Sequential Analysis, 29: 162–175, 2010

ISSN: 0747-4946 print/1532-4176 online
DOI: 10.1080/07474941003741078

Sequential Change-Point Detection When the Pre- and Post-Change Parameters are Unknown

Tze Leung Lai[1] **and Haipeng Xing**[2]
[1]Department of Statistics, Stanford University, Stanford, California, USA
[2]Department of Applied Mathematics and Statistics,
State University of New York, Stony Brook, New York, USA

Abstract: We describe asymptotically optimal Bayesian and frequentist solutions to the problem of sequential change-point detection in multiparameter exponential families when the pre- and post-change parameters are unknown. In this connection we also address certain issues recently raised by Mei (2008) concerning performance criteria for detection rules in this setting.

Keywords: Asymptotic optimality; Bayes; Detection delay; False alarm rate; Generalized likelihood ratio.

Subject Classifications: 62L12; 62F15; 62M05.

1. INTRODUCTION

Let $\mathbf{X}_1, \mathbf{X}_2, \ldots$ be independent random vectors with a common density function f_0 for $t < v$ and with another common density function f_1 for $t \geq v$. Shiryaev (1978) formulated the problem of optimal sequential detection of the change-time v in a Bayesian framework, by putting a geometric prior distribution on v and assuming a loss of c for each observation taken at or after v and a loss of 1 for a false alarm before v. He used optimal stopping theory to show that the Bayes rule triggers an alarm as soon as the posterior probability that a change has occurred exceeds some fixed level. Since

$$P\{v \leq n \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n\} = R_{p,n}/(R_{p,n} + p^{-1}), \tag{1.1}$$

Received April 13, 2009, Revised June 20, 2009, Accepted June 26, 2009
Recommended by A. G. Tartakovsky
Address correspondence to Haipeng Xing, Department of Applied Mathematics and Statistics, State University of New York, Stony Brook, New York 11794, USA; E-mail: xing@ams.sunysb.edu

where p is the parameter of the geometric distribution $P\{\nu = n\} = p(1-p)^{n-1}$ and

$$R_{p,n} \sum_{k=1}^{n} \prod_{i=k}^{n} \{f_1(\mathbf{X}_i)/(1-p)f_0(\mathbf{X}_i)\}, \tag{1.2}$$

the Bayes rule declares at time

$$N(\gamma) = \inf\{n \geq 1 : R_{p,n} \geq \gamma\} \tag{1.3}$$

that a change has occurred. Roberts (1966) proposed to consider the case $p = 0$ in (1.3) as well, yielding

$$\widetilde{N}(\gamma) = \inf\left\{n \geq 1 : \sum_{k=1}^{n} \prod_{i=k}^{n} (f_1(\mathbf{X}_i)/f_0(\mathbf{X}_i)) \geq \gamma\right\}, \tag{1.4}$$

which was called the Shiryaev–Roberts rule by Pollak (1985), who also proved that it is asymptotically Bayes risk efficient as $p \to 0$.

Instead of the Bayesian approach, Lorden (1971) used the minimax approach of minimizing the worst-case expected delay

$$\overline{E}_1(T) = \sup_{\nu \geq 1} \operatorname{ess\,sup} E[(T - \nu + 1)^+ \mid \mathbf{X}_1, \ldots, \mathbf{X}_{\nu-1}] \tag{1.5}$$

over the class $\mathcal{F}_\gamma$ of all rules T satisfying the constraint $E_0(T) \geq \gamma$ on the expected duration to false alarm. He showed that as $\gamma \to \infty$, Page's (1954) CUSUM rule

$$\tau = \inf\left\{n \geq 1 : \max_{1 \leq k \leq n} \sum_{i=k}^{n} \log(f_1(\mathbf{X}_i)/f_0(\mathbf{X}_i)) \geq c\right\}, \tag{1.6}$$

with c so chosen that $E_0(\tau) = \gamma$, is asymptotically minimax in the sense that

$$\overline{E}_1(\tau) \sim \inf_{T \in \mathcal{F}_\gamma} \overline{E}_1(T) \sim (\log \gamma)/\mathcal{I}(f_1, f_0) \quad \text{as } \gamma \to \infty, \tag{1.7}$$

where $\mathcal{I}(f_1, f_0) = E_1\{\log(f_1(\mathbf{X}_t)/f_0(\mathbf{X}_t))\}$ is the Kullback–Leibler information number. Note that (1.6) essentially replaces $\sum_{k=1}^{n}$ in (1.4) by $\max_{1 \leq k \leq n}$, which can be regarded as using maximum likelihood to estimate the unknown change-point.

In this paper we consider the case in which the pre- and post-change density functions are not known in advance, but belong to a multivariate exponential family

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = \exp\{\boldsymbol{\theta}'\mathbf{x} - \psi(\boldsymbol{\theta})\} \tag{1.8}$$

with respect to some measure ω on $\mathbb{R}^d$. The generalized likelihood ratio (GLR) statistic for testing the null hypothesis of no change-point, based on $\mathbf{X}_1, \ldots, \mathbf{X}_n$, versus the alternative hypothesis of a single change-point prior to n but not before n_0 is

$$\begin{aligned}
&\max_{n_0 \leq k < n} \left\{\sup_{\boldsymbol{\theta}} \sum_{i=1}^{k} \log f_{\boldsymbol{\theta}}(\mathbf{X}_i) + \sup_{\tilde{\boldsymbol{\theta}}} \sum_{i=k+1}^{n} \log f_{\tilde{\boldsymbol{\theta}}}(\mathbf{X}_i) - \sup_{\boldsymbol{\lambda}} \sum_{i=1}^{n} \log f_{\boldsymbol{\lambda}}(\mathbf{X}_i)\right\} \\
&\quad = \max_{n_0 \leq k < n} \left\{kI(\overline{\mathbf{X}}_{1,k}) + (n-k)I(\overline{\mathbf{X}}_{k+1,n}) - nI(\overline{\mathbf{X}}_{1,n})\right\},
\end{aligned} \tag{1.9}$$

where

$$\overline{\mathbf{X}}_{m,n} = \sum_{i=m}^{n} \mathbf{X}_i/(n-m+1), \quad I(\boldsymbol{\mu}) = \sup_{\boldsymbol{\theta}}\{\boldsymbol{\theta}^T\boldsymbol{\mu} - \psi(\boldsymbol{\theta})\} = \boldsymbol{\theta}_{\boldsymbol{\mu}}^T\boldsymbol{\mu} - \psi(\boldsymbol{\theta}_{\boldsymbol{\mu}}), \tag{1.10}$$

and $\boldsymbol{\theta}_{\boldsymbol{\mu}} = (\nabla\psi)^{-1}(\boldsymbol{\mu})$, noting that $\sup_\lambda$ is related to maximizing the likelihood under the null hypothesis and that $\sup_{\boldsymbol{\theta}}$ and $\sup_{\tilde{\boldsymbol{\theta}}}$ arise from maximizing the likelihood under the hypothesis that a single change-point occurs at $k+1$. Let

$$g(\alpha, x, y) = \alpha I(x) + (1-\alpha)I(y) - I(\alpha x + (1-\alpha)y). \tag{1.11}$$

Replacing the likelihood ratio statistics in the CUSUM rule (1.6) by the GLR statistics (1.9) yields the GLR rule

$$\widehat{N} = \inf\left\{n > n_0 : \max_{n_0 \le k < n} ng(k/n, \overline{\mathbf{X}}_{1,k}, \overline{\mathbf{X}}_{k+1,n}) \ge c\right\} \tag{1.12}$$

for detecting a change in $\boldsymbol{\theta}$ when the pre- and post-parameters are unknown. For the special case of the $N(\theta, 1)$ family, $ng(k/n, \overline{\mathbf{X}}_{1,k}, \overline{\mathbf{X}}_{k+1,n})$ in (1.12) can be expressed as

$$k(n-k)\big(\overline{\mathbf{X}}_{k+1,n} - \overline{\mathbf{X}}_{1,k}\big)^2/n = nk(\overline{\mathbf{X}}_{1,n} - \overline{\mathbf{X}}_{1,k})^2/(n-k). \tag{1.13}$$

Therefore, in this normal case, $\widehat{N}$ is the same as the rule (5.3) in Siegmund and Venkatraman (1995), which is related to an earlier rule proposed by Pollak and Siegmund (1991) under the assumption of known $\delta := \theta_1 - \theta_0$ and having the form

$$\widetilde{N}_\delta = \inf\left\{n > n_0 : \sum_{k=n_0}^{n} \exp[\delta k(\overline{\mathbf{X}}_{1,n} - \overline{\mathbf{X}}_{1,k}) - \delta^2 k(1-k/n)/2] \ge \gamma\right\}, \tag{1.14}$$

with $\log\gamma \sim c$. In (1.12) or (1.14), it is assumed that the actual change-time, defined as ∞ if no change ever occurs, is larger than the initial sample size n_0. Note that if we replace $\sum_{k=k_0}^{n}$ by $\max_{n_0 \le k \le n}$ and maximize the exponential term in (1.14) over δ, then we obtain (1.12) in this normal case in view of (1.13).

Mei (2006, 2008) has recently introduced two alternative approaches to sequential change-point detection when the pre- and post-change values of the parameter are not completely specified. The first approach is to specify a detection delay at a given post-change parameter value θ_1 and to maximize the ARLs to false alarm for all possible values of the pre-change parameter; this approach has also been extended to the case where both the pre- and post-change parameters are only partially specified. The second approach uses a mixing distribution to handle the unknown pre-change parameter; this is akin to a Bayesian approach, as pointed out by Lai and Chan (2008, p. 387) who "find it much more natural to adopt a full Bayesian change-point model for the unknown pre- and post-change parameters and the unknown change-time." In this paper we introduce a stopping rule, similar to Shiryaev's rule but with $P\{\nu \le n \mid \mathbf{X}_1, \ldots, \mathbf{X}_n\}$ in (1.1) replaced by a modified extension to the case of unknown pre- and post-change parameters. Section 2 describes this modified extension and the associated change-point detection rule. Section 3 develops an asymptotic optimality theory for sequential change-point detection when the pre- and post-change parameters are unknown, from both Bayesian and frequentist viewpoints, and Section 4 reports some simulation studies.

2. EXTENSION OF SHIRYAEV'S BAYESIAN CHANGE-POINT MODEL AND DETECTION RULE

For the multiparameter exponential family (1.8), let π be a prior density function (with respect to Lebesgue measure) on $\boldsymbol{\Theta} := \{\boldsymbol{\theta} : \int e^{\boldsymbol{\theta}'\mathbf{X}} d\omega(\mathbf{X}) < \infty\}$ given by

$$\pi(\boldsymbol{\theta}; a_0, \boldsymbol{\mu}_0) = c(a_0, \boldsymbol{\mu}_0)\exp\{a_0\boldsymbol{\mu}_0'\boldsymbol{\theta} - a_0\psi(\boldsymbol{\theta})\}, \quad \boldsymbol{\theta} \in \boldsymbol{\Theta}, \tag{2.1}$$

where

$$1/c(a_0, \boldsymbol{\mu}_0) = \int_{\boldsymbol{\Theta}} \exp\{a_0\boldsymbol{\mu}_0'\boldsymbol{\theta} - a_0\psi(\boldsymbol{\theta})\}d\boldsymbol{\theta}, \quad \boldsymbol{\mu}_0 \in (\boldsymbol{\nabla}\psi)(\boldsymbol{\Theta}), \tag{2.2}$$

in which $\boldsymbol{\nabla}$ denotes the gradient vector of partial derivatives. The posterior density of $\boldsymbol{\theta}$ given the observations $\mathbf{X}_1, \ldots, \mathbf{X}_m$ drawn from $f_{\boldsymbol{\theta}}$ is

$$\pi\left(\boldsymbol{\theta}; a_0 + m, \left(a_0\boldsymbol{\mu}_0 + \sum_{i=1}^{m}\mathbf{X}_i\right)\Big/(a_0 + m)\right); \tag{2.3}$$

see Diaconis and Ylvisaker (1979, p. 274). Moreover,

$$\int_{\boldsymbol{\Theta}} f_{\boldsymbol{\theta}}(\mathbf{X})\pi(\boldsymbol{\theta}; a, \boldsymbol{\mu})d\boldsymbol{\theta} = \frac{c(a, \boldsymbol{\mu})}{c(a + 1, (a\boldsymbol{\mu} + \mathbf{X})/(a + 1))}. \tag{2.4}$$

Suppose that the parameter $\boldsymbol{\theta}$ takes the value $\boldsymbol{\theta}_0$ for $t < \nu$ and another value $\boldsymbol{\theta}_1$ for $t \geq \nu$, and that the change-time ν and the pre- and post-change values $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ are unknown. Following Shiryaev (1963, 1978), we use the Bayesian approach that assumes ν to be geometric with parameter p but constrained to be larger than n_0 and that $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1$ are independent, have the same density function (2.1), and are also independent of ν. Let $\pi_n = P\{\nu \leq n \mid \mathbf{X}_1, \ldots, \mathbf{X}_n\}$. Whereas π_n is a Markov chain in the case of known $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$, it is no longer Markovian in the present setting of unknown pre- and post-change parameters and Shiryaev's rule that triggers an alarm when π_n exceeds some threshold is no longer optimal; see Zacks (1991, pp. 540–541) who suggests using dynamic programming to find the optimal stopping rule but also notes that "it is generally difficult to determine the optimal stopping boundaries."

Because of the complexity of the Bayes rule, Zacks and Barzily (1981) introduced a more tractable myopic (two-step-ahead) policy in the univariate Bernoulli case ($\mathbf{X}_i = 0$ or 1). Here we introduce a modification of Shiryaev's rule, which will be shown in Section 3.2 to be asymptotically Bayes as $p \to 0$. Let $\mathcal{F}_t$ denote the σ-field generated by $\mathbf{X}_1, \ldots, \mathbf{X}_t$.

Let $\pi_{0,0} = c(a_0, \boldsymbol{\mu}_0)$ and $\pi_{i,j} = c(a_0 + j - i + 1, (a_0\boldsymbol{\mu}_0 + \sum_{t=i}^{j}\mathbf{X}_t)/(a_0 + j - i + 1))$. Note that for $n_0 < i < n$,

$$P\{\nu = i \mid \mathcal{F}_n\} \propto p(1 - p)^{i-1}\pi_{0,0}^2/\pi_{1,i-1}\pi_{i,n}, \quad P\{\nu > n \mid \mathcal{F}_n\} \propto p(1 - p)^n\pi_{0,0}/\pi_{1,n}. \tag{2.5}$$

The normalizing constant is determined by the fact that all the probabilities in (2.5) sum to 1. In view of (2.1), we can express $P(n_0 < \nu \leq n \mid \mathcal{F}_n) = \sum_{i=n_0+1}^{n} P\{\nu = i \mid \mathcal{F}_n\}$ in terms of the $\pi_{i,j}$. Therefore Shiryaev's stopping rule in the present setting of unknown pre- and post-change parameters can again be written in the form of

(1.3) with $R_{p,n} = \sum_{i=n_0+1}^{n} \pi_{0,0}\pi_{1,n}/\{(1-p)^{n-i}\pi_{1,i}\pi_{i,n}\}$ in place of (1.2). Although this stopping rule is no longer optimal, as noted earlier, it can be modified to give

$$N = \inf\{n > n_p : P(v \le n \mid v \ge n - k_p, \mathcal{F}_n) \ge \eta_p\}, \tag{2.6}$$

which is shown in Section 3.2 to be asymptotically optimal as $p \to 0$, for suitably chosen k_p, η_p and $n_p \ge n_0$. Since

$$P(v \le n \mid v \ge n - k_p, \mathcal{F}_n) \sum_{i=n-k_p}^{n} P(v = i \mid \mathcal{F}_n) \Big/ \left\{ \sum_{i=n-k_p}^{n} P(v = i \mid \mathcal{F}_n) + P(v > n \mid \mathcal{F}_n) \right\},$$

we can use (2.5) to rewrite (2.6) in the form

$$N = \inf\left\{ n > n_p : \sum_{i=n-k_p}^{n} \frac{\pi_{0,0}\pi_{1,n}}{(1-p)^{n-i+1}\pi_{1,i-1}\pi_{i,n}} \ge \gamma_p \right\}. \tag{2.7}$$

This has essentially the same form as Shiryaev's rule (1.3) with the obvious changes to accommodate the unknown $f_{\boldsymbol{\theta}_0}$ and $f_{\boldsymbol{\theta}_1}$, and with the important sliding window modification $\sum_{i=n-k_p}^{n}$ of Shiryaev's sum $\sum_{i=n_0+1}^{n}$, which has too many summands to trigger false alarms when $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ are estimated sequentially from the observations.

Lai et al. (2009) have modified the above Bayesian model and detection rule to handle the case where multiple change-points with unknown pre- and post-change parameters can occur, as in sequential surveillance applications. The key idea is still to use sliding windows $\sum_{i=n-k_p}^{n}$ as in (2.7) but with the summands modified to be the posterior probabilities p_{in} that the most recent change-point up to time n occurs at i.

3. ASYMPTOTICALLY OPTIMAL DETECTION RULES WHEN THE PRE- AND POST-CHANGE PARAMETERS ARE UNKNOWN

In this section we first develop an asymptotic optimality theory for sequential change-point detection when the pre- and post-change parameters $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ are unknown. As noted in Section 1, Mei (2008) and the discussants of his paper have pointed out various issues on how performance should be evaluated in this case. In particular, Mei (2008) suggests using a weight function (prior density) w on the unknown pre-change value θ_0, which he assumes to be univariate, and putting a constraint on $\int E_\theta(T)w(\theta)d\theta$ (integrated average run length to false alarm) or on $\int [E_\theta(T)]^{-1} w(\theta)d\theta$ (which he interprets as an expected false alarm rate). He still uses the worst-case conditional expectation (1.5) evaluated at a nominal value λ of the post-change parameter, which may not be equal to the actual θ_1, as the performance measure for detection delay. Although not explicitly stated, he assumes that there is prior knowledge of some $\tilde{\lambda}$ (which is 0 in his example) such that $\tilde{\lambda} > \theta_0$ so that w is supported on $(-\infty, \tilde{\lambda}]$ and $\lambda > \tilde{\lambda}$. As will be shown in Section 3.1, the expected detection delay depends heavily on v because the maximal amount of information to estimate $\boldsymbol{\theta}_0$ comes from the baseline data $\mathbf{X}_1, \ldots, \mathbf{X}_{v-1}$. Hence the worst-case scenario $\sup_{v>1}$ in (1.5) basically rules out any baseline information for inference on $\boldsymbol{\theta}_0$ in this worst case, which therefore has to rely heavily on the prior density

w of $\boldsymbol{\theta}_0$. The performance measures proposed in Lai and Chan's (2008) discussion provide more flexible criteria to evaluate detection rules in this setting. Mei (2008, p. 416) does not agree to use these performance measures because they only yield "*individual* detection delay" at each change-point v but not "taking the supremum over all change-points $v \geq 1$" as in (1.5). However, "individual" rather than "worst-case" detection delay is what we should consider in this case, since asymptotic optimality (under individual detection delay) for large γ entails consistent estimation of the value of $\boldsymbol{\theta}_0$ from the initial segment of the data when v is also large enough, whereas the case $v = 2$ in $\sup_{v\geq 1}$ for the worst-case detection delay allows only one observation for learning the value of $\boldsymbol{\theta}_0$.

In Section 3.1, we develop an asymptotic lower bound for the expected detection delay $E_{\boldsymbol{\theta}_0,\boldsymbol{\theta}_1}(T - v)^+$ that depends on how large v is, subject to a false alarm probability constraint at the actual (but unknown) $\boldsymbol{\theta}_0$. We also show how this asymptotic lower bound can be attained by a sliding-window modification of the GLR rule (1.9). Section 3.2 considers the Bayesian formulation of the detection problems using the model described in Section 2 and a cost of 1 for a false alarm before v and of c each observations taken at or after v. We show that as $p \to 0$, the rule (2.6) is asymptotically Bayes when $\gamma_p \to \infty$, $v \to \infty$ and $c \to 0$ at certain rates depending on p.

3.1. Asymptotically Efficient Window-Limited GLR Detection Rules

The performance measures proposed by Lai and Chan (2008) in their discussion of Mei's (2008) paper include the expected detection delay $E_{\boldsymbol{\theta}_0,\boldsymbol{\theta}_1}(T - v)^+$ of a detection rule with stopping time T and the (long-run) probability of false alarm per unit time

$$\mathrm{PFA}_{m,k} = m^{-1} P_{\boldsymbol{\theta}_0}\{k \leq T < k + m\} \tag{3.1}$$

for some "sufficiently large (but not too large) m," where $P_{\boldsymbol{\theta}_0}$ denotes the probability measure under which all $\mathbf{X}_i$ have parameter $\boldsymbol{\theta}_0$. The rationale underlying (3.1) is explained in Lai (1995, p. 631) in the case of known $\boldsymbol{\theta}_0$. Tartakovsky (2005, 2008) have proposed to impose a constraint on a stronger version than (3.1), called *conditional false alarm* and given by

$$\mathrm{CPFA}_{m,k} = P_{\boldsymbol{\theta}_0}\{k \leq T < k + m \mid T \geq k\} = P_{\boldsymbol{\theta}_0}(k \leq T < k + m)/P_{\boldsymbol{\theta}_0}(T \geq k),$$

in which m is a fixed constant. Although the constraint $\sup_k \mathrm{CPFA}_{m,k} \leq \alpha$ has been demonstrated to hold for the CUSUM rule based on i.i.d. observations (under $P_{\boldsymbol{\theta}_0}$) by Tartakovsky (2005, p. 321), it fails to hold for the Siegmund and Venkatraman (1995) GLR rule for detecting shifts in a normal mean from a known baseline value 0 to an unknown post-change value θ, when α is sufficiently small. This can be shown by choosing k sufficiently large (in $\sup_k \mathrm{CPFA}_{m,k}$) so that $\{T \geq k\}$ is a rare event and the large deviation principle for the sample paths of S_n implies that on $\{T \geq k\}$, $\max_{n\leq k} S_n/\sqrt{n}$ and $S_k/\sqrt{k}$ (or $\max_{n\leq k}(-S_n)/\sqrt{n}$ and $-S_k/\sqrt{k}$) are concentrated relatively close to c. Here $S_n = \sum_{i=1}^n X_i$, $S_{n,j} = S_n - S_j$ for $n > j$ and $T = \inf\{n \geq 1 : |S_{n,j}|/\sqrt{n-j} \geq c$ for some $1 \leq j \leq n\}$, so $k \gg \exp(c^2/2)$ yields $\{T \geq k\}$ as a rare event; see Siegmund and Venkatraman (1995) and Chan and Lai (2003).

When $\boldsymbol{\theta}_0$ is unknown but it is known that $v > n_0$, one can use a training sample of size n_0 to estimate $\boldsymbol{\theta}_0$ and thereby to implement (3.1) for GLR-type detection rules. Such implementation issues will be addressed in Section 4.3, and we focus here on developing an asymptotic lower bound for $E_{\boldsymbol{\theta}_0,\boldsymbol{\theta}_1}(T-v)^+$ subject to an upper bound on (3.1) and showing how this asymptotic lower bound can be attained in the framework of independent observations from the multivariate exponential family (1.8). Let

$$\mathcal{I}(\boldsymbol{\mu}, \gamma) = (\boldsymbol{\theta}_{\boldsymbol{\mu}} - \boldsymbol{\theta}_\gamma)'\boldsymbol{\mu} - (\psi(\boldsymbol{\theta}_{\boldsymbol{\mu}}) - \psi(\boldsymbol{\theta}_\gamma)). \tag{3.2}$$

Note that $I(\boldsymbol{\mu})$ defined in (1.10) is the rate function in large deviations theory and $\mathcal{I}(\boldsymbol{\mu}, \gamma)$ is the Kullback–Leibler information number for the exponential family. Let $\boldsymbol{\mu}_i = \nabla\psi(\boldsymbol{\theta}_i)$ for $i = 0, 1$.

Theorem 3.1. *Suppose the stopping rule T satisfies the constraint*

$$\sup_{k\geq 1} P_{\boldsymbol{\theta}_0}\{k \leq T < k+m\}/m \leq \alpha \tag{3.3}$$

for some $m = m_\alpha$ such that as $\alpha \to 0$,

$$\liminf_{\alpha\to 0} m_\alpha/|\log\alpha| > 1/\kappa(\boldsymbol{\theta}_1, \boldsymbol{\theta}_0) \quad \textit{but} \quad \log m_\alpha = o(\log\alpha), \tag{3.4}$$

where $\kappa(\boldsymbol{\theta}_1, \boldsymbol{\theta}_0)$ is defined below.

(a) *If $v/|\log\alpha| \to \infty$ as $\alpha \to 0$, let $\kappa(\boldsymbol{\theta}_1, \boldsymbol{\theta}_0) = \mathcal{I}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_0)$. Then as $\alpha \to 0$,*

$$E_{\boldsymbol{\theta}_0,\boldsymbol{\theta}_1}(T-v)^+ \geq \{P_{\boldsymbol{\theta}_0}(T \geq v)/\kappa(\boldsymbol{\theta}_1, \boldsymbol{\theta}_0) + o(1)\}|\log\alpha|. \tag{3.5}$$

(b) *If $v \sim b|\log\alpha|$ for some $b > 0$, let $\kappa_b > b$ be the solution of the equation*

$$\frac{b}{\kappa_b} = \frac{I(\boldsymbol{\mu}_1) - I([b\boldsymbol{\mu}_0 + (\kappa_b - b)\boldsymbol{\mu}_1]/\kappa_b)}{I(\boldsymbol{\mu}_1) - I(\boldsymbol{\mu}_0)}, \tag{3.6}$$

and let $\kappa(\boldsymbol{\theta}_1, \boldsymbol{\theta}_0) = \kappa_b$. Then (3.5) still holds.

Proof. Theorem 2 of Lai (1998) gives the lower bound (3.5) with $\kappa(\boldsymbol{\theta}_1, \boldsymbol{\theta}_0) = \mathcal{I}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_0)$ subject to the constraint (3.3), when $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ are both assumed to be known so that the change of measure from $P_{\boldsymbol{\theta}_0,\boldsymbol{\theta}_1}$ involves the likelihood ratio statistic $\prod_{i=v}^T (f_{\boldsymbol{\theta}_1}(\mathbf{X}_i)/f_{\boldsymbol{\theta}_0}(\mathbf{X}_i))$. We shall modify the proof of Theorem 2 of Lai (1998) by using the inequality

$$\prod_{i=v}^n (f_{\boldsymbol{\theta}_1}(\mathbf{X}_i)/f_{\boldsymbol{\theta}_0}(\mathbf{X}_i)) \geq \left\{\prod_{i=1}^{v-1} f_{\boldsymbol{\theta}_0}(\mathbf{X}_i) \prod_{i=v}^n f_{\boldsymbol{\theta}_1}(\mathbf{X}_i)\right\} \Big/ \sup_{\boldsymbol{\theta}} \prod_{i=1}^n f_{\boldsymbol{\theta}}(\mathbf{X}_i) \tag{3.7}$$

for $n \geq v$. Letting $\kappa = \kappa(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$,

$$\ell_n = \sum_{i=1}^{v-1} \log f_{\boldsymbol{\theta}_0}(\mathbf{X}_i) + \sum_{i=v}^n \log f_{\boldsymbol{\theta}_1}(\mathbf{X}_i) - nI(\overline{\mathbf{X}}_{1,n}),$$
$$C_\delta = \{0 \leq T - v < (1-\delta)\kappa^{-1}|\log\alpha|, \ell_r < (1-\delta^2)\mathrm{mod}\log\alpha|\}$$

for $0 < \delta < 1$, we can use the likelihood-ratio identity (associated with the change of measures from $P_{\boldsymbol{\theta}_0}$ to $P_{\boldsymbol{\theta}_0,\boldsymbol{\theta}_1}$ together with (3.3) and (3.7) to conclude that $\alpha \geq E_{\boldsymbol{\theta}_0,\boldsymbol{\theta}_1}(e^{\ell_T} 1_{C_s})$, and then proceed as in the proof of Theorem 2 in Lai (1998, p. 2920) to derive (3.5).

To prove Theorem 3.1, we can in fact apply Theorem 2 of Lai (1998), since the asymptotic lower bound in (a) is the same as (and that in (b) can be shown to be smaller than) the asymptotic lower bound in Theorem 2 of Lai (1998). The reason why we give a direct proof via (3.7) is to use it to shed light on how this asymptotic lower bound can be attained. Since ν, $\boldsymbol{\theta}_0$, and $\boldsymbol{\theta}_1$ on the right hand side of (3.7) are unknown, we can replace them by the maximum likelihood estimate that maximizes the likelihood over $n_0 < k < n$, or over $\boldsymbol{\theta}_0$ (based on $\mathbf{X}_1, \ldots, \mathbf{X}_{k-1}$) or over $\boldsymbol{\theta}_1$ (based on $\mathbf{X}_k, \ldots, \mathbf{X}_n$). This is tantamount to using the sequential GLR rule (1.9), which, however, does not have a sliding window feature. To satisfy the probability of false alarm constraint (3.3), we use the ideas of Lai (1995, p. 626) and Chan and Lai (2003, pp. 416–417). Take $a > 0$, $A > 1$, $B > 1$, and define

$$M = \lfloor a|\log\alpha| \rfloor, \quad \mathcal{N} = \{1, \ldots, M\} \cup \{\lfloor B^j M \rfloor : j = 1, \ldots, J\}, \tag{3.8}$$

$$\Lambda_n \begin{cases} \max\limits_{n_0 \leq k < n} ng(k/n, \overline{\mathbf{X}}_{1,k}, \overline{\mathbf{X}}_{k+1,n}) & \text{if } n_0 < n \leq An_0, \\ \max\limits_{k \geq An_0 : n-k \in \mathcal{N}} n\mathcal{I}(\overline{\mathbf{X}}_{k+1,n}, \overline{\mathbf{X}}_{1,n}) & \text{if } n > An_0, \end{cases} \tag{3.9}$$

$$\widetilde{N} = \inf\{n > n_0 : \Lambda_n \geq c_\alpha\}. \tag{3.10}$$

The set $\mathcal{N}$ of window sizes for window-limited GLR detection rules was introduced by Lai (1995) in the case of known $\boldsymbol{\theta}_0$ and is extended to the case of unknown $\boldsymbol{\theta}_0$ here by using the scan statistics (3.9). These scan statistics are improvements of the window-limited GLR statistics suggested by Lai (1995, p. 633) for the normal case, as we now use the estimate $\overline{\mathbf{X}}_{1,k\wedge\tilde{n}}$ of $\boldsymbol{\theta}_0$, which is more reliable than $\overline{\mathbf{X}}_{k-M,k}$ used by Lai (1995). The following theorem establishes the asymptotic optimality of $\widetilde{N}$ as $\alpha \to 0$, for suitably chosen n_0, c_α, a, A and B.

Theorem 3.2. *Suppose $n_0 \sim b_0|\log\alpha|$, c_α is so chosen that $\widetilde{N}$ satisfies* (3.3), *and either $\nu/|\log\alpha| \to \infty$ or $\nu \sim b|\log\alpha|$ for some $b > b_0$. Then by letting $A \to \infty$ and $B \to 1$ sufficiently slowly in* (3.9), *the rule $\widetilde{N}$ attains the asymptotic lower bound* (3.5).

Proof. We first show that $c_\alpha \sim |\log\alpha|$. Let $\beta > 0$. Simple Bonferroni inequalities and chi-square approximations to large deviation probabilities of GLR statistics (e.g., Woodroofe, 1978) can be used to show that

$$P_{\boldsymbol{\theta}_0}(T \leq \beta c_\alpha) = \exp\{-c_\alpha(1 + o(1))\}. \tag{3.11}$$

In fact, a straightforward modification of the proof of Theorem 6 of Chan and Lai (2003) can even yield a precise asymptotic formula. Let $\epsilon > 0$. We can choose β sufficiently large such that

$$P_{\boldsymbol{\theta}_0}\{\|\overline{\mathbf{X}}_{1,n} - \boldsymbol{\mu}_0\| \geq \epsilon \text{ for some } n \geq \beta c\} = o(e^{-c}) \tag{3.12}$$

as $c \to \infty$. Let $\Omega = \{\|\overline{\mathbf{X}}_{1,n} - \boldsymbol{\mu}_0\| \geq \epsilon$ for all $n \geq \beta c\}$ and choose $A \geq \beta$ in (3.9). For $k \geq \beta c_\alpha$, we can divide the integral $[k, k + m]$ into $K \sim m/(\beta c_\alpha)$ disjoint sub-intervals with equal length $\beta c_\alpha(1+o(1))$ to analyze $P_{\boldsymbol{\theta}_0}(\Omega \cap \{k \leq T < k+m_\alpha\})$ by following the argument of Chan and Lai (2003, p. 417) and modifying the derivation of (3.11). Since (3.12) holds and ε is arbitrary, it then follows that

$$\sup_k P_{\boldsymbol{\theta}_0}\{k \leq T < k+m_\alpha\}/m_\alpha = \exp\{-c_\alpha(1+o(1))\}, \tag{3.13}$$

and therefore $c_\alpha \sim \log \alpha^{-1}$ can be chosen so that the left hand side of (3.13) is equal to α.

For part (b) of the theorem, by choosing $A > \kappa_b$, we can use the law of large numbers and uniform integrability (via exponential bounds) or an argument similar to the proof of Theorem 4 in Lai (1998, pp. 2927–2928) to show that $\widetilde{N}$ attains the asymptotic lower bound (3.5). For part (a) of the theorem, $\widetilde{N}$ can still be shown to attain the asymptotic lower bound by letting $A \to \infty$ and $d \to 1$ sufficiently slowly.

3.2. Asymptotic Bayes Property of (2.6)

Consider the Bayesian change-point model described in the first paragraph of Section 2. Following Shiryaev (1978), suppose there is a loss of c for each observation taken at or after ν and a loss of 1 for a false alarm before ν. We can regard c as the Lagrange multiplier for the constrained optimization problem of minimizing $E(T-\nu)^+$ subject to the constraint $P(T<\nu) \leq c_\alpha$, where P is the probability measure under the Bayesian model and E is the corresponding expectation. Although we have assumed in Section 2 independent conjugate prior densities for $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ to come up with explicit formulas for the posterior distribution of ν given $\mathcal{F}_t$ the following theorem holds for any prior distribution π of $(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$ that has a positive continuous density on $\Theta \times \Theta$ with respect to Lebesgue measure so that $\int [\kappa(\boldsymbol{\theta}_1, \boldsymbol{\theta}_0)]^{-1} d\pi < \infty$.

Theorem 3.3. *Suppose $\alpha_p \to 0$ as $p \to 0$ such that*

$$\log p = o(\log \alpha_p) \quad \text{and} \quad p \log \alpha_p \to 0. \tag{3.14}$$

Moreover, assume that in (2.6), $n_p \sim b_0|\log \alpha_p|$, $k_p/|\log \alpha_p| \to \infty$ *sufficiently slowly* (*e.g.*, $k_p \sim (\log \alpha_p^{-1})(\log\log \alpha_p^{-1})$), *and η_p is so chosen that $P(N < \nu) = \alpha_p$. Then as* $p \to 0$,

$$E(N-\nu)^+ \sim \inf_{T: P(T \leq \nu) \leq \alpha_p} E(T-\nu)^+ \sim |\log \alpha_p| \int \frac{d\pi(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)}{k(\boldsymbol{\theta}_1, \boldsymbol{\theta}_0)}. \tag{3.15}$$

Proof. Making use of Theorem 3.1, we can modify the proof of Theorem 3 of Lai (1998) to show that

$$\inf_{T: P(T \leq \nu) \leq \alpha_p} E(T-\nu)^+ \geq |\log \alpha_p| \left\{ \int \frac{d\pi(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)}{k(\boldsymbol{\theta}_1, \boldsymbol{\theta}_0)} + o(1) \right\}, \tag{3.16}$$

noting that (3.13) ensures that the geometric distribution with parameter p satisfies the conditions of Theorems 3 and 4(i) of Lai (1998), where $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ are assumed to be known.

To show that (2.7) attains the asymptotic lower bound in (3.16), first consider the case where the prior distribution is the same as that in Section 2, for which the detection rule N can be expressed in the form (2.7). Lai and Xing (2008, Lemma 1) have made use of Laplace's asymptotic integration formula to show that

$$\pi_{i,j}^{-1} \sim (2\pi)^{d/2} e^{(a_0+j-i+1)I(\overline{\mathbf{X}}_{i,j})} \Big/ \left\{(a_0 + j - i + 1)^d h(\overline{\mathbf{X}}_{i,j})\right\}^{1/2}, \tag{3.17}$$

where d is the dimensionality of $\boldsymbol{\theta}, \overline{\mathbf{X}}_{i,j}$ and $I(\boldsymbol{\mu})$ are defined in (1.10), $h(\boldsymbol{\mu}) = \det(\nabla^2\psi(\boldsymbol{\theta}_{\boldsymbol{\mu}}))$ and $\nabla^2(\psi)$ denotes the Hessian matrix of second partial derivative $\partial^2\psi/\partial\theta_i\partial\theta_j$. Since $(1-p)^\ell \to 1$ uniformly in $1 \le \ell \le k_p$, in view of (3.13) and the assumption on k_p, we can use (3.17) and the arguments in the proof of Theorem 3.2 to show that $E(N - \nu)^+$ has the same order of magnitude as the lower bound in (3.16). From (3.17), it follows that (2.7) with suitably chosen k_p is asymptotically equivalent to the window-limited GLR rule. This asymptotic equivalence actually holds more generally for the prior densities assumed in the theorem because Laplace's asymptotic formula for integrals can be used for more general prior densities of $(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$ than conjugate priors that are independent, as pointed out in Section 6 of Lai and Xing (2008).

The preceding proof can be combined with that of Theorem 3.2 to show the window-limited GLR detection rule (3.10) also attains the asymptotic lower bound (3.16) for the Bayes risk if c_α in (3.10) is chosen suitably. This is the content of the following.

Theorem 3.4. *With the same notation and assumptions as in Theorem* 3.3, *define* $\widetilde{N}$ *by* (3.10), *in which* c_α *is so chosen that* $P(\widetilde{N} < \nu) = \alpha_p$. *Then* $c_\alpha \sim |\log \alpha_p|$ *and* (3.15) *still holds with* N *replaced by* $\widetilde{N}$.

4. IMPLEMENTATION AND SIMULATION STUDIES

4.1. Performance of Window-Limited GLR Rules

In this section we simulate the performance of the window-limited GLR rule in the case where the exponential family (1.8) is the normal family $N(\theta, 1)$. Since the problem is translation invariant, we shall assume without loss of generality that $\theta_0 = 0$. Siegmund and Venkatraman (1995) have provided asymptotic formulas for the average run length (ARL) $E_0(\widehat{N})$ and $E_{0,\theta_1}(\widehat{N} - \nu)^+$, where $\widehat{N}$ is the GLR rule (1.12). Since θ_0 is assumed to be 0, we can simulate the false alarm ARL $E_0(\widehat{N})$, or $E_0(\widetilde{N})$, and thereby choose c in (1.12) for which the false alarm ARL is 1000; see Lai (1995, p. 628). After determining the value of the threshold c in this way, we can simulate $E_{0,\theta_1}(\widehat{N} - \nu)^+$, or $E_{0,\theta_1}(\widetilde{N} - \nu)^+$ for different values of θ_1 and ν. Table 1 compares the expected detection delays of $\widetilde{N}$, for which we have chosen $n_0 = 2$, $M = 50$, $B = 1.5$, and $J = 7$ in (3.10), with those of the GLR rule $\widehat{N}$. Each result is based on 2000 simulations; standard errors are given in parentheses. Also included as a benchmark to show the role played by the magnitude of ν is Lai's (1995) W-L GLR rule that assumes the baseline value θ_0 to be known, using the same choice of M, B and J.

Table 1. Expected detection delays for different values of ν and θ_1

θ_1	Detection rule	$\nu = 100$	$\nu = 500$	$\nu = 1000$
0.5	GLR (baseline unknown)	104.5 (6.0)	31.3 (0.8)	17.7 (0.6)
	WL-GLR (baseline unknown)	122.2 (6.8)	33.2 (0.8)	18.9 (0.7)
	WL-GLR (baseline known)	41.5 (0.7)	30.6 (0.8)	19.5 (0.7)
1	GLR (baseline unknown)	12.7 (0.2)	7.8 (0.2)	4.8 (0.2)
	WL-GLR (baseline unknown)	12.8 (0.2)	7.9 (0.2)	4.9 (0.2)
	WL-GLR (baseline known)	11.0 (0.2)	7.8 (0.2)	5.1 (0.2)

4.2. Performance of (2.6) in Bayes Model

For the normal family considered in the preceding section, we put independent standard normal priors on θ_0 and θ_1 and assume that ν is geometric with parameter p, as in Section 2, to study the Bayesian detection delay of (2.6), in which we choose $n_p = 5$ and $k_p = 2000$. For $p = 0.005, 0.01, 0.02$ and different values of α, we determine the threshold η_p such that $P(N < \nu) = \alpha$ by using Monte Carlo to evaluate the probability in this Bayesian model that $N < \nu$. Table 2 gives the expected detection delay $E(N - \nu)^+$ in this Bayesian model; each result is based on 10,000 simulations and is represented as mean $\pm$ standard error. Also given for comparison are the following oracle-type rules that we use as benchmarks.

To begin with, consider the *oracle rule* that discloses the values of θ_0 and θ_1 once they are generated from the Bayesian model so that Shiryaev's rule (1.3) can be used, with γ in (1.3) to be chosen that $P\{N(\gamma) < \nu\} = \alpha$, in which P is the probability in the Bayesian data-generating mechanism for ν, θ_0, and θ_1. Table 2 shows that this oracle rule has markedly shorter detection delays than (2.6), indicating that there is considerable cost due to ignorance of θ_0 and θ_1 for detection of the occurence of ν. Since the normal model is translation invariant, we next consider a *semi-oracle rule* that discloses only the value of $\delta := \theta_1 - \theta_0$ once θ_0 and θ_1 are generated from the Bayesian model. This rule is based on $Y_i := X_i - X_1$ having known pre-change mean 0 and known post-change mean δ. Since the Y_i are no longer independent, the optimal stopping problem is not stationary Markov and the Shiryaev-type rule that stops when the posterior probability $P(\nu \leq n \mid \mathcal{F}_n)$ exceeds a threshold is no longer optimal. The rule (1.14) can be used as an approximation to the Bayes rule in this

Table 2. Expected detection delays in Bayesian model

p	α	Oracle	Semi-oracle	Rule (2.6)
0.005	0.003	118.5 ± 2.5	277.0 ± 4.5 (9.88%)	307.2 ± 4.7
	0.002	126.7 ± 2.7	283.6 ± 4.8 (10.07%)	347.1 ± 5.3
	0.001	141.7 ± 3.1	317.4 ± 5.6 (11.42%)	371.3 ± 5.6
0.01	0.007	66.6 ± 1.2	157.7 ± 3.2 (12.16%)	169.7 ± 2.2
	0.005	74.4 ± 1.3	176.1 ± 3.9 (13.23%)	182.7 ± 2.3
	0.003	83.7 ± 1.5	194.6 ± 4.4 (14.49%)	217.2 ± 2.7
0.02	0.015	35.3 ± 0.5	93.3 ± 2.6 (14.65%)	95.4 ± 1.0
	0.01	40.8 ± 0.6	102.5 ± 2.9 (15.38%)	108.8 ± 1.1
	0.005	51.8 ± 0.8	122.1 ± 3.5 (17.35%)	123.7 ± 1.3

case of known δ, and it can be shown by modifying the arguments of Pollak (1985) and Pollak and Siegmund (1991) that (1.14) is Bayes risk efficient as $p \to 0$. The semi-oracle rule, therefore, is the rule $\widetilde{N}_\delta$ defined by (1.14) in which the threshold γ is chosen such that $P\{\widetilde{N}_\delta < v\} = \alpha$ and n_0 is chosen to be 10.

The simulation study of $\widetilde{N}_\delta$ reveals a difficulty that we have not encountered in generating the results of Table 2 for (2.6) or the oracle rule. Although most of the time $\widetilde{N}_\delta$ is of manageable size, there are instances where the program keeps running without stopping to give the value of $(\widetilde{N}_\delta - v)^+$ in a single simulation. We therefore impose an upper bound $8000 + v$ on the maximum length of a simulated sequence $X_1, X_2, \ldots, X_{\widetilde{N}_\delta \wedge (8000+v)}$ and count the proportion of times in the 10,000 simulations that $\widetilde{N}_\delta \geq 8000 + v$. Table 2 gives this proportion (in %) in parentheses; also given are $E((\widetilde{N}_\delta - v)^+ \mid \widetilde{N}_\delta < 8000 + v)$ and the corresponding standard error. The table shows that (2.6) compares favorably with this semi-oracle rule. It also demonstrates the advantage of using $\sum_{i=n-k_p}^{n}$ in (2.6) (instead of $\sum_{i=n_0}^{n}$ in (1.14)) because the sliding window generates a "block-geometric" stopping time whose distribution has an exponential tail, whereas $\sum_{i=n_0}^{n}$ requires a considerably larger threshold that results in a very long detection delay when δ is small.

4.3. Implementation in General Exponential Families

The preceding simulation studies have exploited the translation invariance in the normal family to reduce the detection problem to that with known pre-change mean 0 for $Y_i = X_i - X_1$. This technique is no longer applicable in general exponential families, for which we propose to use the training sample $\{X_1, \ldots, X_{n_0}\}$ to estimate the baseline parameter $\boldsymbol{\theta}_0$, yielding the maximum likelihood estimate $\hat{\boldsymbol{\theta}}_0$. Thus we simulate the false alarm rate (3.1) with $\boldsymbol{\theta}_0$ replaced by $\hat{\boldsymbol{\theta}}_0$; this is basically the parametric bootstrap approach. It can be shown by modifying the arguments of Chan and Lai (2003, Theorem 6) that the bootstrap estimate of the false alarm rate has asymptotically the same order of magnitude as the actual false alarm rate when $n_0 \geq (\xi + o(1))|\log \alpha|$ for some $\xi > 0$. Since Mei (2008) has proposed to put a prior distribution on $\boldsymbol{\theta}_0$ in evaluating the false alarm detection rate or ARL, we can include his suggestion in our approach by sampling $\boldsymbol{\theta}_0$ from its posterior distribution given $\mathbf{X}_1, \ldots, \mathbf{X}_{n_0}$. In this way we can also incorporate the uncertainties in the estimate of $\boldsymbol{\theta}_0$ based on $\mathbf{X}_1, \ldots, \mathbf{X}_{n_0}$.

5. CONCLUDING REMARKS

The Bayesian model in Section 2 is an obvious extension of Shiryaev's model to the more general setting in which the pre- and post-change parameters are unknown, by putting conjugate priors on these unknown parameters of an exponential family. Although explicit formulas for the posterior distributions are available in this Bayesian model, the optimal stopping problem associated with the Bayes rule in this setting becomes much more difficult than Shiryaev's problem. On the other hand, the sliding window idea introduced in Lai (1995) can be used to modify Shiryaev-type rules, as we have done in Section 2, thereby providing an asymptotically optimal solution to the Bayes problem in Section 3.2. In Lai et al. (2009) we have further extended the Bayesian model from a single change-time ν to a sequence of change-times $\nu_1, \nu_2, \ldots$ with independent and geometrically distributed increments that have a common parameter p, thereby extending the sequential changepoint detection rule to a sequential surveillance rule. The scenario of unknown pre- and post-change parameters is of common occurence in surveillance applications, whereas the baseline (in-control) parameter value is usually specified in quality control or fault detection.

ACKNOWLEDGMENTS

Lai's research is supported by the National Science Foundation under grant DMS-0805879. Xing's research is supported by the National Science Foundation under grant DMS-0906593.

REFERENCES

Chan, H. P. and Lai, T. L. (2003). Saddlepoint Approximations and Nonlinear Boundary Crossing Probabilities of Markov Random Walks, *Annals of Applied Probability* 13: 395–429.

Diaconis, P. and Ylvisaker, D. (1979). Conjugate Priors for Exponential Families, *Annals of Statistics* 7: 269–281.

Lai, T. L. (1995). Sequential Changepoint Detection in Quality Control and Dynamical Systems, *Journal of Royal Statistical Society, Series B* 57: 613–658.

Lai, T. L. (1998). Information Bounds and Quick Detection of Parameter Changes in Stochastic Systems, *IEEE Transactions on Information Theory* 44: 2917–2929.

Lai, T. L. (2000). Sequential Multiple Hypothesis Testing and Efficient Fault Detection-isolation in Stochastic Systems, *IEEE Transactions on Information Theory* 46: 595–608.

Lai, T. L. and Chan, H. P. (2008). Discussion on "Is Average Run Length to False Alarm Always an Information Criterion?" by Yajun Mei, *Sequential Analysis* 27: 385–388.

Lai, T. L. and Xing, H. (2008). A Simple Bayesian Approach to Multiple Change-Points, *Technical Report*, Department of Statistics, Stanford University.

Lai, T. L., Liu, T., and Xing, H. (2009). A Bayesian Approach to Sequential Surveillance in Exponential Families, *Communications in Statistics: Theory and Methods* (Special issue in honor of Shelemyahu Zacks) 38: 2958–2968.

Lorden, G. (1971). Procedures for Reacting to a Change in Distribution, *The Annals of Mathematical Statistics* 41: 520–527.

Lorden, G. and Pollak, M. (2008). Sequential Change-point Detection Procedures that are Nearly Optimal and Computationally Simple, *Sequential Analysis* 27: 476–512.

Mei, Y. (2006). Suboptimal Properties of Page's CUSUM and Shiryaev-Roberts Procedures in Change-Point Problems with Dependent Observations, *Statistica Sinica* 16: 883–897.

Mei, Y. (2008). Is Average Run Length to False Alarm Always an Information Criterion? *Sequential Analysis* 27: 351–376.

Page, E. S. (1954). Continuous Inspection Schemes, *Biometrika* 41: 100–114.

Pollak, M. (1985). Optimal Detection of a Change in Distribution, *Annals of Statistics* 18: 1464–1469.

Pollak, M. and Siegmund, D. (1991). Sequential Detection of a Change in a Normal Mean When the Initial Value is Unknown, *Annals of Statistics* 19: 394–416.

Roberts, S. W. (1966). A Comparison of Some Control Chart Procedures, *Technometrics* 8: 411–430.

Shiryaev, A. N. (1963). On Optimum Methods in Quickest Detection Problems, *Theory of Probability and Its Applications* 8: 22–46.

Shiryaev, A. N. (1978). *Optimal Stopping Rules*, New York: Springer-Verlag.

Siegmund, D. and Venkatraman, E. S. (1995). Using the Generalized Livelihood Ratio Statistic for Sequential Detection of a Change-Point, *Annals of Statistics* 23: 255–271.

Tartakovsky, A. G. (2005). Asymptotic Performance of a Multichart CUSUM Test Under False Alarm Probability Constraint, in *Proceedings of the 44th IEEE Conference on Decision and Control and the European Control Conference*, pp. 320–325, Seville, Spain, December 12–15.

Tartakovsky, A. G. (2008). Discussion on "Is Average Run Length to False Alarm Always an Informative Criterion?" by Yajun Mei, *Sequential Analysis* 27: 396–405.

Woodroofe, M. (1978). Large Deviations of Likelihood Ratio Statistics with Applications to Sequential Testing, *Annals of Statistics* 6: 72–84.

Zacks, S. (1991). Detection and Change-Point Problems, in *Handbook of Sequential Analysis*, B. K. Ghosh and P. K. San, eds., pp. 531–562, New York: Dekker.

Zacks, S. and Barzily, Z. (1981). Bayes Procedures for Detecting a Shift in the Probability of Success in a Series of Bernoulli Trials, *Journal of Statistical Planning and Inference* 5: 107–119.