# Sequential Detection of Change-Points in Counting Processes

Allan Gut[1] and Josef G. Steinebach[2]

[1] Department of Mathematics, Uppsala University,
P.O. Box 480,
SE-751 06 Uppsala, Sweden
allan.gut@math.uu.se

[2] Mathematical Institute, University of Cologne
Weyertal 86-90,
D-50931 Köln, Germany
jost@math.uni-koeln.de

**Abstract.** Instead of the standard approach in change-point theory to perform the statistical analysis based on a sample of fixed size, we have, in a series of papers—Gut and Steinebach (2002, 2004, 2005, 2009)—introduced some (truncated) sequential testing procedures for detecting a possible change-point in a sequence of renewal counting data (based on random walks). The main focus of the present paper is on "detection of an early change" and on EWMA charts.

**Keywords.** Change-point, early change, EWMA chart, extreme value asymptotics, first passage time, increment, renewal counting process, sequential test, stationary process, stopping time, strong approximation, training period, Wiener process

## 1 Introduction

The standard situation in a series of observations is that if everything is in order, then the observations follow some kind of common pattern, whereas if something goes astray at some time point, then, from there on, the observations follow a different pattern. One obviously wishes to find out as soon as possible if something goes wrong in order to take appropriate action, and, at the same time, minimize the probability of taking action if nothing is wrong. In this setting one talks about, what is called, the AMOC problem (At Most One Change), which was introduced by Page in the mid-1950's in the context of quality control/control charts.

The classical approach in change-point theory is to perform the statistical analysis based on a sample of fixed size.

Alternatively, and perhaps, more realistically, one cannot, for a variety of reasons, observe the process under consideration continuously, only at, say, equidistant time points, such as once a day, once a week, and so on. In Gut and Steinebach (2002) we suggested some truncated sequential monitoring procedures for detecting a structural break, that is, a "change-point" $k_n^*$, in a series of counting data, e.g., the number of claims of an insurance portfolio, which are sequentially observed at equidistant time-points up to a "truncation point" $n$ (say), thus yielding a "closed-end" procedure. Technically, this amounts to constructing a sequential test for, say,

$$H_0 : k_n^* = n \quad (\text{"no change"})$$

vs. the *two-sided* alternative

$$H_1 : 1 \leq k_n^* < n,\ \theta^* \neq \theta \quad (\text{"change in the drift at } k_n^*\text{"}),$$

based on the observed counting data $N(0), N(1), \ldots, N(n)$, where $\theta$ and $\theta^*$ denote the drifts before and after the change, respectively.

Some limiting extreme value asymptotics (as $n \to \infty$) could be derived in the cited paper under the null hypothesis of "no change", via a strong invariance principle (Proposition 1 below), thus allowing for a choice of critical boundaries in the monitoring schemes in order for the false alarm rate (asymptotically) to attain a prescribed level $\alpha$. Moreover, some limiting properties under the alternative could also be proved showing that the statistical procedures have asymptotic power 1.

In our more recent work, Gut and Steinebach (2009), we look in more detail into the behaviour of the relevant stopping times, in particular the time it takes from the (unknown) change-point until one detects that a change actually has occurred, in other words, asymptotics for stopping times under alternatives are proved.

A review of our results so far will be given in Sections 3 and 4 below.

Another generalization is to consider the problem of more than one change-point. A special case of interest is the so-called epidemic change, for which we refer to Gut and Steinebach (2005).

A problem of a slightly different nature is to allow for some kind of discount, a special case of which are the so-called EWMA (exponentially weighted moving average) charts. A survey of some of the results from Gut and Steinebach (2004) is given in Section 5.

## 2 An invariance principle for renewal processes

As mentioned in the introduction, a crucial tool in the proofs is the following strong invariance principle (Gut and Steinebach (2002), Proposition 1), which allows our renewal counting processes to be approximated by Brownian motions.

Consider a renewal counting process that is based on an i.i.d. sequence $X_1, X_2, \ldots$, up to (say) time $k_n^*$, and then on another i.i.d. sequence $X_1^*, X_2^*, \ldots$, independent of the first one. Suppose that $E\,X = \mu > 0$, that $E\,X^* = \mu^* > 0$, and, further, that the variances $\sigma^2, \sigma^{2*}$ are positive and finite. Let

$$N(t) = \begin{cases} N_0(t), & \text{for} \quad 0 \le t \le k_n^*, \\ N_0(k_n^*) + N_1(t - k_n^*), & \text{for} \quad k_n^* < t \le n, \end{cases}$$

where

$$N_0(t) = \min\{k \ge 1 : \sum_{j=1}^{k} X_j > t\}\,, \qquad N_1(t) = \min\{k \ge 1 : \sum_{j=1}^{k} X_j^* > t\}\,, \quad t \ge 0.$$

**Proposition 1.** *Suppose that $E|X_1|^r < \infty$ and $E|X_1^*|^r < \infty$ for some $r > 2$. Then*

$$\sup_{0 \le t \le n} |N(t) - V(t)| \overset{a.s.}{=} o(n^{1/r}) \quad as \quad n \to \infty,$$

*where*

$$V(t) = \begin{cases} t\theta + \eta W_0(t), & \textit{for} \quad 0 \le t \le k_n^*, \\ V(k_n^*) + (t - k_n^*)\theta^* + \eta^* W_1(t - k_n^*), & \textit{for} \quad k_n^* < t \le n, \end{cases}$$

*with $\theta = 1/\mu$, $\eta^2 = \sigma^2/\mu^3$, $\theta^* = 1/\mu^*$, $\eta^* = \sigma^{*2}/\mu^{*3}$, and two independent (standard) Wiener processes $\{W_i(t),\, t \ge 0\}$, $i = 1, 2$.*

*Remark 1.* The inspiration for the proposition was the strong invariance principle for renewal processes due to Csörgő et al. (1987).

*Remark 2.* Since our results are based on the strong invariance principle above, we tacitly assume throughout in the following that the conditions required for the application of Proposition 1 are fulfilled.

*Remark 3.* In most cases a weak invariance principle, which, in turn, is available for wider classes of stochastic processes, would be sufficient (cf. Horváth and Steinebach (2000), Section 1).

## 3 Null asymptotics

From the sequential observations $N(0), N(1), \ldots, N(n)$, we compute the random variables

$$Y_k = Y_{k,n} = \frac{N(k) - N(k - h_n) - h_n\theta}{\eta\sqrt{h_n}}, \quad k = h_n, \ldots, n,$$

$$Z_k = Z_{k,n} = \frac{N(k) - k\theta}{\eta\sqrt{k}}, \quad k = k_n, \ldots, n,$$

and the stopping times

$$\tau_n^{(1)} = \min\{h_n \le k \le n : |Y_k| > c_n^{(1)}\},$$
$$\tau_n^{(2)} = \min\{k_n \le k \le n : |Z_k| > c_n^{(2)}\}$$

$(\min\emptyset := +\infty)$, where $c_n^{(1)}, c_n^{(2)}$ are suitable critical values and $h_n, k_n$ are the lengths of the respective "training periods".

*Remark 4.* For the sake of simplicity, we assume that the "in-control" parameters $\theta, \eta$ are known, but they can also be replaced by "suitable" sequential estimates (see Gut and Steinebach (2002), Section 5).

The critical values $c_n^{(1)}, c_n^{(2)}$ are chosen such that the false alarm rates (asymptotically) attain a prescribed level $\alpha$, i.e.,

$$P_{H_0}(\tau_n^{(1)} < \infty) = P_{H_0}\big(\max_{h_n \le k \le n} |Y_k| > c_n^{(1)}\big) \approx \alpha, \qquad \text{and}$$

$$P_{H_0}(\tau_n^{(2)} < \infty) = P_{H_0}\big(\max_{k_n \le k \le n} |Z_k| > c_n^{(2)}\big) \approx \alpha,$$

which can be achieved via the following extreme value asymptotics from Gut and Steinebach (2002), Section 4:

**Theorem 1.** *If $h_n \ll n$, but $\sqrt{h_n} \gg n^{1/r}$, then, with normalizations*

$$a_n^{(1)} = \sqrt{2\log(n/h_n)} \quad \text{and} \quad b_n^{(1)} = 2\log(n/h_n) + \frac{1}{2}\log\log(n/h_n) - \frac{1}{2}\log\pi\,,$$

*we have, under $H_0$,*

$$a_n^{(1)} \max_{h_n \le k \le n} |Y_k| - b_n^{(1)} \xrightarrow{d} E \quad \text{as} \quad n \to \infty,$$

*where $P(E \le x) = \exp(-2e^{-x})$, $x \in \mathbb{R}$, that is, the critical value $c_n^{(1)}$ can (asymptotically) be chosen as*

$$c_n^{(1)} = \frac{E_{1-\alpha} + b_n^{(1)}}{a_n^{(1)}} \qquad \big(\sim \sqrt{2\log(n/h_n)}\,\big),$$

*where $E_{1-\alpha}$ denotes the $(1-\alpha)$-quantile of the (two-sided) Gumbel distribution.*

**Theorem 2.** *If $k_n \ll n$, but $\sqrt{k_n} \gg n^{1/r}$, then, with normalizations*

$$a_n^{(2)} = \sqrt{2\log\log(n/k_n)} \quad \text{and} \quad b_n^{(2)} = 2\log\log(n/k_n) + \frac{1}{2}\log\log\log(n/k_n) - \frac{1}{2}\log(4\pi)\,,$$

*we have, under $H_0$,*

$$a_n^{(2)} \max_{k_n \le k \le n} |Z_k| - b_n^{(2)} \xrightarrow{d} E \quad \text{as} \quad n \to \infty,$$

*that is, the critical value $c_n^{(2)}$ can (asymptotically) be chosen as*

$$c_n^{(2)} = \frac{E_{1-\alpha} + b_n^{(2)}}{a_n^{(2)}} \qquad \big(\sim \sqrt{2\log\log(n/k_n)}\,\big),$$

*with $E_{1-\alpha}$ as in Theorem 1.*

An obvious question of interest would be how quickly a possible change-point $k_n^*$ can be detected by the monitoring procedure, that is, what can be said about the behaviour of the stopping times $\tau_n^{(1)}, \tau_n^{(2)}$ or the detection delays $\tau_n^{(1)} - k_n^*, \tau_n^{(2)} - k_n^*$ under the alternative $H_1$? This is the topic of the next section, in which it will turn out that the asymptotic behaviour of the stopping times, suitably normalized, is normal.

## 4 Asymptotics under an early-change alternative

Similar to Aue et al. (2008), we consider an "early change" scenario here, that is, we assume that the change-point $k_n^*$ does not occur too late compared to the length of the training period in the following technical sense:

$$k_n^* = \mathcal{O}\big(h_n \log^\gamma(n/h_n)\big) \quad \text{as} \quad n \to \infty \quad (\text{for some } \gamma > 0). \tag{1}$$

Then we have the following asymptotic normality (see Gut and Steinebach (2009), Section 5):

**Theorem 3.** *Assume that* (1) *holds. If $\{h_n\}$ is as in Theorem 1, then, under $H_1$,*

$$\frac{\tau_n^{(1)} - k_n^*}{\frac{\eta}{|\theta^*-\theta|}\sqrt{h_n}} - c_n^{(1)} \xrightarrow{d} N(0,1) \quad \textit{as} \quad n \to \infty.$$

*Remark 5.* It is obvious from the proof that

$$P_{H_1}(\tau_n^{(1)} \geq k_n^*) \;\to\; 1 \quad \text{as} \quad n \to \infty.$$

For the second stopping time we similarly assume that

$$k_n^* = \mathcal{O}\big(k_n \log^\gamma(n/k_n)\big) \quad \text{as} \quad n \to \infty \quad (\text{for some } \gamma > 0). \tag{2}$$

**Theorem 4.** *Assume that* (2) *holds. If $\{k_n\}$ is as in Theorem 2, then, under $H_1$,*

$$\frac{\tau_n^{(2)} - k_n^*}{\frac{\eta}{|\theta^*-\theta|}\sqrt{k_n^*}} - c_n^{(2)} \xrightarrow{d} N(0,1) \quad \textit{as} \quad n \to \infty.$$

*Remark 6.* Here it is also obvious from the proof that

$$P_{H_1}(\tau_n^{(2)} \geq k_n^*) \;\to\; 1 \quad \text{as} \quad n \to \infty.$$

In Gut and Steinebach (2002), Section 5, we also considered

$$\hat{Y}_k = \hat{Y}_{k,n} = \frac{N(k) - N(k-h_n) - h_n\hat{\theta}_k}{\hat{\eta}_k\sqrt{h_n}}, \quad k = \hat{h}_n, \ldots, n,$$

and, for testing $H_0$ against the two-sided alternative $H_1$,

$$\hat{\tau}_n^{(1)} = \min\{\hat{h}_n \leq k \leq n : |\hat{Y}_k| > \hat{c}_n^{(1)}\}$$

($\min\emptyset := +\infty$), where $\hat{c}_n^{(1)}$ again is a critical value and $\hat{\theta}_k$ and $\hat{\eta}_k$ above are "suitable" sequential estimates. The "null asymptotics" from Theorem 1 (in the case of $\theta, \eta$ known) retain, so that $\hat{c}_n^{(1)}$ can also be chosen from an extreme value asymptotic, that is, we have

**Theorem 5.** *If $h_n \ll \hat{h}_n \ll n$, but $\sqrt{h_n} \gg n^{1/r}$, then, under $H_0$, with the same normalizing sequences $\{a_n^{(1)}\}$ and $\{b_n^{(1)}\}$ as in Theorem 1,*

$$a_n^{(1)} \max_{\hat{h}_n \leq k \leq n} |\hat{Y}_k| - b_n^{(1)} \xrightarrow{d} E \quad \textit{as} \quad n \to \infty,$$

*i.e., the critical value $\hat{c}_n^{(1)}$ can (asymptotically) be chosen as*

$$\hat{c}_n^{(1)} = c_n^{(1)} = \frac{E_{1-\alpha} + b_n^{(1)}}{a_n^{(1)}} \quad \big(\sim \sqrt{2\log(n/h_n)}\,\big),$$

*with $E_{1-\alpha}$ as in Theorem 1.*

Moreover, with $\hat{\theta} = \hat{\theta}_{\hat{\tau}_n^{(1)}}, \hat{\eta} = \hat{\eta}_{\hat{\tau}_n^{(1)}}$, we have (see Gut and Steinebach (2009), Section 6):

**Theorem 6.** *If $\{h_n\}$ and $\{\hat{h}_n\}$ are as in Theorem 5, then, under $H_1$,*

$$\frac{\hat{\tau}_n^{(1)} - k_n^*}{\frac{\hat{\eta}}{|\theta^*-\hat{\theta}|}\sqrt{h_n}} - c_n^{(1)} \xrightarrow{d} N(0,1) \quad \textit{as} \quad n \to \infty.$$

## 5 EWMA charts

Let $0 < \lambda = \lambda_n \leq 1$ be a weight parameter (which thus may vary with $n$), and let $Y_k = Y_{k,n}$ be given as before. Set $S_{h_n-1} = S_{h_n-1,n} = 0$, and define, recursively,

$$S_k = S_{k,n} = \lambda Y_k + (1-\lambda)S_{k-1} = \sum_{j=0}^{k-h_n}(1-\lambda)^j Y_{k-j}, \quad k = h_n, \ldots, n.$$

The relevant stopping time in this setting is

$$\tau_n^{(3)} = \min\{h_n \leq k \leq n : |S_k| > c_n^{(3)}\}\,.$$

Vaguely speaking the stochastic process describing these charts are sums of exponentially weighted (normalized) windows of a random walk, so that, in contrast to the previous models, we are facing heavily correlated data.

### 5.1 Null asymptotics

Following is the EWMA analog of Theorem 1.

**Theorem 7.** *Let $\{h_n\}$ be given as in Theorem 1 and choose $\lambda = \lambda_n = a/h_n$ with $a > 0$ fixed. Then, with normalizations*

$$a_n^{(3)} = \sqrt{2\log(n/h_n)} \quad \textit{and} \quad b_n^{(3)} = 2\log(n/h_n) + \frac{1}{2}\log(C/(2\pi^2))\,,$$

*we have, under $H_0$,*

$$a_n^{(3)}\frac{\max_{h_n\leq k\leq n}|S_k|}{v} - b_n^{(3)} \xrightarrow{d} E \quad \textit{as} \quad n\to\infty,$$

*where*

$$v^2 = v^2(a) = 1 - \frac{1}{a}(1-e^{-a}) \quad \textit{and} \quad C = C(a) = \frac{a(1-e^{-a})}{2(1-\frac{1}{a}(1-e^{-a}))} = \frac{a^2}{2}\cdot\frac{1-v^2}{v^2}\,,$$

*from which it follows that the critical value $c_n^{(3)}$ can (asymptotically) be chosen as*

$$c_n^{(3)} = \frac{v(E_{1-\alpha} + b_n^{(3)})}{a_n^{(3)}}\,,$$

*where $E_{1-\alpha}$ denotes the $(1-\alpha)$-quantile of $E$ as before.*

### 5.2 Asymptotics under the alternative

The following asymptotics can be used to get an idea of the (asymptotic) power of our sequential tests based on EWMA charts.

**Theorem 8.** *Let $\{h_n\}$ be as in Theorem 7 with the additional assumption that*

$$\frac{\log((n-k_n^*)/h_n)}{\log\log n} \to \infty \quad \textit{as} \quad n\to\infty,$$

*and set*

$$a_n^{(4)} = \sqrt{2\log((n-k_n^*)/h_n)} \quad \textit{and} \quad b_n^{(4)} = 2\log((n-k_n^*)/h_n) + \frac{1}{2}\log(C/(2\pi^2)),$$

*where $C$ is as given in Theorem 7. Further, choose $\lambda = \lambda_n = a/h_n$, for some fixed $a > 0$. Then, under $H_1$, we have the following asymptotics:*

$$a_n^{(4)}\frac{\eta}{v\eta^*}\Big(\max_{h_n\leq k\leq n}|S_k| - \frac{|\theta-\theta^*|}{\eta}\sqrt{h_n}\Big) - b_n^{(4)} \xrightarrow{d} E \quad \textit{as} \quad n\to\infty,$$

*where $E$ is as in Theorem 7.*

### 5.3 Sketch of the proof of Theorem 7

The proof proceeds in steps.

**Step 1 (Invariance)** The (obvious) first step is strong approximation, which means that we replace the $Y$-increments and the EWMA chart with the corresponding Wiener analogs.

**Step 2 (Partial maxima)** In this step one shows that a first portion of the chart may be discarded.

**Step 3 (Stationarity)** One replaces the Wiener-EWMA by a stationary Gaussian array.

**Step 4 (Extreme value asymptotics)** The array is replaced by its continuous extension, that is, a sequence of *stationary Gaussian* processes with autocorrelation functions $r_n(t)$, satisfying certain asymptotics, which, together with a series of (somewhat lengthy) computations in order to determine variances, covariances and autocorrelation functions for the Wiener-EWMA charts, allows us to apply an extension of Leadbetter et al. (1983), Theorem 12.3.5, concerning asymptotics for a *single* Gaussian process, to our context, viz., a *sequence of such processes*; cf. Gut and Steinebach (2004), Theorem 5.1.

Having come so far we have established asymptotics for the one-sided extreme, from which the two-sided extreme follows in view of the symmetry of the processes in the stationary array and the asymptotic independence of maxima and minima in the corresponding extreme value asymptotics.

The proof of Theorem 8 follows the same procedure with some additional complications.

## 6 Some concluding remarks

In the present note we have reviewed some of our recent results concerning the asymptotic behaviour (as $n \to \infty$) of some monitoring procedures for detecting structural breaks ("change-points") in a sequence of counting data, which are sequentially observed at equidistant time points $t = 0, 1, \ldots,$ up to a truncation point $n$. It has been shown that critical boundaries can be chosen via extreme value asymptotics, which, via invariance, are based on corresponding results for Gaussian processes.

We wish to conclude our discussion by making the following remarks:

- Although we confined ourselves here to the case of two-sided alternatives, analogous results are available for the one-sides cases too. In fact, typically the one-sided case is discussed first, and the two-sided one follows from the latter via symmetry and asymptotic independence of maxima and minima.
- It is essentially sufficient to discuss the case of the "in-control" parameters $\theta, \eta$ being known. In case of unknown parameters it turns out that suitable sequential estimates can be plugged in.
- As hinted at in the Introduction, all procedures based on the respective stopping rules are designed such that the false alarm rates (asymptotically) attain a prescribed level $\alpha$ and that the tests possess (asymptotic) power 1. In case of "early change" alternatives even more precise asymptotics are available, in that asymptotic normality of stopping times can be proved.

## References

Aue, A., Horváth, L., Kokoszka, P. and Steinebach J. (2008). Monitoring shifts in mean: Asymptotic normality of stopping times. *Test*, **17**, 515–530.

Csörgő, M., Horváth, L. and Steinebach, J. (1987). Invariance principles for renewal processes. *Annals of Probability*, **15**, 1441–1460.

Gut, A. and Steinebach, J. (2002). Truncated sequential change-point detection based on renewal counting processes. *Scandinavian Journal of Statistics*, **29**, 693–719.

Gut, A. and Steinebach, J. (2004). EWMA charts for detecting a change-point in the drift of a stochastic process. *Sequential Analysis*, **23**, 195–237.

Gut, A. and Steinebach, J. (2005). A two-step sequential procedure for detecting an epidemic change. *Extremes*, **8**, 311–326.

Gut, A. and Steinebach, J. (2009). Truncated sequential change-point detection based on renewal counting processes II. *Journal of Statistical Planning and Inference*, **139**, 1921-1936.

Horváth, L. and Steinebach, J. (2000). Testing for changes in the mean or variance of a stochastic process under weak invariance. *Journal of Statistical Planning and Inference*, **91**, 365–376.

Leadbetter, M. R., Lindgren, G. and Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Processes.* Springer-Verlag, New York.