However, for the one-sided alternative we can approximate the function $g(z)$ by the smooth function $\gamma(z) = \gamma_n(z) = \delta \int_{-\infty}^{z-\delta/2} \Phi\big(yn/\epsilon\big)dy$ and a similar approximation can be used for two-sided alternatives. Observe that $\max_z |g(z) - \gamma(z)| = \gamma(\delta/2) = \delta\epsilon/\big(n\sqrt{2\pi}\big)$. It follows that the local process associated with $g(z)$ that enters into the definition of the local terms M_t and S_t may be approximated by the parallel local terms created by the use of γ.

Following the proof in Section 5.1 of Siegmund, Yakir and Zhang (2008) for t of length proportional to n and taking ϵ to zero will produce

$$\sqrt{n}\,\mathbb{E}_t\left[(M_t/S_t)e^{-\tilde\ell_t - \log M_t}; \tilde\ell_t + \log M_t \ge 0\right] \approx \{2\pi\theta^2\ddot\psi(\theta)\}^{-1/2}\Big[\mu(t)\nu\big([2\mu(t)]^{1/2}\big)\Big]^2 ,$$

where

$$\mu(t) = \frac{\theta^2 n}{2(t_2 - t_1)} \int_{\delta/2}^\infty e^{\theta z - \psi(\theta)}\phi(z)dz ,$$

$\phi(z)$ is the density of the standard normal distribution and $\nu(\cdot)$ is a function associated with the Laplace transform of the overshoot of a normal random walk crossing a threshold. See page 112 of Siegmund and Yakir (2007).

The summation with respect to t in the representation of the probability of detection may be approximated by an integral of the approximation of the local expectation. Changing the variable of integration to $x = \mu(t)$ will result in the approximation

$$\mathbb{P}\big(\max_t Y_t \ge \log A\big) = \mathbb{P}\big(\max_{t_2 \le m} Y(t_2) \ge \log A\big) \approx \lambda_{\text{CS}} \cdot [m/f(A)] \tag{1}$$

where $f(A) = A^{\theta - \psi(\theta)/\dot\psi(\theta)}\{\log A\}^{-1/2}$, $x_0(w) = \mu(t) = (\theta^2 n/2w)\int_{\delta/2}^\infty e^{\theta z - \psi(\theta)}\phi(z)dz$, and

$$\lambda_{\text{CS}} = \frac{\theta}{\{8\pi\dot\psi(\theta)\ddot\psi(\theta)\}^{1/2}} \times \int_{\delta/2}^\infty e^{\theta z - \psi(\theta)}\phi(z)dz \times \int_{x_0(w)}^\infty \big[\nu\big(\{2x\}^{1/2}\big)\big]^2 dx .$$

Essentially the same type of analysis will lead to a similar type of approximation for the Shiryaev-Roberts procedure:

$$\mathbb{P}\big(\max_{t_2 \le m} R(t_2) \ge A\big) \approx \lambda_{\text{SR}} \cdot [m/f(A)] \tag{2}$$

for

$$\lambda_{\text{SR}} = \frac{\theta}{\{8\pi\dot\psi(\theta)\ddot\psi(\theta)\}^{1/2}} \times \int_{\delta/2}^\infty e^{\theta z - \psi(\theta)}\phi(z)dz \times \int_{x_0(w)}^\infty x^{-1}\nu\big(\{2x\}^{1/2}\big)dx .$$

4 Approximating the null distribution of the stopping rule

Let us define the (window-restricted) Cusum changepoint detection rule $N_{\text{CS}} = \inf\{t_2 : Y(t_2) \ge \log A\}$, to which we attach the asymptotic rate λ_{CS}. Likewise, define the Shiryaev-Roberts detection rule $N_{\text{SR}} = \inf\{t_2 : R(t_2) \ge A\}$ with the attached rate λ_{SR}. Most of the statements below will be valid to both stopping rules and will use an abstract N_A and an abstract rate λ in their formulation.

The goal of this section is to demonstrate that for a large A the null distribution of $N_A/f(A)$ is approximately exponential with rate λ and the collection of stopping rules $\{N_A/f(A)\}$, indexed by A, is uniformly integrable. The combination of these two statements implies that $\mathbb{E}[N_A] \approx f(A)/\lambda$. Consequently, if a stopping rule with a target null expectation B is required one may use the threshold $A = f^{-1}(\lambda B)$, which will asymptotically produce the target.

The analysis of the distribution of the stopping time is based on the discussion in the previous section and on the application a Poisson approximation. Denote the indicator of the event $B(jm,(j+1)m)$ by X_j. This event is defined to be $\{\max_{jm<t_2\le(j+1)m} Y(t_2) \ge \log A\}$ for the Cusum procedure and for the Shiryaev-Roberts procedure it is $\{\max_{jm<t_2\le(j+1)m} R(t_2) \ge A\}$. Notice that the stopping time is not activated in the interval $[0, xf(A)]$ if, and only if, all the relevant indicators equal zero. Hence,

$$\left\{ \sum_{j=0}^{\lfloor xf(A)/m \rfloor} X_j = 0 \right\} \subset \left\{ N_A > \lfloor xf(A) \rfloor \right\} \subset \left\{ \sum_{j=0}^{\lceil xf(A)/m \rceil} X_j = 0 \right\} . \tag{3}$$

The assessment of the survival function of the rescaled stopping time is obtained by showing that both the sum on the right-hand side event and the sum on the left-hand side converge to the same Poisson distribution with rate given by λx.

Using a somewhat loose notation let us consider the random variable $W = \sum_{j=1}^{xA} X_j$, to which we apply the Poisson approximation of Arratia et al. (1989), with the "neighborhood of dependence" $J(j)$ composed of $\{j-1, j, j+1\}$, with trivial modifications for the first and last j. The conclusion of their Theorem 1 becomes:

Theorem 1. *Consider the window-restricted Cusum or Shiryaev-Roberts stopping rule and define indicators X_j with $m = m(A)$ as above, $\log A \ll m \ll f(A)$. Let W be the sum of indicators then:*

$$\lim_{A\to\infty} \left| \mathbb{P}\left(N_A/f(A) > x \right) - e^{-\mathbb{E}(W)} \right| = 0 . \tag{4}$$

Proof. For any $i \notin J(j)$, X_j and X_i are computed on the basis of disjoint of observations and are therefore independent. Thus, the term b_3 that measures dependence between remote elements vanishes.

Let $J = \{j\}$. The term that measures the neighborhood size becomes:

$$b_1 = \sum_{j\in J} \sum_{i\in J(j)\setminus\{j\}} \mathbb{P}(X_j = 1)\, \mathbb{P}(X_i = 1) \le 2|J| \left\{ \mathbb{P}(X_1 = 1) \right\}^2 + \mathbb{P}(X_1 = 1) ,$$

and the term that measures the expected number of neighbors is:

$$b_2 = \sum_{j\in J} \sum_{i\in J(j)\setminus\{j\}} \mathbb{P}(X_j = 1, X_i = 1) \le 2|J|\, \mathbb{P}(X_1 = 1, X_2 = 1) + \mathbb{P}(X_1 = 1) .$$

However, $\{X_1 = 1, X_2 = 1\} \subset B(2m-w, 2m+w) \cup \{B(m, 2m-w) \cap B(2m+w, 3m)\}$. Thus,

$$\mathbb{P}(X_2 = 1, X_3 = 1) \le \mathbb{P}(B(2m-w, 2m+w)) + \{B(m, 2m-w)\}^2,$$

The term $|J|$, the total number of indicators, is at most $\lceil xf(A)/m \rceil$ and the probabilities of the events B are proportional to the length of the interval over which they are defined, divided by $f(A)$. Therefore,

$$b_1 + b_2 \le \lceil xf(A)/m \rceil \left\{ \frac{c^2(m+w)^2 + c^2m^2}{[f(A)]^2} + \frac{3cw}{f(A)} \right\} + \frac{2cm + cw}{f(A)} ,$$

for some constant c and the proof of the theorem follows. □

Remark 1. By the analysis given in the previous section it can be shown that $\{f(A)/m\}\, \mathbb{P}(X_2 = 1)$ converges to λ. As a result $\mathbb{E}[W] \to \lambda x$ that establishes λ as the rate of the limit exponential distribution.

Next let us deal with the issue of uniform integrability of the sequence $\{N_A/A\}$:

Theorem 2. *The collection* $\{N_A/f(A)\}$ *of re-scaled window-restricted stopping rules index by* A*. is uniformly integrable.*

Proof. Consider again the auxiliary sequence of indicators $\{X_j\}$ and define the random τ that identifies the index of the first even element in the sequence that obtains the value one:

$$\tau = \inf\{k : X_{2k} = 1\} \ .$$

Note that τ has a geometric distribution since the even elements are independent of each other. Moreover, since $N_A \leq 2m\tau$ we get that

$$\mathbb{P}(N_A/f(A) > x) \leq \mathbb{P}\left(\tau > xf(A)/(2m)\right) = \left\{1 - \mathbb{P}(X_2 = 1)\right\}^{\lfloor xf(A)/(2m)\rfloor}$$

and the proof follows once more from the convergence of $\{f(A)/m\}\,\mathbb{P}(X_2 = 1)$. □

Remark 2. The scanning statistic declares at least one detection within the interval $[1, x]$ if and only if the Cusum stopping time is activated by time x. Therefore

$$\mathbb{P}\left(\max_t Y_t \geq \log A\right) \approx 1 - \exp\left\{-\lambda_{\text{CS}} \cdot x/f(A)\right\}$$

Acknowledgment. This paper has benefited from discussions the author had with David Siegmund and Nancy Zhang from Stanford University. This research has been supported by the US-Israel Binational Science Foundation Grant 2006101.

References

Arratia R., Goldstein L., and Gordon L. (1989), Two moments suffice for Poisson approximation: the Chen-Stein method, *Ann. Prob.*, 17, 9-25.

Mei J. (2008). Scalable robust schemes for monitoring multiple data streams, A Georgia Institute of Technology manuscript.

Siegmund D., Yakir B. (2007). *The Statistics of Gene Mapping*, Springer, New-York.

Siegmund D., Yakir B. and Zhang N.R. (2008). Tail approximations for maxima of random fields by likelihood ratio transformations. A manuscript.

Zhang N.R, Ji H. and Li J, Siegmund D.O. (2008). Detecting simultaneous change-points in multiple sequences. A Stanford University manuscript.