Nonlinear Analysis: Modelling and Control, 2005, Vol. 10, No. 2, 171–192

Detecting and Locating a Changed Segment in a Binomial Sequence: Comparison of Tests

D. Zuokas

Department of Econometrical Analysis, Faculty of Mathematics and Informatics
Vilnius University, Naugarduko st. 24, LT-03225 Vilnius, Lithuania
danaz78@one.lt

Received: 19.01.2005
Accepted: 17.02.2005

Abstract. A class of tests for testing a changed segment in a binomial sequence is proposed and an asymptotic behavior is established. A consistent procedure of estimating the length of a changed segment is proposed. The performance of two tests from the given class is compared by Monte-Carlo simulations. The results are applied for the non-coding deoxyribonucleic acid (DNA) sequence analysis.

Keywords: binomial random variables, changed segment, cumulative sums, DNA sequence analysis, epidemic alternative.

1 Introduction

Let $X_1, \dots, X_n$ be independent binomial random variables with

$$\mathrm{P}(X_i = 1) = \mu_i, \quad \mathrm{P}(X_i = 0) = 1 - \mu_i,$$
$$0 < \mu_i < 1, \quad i = 1, \dots, n.$$

We want to test the null hypothesis of a constant occurrence probability

$$\mathrm{H}_0\colon \mu_1 = \dots = \mu_n = \mu_0,$$

against the following so called epidemic (or changed segment) alternative

H_A: there exist integers k^* and m^*, $0 \le k^* < m^* \le n$, such that

$$\mathrm{P}(X_i = 1) = \begin{cases} \mu_1, & i \in \{k^* + 1, \dots, m^*\}, \\ \mu_0, & i \in \{1, \dots, n\} \setminus \{k^* + 1, \dots, m^*\}. \end{cases} \tag{1}$$

Here k^* stands for the beginning, m^* for the end and $l^* = m^* - k^*$ for the length of epidemic. The quantity $s = |\mu_1 - \mu_0|$ is referred to the size of epidemic. If H_0 is rejected, next step is to estimate l^*, k^*, m^*, μ_0 and μ_1. (Note that the problem of epidemic change in occurrence probability can also be reformulated in terms of epidemic change in the mean, because $\mathrm{E}X_i = \mathrm{P}(X_i = 1) = \mu_i$.)

The problem of testing H_0 against the epidemic type alternative and then locating an epidemic has applications in the non-coding deoxyribonucleic acid (DNA) sequence analysis (for details see Avery and Henderson [1,2]) among other applications. Most of the DNA consists of the non-coding DNA. But it is believed that non-coding DNA still has some functional importance. So it is of great value to find locations in the non-coding DNA which may contain some information. One way of approach to this problem is analysis of occurrence probabilities of the four main nucleic acids (marked by A, C, G, T), separately for every acid. The acid which is analyzed is marked by 1 and the other three by 0. Thus the original sequence of nucleic acids is replaced by a binomial sequence. The problem is to answer whether there is a change in an occurrence probability of that base and then to locate the segment where this probability has changed. Different methods are used to tackle this problem. The most common tools are the maximum likelihood method and those based on cumulative sums.

For a short survey of epidemic change problem we refer to Csörgő and Horváth [3], where mainly the cumulative sum type test statistics for testing the epidemic change in the mean of random variables are discussed. Also refer to [4], where different type statistics are analyzed in the case of normally distributed observations. The problem of a changed segment in a binomial sequence was considered by Curnow and Fu [5]. They assumed that μ_0, μ_1 and the length of epidemic are known, what is too restrictive for the most practical applications. Avery and Henderson [1] introduced a test for zero-one observations and obtained the limit distribution for test statistic under null hypothesis. They also applied the test to the DNA sequence analysis. Another type of cumulative sum tests was introduced by Račkauskas and Suquet [6,7] for the sequences of random elements with values in abstract measurable spaces.

In this paper (Section 2), following Račkauskas and Suquet [6, 7], a class of tests that are identified by a certain weight function ρ is proposed for the problem of a changed segment in occurrence probability of binomial sequence. It is then argued that the test introduced by Avery and Henderson [1] can be regarded as a particular case of the latter class of tests. For the introduced class of tests we establish asymptotic behavior under null hypothesis and prove their consistency under epidemic alternative. We propose the estimate of the epidemic length and establish its consistency in probability as well as almost surely. All proofs are collected in the Appendix. We chose two tests from the given class and run a number of Monte-Carlo simulations to compare their performance. In Section 3 we investigate performance of the test statistics under H_0. In Section 4 we compare empirical power of the test statistics. In Section 5 we present results for the tests when locating the changed segment and estimating epidemic mean. In Section 6 we then perform an analysis of the nucleotide acids' sequence of the human glucagon gene's introns 2, 3 and 4 (the same as in Avery and Henderson [1]). We end up with conclusions.

2 Cumulative sum type tests

Cumulative sum type statistics are based on differences between the mean of observations in a certain sliding window and that of the whole sample, $\overline{X}$. For a random binomial sequence $X_1, \dots, X_n$ of length n, denote

$$S(k,m) = \sum_{i=k+1}^{m} (X_i - \overline{X}), \quad 0 \le k < m \le n, \tag{2}$$

where k can be regarded as the beginning of the sliding window and $l = m - k$ as its length. Now for every length $0 < l < n$ set

$$V_\rho(l) = \frac{1}{\varrho(l/n)} \max_{0 \le k \le n-l} \bigl|S(k, k+l)\bigr|, \tag{3}$$

where $\varrho(h) = \rho\bigl(h(1-h)\bigr)$ and $\rho(h)$, $0 < h \le 1$, is a certain weight function to be defined later. Following Račkauskas and Suquet [6], we consider a class of statistics

$$\mathrm{UI}(n,\rho) = \frac{\max_{0<l<n} V_\rho(l)}{\sqrt{(S(0,n)/n)(n - S(0,n))}}$$

to test for a changed segment in a sequence of binomial variables. In the special case $\rho \equiv 1$, we have the test statistic $\mathrm{UI}(n, 1)$, which was considered by Avery and Henderson [1]. To be precise they proposed the following test statistic

$$K_n^* = \max_{i<j} \left| \sum_{k_1=1}^{i} \sum_{k_2=i+1}^{j} \operatorname{sgn}(X_{k_1} - X_{k_2}) + \sum_{k_1=j+1}^{n} \sum_{k_2=i+1}^{j} \operatorname{sgn}(X_{k_1} - X_{k_2}) \right|, \quad (4)$$

and normalized it by $\sqrt{nS(0,n)\bigl(n - S(0,n)\bigr)}$. In (4) $\operatorname{sgn}(x)$ is a sign function. In a binomial case $\operatorname{sgn}(x) = x$ and K_n^* can be simplified to

$$\begin{aligned} K_n^* &= n \max_{0\le i<j\le n} \bigl|S(i,j)\bigr| = n \max_{0<l<n} \max_{0\le k\le n-l} \bigl|S(k, k+l)\bigr| \\ &= n\bigl(\max_{0<i<n} S(0,i) - \min_{0<i<n} S(0,i)\bigr). \end{aligned}$$

We see that $\mathrm{UI}(n,1) = K_n^* / \sqrt{nS(0,n)\bigl(n - S(0,n)\bigr)}$.

To obtain the limiting behavior of $\mathrm{UI}(n, \rho)$ we need to determine an admissible class of weights ρ (see [6] for more details).

Definition 1. *By $\mathcal{R} = \bigl\{\rho\colon [0,1] \mapsto \mathbb{R}_+\bigr\}$ denote the class of non-decreasing functions satisfying:*

(i) $\rho(h) = h^\alpha L(1/h)$, $0 < h \le 1$ *for some* $\alpha \in (0, 1/2]$ *and positive on* $[1, \infty)$, *normalized, slowly varying at infinity function* L*;*

(ii) $\theta(t) = t^{1/2}\rho(1/t)$ *is continuously differentiable on* $[1, \infty)$*;*

(iii) $\theta(t)\log^{-\beta}(t)$ *is non-decreasing on* $[a, \infty)$ *for some* $\beta > 1/2$ *and* $a > 0$.

Function L is normalized, slowly varying at infinity if and only if for every $\delta > 0$ $t^\delta L(t)$ is ultimately increasing and $t^{-\delta}L(t)$ is ultimately decreasing. In the special case where $L(h) = \log^\beta(\gamma/h)$,

$$\rho(h) = \rho(h, \alpha, \beta, \gamma) = h^\alpha \log^\beta(\gamma/h), \quad (5)$$

which belongs to $\mathcal{R}$ if either $\alpha \in (0, 1/2)$ and $\beta \in \mathbb{R}$, or $\alpha = 1/2$ and $\beta > 1/2$. Parameter $\gamma = \gamma(\alpha, \beta) > 0$ is chosen properly in such a way, that the weight function is non-decreasing on $[0, 1]$.

Let $(W(t), t \in [0,1])$ be a standard Wiener process and $(B(t), t \in [0,1])$ the corresponding Brownian bridge, $B(t) = W(t) - tW(1)$, $t \in [0,1]$. Denote by $\xrightarrow[n\to\infty]{\mathcal{D}}$ the convergence in distribution. Let

$$\mathrm{UI}(\rho) = \sup_{0<h<1} \frac{1}{\varrho(h)} \sup_{0\le t\le 1-h} \big|B(t+h) - B(t)\big|, \tag{6}$$

which in the case $\rho \equiv 1$ reduces to

$$\mathrm{UI}(1) = \sup_{0<t<1} B(t) - \inf_{0<t<1} B(t). \tag{7}$$

Under the null hypothesis Theorem 1 (presented below) establishes the convergence in distribution of the test statistics $\mathrm{UI}(n, \rho)$, when either $\rho \in \mathcal{R}$ or $\rho \equiv 1$. In the case $\rho \in \mathcal{R}$ Theorem 1 is a special case of a more general result proved in Račkauskas and Suquet [6] for any independent identically distributed random variables. Using the Donsker-Prokhorov invariance principle, Slutsky's lemma and continuous mapping theorem, one can easily obtain the result when $\rho \equiv 1$.

Theorem 1. *Assume* H_0 *holds and either* $\rho \in \mathcal{R}$ *or* $\rho \equiv 1$. *Then*

$$\mathrm{UI}(n,\rho) \xrightarrow[n\to\infty]{\mathcal{D}} \mathrm{UI}(\rho). \tag{8}$$

In general case the explicit form of distribution function of $\mathrm{UI}(\rho)$ is not known. Thus we use Monte-Carlo simulations to get approximate critical values. In the case $\rho \equiv 1$ one can use approximation as pointed out in [1], namely the first member, $2(4x^2 - 1)\exp(-2x^2)$, of the following series

$$\mathrm{P}\big(\mathrm{UI}(n,1) \ge x\big) \simeq 2\sum_{i=1}^{\infty} \big(4i^2x^2 - 1\big)\exp\big(-2i^2x^2\big). \tag{9}$$

When $\mathrm{H_A}$ holds, we consider cases where $l^*/n \to 0$ or $l^*/n \to 1$. If $l^*/n \to \theta \in (0,1)$, weight function ρ has no influence on the power of $\mathrm{UI}(n,\rho)$ and problem of a changed segment can be solved by existing tests for multiple change points. Next assume that l^* and $n - l^*$ tend to infinity as $n \to \infty$. Denote by $\xrightarrow[n\to\infty]{\mathrm{P}}$ the convergence in probability.

Theorem 2. *Suppose that* $\mathrm{H_A}$ *holds and either* $\rho \in \mathcal{R}$ *or* $\rho \equiv 1$. *Moreover, let*

$$\lim_{n\to\infty} \frac{n^{1/2} h_n s}{\rho(h_n)} = \infty, \tag{10}$$

where $h_n = (l^*/n)(1 - l^*/n)$. *Then* $\mathrm{UI}(n,\rho) \xrightarrow[n\to\infty]{\mathrm{P}} \infty$.

The proof is given in the Appendix.

Remark 1. *Note that for binomial observations,* $\mathrm{UI}(n,\rho)$ *has the same value, if* X_i *is replaced by* $Y_i = (X_i - \overline{X})^2$ *and* $\overline{X}$ *by* $\overline{Y}$. *This means that, no matter what problem we solve, epidemic change in the mean or epidemic change in variance, for binomial observations test stays invariant.*

The motivation for using weight function is the following. Assume for a moment that $l^*/n \to 0$ and s is fixed. If $\rho \equiv 1$, condition (10) reduces to $l^*/n^{1/2} \to \infty$, that is the epidemic length should tend to infinity faster than $n^{1/2}$ to ensure the consistency of the test. Similarly, when $\alpha < 1/2$, $\beta = 0$, l^* should be larger than $n^{(1-2\alpha)/(2-2\alpha)}$. For example, taking $\alpha = 1/4$, the length of epidemic should be such that $n^{1/3} = o(l^*)$. However, the problem with using the parametric weight functions is that there is no strict rule for assigning certain values to parameters. It therefore remains interesting and open theoretical question of data driven choice of parameters.

To estimate the length and the beginning of a changed segment we use the procedure proposed by Račkauskas and Suquet [7]. Using (3) we estimate the length of epidemic by

$$\widehat{l^*} = \min\bigl\{j \colon V_\rho(j) = \max_{0<l<n} V_\rho(l)\bigr\}. \tag{11}$$

To estimate k^*, we go back through differences $\bigl|S(k, k+\widehat{l^*})\bigr|$ and find such index k, which corresponds to the maximal one. So we define

$$\widehat{k^*} = \min\bigl\{i \colon \bigl|S(i, i+\widehat{l^*})\bigr| = \max_{0\le k\le n-\widehat{l^*}} \bigl|S(k, k+\widehat{l^*})\bigr|\bigr\},$$

where $\widehat{l^*}$ is given by (11). To estimate the end of epidemic we take $\widehat{m}^* = \widehat{k^*} + \widehat{l^*}$. Next we estimate μ_1 as sample mean over the integer set $\{\widehat{k^*}+1, \dots, \widehat{m}^*\}$, and μ_0 as sample mean of observations with indices $\{1, \dots, \widehat{k^*}, \widehat{m}^*+1, \dots, n\}$.

Avery and Henderson [1] suggest the following estimates for k^* and m^*,

$$\widehat{k}^* = \min\bigl\{k_1, k_2 \colon S(0,k_1) = \max_{0<i<n} S(0,i), S(0,k_2) = \min_{0<i<n} S(0,i)\bigr\},$$
$$\widehat{m}^* = \max\bigl\{k_1, k_2 \colon S(0,k_1) = \max_{0<i<n} S(0,i), S(0,k_2) = \min_{0<i<n} S(0,i)\bigr\}.$$

One can see that these estimates coincide with those defined above in the special case $\rho \equiv 1$.

Next we investigate the rate of convergence $\widehat{l}^*/l^* \xrightarrow[n\to\infty]{\mathrm{P}} 1$ and give the conditions for almost sure convergence when $\rho(h) = h^\alpha$. Throughout we assume that s is such that

$$l^* s^2 / \log(n) \to \infty. \tag{12}$$

Denote by $\xrightarrow[n\to\infty]{\text{a.s.}}$ the almost sure convergence.

Theorem 3. *Assume that* $\mathrm{H_A}$ *and* (12) *hold,* $\rho(h) = h^\alpha$, $\alpha \in (0, 1/2)$ *and* $l^* \to \infty$ *as* $n \to \infty$.

(i) *If* $l^*/n \to 0$ *and*

$$l^*(l^*/n)^{1-2\alpha} s^2 \to \infty, \tag{13}$$

then $\widehat{l}^*/l^* \xrightarrow[n\to\infty]{\mathrm{P}} 1.$

(ii) *If* $l^*/n \to 0$ *and for each* $\varepsilon > 0$

$$\sum_{n=1}^{\infty} \exp\bigl(-\varepsilon l^*(l^*/n)^{1-2\alpha} s^2\bigr) < \infty, \tag{14}$$

then $\widehat{l}^*/l^* \xrightarrow[n\to\infty]{\textit{a.s.}} 1.$

We present the proof of this theorem in the Appendix.

Remark 2. *When* $l^*/n \to 1$, *the consistency can be proved similarly but now variables* X_i *with* $i \in \{1, \ldots, n\} \setminus \{k^*+1, \ldots, k^*+l^*\}$ *should be viewed as variables having epidemic probability* μ_1. *Epidemic length in this case is* $n - l^*$ *and all the conditions in Theorem* 3 *should be rewritten in such a way that* l^* *is replaced by* $n - l^*$ *and* l^*/n *by* $1 - l^*/n$.

The rest of the paper is intended to compare the performance of two test statistics. Namely, we consider

$$T_1 = \mathrm{UI}(n,1) \quad \text{and} \quad T_2 = \mathrm{UI}(n,\rho) \quad \text{with} \quad \rho(h) = h^{1/4}. \tag{15}$$

We will write UI(1) and UI(ρ) for the limiting statistics of T_1 and T_2 respectively. The motivation of such parameter choice in (15) is the following. Recall that for the weight function of the parametric form as in (5), parameter $\alpha \in [0,1/2)$ (we choose $\beta = 0$). In the problem under investigation statistics UI(n,ρ) with $\rho(h) = h^\alpha$ and α close to 0 behave quite similarly to UI$(n,1)$. On the other hand, when α is close to $1/2$, the behavior of test statistic strongly depends on the distribution of observations. Therefore we chose T_2 as a representative of the set UI(n,ρ) with $\rho(h) = h^\alpha$ and α separated from 0 and $1/2$.

3 The performance under the null hypothesis

In this section we investigate statistics T_1 and T_2 under H$_0$ and perform the p-value analysis. First we find approximations of critical values associated with the certain significance level α_s. We randomly generate $N = 10000$ values of the limiting statistics UI(1) (using (7)) and UI(ρ) (according to (6)) and take empirical quantiles as an approximation for the critical values[1]. Brownian bridge in each replication of UI(1) and UI(ρ) is approximated by partial sum process $\xi(t) = (1/\sqrt{m})(\sum_{i=1}^{[mt]} Z_i - t\sum_{i=1}^{m} Z_i)$, $t \in [0,1]$, $\xi(0) = 0$. Here $Z_i \sim N(0,1)$, $i = 1,\dots,m$, $m = 10000$ and $[\cdot]$ is an integer part of the number. For UI(1) we have also computed critical values using (9). Table 1 gives the results.

Table 1. The critical values

	$\alpha_s = 0.05$	$\alpha_s = 0.01$	$\alpha_s = 0.001$
UI(1) using (9)	1.74726	2.00092	2.30297
UI(1) using (7)	1.73459	1.98175	2.22504
UI(ρ)	2.52019	2.86686	3.33042

We see that the critical values for UI(1) computed in two ways (we took only first member of the series in (9)) differ in the second digit after the point, except for

[1] In further considerations and conclusions we use critical values computed this way.

$\alpha_s = 0.001$. Considering not large replication number to estimate 0.999 quantile we can say that both approximations agree well.

For any statistic Y, assuming only non-negative values, the p-value is $p = 1 - F_0(Y)$, where F_0 is the null distribution function of the statistic. In our case F_0 is not known therefore we use empirical approximation $\widehat{F}_0$. When H_0 holds, we compute R realizations of both statistics T_1 and T_2 (we will denote Y_j for the j-th realization of either of statistics) and the corresponding estimates for p-values (denoted by $\widehat{p}_j$)

$$\widehat{p}_j = 1 - \widehat{F}_0(Y_j) = \frac{1}{N}\sum_{k=1}^{N} \mathbf{1}\{L_k > Y_j\}, \quad j = 1, \ldots, R. \tag{16}$$

Here L_k, $k = 1, \ldots, N$, stands for a sequence of the limiting statistics' values. The random variable $F_0(Y)$ as well as $1 - F_0(Y) = p$ is uniformly distributed on $[0, 1]$, if Y is distributed according to F_0. Having the set $\{\widehat{p}_j, j = 1, \ldots, R\}$, we compare the empirical cumulative distribution function for $\widehat{p}$ with the distribution function of true p-value, $F_p(x) = x$. The convenient way for such analysis is p-value discrepancy plot (Davidson and MacKinnon [8]), representing the difference $\widehat{F}_{\widehat{p}}(x) - F_p(x)$ on y-axis (we will denote $d(x)$) against x on x-axis. For six different parameter sets $R = 6000$ realizations of p-value estimates were computed. In Fig. 1 the results are provided for $N = 10000$, $x \in [0, 0.2]$, $\mu_0 = 0.1$, $\mu_0 = 0.2$ and $n = 200$, 500, 1000.

For all n and μ_0, both tests generally are a bit conservative (in average accept the null hypothesis too often). This discrepancy naturally diminishes when n increases. In all cases the p-value difference $d(x)$ for T_1 is smaller than for T_2. When $\mu_0 = 0.1$, T_2 behaves considerably better than T_1, but passing to $\mu_0 = 0.2$ p-value discrepancy for T_2 increases, nevertheless remaining slightly less than for T_1. For T_1, when passing from $\mu_0 = 0.1$ to $\mu_0 = 0.2$, $d(x)$ slightly decreases. Concluding the p-value analysis, we might say that $d(x)$ for $x \leq 0.05$ is acceptable in all six cases for both statistics.

4 The power analysis

In this section we present the results of simulations when comparing the power of test statistics T_1 and T_2. For every parameter set we have $R = 1000$ replications

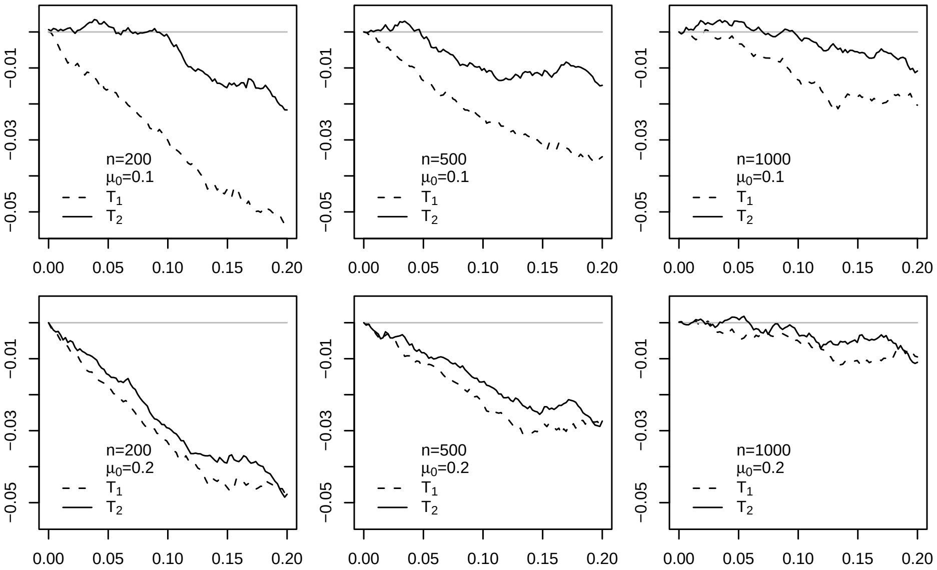

Fig. 1. The p-value discrepancy plots.

of every statistic when $\mathrm{H_A}$ holds and count how much of them are greater than critical value associated with the certain α_s. In other words, we find values of empirical power functions of tests at the point α_s. Table 2 gives the values at $\alpha_s = 0.05$ for several values of n, l^*, μ_0 and μ_1.

Fix n, μ_0, μ_1 and let l^* increase. From Table 2 we see that in all cases the power increases quite rapidly for both statistics. Fix l^* and let n increase. For $l^* = 20$ and 50 the power of both tests gradually decrease except when $\mu_0 = 0.1$, $\mu_1 = 0.2$, $l^* = 50$ in the T_2 case. When $l^* = 100$, both tests reach maximum power for $n = 500$. For fixed n and l^* increase $|\mu_1 - \mu_0|$. We see that power increases and again very quickly. Now let n and l^* increase but the ratio l^*/n keep constant. In this case again the power of both tests increase. For both tests we observe rather interesting effect, which was mentioned in Avery and Henderson [1]. Namely, that shifting both μ_0 and μ_1 but not changing $|\mu_1 - \mu_0|$ decreases the power. This effect can be explained by the fact that, on average, this shift in probabilities has no impact on statistics themselves. But it alters sample variance $\overline{X} - (\overline{X})^2$ and so the value of statistic. So if both μ_0 and μ_1 increase by some $a > 0$ to $\mu_0 + a$ and $\mu_1 + a$, sample variance also increases (only for some values

Table 2. Empirical power at the significance level α_s

		T_1			T_2		
$\alpha_s = 0.05$	$n\backslash l^*$	20	50	100	20	50	100
$\mu_0 = 0.1,\ \mu_1 = 0.2$	200	0.066	0.158	0.241	0.089	0.206	0.264
	500	0.054	0.149	0.372	0.073	0.222	0.445
	1000	0.040	0.101	0.242	0.058	0.154	0.370
$\mu_0 = 0.1,\ \mu_1 = 0.3$	200	0.154	0.590	0.764	0.271	0.648	0.763
	500	0.103	0.450	0.912	0.186	0.646	0.950
	1000	0.078	0.296	0.832	0.126	0.529	0.944
$\mu_0 = 0.2,\ \mu_1 = 0.4$	200	0.100	0.398	0.640	0.142	0.438	0.623
	500	0.067	0.305	0.760	0.092	0.421	0.826
	1000	0.066	0.185	0.616	0.078	0.306	0.794

of a) thus diminishing the value of statistic. But statistic, which under $\mathrm{H_A}$ more often assumes smaller values compared to some critical value, has less power than the statistic which more often assumes larger values.

Comparing the power of T_1 to T_2, from Table 2 we see that T_2 in all cases gains more power except when $|\mu_1 - \mu_0| = 0.2$ for $l^* = 100$ and $n = 200$. When $l^* = 20$, both tests have very little power reaching the biggest value 0.271. The T_2 test shows its advantage for $l^* = 50$, especially when $|\mu_1 - \mu_0| = 0.2$ and $n = 500, 1000$. For example when $\mu_0 = 0.1$, $\mu_1 = 0.3$ and $n = 1000$ it rejects $\mathrm{H_0}$ (when $\mathrm{H_A}$ is true) 529 times out of 1000 compared to 296 for T_1. This case gives the biggest difference. For $l^* = 100$ this difference diminishes and when $n = 200$ both tests behave very alike. When $n = 1000$, T_2 significantly outperforms T_1 and for $n = 500$ the difference is smaller but again in the favor of T_2.

For a more detailed inspection we present the so called size-power curves on a correct size-adjusted (not nominal size) basis (Davidson and MacKinnon [8]). For every parameter set we compute 1000 replications of both statistics and corresponding p-value estimates: first for the sample with no changed segment then for the same sample except for epidemic segment with indexes $\{k^*+1, \dots, m^*\}$. We plot the empirical cumulative distribution function for p-values under $\mathrm{H_A}$ (which is the empirical power function) but on x-axis we have the values of empirical distribution function for p-values under $\mathrm{H_0}$ instead of nominal size α_s. That is we adjust power to true size. In Fig. 2 results are for $n = 500$, 1000, $l^* = 50$, 100 and all three pairs μ_0, μ_1. We exclude $l^* = 20$ cases because of very low power and $n = 200$ cases because the difference in the performance of tests is small.

It is clearly seen from Fig. 2 how for true size values from $[0, 0.2]$ both tests rapidly increase their power when increasing l^* or $|\mu_1 - \mu_0|$, slightly decrease it increasing n or increasing μ_0, μ_1, but keeping $|\mu_1 - \mu_0|$ constant. We can conclude that T_2 displays its advantage for small values of ratio l^*/n ($1/20$ or $1/10$) and the biggest difference being when this ratio is the smallest. For $l^*/n = 1/5$ the advantage of T_2 is minor.

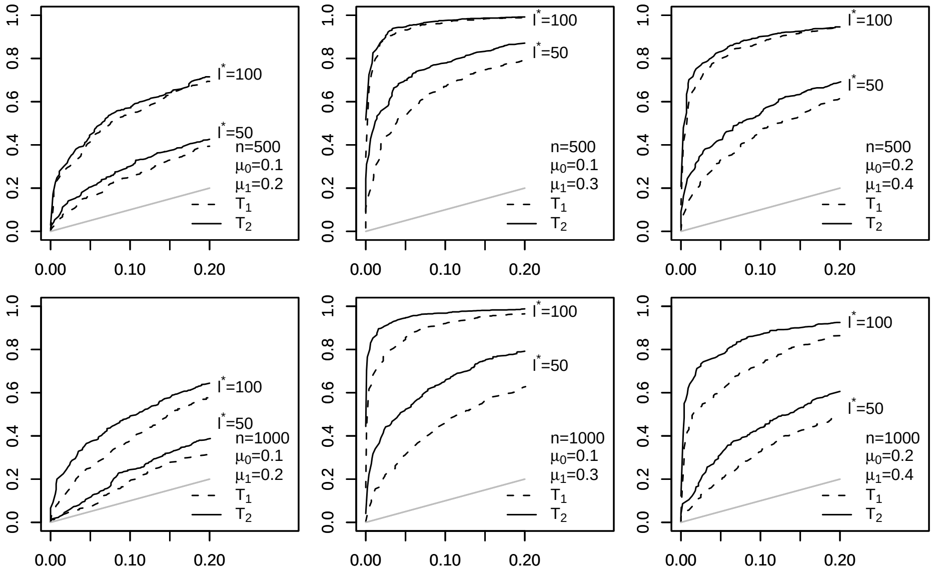

Fig. 2. The adjusted size-power curve plots.

5 Estimating parameters

In this section we investigate the estimates of the beginning, the length and the size of epidemic for both tests. We will rest upon the procedures described in Section 2. For every parameter set we have computed $R = 1000$ replications of estimates. For a sequence of realizations $\widehat{Z} = \{\widehat{Z}_1, \dots, \widehat{Z}_R\}$ of any estimate denote $\mathrm{M}\widehat{Z} = \sum_{i=1}^{R} \widehat{Z}_i / R$, $\mathrm{pw}_{0.05}$ the empirical test power value for significance level $\alpha_s = 0.05$ and

$$\mathrm{SE}\widehat{l^*} = \mathrm{M}\Big(\frac{\widehat{l^*}}{l^*} - 1\Big)^2, \quad \mathrm{SE}\widehat{k^*} = \mathrm{M}\Big(\frac{\widehat{k^*} - k^*}{l^*}\Big)^2, \quad \mathrm{SE}\widehat{\mu}_1 = \mathrm{M}(\widehat{\mu}_1 - \mu_1)^2.$$

In Tables 3 to 5 we present results (we took $k^* = 90, 240, 490$ for sample sizes respectively $n = 200, 500, 1000$).

Table 3. The estimates for k^*, l^* and μ_1 when $\mu_0 = 0.1$ and $\mu_1 = 0.2$

	l^*	n	pw$_{0.05}$	M$\widehat{k^*}$	SE$\widehat{k^*}$	M$\widehat{l^*}$	SE$\widehat{l^*}$	M$\widehat{\mu}_1$	SE$\widehat{\mu}_1$
T_1	50	200	0.158	70.99	0.56	74.55	0.58	0.207	0.0080
		500	0.149	165.25	5.29	185.17	9.85	0.155	0.0063
		1000	0.101	315.05	25.38	386.26	55.91	0.129	0.0072
	100	200	0.241	78.46	0.16	84.39	0.10	0.220	0.0098
		500	0.372	190.48	0.82	170.85	0.98	0.192	0.0037
		1000	0.242	351.09	4.57	339.17	8.11	0.153	0.0048
T_2	50	200	0.206	82.50	0.56	52.94	0.53	0.315	0.0517
		500	0.222	197.46	4.45	123.09	6.29	0.245	0.0287
		1000	0.154	372.25	20.84	257.05	36.25	0.199	0.0171
	100	200	0.264	90.03	0.19	66.24	0.25	0.296	0.0422
		500	0.445	216.19	0.60	128.96	0.69	0.240	0.0130
		1000	0.370	412.00	3.16	227.35	4.93	0.205	0.0097

Table 4. The estimates for k^*, l^* and μ_1 when $\mu_0 = 0.1$ and $\mu_1 = 0.3$

	l^*	n	pw$_{0.05}$	M$\widehat{k^*}$	SE$\widehat{k^*}$	M$\widehat{l^*}$	SE$\widehat{l^*}$	M$\widehat{\mu}_1$	SE$\widehat{\mu}_1$
T_1	50	200	0.590	79.90	0.25	65.22	0.30	0.301	0.0083
		500	0.450	187.50	2.89	143.98	5.39	0.225	0.0124
		1000	0.296	345.99	18.23	320.68	39.20	0.168	0.0212
	100	200	0.764	88.93	0.07	89.56	0.04	0.316	0.0077
		500	0.912	222.04	0.18	133.86	0.31	0.288	0.0037
		1000	0.832	421.49	1.55	234.83	3.26	0.238	0.0085
T_2	50	200	0.648	86.72	0.20	53.97	0.25	0.348	0.0166
		500	0.646	217.23	1.46	89.29	2.39	0.304	0.0121
		1000	0.529	432.36	8.03	156.56	15.41	0.280	0.0140
	100	200	0.763	90.62	0.09	83.90	0.07	0.328	0.0112
		500	0.950	234.47	0.09	109.74	0.15	0.315	0.0041
		1000	0.944	476.40	0.24	132.18	0.62	0.303	0.0045

Table 5. The estimates for k^*, l^* and μ_1 when $\mu_0 = 0.2$ and $\mu_1 = 0.4$

	l^*	n	$\mathrm{pw}_{0.05}$	$\mathrm{M}\widehat{k^*}$	$\mathrm{SE}\widehat{k^*}$	$\mathrm{M}\widehat{l^*}$	$\mathrm{SE}\widehat{l^*}$	$\mathrm{M}\widehat{\mu}_1$	$\mathrm{SE}\widehat{\mu}_1$
T_1	50	200	0.398	75.16	0.38	73.16	0.48	0.390	0.0110
		500	0.305	171.59	4.27	171.02	8.19	0.307	0.0172
		1000	0.185	326.68	21.52	352.80	46.64	0.258	0.0247
	100	200	0.640	81.50	0.11	90.62	0.04	0.401	0.0123
		500	0.760	208.46	0.38	152.40	0.56	0.377	0.0056
		1000	0.616	386.29	2.75	288.32	5.26	0.317	0.0120
T_2	50	200	0.438	83.13	0.32	60.51	0.39	0.434	0.0196
		500	0.421	200.84	2.68	116.81	4.90	0.385	0.0178
		1000	0.306	398.46	13.83	219.69	26.55	0.339	0.0192
	100	200	0.623	83.88	0.13	83.83	0.08	0.410	0.0177
		500	0.826	223.69	0.24	125.86	0.33	0.407	0.0062
		1000	0.794	448.41	1.15	180.30	2.14	0.381	0.0073

From results presented in Tables 3 to 5 we can draw several conclusions.

- For every fixed l^* and all three pairs of μ_0 and μ_1, let n decrease. We observe that the sample means $\mathrm{M}\widehat{l^*}$ approach true values l^* except for T_2 with $\mu_0 = 0.1$, $\mu_1 = 0.3$ and $l^* = 100$. The sample means of squared errors $\mathrm{SE}\widehat{k^*}$ and $\mathrm{SE}\widehat{l^*}$ rapidly approach zero.

- For every fixed n and all pairs μ_0, μ_1, let l^* increase. We see that for both tests $\mathrm{M}\widehat{k^*}$ approach their true values k^*, $\mathrm{SE}\widehat{k^*}$ and $\mathrm{SE}\widehat{l^*}$ decrease.

- In two above cases no explicit conclusion can be drawn about $\mathrm{M}\widehat{\mu}_1$ and $\mathrm{SE}\widehat{\mu}_1$, except that they behave very alike, which means that, when $\mathrm{SE}\widehat{\mu}_1$ decreases, $\mathrm{M}\widehat{\mu}_1$ gets closer to the true value μ_1.

- Fix l^*/n but let l^* and n increase. For all pairs μ_0, μ_1, the means of squared errors decrease for all three parameters under investigation k^*, l^* and μ_1.

- Let $|\mu_1 - \mu_0|$ increase. In all cases $\mathrm{M}\widehat{k^*}$, $\mathrm{SE}\widehat{k^*}$, $\mathrm{M}\widehat{l^*}$, $\mathrm{SE}\widehat{l^*}$ improve. We mean that the empirical means approach their true values and the means of squared errors decrease.

- Now fix $|\mu_1 - \mu_0|$ but let μ_0 and μ_1 increase. Similarly as for the behavior of the power of both tests explained in Section 4, the results for all parameters get worse both in mean and mean square error sense.

Comparing the results of both tests, we see that when estimating the beginning of the epidemic, $\mathrm{M}\widehat{k}^*$ for T_2 are closer to their true values k^* in all cases. Also for T_2, $\mathrm{M}\widehat{l}^*$ are closer to l^*, $\mathrm{SE}\widehat{k}^*$ and $\mathrm{SE}\widehat{l}^*$ are smaller in all cases except for $l^* = 100$ and $n = 200$ and for all μ_0, μ_1. For $n = 200$ and $l^* = 50$, 100, $\mathrm{M}\widehat{\mu}_1$ is closer to μ_1 and $\mathrm{SE}\widehat{\mu}_1$ is smaller for T_1 test. For $n = 1000$ and both l^*, these values are in the favor of T_2 test. The rest of the cases are difficult to classified. The results in this analysis somewhat agree with the results of the power analysis.

6 An application to human glucagon gene data

In this Section we investigate human glucagon gene (GCG), located on chromosome 2, as a sequence of four main bases A, C, G, T. This gene consists of 6 exons and 5 introns and we deal with the introns 2, 3, and 4. We refer to National Center's for Biotechnology Information internet page[2] for more information about this gene and the sequence itself. Every base was analyzed separately. We transformed the initial sequence to that of one's and zero's: the base under analysis was marked by 1 and the other three by 0. Using both tests, T_1 and T_2, we have first tested the null hypothesis of no epidemic against epidemic alternative and computed p-value estimates according to (16). Then we have estimated the unknown parameters of epidemic (also in the cases where the H_0 was not rejected for small α_s values). The same procedure was done for all three introns. We present the results in Table 6.

In Table 6, T stands for either of statistics, first for T_1 and in the next line for T_2. Blank positions in T_2 case means that the values are the same as for T_1 in a line above.

For both statistics the p-value estimates are quite similar except for the intron 2 bases T and A, and intron 4 base C. Both tests significantly reject H_0 for intron 3 and all bases, also for intron 4 base A, intron 2 base G, and with $\alpha_s = 0.1$

[2] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene&cmd=Retrieve&dopt =Graphics&list_uids=2641

Table 6. The results of analysis for GCG introns 2, 3 and 4 (sample sizes are $n = 1572$, 1675 and 1369 respectively)

Intron	Base	$S(0,n)$	T	$\widehat{p}$	$\widehat{l}^*$	$\widehat{k}^*$	$\widehat{m}^*$	$\widehat{\mu}_0$	$\widehat{\mu}_1$
2	T	566	1.503	0.167	701	473	1174	0.327	0.401
			2.131	0.226					
	A	516	1.405	0.254	689	473	1162	0.358	0.290
			1.994	0.343					
	C	263	1.620	0.094	562	709	1271	0.144	0.210
			2.379	0.090	293	842	1135	0.150	0.242
	G	227	2.003	0.008	1059	227	1286	0.199	0.118
			2.925	0.008					
3	T	455	2.366	0.000	501	654	1155	0.308	0.186
			3.587	0.000	312	654	966	0.302	0.141
	A	530	2.166	0.002	723	318	1041	0.273	0.373
			3.107	0.003	319	666	985	0.289	0.433
	C	333	2.630	0.000	699	403	1102	0.243	0.137
			3.745	0.000					
	G	357	2.966	0.000	691	481	1172	0.163	0.285
			4.231	0.000	609	563	1172	0.167	0.294
4	T	446	1.405	0.254	437	253	690	0.352	0.270
			2.057	0.283					
	A	506	2.025	0.007	638	342	980	0.320	0.426
			2.868	0.010					
	C	206	1.243	0.451	316	981	1297	0.135	0.203
			2.254	0.148	126	1171	1297	0.138	0.278
	G	211	1.250	0.442	342	372	714	0.170	0.105
			1.901	0.439					

intron 2 base C. In the cases where both tests do not reject H_0, with small values of α_s, the estimates for the parameters of epidemic are the same for both tests except the case of intron 4 base C. In this case T_2 gives quite smaller $\widehat{p}$ (nearly indicating significant change), shorter the length and bigger the size $|\widehat{\mu}_1 - \widehat{\mu}_0|$. When the tests significantly reject H_A but give different results, again T_2 indicates shorter and bigger epidemics. For $\widehat{p}$ smaller than 0.1 the estimated lengths of epidemics might seem quite big, $\widehat{l}^*/n$ ranging approximately from $1/5$ (corresponding intron 2 base C and intron 3 bases T and A, all in the case of T_2 test) to $1/2$ (intron 4 base A; the case of intron 2 base G may be regarded as the epidemic of length $n - \widehat{l}^* = 513$). But on the other hand the values of $|\widehat{\mu}_1 - \widehat{\mu}_0|$ are quite small. Minimum value 0.066 is in the case of intron 2 base C for test T_2 and maximum

0.161 for test T_1 in the case of intron 3 base T. Thus bigger length somewhat must compensate for smaller size to detect epidemic (see condition (10)).

7 Conclusions

When the means of squared errors (SE) are big, the results of both procedures T_1 and T_2 should be qualified with care. On the other hand, when the power is small, the results are of little value even if the means of squared errors are small. Thus only when the power reaches high levels and the SE are small we might be able to get reliable estimates for k^*, l^* or μ_1 and see the true picture of the behavior of both tests. These cases might be when $|\mu_1 - \mu_0| = 0.2$, $l^* = 100$ and all values of n in the Tables 4 and 5. These cases strengthen the notion that for big values of l^*/n (1/2), T_1 test performs slightly better, for smaller l^*/n (1/5) moderate advantage is for T_2, and for small l^*/n (1/10), test T_2 shows its biggest advantage.

The example of human glucagon gene demonstrates two alternative (as a test statistic using T_1 or T_2) ways to analyze the nucleotide sequences. It shows that, when both tests strongly indicate the presence of an epidemic, often T_2 test estimates shorter epidemic with bigger change in proportion of a certain nucleotide base. This example can be regarded as a template for further applications of methods presented for search and location of epidemic. Not only one certain nucleotide base can be under investigation, but also any codon or amino acid.

Appendix

For the proofs of Theorem 1 and Theorem 2 consider a sequence of i.i.d. random binomial variables $X_1', \ldots, X_n'$ characterized by $\mathrm{P}(X_i' = 1) = \mu_0$, $i \in \{1, \ldots, n\}$. Also for independent but not identically distributed variables $X_1, \ldots, X_n$ assume $\mathrm{P}(X_i = 1) = \mu_1$, $i \in I_1$, $I_1 = \{k^* + 1, \ldots, m^*\}$ and $X_i = X_i'$ when $i \in I_0$, $I_0 = \{1, \ldots, n\} \setminus I_1$.

Proof of Theorem 2. Denote $M_n = n^{1/2} h_n s / \rho(h_n)$. Next expand

$$S(k^*, k^* + l^*) = \left(1 - \frac{l^*}{n}\right) \sum_{i \in I_1} X_i - \frac{l^*}{n} \sum_{i \in I_0} X_i' = l^* \left(1 - \frac{l^*}{n}\right)(\mu_1 - \mu_0) + R_n,$$

$$R_n = \left(1 - \frac{l^*}{n}\right) \sum_{i \in I_1} (X_i - \mu_1) - \frac{l^*}{n} \sum_{i \in I_0} \bigl(X_i' - \mu_0\bigr).$$

Noting that $(\overline{X}(1-\overline{X}))^{1/2} < 1$, we find the lower bound LB for $\mathrm{UI}(n, \rho)$:

$$\begin{aligned} \mathrm{UI}(n,\rho) &> n^{-1/2} \max_{0<l<n} V_\rho(l) \geq \frac{n^{-1/2}}{\rho(h_n)} \bigl|S(k^*, k^* + l^*)\bigr| \\ &\geq M_n \left(1 - \frac{|R_n|}{n h_n s}\right) =: \mathrm{LB}. \end{aligned} \tag{A.1}$$

Since both $\mu_0 - \mu_0^2$ and $\mu_1 - \mu_1^2$ are less or equal $1/4 < 1$, we have

$$\mathrm{E}\left(\frac{|R_n|}{n h_n s}\right)^2 \leq \left(1 - \frac{l^*}{n}\right)^2 \frac{l^*(\mu_1 - \mu_1^2)}{n^2 h_n^2 s^2} + \left(\frac{l^*}{n}\right)^2 \frac{(n - l^*)(\mu_0 - \mu_0^2)}{n^2 h_n^2 s^2} \leq \frac{1}{n h_n s^2},$$

which tends to 0, provided $n^{1/2} h_n^{1/2} s = M_n \rho(h_n)/h_n^{1/2} \to \infty$. But the latter follows from the divergence of M_n. Indeed, if $h_n \to 0$ (when $l^*/n \to 0$ or $l^*/n \to 1$), $\rho(h_n)/h_n^{1/2} \to \infty$. Thus the random element $1 - |R_n|/n h_n s$ is $O_{\mathrm{P}}(1)$ and the lower bound in (A.1) tends to infinity provided $M_n \to \infty$.

Next proof requires more notations. For any k and l, $0 \leq k < k + l \leq n$

$$I_{kl} = \{k+1, \ldots, k+l\}, \quad A_{kl} = I_{kl} \cap I_1, \quad |A_{kl}| = \# A_{kl}.$$

Note that $|A_{kl}| \leq l \wedge l^*$. Use $\overline{X} = \overline{X'} + (1/n) \sum_{i \in I_1} (X_i - X_i')$ and (2) to get

$$\begin{aligned} S(k, k+l) &= \sum_{i \in I_{kl}} X_i' + \sum_{i \in I_{kl}} \bigl(X_i - X_i'\bigr) - l\overline{X'} - \frac{l}{n} \sum_{i \in I_1} \bigl(X_i - X_i'\bigr) \\ &= S'(k, k+l) + Z_{kl} - (l/n) Z_1 \\ &\quad + \bigl(|A_{kl}| - l l^*/n\bigr)(\mu_1 - \mu_0), \end{aligned} \tag{A.2}$$

where $S'(k, k+l) = \sum_{i \in I_{kl}} (X_i' - \overline{X'})$ and

$$Z_{kl} = \sum_{i \in A_{kl}} \eta_i, \quad Z_1 = \sum_{i \in I_1} \eta_i, \quad \eta_i = (X_i - \mathrm{E}X_i) - \bigl(X_i' - \mathrm{E}X_i'\bigr).$$

If $i \in I_0$, then $\eta_i \equiv 0$. When $I_{kl} = I_1$, we see that $Z_{kl} = Z_1$, $|A_{kl}| = l^*$ and

$$S(k^*, k^* + l^*) = S'(k^*, k^* + l^*) + (1 - l^*/n) Z_1 + (1 - l^*/n) l^* (\mu_1 - \mu_0). \tag{A.3}$$

Proof of Theorem 3. We follow the proofs of Theorem 4 and Proposition 13 in Račkauskas and Suquet [7]. Event $\{|\widehat{l}^*/l^* - 1| \geq \varepsilon\}$ is equivalent to $\{\widehat{l}^* \leq (1-\varepsilon)l^* \cup \widehat{l}^* \geq (1+\varepsilon)l^*\}$. On this event we have

$$\Big\{ \max_{0<l\leq(1-\varepsilon)l^*} V_\rho(l) = \max_{0<l\leq l^*} V_\rho(l) \cup \max_{l^*\leq l<n} V_\rho(l) = \max_{(1+\varepsilon)l^*\leq l<n} V_\rho(l) \Big\}.$$

Hence for any upper bounds UB_1 and UB_2 of $\max_{0<l\leq(1-\varepsilon)l^*} V_\rho(l)$ and $\max_{l^*\leq l<n} V_\rho(l)$ and for lower bounds LB_1 and LB_2 of $\max_{0<l\leq l^*} V_\rho(l)$ and $\max_{(1+\varepsilon)l^*\leq l<n} V_\rho(l)$ we have

$$\begin{aligned} \mathrm{P}\big(|\widehat{l}^*/l^* - 1| \geq \varepsilon\big) &\leq \mathrm{P}\big(\mathrm{UB}_1 \geq \mathrm{LB}_1 \cup \mathrm{UB}_2 \geq \mathrm{LB}_2\big) \\ &\leq \mathrm{P}(\mathrm{UB}_1 \geq \mathrm{LB}_1) + \mathrm{P}(\mathrm{UB}_2 \geq \mathrm{LB}_2). \end{aligned} \tag{A.4}$$

We will find upper and lower bounds such that (A.4) converges to zero.

Recall that by assumption $l^*/n \to 0$. This allows us to replace $\varrho(h) = \big(h(1-h)\big)^\alpha$ by $\rho(h) = h^\alpha$ in the rest of the proof. For shortness we will use the following notations

$$\begin{aligned} E_1 &= \max_{0<l<n} \frac{1}{(l/n)^\alpha} \max_{0\leq k\leq n-l} \big|S'(k,k+l)\big|, \quad E_3 = \frac{|Z_1|}{(l^*/n)^\alpha}, \\ E_2 I &= \max_{l\in I} \frac{1}{(l/n)^\alpha} \max_{0\leq k\leq n-l} |Z_{kl}|, \quad I \subset \{1,\dots,n\}. \end{aligned} \tag{A.5}$$

From (3) and (A.3) we get

$$\max_{0<l\leq l^*} V_\rho(l) \geq \frac{|S(k^*,k^*+l^*)|}{(l^*/n)^\alpha} \geq \frac{(1-l^*/n)l^*s}{(l^*/n)^\alpha} - E_1 - (1-l^*/n)E_3 := \mathrm{LB}_1.$$

For $l \leq l^*$ we can use $|A_{kl}| \leq l$ and so $||A_{kl}| - l(l^*/n)| \leq \max\{l(l^*/n), l(1 - l^*/n)\} \leq l(1-l^*/n)$ for large n. Using (A.2) and the fact that $l/(l/n)^\alpha$ is increasing in l, we find an upper bound

$$\max_{0<l\leq(1-\varepsilon)l^*} V_\rho(l) \leq \frac{(1-l^*/n)(1-\varepsilon)l^*s}{\big((1-\varepsilon)l^*/n\big)^\alpha} + E_1 + E_2(0,l^*] + (l^*/n)E_3 =: \mathrm{UB}_1.$$

So we have that

$$\mathrm{P}(\mathrm{UB}_1 \geq \mathrm{LB}_1) \leq \mathrm{P}\big(2E_1 + E_2(0,l^*] + E_3 \geq \lambda_1\big), \tag{A.6}$$

where

$$\lambda_1 = \frac{(1 - l^*/n)l^*\delta_1(\varepsilon)s}{(l^*/n)^\alpha}, \quad \delta_1(\varepsilon) = 1 - (1-\varepsilon)^{1-\alpha}.$$

Similarly we look for upper and lower bounds UB_2 and LB_2. First,

$$\max_{l^* \le l < n} V_\rho(l) \ge \frac{(1 - l^*/n)l^*s}{(l^*/n)^\alpha} - E_1 - E_3 =: \mathrm{LB}_2.$$

To find an upper bound we analyze two cases. In the case where $|A_{kl}| - ll^*/n \ge 0$, we use $|A_{kl}| \le l^*$ to obtain $||A_{kl}| - ll^*/n| \le l^*(1 - l/n)$. When $|A_{kl}| - ll^*/n \le 0$, $||A_{kl}| - ll^*/n| \le ll^*/n$. Then the upper bound is

$$\begin{aligned}\max_{(1+\varepsilon)l^* \le l < n} V_\rho(l) &\le \left\{ \frac{\bigl(1 - (1+\varepsilon)l^*/n\bigr)l^*}{((1+\varepsilon)l^*/n)^\alpha} \vee l^* \right\} s + E_1 + E_2[l^*, n) + E_3 \\ &\le \frac{(1 - (1+\varepsilon)l^*/n)l^*s}{((1+\varepsilon)l^*/n)^\alpha} + E_1 + E_2[l^*, n) + E_3 := \mathrm{UB}_2\end{aligned}$$

(we use $|Z_1| \le |Z_1|/(l^*/n)^\alpha$). Similarly to (A.6), we can now write

$$\mathrm{P}(\mathrm{UB}_2 \ge \mathrm{LB}_2) \le \mathrm{P}\bigl(2E_1 + E_2[l^*, n) + 2E_3 \ge \lambda_2\bigr), \tag{A.7}$$

where, if $\delta_2(\varepsilon) = 1 - (1+\varepsilon)^{-\alpha}$, then

$$\frac{(1 - l^*/n)l^*s}{(l^*/n)^\alpha}\left(1 - \frac{1 - (1+\varepsilon)l^*/n}{(1 - l^*/n)(1+\varepsilon)^\alpha}\right) \ge \frac{(1 - l^*/n)l^*\delta_2(\varepsilon)s}{(l^*/n)^\alpha} =: \lambda_2.$$

Our next step is to obtain the convergence to zero of the probabilities on the right hand sides of (A.6) and (A.7). For either λ_1 or λ_2 we will write λ, and $c(\varepsilon)$ denotes a constant (may be different in different parts of the proof) depending on ε and such that $c(\varepsilon) \to 0$ as $\varepsilon \to 0$.

First we analyze $\mathrm{P}(E_1 \ge c\lambda)$ for some constant $c > 0$. We have

$$\begin{aligned}E_1 \le \max_{0<l<n} \frac{1}{(l/n)^\alpha} \max_{0 \le k \le n-l} \Bigl| \sum_{i \in I_{kl}} \bigl(X_i' - \mathrm{E}X_i'\bigr) \Bigr| \\ + \max_{0<l<n} \frac{l/n}{(l/n)^\alpha} \Bigl| \sum_{i=1}^{n} \bigl(X_i' - \mathrm{E}X_i'\bigr) \Bigr| \le 2 \max_{0<l<n} \frac{1}{(l/n)^\alpha} \max_{0 \le k \le n-l} |S_{k+l} - S_k|,\end{aligned}$$

where $S_i = X_1' - \mathrm{E}X_1' + \cdots + X_i' - \mathrm{E}X_i'$, $i = 1, \ldots, n$. Defining the integer J_n by $2^{J_n} \le n < 2^{J_n+1}$ and using the same technique of dyadic splitting of the

l's and k's indexation ranges as in the proof of Proposition 13 in Račkauskas and Suquet [7], we obtain for some constant $c > 0$

$$\begin{aligned}\mathrm{P}(E_1 \geq c\lambda) &\leq 8 \sum_{j=1}^{J_n+1} 2^{j-1} \exp(-2^{ja} b) \leq 8 \sum_{j=1}^{J_n+1} \int_{2^{j-1}}^{2^j} \exp(-x^a b) dx \\ &\leq 8 \int_1^\infty \exp(-x^a b) dx = 8(1/a)(1/b)^{1/a} \Gamma(1/a, b).\end{aligned} \tag{A.8}$$

Here $\Gamma(1/a, b)$ is the incomplete gamma function and

$$a = 1 - 2\alpha, \quad b = b_n(\varepsilon) = c(\varepsilon) l^* (l^*/n)^a s^2. \tag{A.9}$$

We finally have that $\mathrm{P}(E_1 \geq c\lambda) \to 0$ provided that condition (13) holds.

Next we analyze $E_2(0, l^*]$ and $E_2[l^*, n)$ (see (A.5)). For both cases

$$\mathrm{P}\biggl(\max_l \frac{1}{(l/n)^\alpha} \max_{0 \leq k \leq n-l} |Z_{kl}| \geq c\lambda\biggr) \leq \sum_l \sum_{0 \leq k \leq n-l} \mathrm{P}\biggl(\frac{|Z_{kl}|}{(l/n)^\alpha} \geq c\lambda\biggr)$$

for some constant $c > 0$. Using Hoeffding's inequality we estimate

$$\mathrm{P}\biggl(\frac{|Z_{kl}|}{(l/n)^\alpha} \geq c\lambda\biggr) \leq 2 \exp\biggl(-\frac{c(\varepsilon)(l^*)^2 s^2 (l/l^*)^{2\alpha}}{|A_{kl}|}\biggr) \leq 2 \exp\bigl(-c(\varepsilon) l^* s^2\bigr).$$

When $0 < l \leq l^*$, there are at most $2l^*$ indexes k for which A_{kl} is not empty and so Z_{kl} is a proper sum with non-empty summation index set. When $l^* \leq l < n$, we can find at most $(n + l^*)/2$ such indexes k. Thus

$$\begin{aligned}\sum_{0<l\leq l^*} \sum_{0 \leq k \leq n-l} \mathrm{P}\bigl(|Z_{kl}| \geq c\lambda_1 (l/n)^\alpha\bigr) &\leq 2l^* \sum_{0<l\leq l^*} 2 \exp\bigl(-c(\varepsilon) l^* s^2\bigr) \\ &\leq 4 \exp\bigl(-c(\varepsilon) l^* s^2 + 2\log(l^*)\bigr), \quad \text{(A.10)} \\ \sum_{l^* \leq l < n} \sum_{0 \leq k \leq n-l} \mathrm{P}\bigl(|Z_{kl}| \geq c\lambda_2 (l/n)^\alpha\bigr) &\leq \frac{n + l^*}{2} \sum_{l^* \leq l < n} 2 \exp\bigl(-c(\varepsilon) l^* s^2\bigr) \\ &\leq \exp\bigl(-c(\varepsilon) l^* s^2 + 2\log(n)\bigr). \quad \text{(A.11)}\end{aligned}$$

If condition (12) holds, (A.11) converges to zero; (A.10) approaches zero when $l^* s^2 / \log(l^*) \to \infty$. But the latter follows from the same condition (12).

For E_3 and some constant $c > 0$ we get

$$\mathrm{P}(E_3 \geq c\lambda) = \mathrm{P}\bigl(|Z_1| \geq c\lambda (l^*/n)^\alpha\bigr) \leq 2 \exp\bigl(-c(\varepsilon) l^* s^2\bigr), \tag{A.12}$$

which tends to zero when $l^* s^2 \to \infty$. This condition follows again from (12). Consequently the convergence in probability is proved.

To prove $\widehat{l}^*/l^* \to 1$ almost surely we show that for all $\varepsilon > 0$

$$\sum_{n=1}^{\infty} \mathrm{P}\bigl(\bigl|\widehat{l}^*/l^* - 1\bigr| \geq \varepsilon\bigr) < \infty.$$

Using estimates (A.8), (A.10), (A.11) and (A.12) this reduces in proving the convergence of the following three series

$$\sum_{n=1}^{\infty} \frac{1}{a}\biggl(\frac{1}{b_n(\varepsilon)}\biggr)^{1/a} \Gamma\bigl(1/a, b_n(\varepsilon)\bigr), \quad \sum_{n=1}^{\infty} \exp\bigl(-\varepsilon l^* s^2 + c\log(n)\bigr),$$
$$\sum_{n=1}^{\infty} \exp\bigl(-\varepsilon l^* s^2\bigr),$$

where a and $b_n(\varepsilon)$ are as in (A.9). The convergence of these series follows straightforwardly by conditions (12) and (14).

References

1. P.J. Avery, D.A. Henderson. Detecting a changed segment in DNA sequences, *Appl. Stat.*, **48**, pp. 489–503, 1999.

2. P.J. Avery. The effect of dependence in a binary sequence on tests for a changepoint or a changed segment, *Appl. Stat.*, **50**, pp. 243–246, 2001.

3. M. Csörgő, L. Horváth. *Limit Theorems in Change-Point Analysis*, John Wiley & Sons, Baffins Lane, Chichester, 1997.

4. Q. Yao. Tests for change-points with epidemic alternatives, *Biometrika*, **80**, pp. 179–191, 1993.

5. R.N. Curnow, Y.-X. Fu. Locating a changed segment in a sequence of Bernoulli variables, *Biometrika*, **77**, pp. 295–304, 1990.

6. A. Račkauskas, Ch. Suquet. Hölder norm test statistics for epidemic change, *J. Statist. Plann. Inference*, **126**, pp. 495–520, 2004.

7. A. Račkauskas, Ch. Suquet. Testing epidemic change of infinite dimensional parameters, *Pub. IRMA Lille*, **60-VIII**, 2003.

8. R. Davidson, J.G. MacKinnon. Graphical methods for investigating the size and power of hypothesis tests, *Manchester School*, **66**, pp. 1–26, 1998.