Discussion on "Quickest Detection Problems: Fifty Years Later" by Albert N. Shiryaev

Michael Baron and Sam Efromovich
Department of Mathematical Sciences, University of Texas at Dallas, Richardson, Texas, USA

Abstract: Professor Shiryaev gave an excellent overview of the modern theory and applications of quickest detection problem. This discussion contains possible extensions based on Bayesian and nonparametric curve estimation approaches. Asymptotically pointwise optimal solution, detection of changes in mixing proportions, challenges of settings in the presence of auxiliary variables, and the use of wavelets are discussed.

Keywords: Asymptotically pointwise optimal; Bayesian solution; Change point; Mixture regression; Wavelet.

Subject Classifications: 62G05; 62G07; 62L10; 60G35.

1. INTRODUCTION

Professor Shiryaev is congratulated for his beautiful paper on quickest detection problems with a comprehensive historical overview and further directions of the research. It nicely blends the theory of detection with important applications. We were truly impressed by his approaches to very complicated settings of optimal methods of detection.

The paper is also extremely stimulating. In particular, while reading it, we began to think about possible connections with other settings that may be of interest for the detection theory. In what follows we briefly describe possible topics of interest.

2. A NEARLY BAYES SOLUTION, APO STOPPING RULES

Optimal change-point detection has brought a number of challenges for statisticians, and many of them concern practical forms of the Bayes problem described by Professor Shiryaev in Section 6. In practice, many changes occur unexpectedly, at random times, being caused by a variety of factors. Thus, objectively, the change-point parameter θ often has its own prior distribution that is *governed by a particular application*. On the other hand, the theory of optimal stopping rules requires a *specific form*, namely, memoryless prior distribution like (6.1) in order to guarantee Markovian structure of the posterior distribution of θ.

Development of optimal stopping rules that can take into account any, even rather complex, prior information about the change-point parameter remains an interesting open challenge.

An approximate solution to this problem can be developed based on the fundamental idea of *asymptotically pointwise optimal* (APO) stopping rules proposed by Bickel and Yahav (1968).

Address correspondence to M. Baron, Department of Mathematical Sciences, University of Texas at Dallas, EC-35, Richardson, TX 75080, USA; Fax: 972-883-6622; E-mail: mbaron@utdallas.edu

For classical sequential statistical problems, an APO rule T is defined as a stopping rule satisfying the limiting inequality,

$$\limsup_{c\downarrow 0} \frac{\boldsymbol{E}_X L(\delta(T; X_1, \ldots, X_T), \theta) + cT}{\boldsymbol{E}_X L(\delta(U; X_1, \ldots, X_U), \theta) + cU} \leq 1, \tag{2.1}$$

almost surely with respect to the distribution of $X_1, X_2, \ldots$, where $L = L(\delta, \theta)$ is a loss function of the parameter θ and terminal decision rule δ, c is (presumably constant) cost of each observation, and both expectations are computed under the posterior distribution of θ.

Certainly, the APO idea yields only a nearly Bayes stopping rule that becomes optimal only when the cost of each single observation is small comparing with the magnitude of the loss.

The advantage of APO rules is that they have a clear explicit expression. Bickel and Yahav prove that the stopping rule

$$T^* = \min \left\{ n \mid n^{-1} \boldsymbol{E}_X \left\{ L(\delta, \theta) + cn \right\} \leq \frac{c}{\beta} \right\}$$

is APO, a clear and simple formula, where β is such that $n^\beta \boldsymbol{E}_X L(\delta(X_1, \ldots, X_n), \theta) \to l \in (0, \infty)$, as $n \to \infty$.

At the same time, exact Bayes rules are difficult to find in general, and even when the algorithm exists, its actual implementation often involves numerical computations that will inevitably end up with some approximation error. In particular, the Bayes stopping rule for the change-point detection in Chapter 4 of Shiryaev (1978) requires computation of the payoff function. The latter can be computed by solving a fixed-point integral equation that can be done on a lattice, through a series of fixed-point iterations.

Essentially, the choice is between the exact Bayes stopping rule that can be computed approximately and the approximately Bayes stopping rule that is simple enough to be computed exactly.

The idea of APO rules can be extended to the problem of quickest change-point detection. However, results of Bickel and Yahav cannot be applied directly because of the peculiar nature of change-point analysis that denies the constant cost c. Detection delays are costly, therefore, the cost of *pre-change* data is deemed almost negligible comparing with the cost of *after-change* data, until the change is detected and the process is stopped. Thus, cost c is now replaced by the delay term, and the rule of the loss function $L(\theta, \delta)$ is played by the probability of the false alarm.

An *APO stopping rule for the change-point detection* is then defined by analogy with (2.1) as such a rule that satisfies the limiting inequality

$$\limsup_{\lambda\downarrow 0} \frac{c\boldsymbol{E}_X (T-\nu)^+ - \log^{-1} \boldsymbol{P}_X \left\{ T < \nu \right\}}{c\boldsymbol{E}_X (U-\nu)^+ - \log^{-1} \boldsymbol{P}_X \left\{ U < \nu \right\}} \leq 1, \tag{2.2}$$

a.s., for any stopping rule U.

Noticeably, the form of the risk function

$$R(T) = c\boldsymbol{E}(T-\nu)^+ - \log^{-1} \boldsymbol{P} \left\{ T < \nu \right\}, \tag{2.3}$$

implied in (2.2), differs from the risk function (6.7) of Professor Shiryaev's article, where the expected delay and the probability of a false alarm enter as a linear combination. Our choice of $R(T)$ is dictated by the asymptotically linear relation between the mean delay and the *logarithm* of the probability of a false alarm, as $c \to 0$, causing $T \to \infty$.

Under risk (2.3), the explicit form of an APO rule for the change-point detection is given by

$$T_{\text{APO}} = \inf \left\{ n \mid -n \log S_X(n) \geq \beta / c \right\}, \tag{2.4}$$

where $S_X(t) = \boldsymbol{P}\{\theta > t \mid X_1, \ldots, X_t\}$ is the posterior survival function of the change point θ, and $\beta \geq 1$ is such that $t^{-\beta}|\log S(t)| \to l \in (0, \infty)$ as $t \to \infty$, for the prior survival function $S(t) = \boldsymbol{P}\{\theta > t\}$, or $\beta = 1$ if $t^{-1}|\log S(t)| \to 0$ (Baron, 2004).

Thus, we have an asymptotically optimal solution for the quickest detection that can serve virtually any arbitrary prior distribution of the change point.

3. CHANGES IN MIXING PROPORTIONS

The following quickest detection problem often arises in quality and process control. Suppose there are two stochastic processes U_t and V_t that govern the observed data sequence. For example, U_t is the in-control process resulting in an adequate quality product, and V_t is the out-of-control process resulting in nonconforming product or scrap. The distributions F and G of U_t and V_t may be partially or completely unknown.

Under "normal" circumstances, each observation comes from the process U_t with probability η and from V_t with probability $(1 - \eta)$, until a change occurs. This change will imply a suddenly higher or lower proportion of data coming from the nonconforming process. The quickest detection of such a change is a problem of practical importance.

An APO solution for this problem can be calculated by (2.4). Assuming for a moment that the mixing proportion can only change from value η_0 to η_1, and that F and G have known densities f and g with respect to some reference measure, we calculate the likelihood ratios

$$\rho_1 = \frac{\eta_1 f(X_1) + (1-\eta_1)g(X_1)}{\eta_0 f(X_1) + (1-\eta_0)g(X_1)}, \; \rho_{n+1} = \frac{\eta_1 f(X_{n+1} \mid X_1, \ldots, X_n) + (1-\eta_1)g(X_{n+1} \mid X_1, \ldots, X_n)}{\eta_0 f(X_{n+1} \mid X_1, \ldots, X_n) + (1-\eta_0)g(X_{n+1} \mid X_1, \ldots, X_n)}.$$

Then we compute the posterior survival function $S_X(t) = S(t)(\sum_{k \leq t} \pi_k \rho_{k+1} \cdots \rho_t + S(t))^{-1}$, and an APO stopping rule T_{APO} is given by (2.4).

It is more realistic, however, not to assume completely known pre- and after-change mixing proportions $\eta_{0,1}$ and densities f, g. Rather, we can assume prior distributions of η before and after the change, as well as prior distributions on possible nuisance parameters of f and g. This problem is equivalent to the change-point detection between *marginal* distributions of data before and after the change point. The APO rule will again be given by (2.4), with S_X replaced by the marginal (parameter-free) posterior survival function of the change point (Baron, 2004).

The problem has many interesting and challenging extensions. A change in the mixing proportion can occur more the once in the production process. The process may oscillate between two regimes or change to a new regime every time. Changes can occur abruptly or gradually, when at an unexpected moment the mixing proportion η starts slipping off its "normal" value along a smooth curve whose estimation may also be useful. Then η can either stabilize at a new value after a transition period (which is to be estimated) or continue the skid, and it may be important to differentiate between the two situations.

4. MIXTURES REGRESSION

Now let us consider a specific setting and shed light on the nonparametric curve estimation component of the problem. Suppose that one observes $Z^\tau := (Z_1, Z_2, \ldots, Z_\tau)$, $Z_l = (Y_l, X_l)$ where $Y_l := W_l \zeta_l + (1 - W_l)\eta_l$ and τ is a stopping time. The problem is to estimate a function $f(x) := \boldsymbol{E}\{W|X = x\}$. For simplicity it is assumed that W, η, ζ are independent and their iid sample generates the observations, and that W is Bernoulli. If $f(x)$ is continuous then this is a *gradual change problem.* Further, deterministic and increasing X_l correspond to a time series setting, and a random X describes the case of a casual regression important in many applications. Let us explain what is known about the solution of this problem. Suppose that mean values of ζ and η are known and different (a typical situation in a change-point setting), variances are finite and nothing else is known about underlying distributions. Efromovich (1999, s.4.8) suggested an ad hoc adaptive estimator $\hat{f}(x, Z^n)$ which is asymptotically rate optimal. Namely, for α-fold differentiable f its mean integrated squared error (MISE) decreases proportionally to $n^{-2\alpha/(2\alpha+1)}$, and this rate cannot be improved by any sequential estimator-oracle whose stopping time satisfies a restriction $\boldsymbol{E}\{\tau/n\} \leq 1$. On the other hand, if the problem is to estimate with a fixed MISE then a sequential estimator dominates a fixed-sample estimator; see Efromovich (1999, 2003, 2008). If densities p_ζ and p_η of ζ and η are known and sufficiently smooth then a sharp minimax estimation of f is possible. In particular, a sharp constant depends on the following coefficient of difficulty, $d = \int[\pi(x)I(x)]^{-1}dx$ where $\pi(x)$ is the density of X and $I(x) = \boldsymbol{E}\{(p_\zeta(Y) - p_\eta(Y))^2[f(x)p_\zeta(Y) + (1 - f(x))p_\eta(Y)]^{-2}\}$. A coefficient of difficulty shows how complicated a nonparametric problem is; see a discussion in Efromovich (1999). Further, as usual it is possible to construct a confidence interval for the minimax estimator if f is analytic (infinitely differentiable) which is a reasonable assumption in gradual changes.

An interesting and open problem is optimal estimation in the case of unknown distributions of ζ and η; this setting has been discussed earlier. Another interesting extension is the case of missing data that may be imputed.

5. AUXILIARY COVARIATES

As it is nicely shown by Professor Shiryaev, a large noise may significantly slow down a detection process. Sometimes auxiliary (lurking) covariates, via modeling the noise, can be helpful in noise reduction.

This approach was undertaken in Efromovich (2005) where in a regression setting $Y = f(X) + \epsilon$, $f(X) = \boldsymbol{E}\{Y|X\}$ the noise ϵ was modeled using auxiliary covariates. Namely, if $\mathbf{Z}$ is an auxiliary vector-covariate then $\epsilon = g(X, \mathbf{Z}) + \eta$ where $g(X, \mathbf{Z}) = \boldsymbol{E}\{Y|X, \mathbf{Z}\} - \boldsymbol{E}\{Y|X\}$. The latter yields $\mathrm{Var}(\eta) \leq \mathrm{Var}(\epsilon)$ with the equality if $g(X, \mathbf{Z}) = 0$ almost surely. This inequality justifies the approach. It is shown that under a mild assumption an estimator can match performance of an oracle that knows the nuisance function $g(X, \mathbf{Z})$. Further, an industial example shows that this approach is feasible for small data sets.

An interesting modification of the problem is the optimal sequential estimation in controlled regression settings discussed in Efromovich (2008).

We believe that this nonparametric approach may be of a special interest for quickest detection problems, specifically in biomedical and economics applications where noise is always large and nuisance/lurking covariates are common.

6. WAVELETS

Feasible solution of quickest detection problems is not just about theory, it is also about implementation of new and exciting tools developed by mathematicians. Here we would like to mention exciting possibilities that wavelets and multiwavelets can produce in recovering underlying signals and their derivatives.

Different settings and applications are discussed in Efromovich et al. (2004) and Efromovich (2009).

They include statistical analysis of ChIP-on-Chip microarrays, which allows to shed a new light on interactions between proteins and DNA. Here a rapid change (a peak) in a noisy signal indicates a possible place where protein binds to sites in the promoter of genes. Another specific application is the analysis of ultra-fast functional magnetic resonance images which allow statisticians to analyze changes in brain activities within 50 msc time intervals! Using special statistical procedures allows even to visualize spikes in neurons activity that last for just several msc.

It is well expected that combination of new wavelet techniques with theoretical results and new research areas, outlined by Professor Shiryaev, can lead to feasible practical solutions in vast areas of statistical applications.

ACKNOWLEDGMENTS

Research of M. Baron is supported by NSF Grant DMS-1007775. Research of S. Efromovich is supported in part by NSF Grants DMS-0604558 and DMS-0906790.

REFERENCES

Baron, M. (2004). Detection of Epidemics As a Sequential Change-Point Problem, in *Longevity, Aging and Degradation Models in Reliability, Public Health, Medicine and Biology*, volume 2, V. Antonov, C. Huber, M. Nikulin, and V. Polischook, eds., pp. 31–43, St. Petersburg: Polytechnical University.

Bickel, P. J., and Yahav, J. A. (1968). Asymptotically Optimal Bayes and Minimax Procedures in Sequential Estimation, *Annals of Mathematical Statistics* 39: 442–456.

Efromovich, S. (1999). *Nonparametric Curve Estimation: Methods, Theory, and Applications*, New York: Springer.

Efromovich, S. (2004). Financial Applications of Sequential Nonparametric Curve Estimation, in *Applications of Sequential Methodologies*, N. Mukhopadhyay, S. Datta, and S. Chattopadhyay, eds., pp. 171–192, New York: Marcel Dekker.

Efromovich, S. (2005). Univariate Nonparametric Regression in the Presence of Auxiliary Covariates, *Journal of American Statistical Association* 100: 1185–1201.

Efromovich, S. (2008). Optimal Sequential Design in a Controlled Nonparametric Regression, *Scandinavian Journal of Statistics* 35: 266–285.

Efromovich, S. (2009). Multiwavelets: Theory and Bioinformatic Applications, *Communications in Statistics - Theory & Methods* 38: 2829–2842.

Efromovich, S., Lakey, J., Pereyra, M. C., and Tymes, N. (2004). Data-Driven and Optimal Denoising of a Signal and Recovery of Its Derivative Using Multiwavelets, *IEEE Transactions on Signal Processing* 52: 628–635.

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