The Annals of Statistics
1999, Vol. 27, No. 6, 1896–1913

DETECTING A CHANGE IN REGRESSION: FIRST-ORDER OPTIMALITY

BY BENJAMIN YAKIR,[1] ABBA M. KRIEGER AND MOSHE POLLAK[2]

Hebrew University and University of Pennsylvania

Observations are generated according to a regression with normal error as a function of time, when the process is in control. The process potentially changes at some unknown point of time and then the ensuing observations are normal with the same mean function plus an arbitrary function under suitable regularity conditions. The problem is to obtain a stopping rule that is optimal in the sense that the rule minimizes the expected delay in detecting a change subject to a constraint on the average run length to a false alarm. A bound on the expected delay is first obtained. It is then shown that the cusum and Shiryayev–Roberts procedures achieve this bound to first order.

1. Introduction. A new product has entered the market recently (e.g., cellular phones). Industry sales are measured monthly. If the product catches on, one may anticipate a sharp and sustained increase in the growth of monthly sales. For a variety of reasons (e.g., production, inventory control and advertising strategy), it is important to determine as soon as possible that a marked increase has occurred in the growth of monthly sales. For similar reasons, it is costly to claim that such a change has occurred when it has not.

The problem can be modeled by a sequence of independent, homoscedastic, normally distributed observations Y_i (where i indexes time). Initially, the mean grows according to a function $\beta(i)$. If a change occurs at time ν, then the mean of Y_i, $i \geq \nu$, grows according to $\beta(i) + \gamma(i - \nu + 1)$.

In the first part of this paper, it is assumed that the variance σ^2 of Y_i is known and that $\beta(i)$ and $\gamma(i)$ are also known. (Without loss of generality, $\beta(i)$ can be assumed to be 0 and σ^2 can be assumed to be 1. Some regularity conditions on γ will be imposed.) In the latter part of the paper, the case of unknown parameters is studied.

In formal terms, a detection scheme is characterized by a stopping time N at which an alarm is raised. The problem addressed in this paper is that of minimizing the expected delay until detection, $\mathrm{E}^{(\nu)}(N - \nu + 1 \mid N \geq \nu)$, subject to a lower bound on the average run length (ARL) to false alarm $\mathrm{E}^{(\infty)}N$. The expectation $\mathrm{E}^{(\infty)}$ corresponds to $\nu = \infty$, or, in other words, the case of no change.

Received December 1996; revised May 1999.
[1]Supported by a grant from the Israel Science Foundation.
[2]Supported by a grant from the Israel Science Foundation.
AMS 1991 *subject classifications*. 62L10, 62N10.
Key words and phrases. Change point detection, regression, stopping rules, information bound.

Much has been written on this problem in the context of a change from one fixed distribution to another fixed distribution. In that context, Lorden (1971) found an asymptotic lower bound on the expected delay to detection as a function of the ARL to false alarm (as the latter tends to ∞). Other and more precise optimality results were obtained later [Pollak (1985), Pollak and Siegmund (1985), Moustakides (1986), Ritov (1990), Lai (1995), Yakir (1997)]. Yao (1993) extended Lorden's results to detecting a change in regression in the case of bounded information per observation (i.e., essentially $\sup_{1\le i<\infty} \mathrm{E}^{(1)} \log[d\mathrm{P}^{(1)}(Y_i)/d\mathrm{P}^{(\infty)}(Y_i)] < \infty$). Here we study the case of possibly unbounded information, which includes the problem of detecting a change of a slope with respect to time.

There are two basic approaches to proving optimality statements. One approach is Bayesian or decision theoretic [cf. Pollak (1985), Ritov (1990), Yakir (1997)]. The other approach is classical and uses the theory of optimal stopping [cf. Lorden (1971), Moustakides (1986), Yao (1993), Lai (1995)]. It is this latter approach that is employed here, though our method of proof is different. [See also Lai (1993) and Yakir (1996).]

This paper has three main results. In Section 2 a lower bound on the expected delay is developed. In Section 3, it is shown that, asymptotically, the cusum and the Shiryayev–Roberts procedures achieve this bound and are therefore asymptotically optimal (to first order). In Section 4, a procedure which asymptotically attains these bounds is developed for the more practical case where baseline and post-change parameters are unknown. The paper is concluded with remarks and a discussion of extensions.

2. A lower bound. Our main concern in this section is the investigation of the optimal rate of detection for various post-change structures, and it will be assumed that it is known that $\sigma^2 = 1$ and $\beta(\cdot) \equiv 0$. The distribution of the sequence of observations is denoted by $\mathrm{P}^{(\nu)}$. If $\nu = k$, then

$$Y_i \sim N(\gamma(i-k+1)\mathbb{1}(i \ge k), 1)$$

for some given function $\gamma(\cdot)$.

Denote by $l^{(k)}(i)$ the log-likelihood ratio of the observation Y_i for the $\mathrm{P}^{(k)}$ relative to the $\mathrm{P}^{(\infty)}$ measures when $\nu = k$. It follows that if $i < k$, then $l^{(k)}(i) = 0$, and if $i \ge k$, then

$$l^{(k)}(i) = \gamma(i-k+1)Y_i - \gamma^2(i-k+1)/2.$$

Let $S^{(k)}(n) = \sum_{i=k}^{n} l^{(k)}(i)$ be the log-likelihood ratio of the first n observations.

Let N be any change point detection policy. In the following theorem the rate of detection of N, $\mathrm{E}^{(k)}(N-k+1 \mid N \ge k)$, is related to the information on $\gamma(\cdot)$ carried by the sequence of observations. The Kullback–Leibler information function is $\mathrm{E}^{(k)}S^{(k)}(n) = \sum_{i=k}^{n} \gamma^2(i-k+1)/2$. In particular, when $k = 1$, the information becomes

$$I(n) = \sum_{i=1}^{n} \frac{\gamma^2(i)}{2}.$$

For the sake of convenience, regard $I(x)$ as a piecewise linear continuous function of real x. Moreover, for the sake of clarity of exposition, we will assume that $I(x)$ is strictly increasing in x. We show that if $I(n)$ behaves as a power function in n and if the ARL to false alarm of N is not less than A, then the rate of detection is bounded from below by the inverse of the information function, calculated at $\log A$.

THEOREM 1. *Assume that $I(n)$ is of the form $L(n)n^r$, for some $r > 0$ and some slowly changing function L. Then*

$$\inf_{\{N:\ \mathrm{E}^{(\infty)}N \geq A\}} \sup_{1 \leq k < \infty} \mathrm{E}^{(k)}(N - k + 1 \mid N \geq k) \geq I^{-1}(\log A)(1 + o(1)), \tag{2.1}$$

where $o(1) \to 0$ as $A \to \infty$.

Note that $I(n)$ must be monotone in n. Recall that $L(n)$ is slowly changing if $\lim_{n\to\infty} L(bn)/L(n) = 1$, for all $b > 0$.

A key ingredient in the proof of the theorem is a lemma that relates the distribution of the detection policy N to the distribution of the stopping time of a one-sided sequential probability ratio test (SPRT). Given a boundary $b > 0$, the one-sided SPRT of H_0: $\nu = \infty$ versus H_1: $\nu = k$ is defined by

$$M_b^{(k)} = \inf\{n : S^{(k)}(n) \geq b\}.$$

LEMMA 1. *Let A, c and t be positive numbers such that $a = \log A > 2c$. If N is a stopping time for which $\mathrm{E}^{(\infty)}N \geq A$, then there exists an integer k such that*

$$\mathrm{P}^{(k)}\big(N - k + 1 < M_{a-2c}^{(k)} - k + 1 \mid N \geq k\big) \leq \mathrm{P}^{(1)}\big(M_{a-2c}^{(1)} > t\big) + (8t + 9)e^{-c}.$$

PROOF. Define the (truncated) Shiryayev–Roberts statistics by

$$R(n; t, m) = \sum_{k=(n-t)\vee m}^{n} \exp\{S^{(k)}(n)\}.$$

For any $j \geq 0$,

$$\begin{aligned}
\mathrm{P}^{(\infty)}(N \geq jA/2) &= \sum_{n=jA/2}^{\infty} \mathrm{P}^{(\infty)}(N = n) \\
&= \sum_{n=jA/2}^{\infty} \sum_{k=(n-t)\vee jA/2}^{n} \mathrm{E}^{(k)}\left[\frac{\mathbb{1}(N = n)}{R(n; t, jA/2)}\right] \\
&= \sum_{k=jA/2}^{\infty} \sum_{n=k}^{k+t} \mathrm{E}^{(k)}\left[\frac{\mathbb{1}(N = n)}{R(n; t, jA/2)}\right] \\
&= \sum_{k=jA/2}^{\infty} \mathrm{E}^{(k)}\left[\frac{\mathbb{1}(k \leq N \leq k + t)}{R(N; t, jA/2)}\right]
\end{aligned}$$

$$\geq \sum_{k=jA/2}^{(j+1)A/2} \mathrm{E}^{(k)}\left[\frac{\mathbb{1}(N \leq k+t)}{R(N;t,\,jA/2)} \,\middle|\, N \geq k\right]\mathrm{P}^{(\infty)}(N \geq k)$$

$$\geq \mathrm{P}^{(\infty)}(N \geq (j+1)A/2)$$

$$\times \sum_{k=jA/2}^{(j+1)A/2} \mathrm{E}^{(k)}\left[\frac{\mathbb{1}(N \leq k+t)}{R(N;t,\,jA/2)} \,\middle|\, N \geq k\right].$$

Define, for A, the given N and all $0 \leq j < \infty$, the conditional probability

$$\gamma_j = \mathrm{P}^{(\infty)}(N \geq (j+1)A/2 \,|\, N \geq jA/2) = \frac{\mathrm{P}^{(\infty)}(N \geq (j+1)A/2)}{\mathrm{P}^{(\infty)}(N \geq jA/2)}.$$

It follows that

$$\frac{1}{\gamma_j} \geq \frac{A}{2} \min_{jA/2 \leq k \leq (j+1)A/2} \mathrm{E}^{(k)}\left[\frac{\mathbb{1}(N \leq k+t)}{R(N;t,\,jA/2)} \,\middle|\, N \geq k\right]. \tag{2.2}$$

Since $A \leq \mathrm{E}^{(\infty)}N \leq [A/2]\sum_{n=0}^{\infty}\prod_{j=0}^{n-1}\gamma_j$ (with $\prod_{j=0}^{-1}\gamma_j = 1$) it can be concluded that there must exist j for which $\gamma_j \geq 1/2$. Hence, there must exist k, $jA/2 \leq k \leq (j+1)A/2$, such that

$$\mathrm{E}^{(k)}\left[\frac{\mathbb{1}(N \leq k+t)}{R(N;t,\,jA)} \,\middle|\, N \geq k\right] \leq 4/A. \tag{2.3}$$

In the rest of the proof we fix these j and k.

Let $c > 0$ be given and consider the log-likelihood ratio $S^{(k)}(N)$. It is straightforward to show that

$$\begin{aligned}\mathrm{P}^{(k)}(S^{(k)}(N) < \log A - 2c;\, N \leq k+t \,|\, N \geq k) \\ \leq \mathrm{P}^{(k)}(R(N;t,\,jA/2) \leq e^{-c}A/4;\, N \leq k+t \,|\, N \geq k)\end{aligned} \tag{2.4}$$

$$\begin{aligned}+\mathrm{P}^{(k)}\big(R(N;t,\,jA/2)\exp(-S^{(k)}(N)) \\ \geq e^{c}/4;\, N \leq k+t \,|\, N \geq k\big).\end{aligned} \tag{2.5}$$

Inequality (2.3) can be used to show for (2.4) that

$$\mathrm{P}^{(k)}(R(N;t,\,jA) \leq e^{-c}A/4;\, N \leq k+t \,|\, N \geq k) \leq e^{-c}.$$

In order to bound the term in (2.5) notice that over the event $\{k \leq N \leq k+t\}$,

$$R(N;t,\,jA/2)\exp(-S^{(k)}(N)) \leq \max_{k \leq n \leq k+t} \sum_{i=(k-t)\vee jA/2}^{n} \exp(S^{(i)}(n) - S^{(k)}(n)).$$

Hence,

$$\begin{aligned}&\mathrm{P}^{(k)}(R(N;t,\,jA/2)\exp(-S^{(k)}(N)) \geq e^{c}/4;\, N \leq k+t \,|\, N \geq k) \\ &\leq \mathrm{P}^{(k)}\left(\max_{k \leq n \leq k+t} \sum_{i=(k-t)\vee jA/2}^{n} \exp(S^{(i)}(n) - S^{(k)}(n)) \geq e^{c}/4 \,|\, N \geq k\right)\end{aligned}$$

$$\le \mathrm{P}^{(k)}\Bigg(\max_{k\le n\le k+t}\sum_{i=(k-t)\vee jA/2}^{n}\exp(S^{(i)}(n)-S^{(k)}(n))\ge e^{c}/4 \,|\, N\ge jA/2\Bigg)$$
$$\times \mathrm{P}^{(\infty)}(N\ge jA/2)/\mathrm{P}^{(\infty)}(N\ge (j+1)A/2)$$
$$= \frac{1}{\gamma_j}\mathrm{P}^{(k)}\Bigg(\max_{k\le n\le k+t}\sum_{i=(k-t)\vee jA/2}^{n}\exp(S^{(i)}(n)-S^{(k)}(n))\ge e^{c}/4\Bigg)$$
$$\le 8(t+1)e^{-c},$$

since $\gamma_j \ge 1/2$ and since Doob's inequality can be applied to the $\mathrm{P}^{(k)}$-submartingale

$$\sum_{i=(k-t)\vee jA}^{n}\exp(S^{(i)}(n)-S^{(k)}(n)), \qquad k\le n\le k+t.$$

To conclude the above discussion,

$$\mathrm{P}^{(k)}(S^{(k)}(N)<a-2c;\ N\le k+t \,|\, N\ge k)\le (8t+9)e^{-c}. \tag{2.6}$$

It follows from (2.6) that

$$\mathrm{P}^{(k)}\big(N<M^{(k)}_{a-2c};\ N\le k+t \,|\, N\ge k\big)\le (8t+9)e^{-c}.$$

Therefore,

$$\mathrm{P}^{(k)}\big(N<M^{(k)}_{a-2c};\ M^{(k)}_{a-2c}\le k+t \,|\, N\ge k\big)\le (8t+9)e^{-c}.$$

The event $\{N\ge k\}$ is independent of the stopping time $M^{(k)}_{a-2c}$ and the $\mathrm{P}^{(k)}$-distribution of $M^{(k)}_{a-2c}-k+1$ is independent of k. Hence

$$\mathrm{P}^{(k)}\big(N<M^{(k)}_{a-2c} \,|\, N\ge k\big)\le \mathrm{P}^{(k)}\big(M^{(k)}_{a-2c}-k+1>t\big)+(8t+9)e^{-c}$$
$$= \mathrm{P}^{(1)}\big(M^{(1)}_{a-2c}>t\big)+(8t+9)e^{-c}. \qquad \square$$

PROOF OF THEOREM 1. Let I^{-1} be the inverse function of I. Observe first that for b fixed, $b>0$, and for a, $a\to\infty$,

$$I(bI^{-1}(a)) = ab^{r}\frac{L(bI^{-1}(a))}{L(I^{-1}(a))}\sim ab^{r},$$

since L is slowly changing. It follows that $I^{-1}(ba)\sim b^{1/r}I^{-1}(a)$.

Let $\varepsilon>0$ be given. It will be shown that for any A large,

$$\inf_{\{N:\ \mathrm{E}^{(\infty)}N\ge A\}}\sup_{1\le\nu<\infty}\mathrm{E}^{(\nu)}(N-\nu+1 \,|\, N\ge\nu)\ge e^{-\varepsilon}I^{-1}(\log A). \tag{2.7}$$

Define $t=\lceil I^{-1}(\log A)\rceil=\lceil I^{-1}(a)\rceil$, and for a small, but positive ε_1, define $c=\varepsilon_1 a$. It follows from Lemma 1 and from Chebyshev's inequality that for

any stopping rule N, for which $\mathrm{E}^{(\infty)}N \geq A$, there exists an integer k such that

$$\begin{aligned}
&\mathrm{P}^{(k)}\big(N-k+1 < M^{(k)}_{(1-2\varepsilon_1)a} - k + 1 \,|\, N \geq k\big) \\
&\quad \leq \mathrm{P}^{(1)}\big(M^{(1)}_{(1-2\varepsilon_1)a} > t\big) + (8t+9)\exp\{-\varepsilon_1 a\} \\
&\quad \leq \mathrm{P}^{(1)}\big(S^{(1)}(t) < (1-2\varepsilon_1)a\big) + (8I^{-1}(a)+9)\exp\{-\varepsilon_1 a\} \\
&\quad \leq \frac{1}{2\varepsilon_1^2 a} + (8I^{-1}(a)+9)\exp\{-\varepsilon_1 a\},
\end{aligned}$$

since $\mathrm{E}^{(1)}\big(S^{(1)}(t)\big) = I(t) \sim a$ and $\mathrm{Var}^{(1)}\big(S^{(1)}(t)\big) = 2I(t)$. This expression converges to zero as $a \to \infty$.

Moreover,

$$\begin{aligned}
&\mathrm{P}^{(k)}\big(M^{(k)}_{(1-2\varepsilon_1)a} - k + 1 < \exp(-\varepsilon_1)I^{-1}((1-2\varepsilon_1)a)\big) \\
&\qquad \leq \mathrm{P}^{(1)}\Bigg(\sup_{n \leq \exp(-\varepsilon_1)I^{-1}((1-2\varepsilon_1)a)} S^{(1)}(n) > (1-2\varepsilon_1)a\Bigg).
\end{aligned}$$

However, from Doob's inequality, the monotonicity of $I(\cdot)$ and for a_1, such that $a_1 = (1-2\varepsilon_1)a$,

$$\begin{aligned}
&\mathrm{P}^{(1)}\Bigg(\sup_{n \leq \exp(-\varepsilon_1)I^{-1}(a_1)} S^{(1)}(n) > a_1\Bigg) \\
&\quad \leq \mathrm{P}^{(1)}\Bigg(\sup_{n \leq \exp(-\varepsilon_1)I^{-1}(a_1)} (S^{(1)}(n) - I(n))^2 > (a_1 - I(\exp(-\varepsilon_1)I^{-1}(a_1))^2\Bigg) \\
&\quad \leq \frac{2I(\exp(-\varepsilon_1)I^{-1}(a_1))}{(a_1 - I(\exp(-\varepsilon_1)I^{-1}(a_1))^2},
\end{aligned}$$

which, again, converges to zero as $a \to \infty$.

Therefore,

$$\mathrm{P}^{(k)}\big(N-k+1 \geq \exp(-\varepsilon_1)I^{-1}((1-2\varepsilon_1)a) \,|\, N \geq k\big) \geq \exp(-\varepsilon/2),$$

provided that a is large. A choice of a small enough ε_1 would lead to inequality (2.7) since $\exp(-\varepsilon_1)I^{-1}((1-2\varepsilon_1)a) \sim \exp(-\varepsilon_1)(1-2\varepsilon_1)^{1/r}I^{-1}(a)$. The proof of the theorem thus follows. □

3. Asymptotically optimal detection policies. In this section we relax the assumption that γ is known and consider the construction of asymptotically optimal detection stopping times. Natural candidates, such as the cusum or the Shiryayev–Roberts procedure, are based on log-likelihood ratios. These log-likelihood ratios, however, involve unknowns—the function γ—which can be estimated. Martingale consideration would suggest estimation which is based only on observations which are prior to the current one. A simple estimate of $\gamma(i-k+1)$, which is the $\mathrm{P}^{(k)}$-mean of the observation Y_i, is Y_{i-1} (or

0 if $i = k$). Substituting estimates leads to the (approximated) log-likelihood ratios

$$\tilde{S}^{(k)}(n) = \sum_{i=k+1}^{n} \big(Y_{i-1}Y_i - Y_{i-1}^2/2\big).$$

Note that $\tilde{S}^{(k)}(n)$ is itself a log-likelihood ratio.

Consider the Shiryayev–Roberts stopping rule

$$N_{\mathrm{SR}} = \inf\Big\{n: \sum_{k=1}^{n} \exp\{\tilde{S}^{(k)}(n)\} \geq A\Big\},$$

and the cusum stopping rule

$$N_{\mathrm{CS}} = \inf\Big\{n: \max_{k\leq n} \tilde{S}^{(k)}(n) \geq \log A\Big\}.$$

THEOREM 2. *Let N be either N_{SR} or N_{CS}.*

(i) $\mathrm{E}^{(\infty)}N \geq A$.

(ii) *If $\gamma(\cdot)$ is a positive and increasing function and if $I(n) = L(n)n^r$ for some $r > 1$ and some slowly changing function L, then*

$$\sup_{1\leq k<\infty} \mathrm{E}^{(k)}(N - k + 1 \,|\, N \geq k) = I^{-1}(\log A)(1 + o(1)),$$

where $o(1) \to 0$ as $A \to \infty$.

REMARK. Note that $\gamma(n) \to \infty$, $\gamma(n)/\gamma(n-1) \to 1$ and $Y_n/\gamma(n-k+1) \to_p 1$ under $\mathrm{P}^{(k)}$ as $n \to \infty$, providing for intuitive plausibility of the success of N_{SR} and N_{CS}.

PROOF OF THEOREM 2(i). The stopping time N_{CS} dominates the stopping time N_{SR}. Hence it is sufficient to prove the claim for $N = N_{\mathrm{SR}}$.

If $\lim_{m\to\infty} \mathrm{P}^{(\infty)}(N < m) < 1$, then $\mathrm{E}^{(\infty)}N = \infty$ and the claim is trivial. Assume that the probability converges to 1. Obviously, $\mathrm{E}^{(\infty)}N \geq \mathrm{E}^{(\infty)}(N \wedge m)$, for $1 \leq m < \infty$. The process $\sum_{k=1}^{n} \exp\{\tilde{S}^{(k)}(n)\} - n$ is a $\mathrm{P}^{(\infty)}$-martingale. Hence, by the optional sampling theorem,

$$\mathrm{E}^{(\infty)}(N \wedge m) = \mathrm{E}^{(\infty)} \sum_{k=1}^{N\wedge m} \exp\{\tilde{S}^{(k)}(N \wedge m)\}.$$

The statistic $\sum_{k=1}^{N\wedge m} \exp\{\tilde{S}^{(k)}(N \wedge m)\}$ is positive and is larger than A on the event $\{N < m\}$. Therefore,

$$\mathrm{E}^{(\infty)}N \geq A\mathrm{P}^{(\infty)}(N < m).$$

The proof follows since $\mathrm{P}^{(\infty)}(N < m) \longrightarrow_{m\to\infty} 1$ by assumption. □

PROOF OF THEOREM 2(ii). Let N be either N_{SR} or N_{CS}. For any k, $1 \le k < \infty$, let $M^{(k)}_{\log A}$ be the SPRT stopping time as in the previous section with $S^{(k)}(n)$ replaced by $\tilde{S}^{(k)}(n)$. It is easy to see that

$$\begin{aligned}\mathrm{E}^{(k)}(N-k+1 \mid N \ge k) &\le \mathrm{E}^{(k)}\bigl(M^{(k)}_{\log A} - k + 1 \mid N \ge k\bigr)\\ &= \mathrm{E}^{(1)} M^{(1)}_{\log A}.\end{aligned}$$

In the sequel we bound this last expectation.

From the definition of $M^{(1)}_{\log A} = M$ it follows that

$$\begin{aligned}\log A &\ge \sum_{i=2}^{M-1}\bigl(Y_{i-1}Y_i - Y_{i-1}^2/2\bigr)\\ &= I(M-1) + \sum_{i=2}^{M-1}\bigl(Y_{i-1}Y_i - Y_{i-1}^2/2 - \gamma^2(i)/2\bigr) - \frac{\gamma^2(1)}{2}.\end{aligned} \tag{3.1}$$

The expectation of the second term in the last line of (3.1) can be bounded from below by

$$\begin{aligned}&\mathrm{E}^{(1)} \sum_{i=2}^{M-1}\bigl(Y_{i-1}Y_i - Y_{i-1}^2/2 - \gamma^2(i)/2\bigr)\\ &= \mathrm{E}^{(1)} \sum_{i=2}^{\infty}\bigl(Y_{i-1}Y_i - Y_{i-1}^2/2 - \gamma^2(i)/2\bigr)\,\mathbb{1}(M \ge i+1)\\ &\ge -\mathrm{E}^{(1)} \sum_{i=2}^{\infty}\bigl|Y_{i-1}Y_i - Y_{i-1}^2/2 - \gamma^2(i)/2\bigr|\,\mathbb{1}(M \ge i+1)\\ &\ge -\mathrm{E}^{(1)} \sum_{i=2}^{\infty}\bigl|Y_{i-1}Y_i - Y_{i-1}^2/2 - \gamma^2(i)/2\bigr|\,\mathbb{1}(M \ge i-1)\\ &= -\sum_{i=2}^{\infty}\mathrm{E}^{(1)}\bigl|Y_{i-1}Y_i - Y_{i-1}^2/2 - \gamma^2(i)/2\bigr|\mathrm{P}^{(1)}(M \ge i-1)\\ &= -\mathrm{E}^{(1)} \sum_{i=2}^{M+1}\mathrm{E}^{(1)}\bigl|Y_{i-1}Y_i - Y_{i-1}^2/2 - \gamma^2(i)/2\bigr|.\end{aligned}$$

Straightforward calculations and the assumptions of the theorem show that $\gamma^2(n)/2 = o(I(n))$, implying $I(n-1) = (1+o(1))I(n)$, and that

$$\begin{aligned}&\bigl|Y_{i-1}Y_i - Y_{i-1}^2/2 - \gamma^2(i)/2\bigr|\\ &\quad\le \bigl|Y_{i-1} - \gamma(i-1)\bigr|\bigl|Y_i - \gamma(i)\bigr| + \gamma(i)\bigl|Y_i - \gamma(i)\bigr|\\ &\qquad + (Y_i - \gamma(i-1))^2/2 + (\gamma^2(i) - \gamma^2(i-1))/2,\end{aligned}$$

so that

$$\sum_{i=2}^{n+1}\mathrm{E}^{(1)}\bigl|Y_{i-1}Y_i - Y_{i-1}^2/2 - \gamma^2(i)/2\bigr| = o(I(n)).$$

Therefore, for any $\varepsilon > 0$ a constant C can be found such that

$$I(n-1) - \sum_{i=2}^{n+1} \mathrm{E}^{(1)} \left| Y_{i-1} Y_i - Y_{i-1}^2/2 - \gamma^2(i)/2 \right| - \frac{\gamma^2(1)}{2} \geq (1-\varepsilon) I(n) - C,$$

for all $n \geq 1$. Moreover, the function $I(n)$ is convex since the function $\gamma(\cdot)$ is increasing. It can be concluded that

$$\begin{aligned} \log A &\geq (1-\varepsilon)\mathrm{E}^{(1)} I(M) - C \\ &\geq (1-\varepsilon) I(\mathrm{E}^{(1)} M) - C. \end{aligned}$$

The proof of the theorem thus follows. □

4. A case of unknown baseline parameters. Often, β and σ^2 are unknown and need to be estimated from the data. In this section, we construct a first-order asymptotically optimal procedure for this case. In order for the change point problem to be well defined, β should be structured, since γ is not. In this section we assume that $\beta(i) = \beta_0 + i\beta_1$ for some unknown scalars β_0 and β_1. Given the first n observations, the standard estimators of $(\beta_0, \beta_1, \sigma^2)$ are (when all observations are prechange)

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n i Y_i - ((n+1)/2) \sum_{i=1}^n Y_i}{n(n^2-1)/12} \\ \hat{\beta}_0 &= \frac{\sum_{i=1}^n Y_i}{n} - \hat{\beta}_1 \frac{n+1}{2} \\ \hat{\sigma}^2 &= \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 i)^2. \end{aligned}$$

Note that $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$ and $\hat{\sigma}^2$ implicitly depend on n, the number of observations at the current inspection period. Define $\hat{l}^{(k)}(i,n) = \hat{Y}_{i-1,n} \hat{Y}_{i,n} - \hat{Y}_{i-1,n}^2/2$, for all $3 < k < i \leq n$, with $\hat{Y}_{i,n} = (Y_i - \hat{\beta}_0 - \hat{\beta}_1 i)/\hat{\sigma}$. Consider the pseudo-log-likelihood ratio

$$\hat{S}^{(k)}(n) = \sum_{i=k+1}^{n} \hat{l}^{(k)}(i,n), \tag{4.1}$$

and define the (window truncated) Shiryayev–Roberts stopping rule,

$$N_{\mathrm{SR}} = \inf \left\{ n: \sum_{k=n-t}^{n} \exp\{\hat{S}^{(k)}(n)\} \geq dA \right\},$$

and the (window truncated) cusum stopping rule,

$$N_{\mathrm{CS}} = \inf \left\{ n: \max_{n-t \leq k \leq n} \hat{S}^{(k)}(n) \geq \log(dA) \right\},$$

for some $t = t(A)$, $t = o(A)$, $t/\log A \to \infty$, and $d > 2$. This d, together with $c > 1$, solves the equation

$$A = \frac{1/2 + [1/4 - (2t)/(dA)]^{1/2}}{1 - \{(cA-1)/(cA)\}\{1/2 + [1/4 - (2t)/(dA)]^{1/2}\}^{1/t}}.$$

Note that $c = c(A)$ and $d = d(A)$ can be taken to be bounded. Let G be a geometric random variable, independent of everything else, such that $\mathrm{E}(G) = cA$ and denote $N = N_{\mathrm{SR}} \wedge G$ or $N = N_{\mathrm{CS}} \wedge G$.

It will be shown below that, under mild regularity conditions, N attains the optimal rate of detection, at least when the change does not occur too soon after initiation of the monitoring. (This restriction is unavoidable; for example, if the change is a change of slope, the state {no change} is indistinguishable from the state {change from the very beginning}.) Before considering the rate of detection of a policy N, however, one needs to demonstrate that $\mathrm{E}^{(\infty)}N \geq A$. It should be noted that $\hat{S}^{(k)}(n)$ is no longer a log-likelihood ratio, hence standard martingale results cannot be applied. Nonetheless, as shown in the proof of the following lemma, $\hat{S}^{(k)}(n)$ is smaller than another (true) log-likelihood ratio. It follows that the expectation constraint on the rate of false alarms is satisfied for the proposed procedure.

LEMMA 2. *Let N be defined as above. Then* $\mathrm{E}^{(\infty)}N \geq A$.

The optimality claim is stated next as a theorem. The proof of both this theorem and of Lemma 2 is relegated to the Appendix.

THEOREM 3. *Let N be defined as above. If $I(n) = L(n)n^r$ for some $r > 1$ and some slowly changing function L and if $\sum_{i=1}^{n}[\gamma(i) - \gamma(i-1)]^2 = o(I(n))$, then there exists a constant η such that*

$$\sup_{\eta \log A \leq k < \infty} \mathrm{E}^{(k)}(N - k + 1 \mid N \geq k) = I^{-1}(\log A)(1 + o(1)).$$

REMARK. Note that one need not know the value of r to apply this procedure and that this procedure is first-order asymptotically optimal whatever the value of $r > 1$ is.

5. Comments and extensions. In the first part of this paper, the first-order optimality of the cusum and the Shiryayev–Roberts procedures was established. This result is in a regression context where the observations are independent and normally distributed with known variance and the mean level grows according to a known function of time before a change has taken place and according to another unknown function once a change has occurred. It is interesting to note that an optimal procedure is still based on the likelihood ratio's crossing of a constant boundary. We conjecture that this may be the case in a wide variety of contexts, as hinted by the following considerations.

Regard a sequential hypothesis testing problem of a simple H_0 versus a simple H_1 based on a sequence of (not necessarily independent) observations

$Y_1, Y_2, \ldots$. Without loss of generality, regard procedures based on $L_1, L_2, \ldots$, where L_n is the likelihood ratio of the first n observations. Note that L_n equals the likelihood ratio of $(L_1, \ldots, L_n)$. Letting f_n be the density of L_n, one can rephrase H_0 and H_1 as H_0: $f_n = f_n^0$; $n = 1, 2, \ldots$ and H_1: $f_n = f_n^1$; $n = 1, 2, \ldots$. Note that $f_n^1(y)/f_n^0(y) = y$. For a power one test, the operating characteristics of interest are $\alpha = \mathrm{P}_0(N < \infty)$ and $\mathrm{E}_1 N$, where N is a stopping time at which (if $N < \infty$) one stops and rejects H_0. The optimal stopping problem is to minimize $\mathrm{E}_1 N$ subject to a given α.

Consider stopping times of the form $N = \min\{n\colon\ L_n \geq C_n\}$ ($N = \infty$ if no such n exists), where $\{C_n\}$ is a sequence of constants. Regard the surrogate problem of minimizing the H_1-expected number of times L_n is below C_n, subject to a fixed value δ of the H_0-expected number of visits of L_n above C_n. Intuitively, the H_1-expected number of times L_n is below C_n differs from $\mathrm{E}_1 N$ by an additive constant, and δ differs from α by a multiplicative constant, so that an optimal procedure for the surrogate problem is almost (asymptotically, $\alpha \to 0$) optimal for the original one. Now use a Lagrangian multiplier argument to solve the surrogate problem,

$$\frac{\partial}{\partial C_j}\left[\sum_{n=1}^{\infty} \mathrm{P}_1(L_n \leq C_n) + \lambda\left(\sum_{n=1}^{\infty} \mathrm{P}_0(L_n > C_n) - \delta\right)\right] = f_j^1(C_j) - \lambda f_j^0(C_j)$$

which equal zero iff $f_n^1(C_n)/f_n^0(C_n) = \lambda$ for each n. Since $f_n^1(C_n)/f_n^0(C_n) = C_n$, it follows that the rule which calls for stopping the first time that the likelihood ratio exceeds a constant is optimal for the surrogate problem. The close relationship between power one tests and change point problems leads one to conjecture that the analogous rule (cusum, or Shiryayev–Roberts) is asymptotically almost optimal. [Actually, these considerations lead one to suspect that the optimality is much stronger than that claimed in Sections 2 and 3; $I^{-1}(\log A)(1 + o(1))$ may perhaps turn out to be $I^{-1}(\log A) + O(1)$. However, we have so far been unable to provide a full proof along these lines.]

The results of this paper apply to cases where the information contained in the first n observations increases according to $L(n)n^r$. Note that a change in a linear trend can be related to $r = 3$. The example of the success of a new product (mentioned in the introduction) may correspond to $r > 3$. The classical change point problem (change from one constant mean to another constant mean level) corresponds to $r = 1$ and the case $r < 1$ corresponds to a situation where the mean tends to revert back to its original prechange level (at not too fast a rate), and the results of Section 2 apply, although those of Section 3 and 4 do not. The results of this paper, however, do not apply to regression problems where the information accumulates exponentially.

Parallel results can be obtained when the errors are not normally distributed. It is important to note that the main lemma in Section 2 is not based on normality. In this context it is useful to compare our results with the results in Robbins and Zhang (1993). They considered a change point detection problem in an exponential family context with a probability constraint imposed on the rate of false alarm. Mixture-type stopping rules were inves-

tigated in terms of their ARL to detection. It can be shown, however, that mixture-type stopping rule are suboptimal when $r > 1$.

APPENDIX

This Appendix contains the proofs of Lemma 2 and Theorem 3.

The basic idea in proving Lemma 2 is to show that $\hat{S}^{(k)}(n)$ is dominated by a true log-likelihood. This is done in five steps:

1. A true log-likelihood ratio, which involves invariant statistics, is defined. The likelihood ratio is formed by dividing the joint density of these statistics under $\mathrm{P}^{(k)}$ by their joint density under $\mathrm{P}^{(\infty)}$.
2. It is shown that the true log-likelihood ratio is greater than a computable expression.
3. The above computable expression is presented as a sum of $\hat{S}^{(k)}(n)$ and an additional term.
4. The additional term in (3) is proven to be positive.
5. The stopping time involving the true log-likelihood ratio of (1) is shown to satisfy the constraint on false alarms.

Step 1. Initially, let Y_i, $i \geq 1$, be independent homoscedastic normally distributed observations. It will be convenient to use vector and matrix notation. Let m be the number of observations that is known to be before the change has taken place. (Note that in the paper m is assumed to be 3.) Let

$$\mathbf{Y}' = (\mathbf{Y}_{-1}, \mathbf{Y}_0, \mathbf{Y}_1)' = ((Y_1, \ldots, Y_m), (Y_{m+1}, \ldots, Y_{k-1}), (Y_k, \ldots, Y_n))$$

and let $\mathbf{X} = (\mathbf{X}_{-1}, \mathbf{X}_0, \mathbf{X}_1)$, where

$$\mathbf{X}' = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 2 & 3 & \cdots & n \end{pmatrix},$$

is the associated design matrix. Note that in the text $m = 3$ observations are assumed to be prechange. Let $\hat{\mathbf{b}}$ and $\hat{v}$ be estimates of $(\beta_0, \beta_1)'$ and σ^2, based on $\mathbf{Y}_{-1}$,

$$\hat{\mathbf{b}} = (\mathbf{X}'_{-1}\mathbf{X}_{-1})^{-1}\mathbf{X}'_{-1}\mathbf{Y}_{-1},$$
$$\hat{v} = \mathbf{Y}'_{-1}[\mathbf{I} - \mathbf{X}_{-1}(\mathbf{X}'_{-1}\mathbf{X}_{-1})^{-1}\mathbf{X}'_{-1}]\mathbf{Y}_{-1}/(m-2).$$

Define the statistics

$$\mathbf{T} = (\mathbf{T}_0, \mathbf{T}_1) = \hat{v}^{-1/2}\big((\mathbf{Y}_0, \mathbf{Y}_1) - (\mathbf{X}_0, \mathbf{X}_1)\hat{\mathbf{b}}\big). \tag{A.1}$$

Regard $\mathbf{Y}_{-1}$ as a learning sample, taken from the prechange distribution. A change of the regression parameters, if it occurs at all, is assumed to occur at some (unknown) time $\nu = k > m$, in which case $\mathscr{L}(Y_i \mid Y_1, \ldots, Y_{i-1}) = N(Y_{i-1}, \sigma^2)$ for $i \geq k+1$. (If there is no change at all, denote $\nu = \infty$.) Monitoring for a change will be based on the the invariant sequence of $|\mathbf{T}|$'s (where $|\mathbf{T}|$ is the vector of absolute values of the coordinates of $\mathbf{T}$). Let $S^{(k)}(\mathbf{T})$ be the marginal log-likelihood ratio of the sequence of invariant statistics.

The distributions of both $\hat{S}^{(k)}(n)$ and $S^{(k)}(\mathbf{T})$ are invariant with respect to an affine transformation of time and multiplication by a positive scalar. Hence, one may assume, without loss of generality, that $\beta_0 = \beta_1 = 0$ and $\sigma^2 = 1$.

Step 2. Denote: $\Theta = (\hat{\mathbf{b}}, \hat{v})$. Let $S^{(k)}(\mathbf{T} \mid \Theta)$ be the conditional log-likelihood ratio of $\mathbf{T}$, for $\{\nu = k\}$ versus $\{\nu = \infty\}$. Let f_k denote the density under $\nu = k$. Now,

$$\begin{aligned} S^{(k)}(t) \stackrel{\text{def}}{=} \log \frac{f_k(t)}{f_\infty(t)} &= \log \frac{\int f_k(t, \theta)\, d\,\theta}{\int f_\infty(t, \theta)\, d\,\theta} \\ &= \log \int \frac{f_k(t, \theta)}{f_\infty(t, \theta)} \left[\frac{f_\infty(t, \theta)}{\int f_\infty(t, u)\, d\, u} \right] d\,\theta \\ &= \log \int \frac{f_k(t|\theta)}{f_\infty(t|\theta)} f_\infty(\theta|t)\, d\,\theta \\ &\geq \int \log \left[\frac{f_k(t|\theta)}{f_\infty(t|\theta)} \right] f_\infty(\theta|t)\, d\,\theta. \end{aligned}$$

□

It follows that $\tilde{S}^{(k)}(\mathbf{T}) = \mathrm{E}_\infty[S^{(k)}(\mathbf{T} \mid \Theta) \mid \mathbf{T}]$ is smaller than the (true) log-likelihood ratio $S^{(k)}(\mathbf{T})$.

Step 3. Here it is shown that

$$\tilde{S}^{(k)}(\mathbf{T}) = \hat{\sigma}^{-2}(\mathbf{Y}_1 - \mathbf{X}_1\hat{\beta})'\mathbf{A}(\mathbf{Y}_1 - \mathbf{X}_1\hat{\beta}) + \operatorname{trace}(\mathbf{X}_1'\mathbf{A}\mathbf{X}_1(\mathbf{X}'\mathbf{X})^{-1}), \tag{A.2}$$

with

$$\begin{aligned} \hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}, \\ \hat{\sigma}^2 &= \frac{\mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}}{n-2}. \end{aligned}$$

Note that the first term on the right-hand side of (A.2) is what is denoted by $\hat{S}^{(k)}(n)$. Throughout this step the superscript k is omitted in order to simplify the notation.

Observe that

$$\begin{aligned} \hat{\mathbf{b}} &\sim N(0, (\mathbf{X}_{-1}'\mathbf{X}_{-1})^{-1}), \\ \hat{v} &\sim \Gamma\big((m-2)/2, (m-2)/2\big) \end{aligned}$$

are independent. Furthermore,

$$\mathscr{L}(\mathbf{T} \mid \theta) = \mathscr{L}(\mathbf{T} \mid \hat{\mathbf{b}}, \hat{v}) = N(-\hat{v}^{-1/2}(\mathbf{X}_0, \mathbf{X}_1)\hat{\mathbf{b}}, \hat{v}^{-1}\mathbf{I}).$$

Standard Bayesian arguments can be used to show that

$$\mathscr{L}(\hat{\mathbf{b}} \mid \hat{v}, \mathbf{T}) = N(-\hat{v}^{1/2}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}_0, \mathbf{X}_1)'\mathbf{T}, (\mathbf{X}'\mathbf{X})^{-1}), \tag{A.3}$$

$$\mathscr{L}(\hat{v} \mid \mathbf{T}) = \Gamma\left(\frac{n-2}{2}, \frac{\mathbf{T}'\mathbf{T} + m - 2 - \mathbf{T}'(\mathbf{X}_0, \mathbf{X}_1)(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}_0, \mathbf{X}_1)'\mathbf{T}}{2} \right). \tag{A.4}$$

Let $\mathbf{A} = \mathbf{A}_{k,n}$ be, for general k and n, an $(n-k+1)$ by $(n-k+1)$ matrix of the following form (demonstrated for $n-k+1=5$):

$$\mathbf{A} = \begin{pmatrix} -0.5 & 1 & 0 & 0 & 0 \\ 0 & -0.5 & 1 & 0 & 0 \\ 0 & 0 & -0.5 & 1 & 0 \\ 0 & 0 & 0 & -0.5 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Recall that if a change occurs at time $\nu = k$, then $\mathcal{L}(Y_i \mid Y_1, \ldots Y_{i-1}) = N(Y_{i-1}, 1)$ for $i \geq k+1$. Thus the log-likelihood ratio of $\mathbf{Y}$ for $\{\nu = k\}$ versus $\{\nu = \infty\}$ is $\sum_{i=k+1}^{n}(Y_i Y_{i-1} - Y_{i-1}^2/2) = \mathbf{Y}_1'\mathbf{A}\mathbf{Y}_1$.

The term $S(\mathbf{T} \mid \hat{\mathbf{b}}, \hat{v})$ is the conditional log-likelihood ratio of $\mathbf{T}$ for $\nu = k$ versus $\nu = \infty$. From (A.1) it follows that $S(\mathbf{T} \mid \hat{\mathbf{b}}, \hat{v})$ equals the conditional log-likelihood ratio of $\mathbf{Y}_1 = \hat{v}^{1/2}\mathbf{T}_1 + \mathbf{X}_1\hat{\mathbf{b}}$, so that

$$S(\mathbf{T} \mid \hat{\mathbf{b}}, \hat{v}) = [\hat{v}^{1/2}\mathbf{T}_1 + \mathbf{X}_1\hat{\mathbf{b}}]'\mathbf{A}[\hat{v}^{1/2}\mathbf{T}_1 + \mathbf{X}_1\hat{\mathbf{b}}].$$

Regarding this as a function of $\hat{\mathbf{b}}$, obtain from (A.3) that

$$\mathrm{E}_\infty[S(\mathbf{T} \mid \hat{\mathbf{b}}, \hat{v}) \mid \hat{v}, \mathbf{T}] = \hat{v}(\mathbf{T}_1 - \mathbf{X}_1\tilde{\beta})'\mathbf{A}(\mathbf{T}_1 - \mathbf{X}_1\tilde{\beta}) + \mathrm{trace}(\mathbf{X}_1'\mathbf{A}\mathbf{X}_1(\mathbf{X}'\mathbf{X})^{-1}),$$

where $\tilde{\beta} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}_0, \mathbf{X}_1)'\mathbf{T}$. Therefore,

$$\tilde{S}(\mathbf{T}) = \mathrm{E}_\infty[S(\mathbf{T} \mid \hat{\mathbf{b}}, \hat{v}) \mid \mathbf{T}] = \mathrm{E}_\infty[\mathrm{E}_\infty(S(\mathbf{T} \mid \hat{\mathbf{b}}, \hat{v}) \mid \hat{v}, \mathbf{T}) \mid \mathbf{T}] \tag{A.5}$$

$$= \tilde{\sigma}^{-2}(\mathbf{T}_1 - \mathbf{X}_1\tilde{\beta})'\mathbf{A}(\mathbf{T}_1 - \mathbf{X}_1\tilde{\beta}) + \mathrm{trace}(\mathbf{X}_1'\mathbf{A}\mathbf{X}_1(\mathbf{X}'\mathbf{X})^{-1}), \tag{A.6}$$

where

$$\tilde{\sigma}^2 = \frac{\mathbf{T}'\mathbf{T} + m - 2 - \mathbf{T}'(\mathbf{X}_0, \mathbf{X}_1)(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}_0, \mathbf{X}_1)'\mathbf{T}}{n-2}.$$

However, from

$$(\mathbf{T}_1 - \mathbf{X}_1\tilde{\beta}) = \frac{\mathbf{Y}_1 - \mathbf{X}_1\hat{\mathbf{b}}}{\hat{v}^{1/2}} - \frac{\mathbf{X}_1(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}_0, \mathbf{X}_1)'[(\mathbf{Y}_0, \mathbf{Y}_1) - (\mathbf{X}_0, \mathbf{X}_1)\hat{\mathbf{b}}]}{\hat{v}^{1/2}}$$

obtain that [since $(\mathbf{X}_{-1}'\mathbf{X}_{-1})\hat{\mathbf{b}} = \mathbf{X}_{-1}'\mathbf{Y}_{-1}$]

$$\begin{aligned} &\hat{v}^{1/2}(\mathbf{T}_1 - \mathbf{X}_1\tilde{\beta}) \\ &\quad = \mathbf{Y}_1 - \mathbf{X}_1\{(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}_{-1}, \mathbf{X}_0, \mathbf{X}_1)'(\mathbf{Y}_{-1}, \mathbf{Y}_0, \mathbf{Y}_1)\} \\ &\qquad + \mathbf{X}_1\{(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}_0, \mathbf{X}_1)'(\mathbf{X}_0, \mathbf{X}_1)\hat{\mathbf{b}} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_{-1}'\mathbf{Y}_{-1} - \hat{\mathbf{b}}\} \\ &\quad = \mathbf{Y}_1 - \mathbf{X}_1\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\} \\ &\qquad + \mathbf{X}_1\{(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}_0, \mathbf{X}_1)'(\mathbf{X}_0, \mathbf{X}_1)\hat{\mathbf{b}} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_{-1}'\mathbf{Y}_{-1} - \hat{\mathbf{b}}\} \\ &\quad = \mathbf{Y}_1 - \mathbf{X}_1\hat{\beta} + \mathbf{X}_1\{(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\hat{\mathbf{b}} - \hat{\mathbf{b}}\} \\ &\quad = \mathbf{Y}_1 - \mathbf{X}_1\hat{\beta} \end{aligned}$$

and likewise

$$\hat{v}^{1/2}(\mathbf{T}_0 - \mathbf{X}_0\tilde{\beta}) = \mathbf{Y}_0 - \mathbf{X}_0\hat{\beta}.$$

Therefore, this step will be completed if it can be shown that $\hat{v}\tilde{\sigma}^2 = \hat{\sigma}^2$. Indeed,

$$\begin{aligned}(n-2)\tilde{\sigma}^2 &= \mathbf{T}_0'(\mathbf{T}_0 - \mathbf{X}_0\tilde{\beta}) + \mathbf{T}_1'(\mathbf{T}_1 - \mathbf{X}_1\tilde{\beta}) + \hat{v}^{-1}\hat{v}(m-2)\\ &= \hat{v}^{-1}(\mathbf{Y}_0 - \mathbf{X}_0\hat{\mathbf{b}})'(\mathbf{Y}_0 - \mathbf{X}_0\hat{\beta}) + \hat{v}^{-1}(\mathbf{Y}_1 - \mathbf{X}_1\hat{\mathbf{b}})'(\mathbf{Y}_1 - \mathbf{X}_1\hat{\beta})\\ &\quad + \hat{v}^{-1}(\mathbf{Y}_{-1}'\mathbf{Y}_{-1} - \mathbf{Y}_{-1}'\mathbf{X}_{-1}\hat{\mathbf{b}})\end{aligned}$$

so that [since$\mathbf{Y}_{-1}'\mathbf{X}_{-1} = \hat{\mathbf{b}}'(\mathbf{X}_{-1}'\mathbf{X}_{-1})$]

$$\begin{aligned}&\hat{v}(n-2)\tilde{\sigma}^2\\ &\quad= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\hat{\beta} - (\hat{\mathbf{b}}'\mathbf{X}'\mathbf{Y} - \hat{\mathbf{b}}'\mathbf{X}_0'\mathbf{X}_0\hat{\beta} - \hat{\mathbf{b}}'\mathbf{X}_1'\mathbf{X}_1\hat{\beta} - \mathbf{Y}_{-1}'\mathbf{X}_{-1}\hat{\beta})\\ &\quad= (n-2)\hat{\sigma}^2 - (\hat{\mathbf{b}}'\mathbf{X}'\mathbf{Y} - \hat{\mathbf{b}}'\mathbf{X}'\mathbf{X}\hat{\beta})\\ &\quad= (n-2)\hat{\sigma}^2.\end{aligned}$$

□

Step 4. trace$(\mathbf{X}_1'\mathbf{A}\mathbf{X}_1(\mathbf{X}'\mathbf{X})^{-1}) > 0$. Straightforward calculations yield

$$\mathbf{X}_1'\mathbf{A}\mathbf{X}_1 = \begin{pmatrix} \dfrac{n-k}{2} & \dfrac{(n-k)(n+k+3)}{4} \\ \dfrac{(n-k)(n+k-1)}{4} & \dfrac{n-k}{12}(2n^2+2kn+2k^2+3n+3k-5) \end{pmatrix}$$

and

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} \dfrac{n(n+1)(2n+1)}{6} & -\dfrac{n(n+1)}{2} \\ -\dfrac{n(n+1)}{2} & n \end{pmatrix}\frac{1}{Q},$$

where $Q = n^2(n+1)(2n+1)/6 - n^2(n+1)^2/4$. Algebra yields (for $n \geq k \geq 2$)

$$\text{trace}(\mathbf{X}_1'\mathbf{A}\mathbf{X}_1(\mathbf{X}'\mathbf{X})^{-1}) = \frac{(n-k)n}{12Q}[n^2 - nk + 2k^2 - 7] > 0.$$

□

Step 5. Now the proof of Lemma 2 is complete. From Steps 3 and 4 obtain that $S^{(k)}(\mathbf{T}) \geq \hat{S}^{(k)}(n)$. Let N_{SR}^S be defined as N_{SR} with $S^{(k)}(\mathbf{T})$ instead of $\hat{S}^{(k)}(n)$. It follows from the above that N_{SR}^S is stochastically smaller than N_{SR}.

We turn to the investigation of the $\mathrm{P}^{(\infty)}$-distribution of N_{SR}^S. For any $j \geq 2$,

$$\begin{aligned}&\mathrm{P}^{(\infty)}((j-1)t < N_{\text{SR}}^S \leq jt)\\ &\quad= \mathrm{P}^{(\infty)}\left((j-1)t < N_{\text{SR}}^S, \max_{(j-1)t<n\leq jt}\sum_{k=n-t}^{n}\exp(S^{(k)}(n)) \geq dA\right)\end{aligned} \tag{A.7}$$

$$\text{(A.8)} \qquad \le \mathrm{P}^{(\infty)}\left((j-2)t < N_{\mathrm{SR}}^{S}, \max_{(j-1)t<n\le jt} \sum_{k=(j-2)t+1}^{n} \exp(S^{(k)}(n)) \ge dA\right)$$

$$\text{(A.9)} \qquad \le \frac{2t}{dA}\mathrm{P}^{(\infty)}((j-2)t < N_{\mathrm{SR}}^{S}),$$

where the last inequality follows from Doob's inequality, after noticing that

$$\exp\{S^{(k)}(\mathbf{T})\} = \frac{f_k(T_4,\dots,T_n)/f_k(T_4,\dots,T_{(j-2)t})}{f_\infty(T_4,\dots,T_n)/f_\infty(T_4,\dots,T_{(j-2)t})}$$

is a likelihood ratio (conditional on $T_4,\dots,T_{(j-2)t}$), making the sum in (A.8) a conditional submartingale.

For all $1 \le n \le t$, note that $\mathrm{P}^{(\infty)}(N_{\mathrm{SR}}^{S} > n) \ge 1 - n/dA$. Divide (A.7) and (A.9) by $\mathrm{P}^{(\infty)}(N_{\mathrm{SR}}^{S} > (j-1)t)$. Apply induction to get the bound

$$\mathrm{P}^{(\infty)}(N_{\mathrm{SR}}^{S} > jt \,|\, N_{\mathrm{SR}}^{S} > (j-1)t) \ge 1/2 + [1/4 - (2t)/(dA)]^{1/2} \approx 1 - (2t)/(dA),$$

which leads to the relation

$$\text{(A.10)} \qquad \mathrm{P}^{(\infty)}(N_{\mathrm{SR}}^{S} > n) \ge \left\{1/2 + [1/4 - (2t)/(dA)]^{1/2}\right\}^{n/t+1}.$$

It follows from the definition of c and d that $\mathrm{E}^{(\infty)}N \ge \mathrm{E}^{(\infty)}(G \wedge N_{\mathrm{SR}}^{S}) \ge A$. This completes the proof of Lemma 2. □

PROOF OF THEOREM 3. The distribution of $\hat{S}^{(k)}(n)$ is invariant with respect to an affine transformation of time and multiplication by a positive scalar. Hence, one may assume, without loss of generality, that $\beta_0 = \beta_1 = 0$ and $\sigma^2 = 1$.

Fix k. Consider the statistic $\mathbf{Z}_k = (\sum_{i=1}^{k-1} Y_i, \sum_{i=1}^{k-1} iY_i, \sum_{i=1}^{k-1} Y_i^2)$ and define the event B_k as

$$\left\{\left|\sum_{i=1}^{k-1} Y_i\right| \le \varepsilon(k-1),\ \left|\sum_{i=1}^{k-1} iY_i\right| \le \varepsilon(k-1)^2,\ \left|\sum_{i=1}^{k-1} Y_i^2 - (k-1)\right| \le \varepsilon(k-1)\right\}.$$

Note that $\mathrm{P}^{(k)}(\bar{B}_k) \le \exp(-\tau(k-1))$, for some $\tau > 0$, where $\bar{B}_k$ is the complement of B_k. Let $k < n$ and consider $\hat{S}^{(k)}(n)$. Note that given $\mathbf{Z}_k$, the event $\{N \ge k\}$ and the statistic $\hat{S}^{(k)}(n)$ are independent.

Let $n_1 = k - 1 + \lfloor (1+\varepsilon)I^{-1}(\log A)\rfloor$ and $n_2 = k + \lceil c_1 I^{-1}(\log A)\rceil$ for some small ε and some large c_1 (to be determined later). It follows that

$$\begin{aligned}
&\mathrm{E}^{(k)}(N - k + 1 \,|\, N \ge k) \\
&\quad \le (1+\varepsilon)I^{-1}(\log A) + c_1 I^{-1}(\log A) \max_{\mathbf{Z}_k \in B_k} \mathrm{P}^{(k)}(\hat{S}^{(k)}(n_1) < \log A \,|\, \mathbf{Z}_k) \\
&\qquad + \mathrm{E}(G)\frac{\mathrm{P}^{(\infty)}(\bar{B}_k)}{\mathrm{P}^{(\infty)}(N \ge k)} + \mathrm{E}(G) \max_{\mathbf{Z}_k \in B_k} \mathrm{P}^{(k)}(\hat{S}^{(k)}(n_2) < \log A \,|\, \mathbf{Z}_k).
\end{aligned}$$

The proof will be complete when, on the right-hand side of this inequality, all terms but the first will be shown to be negligible.

Begin with the term $\mathrm{E}(G)\mathrm{P}^{(\infty)}(\bar{B}_k)/\mathrm{P}^{(\infty)}(N \geq k)$. Note that the ratio of the probabilities is less than $\exp\{t/A - (k-1)[\tau - (\zeta + 1/c)/A]\}$, for some $\zeta > 0$. Thus, if $k \gg (1/\tau)\log A$ then this term is $o(1/A)$, as $A \to \infty$.

For the other two terms we need to investigate $\hat{S}^{(k)}(n)$, both when $n = n_1$ and when $n = n_2$. However, it can be shown that

$$\hat{S}^{(k)}(n) = \frac{1}{2\hat{\sigma}^2}\Bigg[\sum_{i=k+1}^{n}(Y_i - (\hat{\beta}_0 + i\hat{\beta}_1))^2 - \sum_{i=k+1}^{n}(Y_i - Y_{i-1} - \hat{\beta}_1))^2\Bigg].$$

The distribution of $\hat{S}^{(k)}(n)$ depends on $\mathbf{Z}_k$ only via $(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2)$. We first show that the conditional probability of the event

$$C_{k,n} = \{\hat{\sigma}^2 > 1 + 2\varepsilon\} \cup \{\hat{\beta}_0 + |n\hat{\beta}_1| > 2\varepsilon\}$$

is $o(1)$ when $n = n_1$ and is $o(1/A)$ when $n = n_2$. It is then straightforward to see that the probability of the event $\{\hat{S}^{(k)}(n) < \log A\} \cap \bar{C}_{k,n}$, is $o(1)$ when $n = n_1$ and is $o(1/A)$ when $n = n_2$, which is all that is needed in order to prove that the terms in question are small. Indeed,

$$\begin{aligned}\mathrm{P}^{(k)}(\hat{\sigma}^2 > 1 + 2\varepsilon \,|\, \mathbf{Z}_k) &\leq \mathrm{P}^{(k)}\left(\frac{\sum_{i=1}^{k-1} Y_i^2 + \sum_{i=k}^{n} Y_i^2}{n-2} > 1 + \varepsilon \,|\, \mathbf{Z}_k\right) \\ &\leq \mathrm{P}^{(k)}(\textstyle\sum_{i=k}^{n} Y_i^2 \geq \varepsilon(n-2)),\end{aligned}$$

since $\mathbf{Z}_k \in B_k$. The $\mathrm{P}^{(k)}$- distribution of $\sum_{i=k}^{n} Y_i^2$ is a noncentral χ^2 with $n-k+1$ degrees of freedom and $2I(n-k+1)$ as the parameter of noncentrality. An exponential Markov inequality can be used in order to establish the necessary bounds. Similar arguments can be used for the conditional probability of the event $\{\hat{\beta}_0 + |n\hat{\beta}_1| > \varepsilon\}$. □

REFERENCES

LAI, T. L. (1993). Information bounds and quick detection of parameter changes in stochastic systems. Technical report, Stanford Univ.

LAI, T. L. (1995). Sequential changepoint detection in quality control and dynamical systems. *J. Roy. Statist. Soc. Ser. B* **57** 613–658.

LORDEN, G. (1971). Procedures for reacting to a change in distribution. *Ann. Math. Statist.* **42** 1897–1908.

MOUSTAKIDES, G. V. (1986). Optimal stopping times for detecting changes in distributions. *Ann. Statist.* **14** 1379–1387.

POLLAK, M. (1985). Optimal detection of a change in distribution. *Ann. Statist.* **13** 206–227.

POLLAK, M. and SIEGMUND, D. (1985). A diffusion process and its application to detecting a change in the drift of a Brownian motion. *Biometrika* **72** 267–280.

RITOV, Y. (1990). Decision theoretic optimality of the CUSUM procedure. *Ann. Statist.* **18** 1464–1469.

ROBBINS, H. and ZHANG, C.-H. (1993). A change point problem with some applications. Technical report, Rutgers Univ.

YAKIR, B. (1996). A lower bound on the ARL to detection with a probability constraint on false alarm. *Ann. Statist.* **24** 431–435.

YAKIR, B. (1997). A note on optimal detection of a change in distribution. *Ann. Statist.* **25** 2117-2126.

YAO, Q. (1993). Asymptotically optimal detection of a change in a linear model. *Sequential Anal.* **12** 201–210.

DEPARTMENT OF STATISTICS
HEBREW UNIVERSITY
JERUSALEM
ISRAEL 91905
E-MAIL: msby@mscc.huji.ac.il
msmp@mscc.huji.ac.il

DEPARTMENT OF STATISTICS
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104-6302
E-MAIL: abba@compstat.wharton.upenn.edu