The Shiryaev-Roberts Changepoint Detection Procedure in Retrospect - Theory and Practice

Moshe Pollak

Department of Statistics
The Hebrew University of Jerusalem
Mount Scopus
91905 Jerusalem, Israel
msmp@mscc.huji.ac.il

Abstract. A retrospective view of the Shiryaev-Roberts procedure is presented, placing it in a historical context and describing its evolution into a powerful changepoint detection method.

Keywords. changepoint detection, quality control, sequential analysis, Shiryaev-Roberts, spc.

1 Introduction

The ingredients of the changepoint detection problem are a sequence of observations $\{X_i\}$ whose baseline distribution has a density f_0 that may change to a density f_1. The changepoint ν (the serial number of the first post-change observation) is unknown, and can take place at any $1 \leq \nu < \infty$. ($\nu = \infty$ denotes the case where a change never takes place.) A detection scheme is characterized by a stopping time N, at which an alarm is raised (a change is declared to have taken place).

The motivation of attempts to deal with this problem can be classified as either classical or Bayesian. Classical approaches define operating characteristics for false alarm rates and for speed of detection. Bayesian approaches assume a prior distribution on the changepoint, and attempt to minimize the expected value of a loss function. As is often the case in statistics in general, the twain do meet.

2 Some history

The first attempt to deal with the changepoint problem formally is due to Shewhart (1931), who was primarily interested in detecting a shift in the mean of a normal distribution. His frame of reference is a series of independent variables with known baseline distribution. He proposed (ad hoc; it seems to have been satisfactory for the industrial applications of his time) raising an alarm the first time that an observation exceeds the (known) baseline mean by more than three standard deviations. His perspective is classical; ARL's (average run lengths, until a false alarm, and from change to its detection) are the operating characteristics of the method. The method is known to be very good in detecting a large change quickly.

During the next decade, the fact that the Shewhart procedure does not enable information to accumulate brought about ad hoc attempts to correct for this (such as "warning lines" and "action lines", where too many proximate observations in exceedence of a warning line would also be cause for alarm). The perspective of these methods, too, is classical.

The first Bayesian consideration of the problem is due to Girschick and Rubin (1952). They assumed the observations to be independent and f_0 and f_1 to be known. They posited a geometric prior on the changepoint and a gain (or loss) function for each observation. Their objective was to maximize expected gain per observation. Their solution calls for raising an alarm whenever the posterior probability of a change having taken place is large enough. This procedure is a precursor of the Shiryaev-Roberts procedure. With the same probabilistic structure, Shiryaev (1963, 1978) considered minimizing $E(N - \nu | N \geq \nu)$ when each post-change observation costs $c > 0$ units and the penalty for a false alarm is 1 unit. His solution is the same.

The next development was classical: Page (1954) proposed the Cusum scheme, which in essence is a repeated sequence of SPRT's defined by f_0 and f_1 that calls for an alarm the first time that a

SPRT exits on the side of f_1. The ARL to false alarm and the average delay to detection are the relevant operating characteristics. The observations are assumed to be independent, f_0 is assumed known, and although the post-change distribution need not be known to implement the procedure, it is necessary to represent it by a (fixed) density f_1 in order to spell out the SPRT. The definition was ad hoc. Lorden (1971) proved that the minimum over all stopping times N with ARL to false alarm $\geq B$ of $sup_{1\leq k<\infty}$ *ess sup* $E_{\nu=k}(N-k|X_1,X_2,\ldots,X_{k-1})$ is $(1+(o(1))\times(logB)/I$, where I is the Kullback-Leibler information number (of the post-change density vs. the pre-change density) and $o(1)\rightarrow 0$ as $B\rightarrow\infty$, and showed that the Cusum scheme achieves this asymptotic lower limit. Moustakides (1986) went the last leg and proved that (when f_1 is the true post-change density) the appropriate Cusum scheme is strictly optimal in minimizing $sup_{1\leq k<\infty}$ *ess sup* $E_{\nu=k}(N-k|X_1,X_2,\ldots,X_{k-1})$ over all stopping times N with ARL to false alarm $\geq B$. In this context, Ritov (1990) proved the optimality of the Cusum in a game-theoretic setup, with Nature and the statistician being opposing players. Beibel (1996) proved the same in a Brownian motion context.

Roberts (1959) proposed the exponentially weighted moving average (EWMA) method. His approach is classical (motivated by time series). Srivastava and Wu (1993) found this method to be inferior to others.

In a Brownian motion context of detecting a shift in mean, Shiryaev (1961, 1963) considered the problem of detecting an object with the aim of minimizing expected delay (from change to detection), asymptotically after a long run of false alarms raised by successive application of a stopping time N, under the constraint that the ARL to false alarm (in a single application of N) be $\geq B$. He found that the optimal procedure is analogous to that of Girschick and Rubin's when the parameter of the geometric prior tends to zero. (A discrete time analog of Shiryaev's result was derived by Pollak and Tartakovsky, 2009.) Independently, Roberts (1966) was the first to consider this limit in the context of reducing expected delay to detection (of a single application of the stopping time) subject to a lower bound on the ARL to false alarm. Roberts studied the procedure by simulation, comparing it to other procedures (Cusum, EWMA and others), and found it to be good. The procedure is now known as the Shiryaev-Roberts procedure.

Pollak (1985) also considered the changepoint problem in a classical framework. The conditions considered are that the observations are independent, with known f_0 and f_1, and the goal is to minimize $sup_{1\leq k<\infty}E_{\nu=k}(N-k|N\geq k)$ subject to ARL to false alarm $\geq B$. He found that that a method based on starting the Shiryaev-Roberts procedure at a random value is optimal to within an additive $o(1)$ term, with $o(1)\rightarrow 0$ as $B\rightarrow\infty$. The method of proof is Bayesian; he took Shiryaev's (1978) solution of the Bayesian problem a step further, showing that the aforementioned procedure is a limit of Bayes rules. The question of whether this procedure is strictly optimal was open until very recently; Polunchenko and Tartakovsky (2009) produced a counterexample.

3 Generalization

The basic Shiryaev-Roberts statistic and stopping time are respectively

$$R_n=\sum_{k=1}^{n}\prod_{i=k}^{n}\frac{f_1(X_i)}{f_0(X_i)} \qquad \text{and} \qquad N_A=min\{n|R_n\geq A\}$$

where A is a threshold value tuned to satisfy ARL to false alarm $\geq B$ (actually, $=B$), where B is a lower bound on the acceptable rate of false alarms. A generalization of this procedure is to define

$$R_n=\sum_{k=1}^{n}\frac{f_{\nu=k}(X_1,X_2,\ldots,X_n)}{f_{\nu=\infty}(X_1,X_2,\ldots,X_n)} \qquad \text{and} \qquad N_A=min\{n|R_n\geq A\}$$

where $f_{\nu=\infty}$ is the joint density of the observations when no change ever takes place and $f_{\nu=k}$ is a joint density of the observations when $\nu=k$ and the first $k-1$ observations are distributed as they would be under the regime dictated by f_∞. The observations need not be independent.

As mentioned above, optimality properties of the basic Shiryaev-Roberts procedure (as well as that of Cusum) hinge on the true post-change density being f_1. Since f_1 is usually a representative of possible post-change densities while in practice the true post-change density is unknown (generally different from f_1), the regret for ignorance of the true post-change density is of order of magnitude $logB$ - a rather heavy price to pay. To boot, there are many situations where f_0 is unknown, and even a small misspecification of f_0 can result in the true ARL to false alarm being very different from the nominal one (van Dobben de Bruyn, 1968). This can be rectified sometimes by the generalized Shiryaev-Roberts statistic. For example, consider a case where observations are independent, with pre-change known density f_0 and post-change density f_θ, where the family $\{f_\theta\}$ is known but the actual value of θ is not. Positing (arbitrarily) a prior G on θ results in

$$f_{\nu=k}(X_1, X_2, \ldots, X_n) = f_0(X_1)f_0(X_2)\cdots f_0(X_{k-1}) \int f_\theta(X_k)f_\theta(X_{k+1})\cdots f_\theta(X_n)dG(\theta) \ .$$

To fix ideas, suppose one knows that the baseline distribution is standard normal and one monitors for a change of mean, assuming the standard deviation remains the same, and one chooses a standard normal prior on θ. Assuming $\nu = k$, when calculating the likelihood ratio of the observations up to time n the part $f_0(X_1)f_0(X_2)\cdots f_0(X_{k-1})$ in the numerator cancels out with the same in the denominator, so

$$R_n = \sum_{k=1}^{n} \frac{\int e^{-\frac{1}{2}\sum_{i=k}^{n}(X_i-\theta)^2}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\theta^2}d\theta}{e^{-\frac{1}{2}\sum_{i=k}^{n}X_i^2}} = \sum_{k=1}^{n}\int e^{\sum_{i=k}^{n}\theta X_i-\frac{1}{2}(n-k+1)\theta^2}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\theta^2}d\theta = \sum_{k=1}^{n}\frac{e^{\frac{1}{2}\frac{\sum_{i=k}^{n}X_i^2}{n-k+2}}}{\sqrt{n-k+2}} \ .$$

An alternative to utilizing a prior is to use an estimate of the post-change parameter for obtaining a likelihood ratio for the next observation. For instance, in the previous example, the likelihood ratio of observation X_{n+1} is $e^{\theta X_{n+1}-\frac{1}{2}\theta^2}$. When putatively $\nu = k$, an estimate of θ based on the first n observations is $\hat{\theta}_k^n = \frac{1}{n-k+1}\sum_{i=k}^{n}X_i$, so a Shiryaev-Roberts-type statistic could be defined as

$$R_n = \sum_{k=1}^{n} e^{\sum_{i=k}^{n}\hat{\theta}_k^{i-1}X_i-\frac{1}{2}(n-k+1)(\hat{\theta}_k^{i-1})^2}$$

(Lorden and Pollak, 2005. See also Siegmund and Venkatraman, 1995). This method is slightly less efficient than the former, but its implementation is usually much faster when utilizing a prior entails numerical integration.

The generalized Shiryaev-Roberts statistic can be applied sometimes even when f_0 is unknown (cf. Pollak and Siegmund, 1991). For example, suppose observations are independent, $N(\mu, \sigma^2)$ before change, $N(\mu + \delta\sigma, \eta^2\sigma^2)$ post-change, with $\mu, \delta, \eta, \sigma$ unknown. (This is a description of monitoring a process observed for the first time for a change in mean and/or variance, when normality and independence of the observations is deemed appropriate but nothing is known about any of the parameters.) The invariance structure can be exploited in the following way: for starters, choose arbitrary δ and η. Define $Z_i = \frac{X_i - X_1}{|X_2 - X_1|}$. Obviously, the distribution of $Z_1, Z_2, \ldots, Z_n$ does not depend on μ and σ. Although the Z_i's are dependent, their joint density is amenable to calculation. If there is no change, the distribution of $Z_1, Z_2, \ldots, Z_n$ is completely specified, and when $\nu = k$ it depends on δ and η only. Therefore, if one regards δ, η as representatives of the post-change parameters, one can calculate the likelihood ratios that make up R_n if one bases surveillance on the sequence of Z's instead of the original X's. Furthermore, if taking a single representative pair is not palatable, then a prior G on δ, η can be chosen as above. The resulting formulae are complicated and integration with respect to G may have to be numerical, but with today's computing facilities all of this is feasible.

Whenever an invariance structure exists, surveillance based on a sequence of (maximal) invariants can be handled in a similar fashion. Detecting a change in the slope of a regression can be handled this way - all possible combinations of mean, standard deviation, slope - known baseline, partially known baseline, no parameters known - is amenable to this method (cf. Krieger et al., 2003, and references therein).

The generalization of the basic Shiryaev-Roberts procedure can even be applied to a nonparametric setup - observations are independent, with densities f_0 pre-change, f_1 post-change, both unknown (a situation similar to the previous one, without assumption of normality). A reasonable approach would be to base surveillance on the series of sequential ranks $\{Z_n\}$ (i.e., $Z_n = \sum_{i=1}^{n} 1(X_i \leq X_n)$, where $1(\cdot)$ is the indicator function). To construct a generalized Shiryaev-Roberts statistic, all that is needed are likelihood ratios for $Z_1, Z_2, \ldots, Z_n$. The denominator will always be $1/n!$. Although more problematic, the numerator can also be dealt with under certain assumptions (cf. Gordon and Pollak, 1995, and references therein).

When an invariance structure does not exist, the going can get to be quite complicated. See Siegmund and Venkatraman (1995) and references therein.

4 ARL to false alarm

Obviously, to implement a Shiryaev-Roberts procedure (basic or generalized), one has to connect between the desired ARL to false alarm B and the stopping threshold A. Exact expressions are hard to come by, but an asymptotic ($B \to \infty$) approximation of the form $A = B/const$ (as well as the value of $const$) is usually obtainable by renewal-theoretic considerations (Pollak, 1987, Yakir, 1995, 1998 and references therein). The most fruitful approach for proofs employs a change-of measure ploy due to Yakir (1995).

In addition, by the optional sampling theorem, the (simple to show) fact that when a change is not in effect $\{R_n - n\}$ is a martingale with zero expectation translates into $E_\infty(R_{N_A} - N_A) = 0$; since by definition $R_{N_A} \geq A$, this obtains the inequality $E_\infty N_A = E_\infty R_{N_A} \geq A$, meaning that simply setting $A = B$ satisfies (conservatively) ARL to false alarm $\geq B$.

5 Average delay to detection

When the baseline distribution of the X's is known, taking a continuous prior G on post-change parameter values obtains a procedure whose maximal average delay to detection is $\frac{logB+\frac{1}{2}loglogB}{I} + O(1)$ (Pollak, 1987), and up to an additive $O(1)$ term one cannot do better uniformly. Only the $O(1)$ term depends on G, meaning that asymptotically the prior almost "washes out", so that choosing a "comfortable" (appropriate) G is reasonable. There are more accurate asymptotic expressions for the average delay to detection, but they often contain constants that are obtainable only via simulation, and since by design the average delay to detection is hopefully small, asymptotics don't kick in quickly.

Note than in case the distribution of pre-change observations X_i's is unknown and invariance is invoked, the distribution of the (maximal) invariant when there is no change is the same as when the change is in effect from the very beginning. Hence, the maximal average delay to detection will be B. However, it can usually be shown for some constants $\eta_1 > 1, \eta_2 > 1$ that if $O((logB)^{\eta_1}) \leq \nu \leq B^{\eta_2}$ then the regret due to ignorance of the baseline distribution is minor.

6 Techniques of proofs

Most claims of optimality and asymptotic optimality are proved either by finding a lower bound for average delay to detection and showing that a method attains the bound or by Bayesian methods. Asymptotic formulae for ARL to false alarm and for average delay to detection are generally obtained by (nonlinear) renewal theory. Average delay to detection is directly amenable to renewal-theoretic methods, as the post-change distribution of the surveillance statistic has a positive drift. ARL to false alarm is more difficult, as generally the surveillance statistic does not have positive drift. The way around this is by a change-of-variable technique (Yakir, 1995); $dP_\infty = \sum_{k=1}^{j} \frac{dP_k}{R_j}$, enabling calculations under the regime dP_k, where the drift is positive.

7 Practical considerations

The basic Shiryaev-Roberts statistic admits the recursion $R_{n+1} = (R_n + 1)\frac{f_1(X_{n+1})}{f_0(X_{n+1})}$ with $R_0 = 0$, and is therefore easy to implement. If one is worried that a change may have a good chance to be in effect from the very beginning, starting things off at some $R_0 > 0$ enables a fast initial response (Moustakides et al., 2009. See also Lucas and Crosier, 1982).

The generalized Shiryaev-Roberts statistic generally does not admit a recursion, meaning that all past observations must be retained, and the longer the list the lengthier the computation time (the algorithm is of order n^2 at least, and if numerical integration is involved this may become a serious problem). An obvious fix is a form of truncation: usually, calculating the first 100-200 and the last 200-300 likelihood ratios and summing them is hardly different from summing them all.

Generally, the asymptotic constant $c = lim_{A\to\infty}\frac{E_\infty N_A}{A}$ is good enough for setting a threshold $A = B/c$ when B is the required lower bound on the rate of false alarms, even when B is relatively small ($B > 100$). Calculation of c is often possible by renewal theory. In any case, simulation invariably obtains a good approximation. If this is too tedious, due to the martingale structure of the generalized Shiryaev-Roberts statistic one can always set conservatively $A = B$.

Asymptotics aside, when B is of order magnitude $100 - 1000$, differences between schemes that apply representative values for post-change parameters often are not markedly different from the more complicated procedures. Also, differences between Shiryaev-Roberts and Cusum are not marked when B is large (after all, both have optimality properties in their own way).

While the basic Cusum procedure appears in many popular statistical computer packages (Shiryaev-Roberts doesn't (yet)), a Cusum analog of the generalized Shiryaev-Roberts procedure does not.

A data-analytic advantage of the Shiryaev-Roberts control chart over the standard Cusum chart is the linear connection between the ARL to false alarm B and the stopping threshold A, that (when plotting R_n by n) lends R_n the flavor of a p-value (Kenett and Pollak, 1996).

Although obviously important, the problem of estimation after detection has not received extensive attention so far (but see Wu, 2005, and Foster and George, 1993). The difficulty with estimation of post-change parameter values stems from the attempt to minimize the number of post-change observations, which obviously is detrimental for estimation purposes.

8 Concluding remarks

The Shiryaev-Roberts approach provides an extensive arsenal of tools for changepoint detection. Nonetheless, the most popular control chart is still Shewhart. In spite of its lesser efficiency, its simplicity makes it easy to explain to the uninitiated, and it does not require sophisticated computer programs. Next in line of popularity are the EWMA and Cusum basic procedures, which have the fore in terms of their early and wide publication, especially in applied journals. In this respect, Shiryaev-Roberts is a latecomer. Hopefully it will take its rightful place as a leading tool for changepoint detection.

9 Acknowledgments

This work was supported by a grant from the Israel Science Foundation and by the Marcy Bogen Chair of Statistics at the Hebrew University of Jerusalem.

References

Foster, D.P. and George, E.I. (1993). Estimation up to a change point. *The Annals of Statistics*, **21**, 625-644.

Girschick, M.A., and Rubin, H. (1952). A Bayes approach to a quality control model. *The Annals of Mathematical Statistics*, **23**, 114-125.

Gordon, L. and Pollak, M. (1995). A robust surveillance scheme for stochastically ordered alternatives. *The Annals of Statistics*, **23**, 1350-1375.

Kenett, R.S. and Pollak, M. (1996). Data-analytic aspects of the Shiryaev-Roberts control chart: surveillance of a non-homogeneous Poisson process. *The Journal of Applied Statistics* **23**, 125-138.

Krieger, A.M., Pollak, M. and Yakir, B. (2003). Surveillance of a simple linear regression. *The Journal of the American Statistical Association*, **98**, 1-15.

Lorden, G. (1971). Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, **42**, 1897-1908.

Lorden, G. and Pollak, M. (2005). Nonanticipating estimation applied to sequential analysis and changepoint detection. *The Annals of Statistics*, **33**, 1422-1454.

Lucas, J. M. and Crosier, R. B. (1982). Fast initial response for CUSUM quality-control schemes: give your CUSUM a head start. *Technometrics*, **24**, 199-205.

Moustakides, G. (1986). Optimal stopping times for detecting changes in distributions. *The Annals of Statistics*, **14**, 1379-1387.

Moustakides, G., Polunchenko, A. and Tartakovsky, A. G. (2009). A numerical approach to performance analysis of quickest change-point detection procedures. To appear in *Statistica Sinica*

Page, E.S. (1954). Continuous inspection schemes. *Biometrics*, **41**, 100-115.

Pollak, M. (1985). Optimal detection of a change in distribution. *The Annals of Statistics*, **13**, 206-227.

Pollak, M. (1987). Average run lengths of an optimal method of detecting a change in distribution. *The Annals of Statistics*, **15**, 749-779.

Pollak, M. and Siegmund, D. (1991). Sequential detection of a normal mean when the initial value is unknown. *The Annals of Statistics*, **19**, 394-416.

Pollak, M. and Tartakovsky, A.G. (2009). Optimality properties of the Shiryaev-Roberts procedure. *Statistica Sinica*, **19**, (in press).

Polunchenko, A. and Tartakovsky, A. G. (2009). *Personal communication*.

Ritov, Y. (1990). Decision-theoretic optimality of the CUSUM procedure. *The Annals of Statistics*, **18**, 1464-1469.

Roberts, S.W. (1959). Control charts tests based on geometric moving averages. *Technometrics*, **1**, 234250.

Roberts, S.W. (1966). A comparison of some control chart procedures. *Technometrics*, **8**, 411-430.

Shewhart, W.A. (1931). *Economic control of quality of manufactured product*. New York: D. Van Nostrand Company.

Shiryaev, A.N. (1963) (1961 in Russian). On optimum methods in quickest detection problems. *Theory of Probability and Its Applications*, **8**, 22-46.

Shiryaev, A.N. (1978). *Optimal Stopping Rules*. Springer Verlag.

Siegmund, D. and Venkatraman, E.S. (2005). Using the generalized likelihood ratio statistic for sequential detection of a change-point. *The Annals of Statistics*, **23**, 255-271.

van Dobben de Bruyn, C.S. (1968). *Cumulative sum tests: theory and practice*. Griffin.

Wu, Y. (2005). *Inference for Change-Point and Post-Change Means After a Cusum Test*. Springer.

Yakir, B. (1995). A note on the run length to false alarm of a change-point detection policy. *The Annals of Statistics*, **28**, 272-281.

Yakir, B. (1998). On the average run length to false alarm in surveillance problems which possess an invariance structure. *The Annals of Statistics*, **26**, 1198-1214.