One approach to the sequential change-point problem

G.Sofronov, T.Polushina

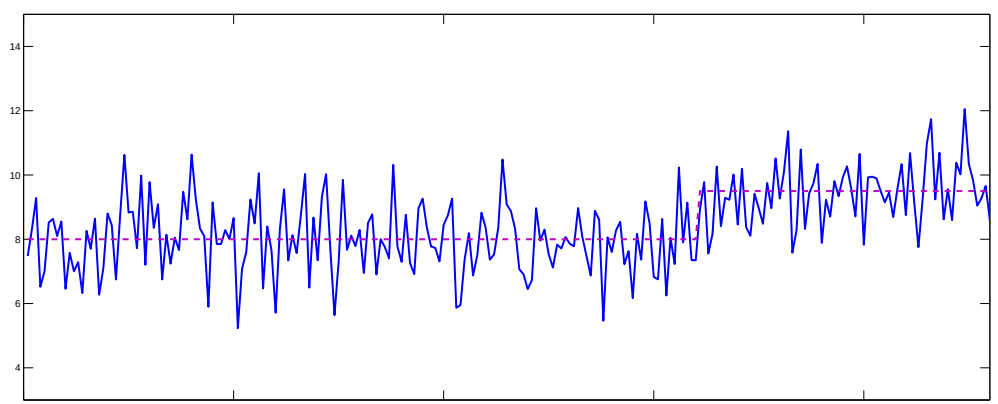
Macquarie University, NTNU

June 27, 2012

Introduction

A sequence $X_1, X_2, X_3, \ldots$ of random variables is observed sequentially, in a one-at-a-time.

We can make a decision based on previous history (stop, buy, repair and so on).



Some applications

- Medicine
  By using an alarm system we can detect the start of an epidemic (Andersson, 2003, Frisen, 1992)
- Environment
  We can analyse and control biodiversity (Barnett, Turkman, 1992, Pettersson, 1998)
  Testing climate regime shifts (Rodionov, 2004)
- Finance
  Early detection of crises in economics and financial systems (Shiryaev, 1999, 2002)
- Technology
  We can repair different breakages in technological processes as soon as possible

Statistical hypotheses

$X_1, X_2, ..., X_t$ — "in-control" with pdf $f_0(X_n|X_1, X_2, \ldots, X_{n-1})$

$X_{t+1}, \ldots, X_n$ — "out of control" with pdf
$f_1(X_n|X_1, X_2, ..., X_{n-1}) \neq f_0(X_n|X_1, X_2, ..., X_{n-1})$

$$H_1 : t = k \geq 0$$

a change occurs at time $t = k \geq 0$

$$H_0 : t = \infty$$

it means that there is never a change

Likelihood ratio

$$p(X_n|H_0) = \prod_{j=1}^{n} f_0(X_j|X_1, \ldots, X_{j-1}),$$

$$p(X_n|H_1) = \prod_{j=1}^{k} f_0(X_j|X_1, \ldots X_{j-1}) \prod_{j=k+1}^{n} f_1(X_j|X_1, \ldots, X_{j-1}), k < n.$$

Likelihood Ratio

$$LR = \prod_{j=k+1}^{n} \frac{f_1(X_n|X_1, \ldots X_{n-1})}{f_0(X_n|X_1, \ldots, X_{n-1})}.$$

Stopping time

A stopping moment with respect to $\{X_n\}, n \geq 1$ is an integer-valued r. v. T: $\forall n$, the event $\{T = n\}$ depends solely on the past observations $\{X_i\}$.

Evaluation:

- quick detection
- few false alarms

Minimal expected delay
for a fixed false alarm probability

Shiryaev-Roberts procedure

The Shiryaev-Roberts procedure is given by the stopping time

$$S_A = \inf\{n \geq 1 : R_n \geq A\},$$

where

$$R_n = (1 + R_{n-1})LR, \quad n = 1, 2, \ldots, \quad R_0 = 0$$

is the Shiryaev-Roberts statistic,

A is a positive threshold, which controls the false alarm rate.

Shiryaev A.N. The problem of the most rapid detection of a disturbance in a stationary process. Dokl. Math., 2. 795–799 (1961)
Roberts S.W. A comparison of some control chart procedures. Technomet., 8. 411–430 (1966)

CUSUM procedure

The stopping moment of the CUSUM procedure is defined by

$$C_A = \inf\{n \geq 1 : W_n \geq A\},$$

where

$$W_n = \max\{1, W_{n-1}\}LR, \quad n = 1, 2, \ldots, \quad W_0 = 1$$

is the CUSUM statistic,

A is a positive threshold, which controls the false alarm rate.

Page E. S. Continuous Inspection Scheme. Biometrika, 41. 100–115 (1954)

Cross-Entropy method

The Cross-Entropy Method allows to estimate rare event probabilities:

Estimate

$$\mathbf{P}(S(X) \geq \gamma)$$

X is random vector, S is real-valued function on X.

- Generate a random sample
- Update the parameters of the sampling distribution on the basis of the best scoring samples

Rubinstein R., Kroese D. The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning. Springer-Verlag, New York (2004)

Algorithm

- Let $v_0 = u$. Set $\tau = 1$.
- Generate a sample $x_1, ..., x_M$ from $g(\cdot; v_{\tau-1})$. Compute the sample $(1-\rho)$-quantile $\hat{\gamma}_\tau$. Calculate

$$\hat{s}_\tau = \frac{1}{M} \sum_{i=1}^{M} I\{S(x_i) \geq \hat{\gamma}_\tau\} W(x_i; u; v_{\tau-1}),$$

  provided this is greater than s, otherwise set $\hat{l}_\tau = l$.
- Use the same sample $x_1, \ldots, x_M$ to solve

$$\max_v \frac{1}{M} \sum_{i=1}^{M} I\{S(x_i)\} W(x_i, u, v_{\tau-1}) \ln g(x_i, v)$$

  Denote the solution as $\hat{v}_\tau$.
- If $\hat{l}_\tau = l$ proceed to the next step. Otherwise let $\tau = \tau + 1$ and return to step 2.

Last step

- (a) Let τ_{last} be the final iteration number. Generate a sample $x_1, \ldots, x_{N_{\text{last}}}$ from density $g(\cdot, \hat{v}_{\tau_{\text{last}}})$ and take as estimator of γ the smallest number $\hat{\gamma}$ such that

$$\frac{1}{N_{\text{last}}} \sum_{i=1}^{N_{\text{last}}} I\{S(x_i) \geq \gamma\} W(x_i, u, \hat{v}_{\tau_{\text{last}}}) \geq l.$$

- (b) Apply the bisection method.

Examples

$X_1, X_2, ...$ are independent and identically distributed (i.i.d.) from $\mathbf{N}(\theta_0, 1)$ before a change-point. $\theta_0 = 0$

After a breakage
θ_1 is the value of the mean of the normal distribution after change-point

- Example 1
  $\theta_1 = 1$

- Example 2
  θ_1 is estimated from previous observations $X_1, X_2, ..., X_n$

Parameters for the Cross-Entpory method

- $M = 1000$
- $\rho = 0.1$
- $N_{last} = 10000$
- $x \sim \mathbf{N}(10, 5)$ for an initial step
- N_1 is the number of simulations
- N_2 is the maximum number of steps

Example 1. CPU time for different simulation parameters for CUSUM procedure

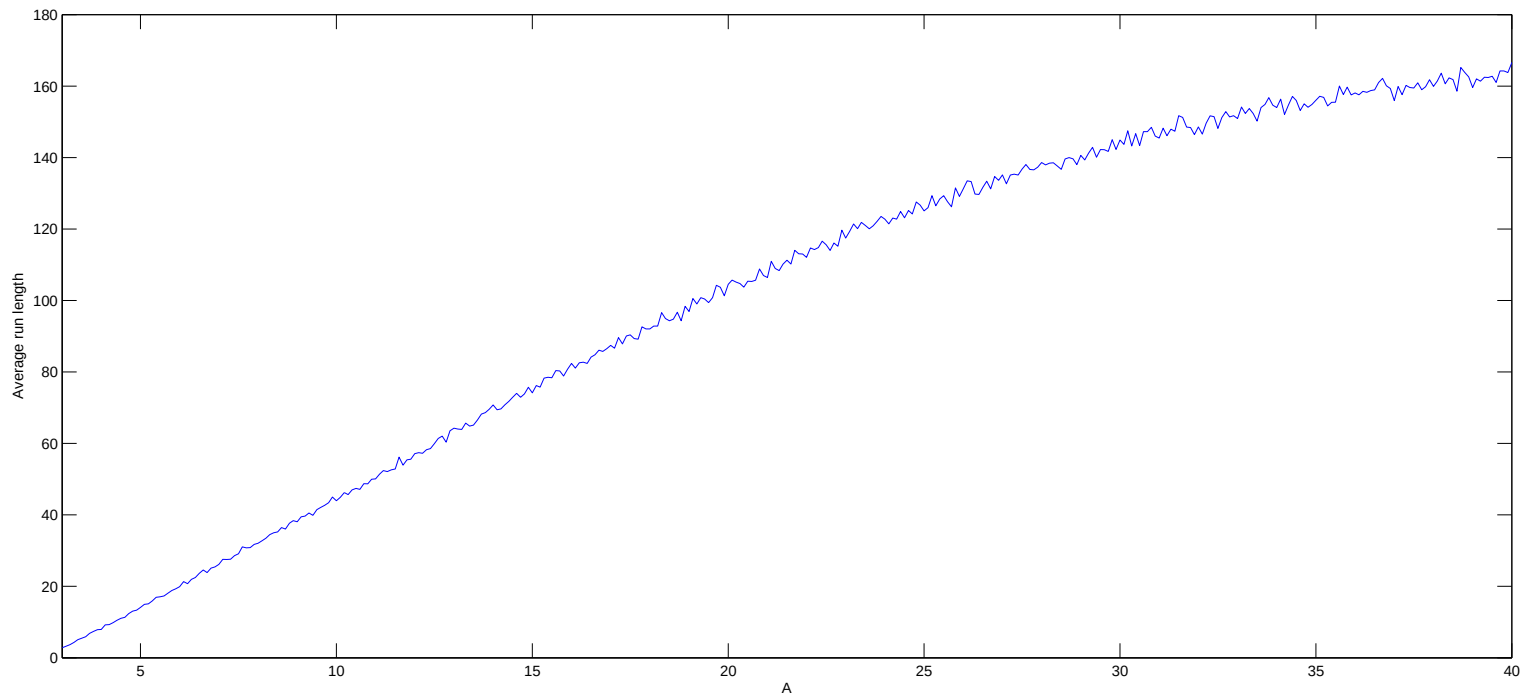
Table : CE-method

N_1	10000	5000	2000	1000	500
$N_2 = 100$	552.583	252.923	93.322	40.787	22.991
$N_2 = 200$	1024.712	478.670	178.011	86.988	41.501
$N_2 = 300$	1320.42	578.345	247.762	135.231	74.301

Table : CE-method + bisection

N_1	10000	5000	2000	1000	500
$N_2 = 100$	85.419	41.758	17.335	5.652	3.233
$N_2 = 200$	174.182	79.835	33.116	11.762	5.339
$N_2 = 300$	250.047	119.398	48.224	17.393	12.429

Average run length as a function of A



Example 1. CPU time for different simulation parameters for Shiryaev-Roberts procedure

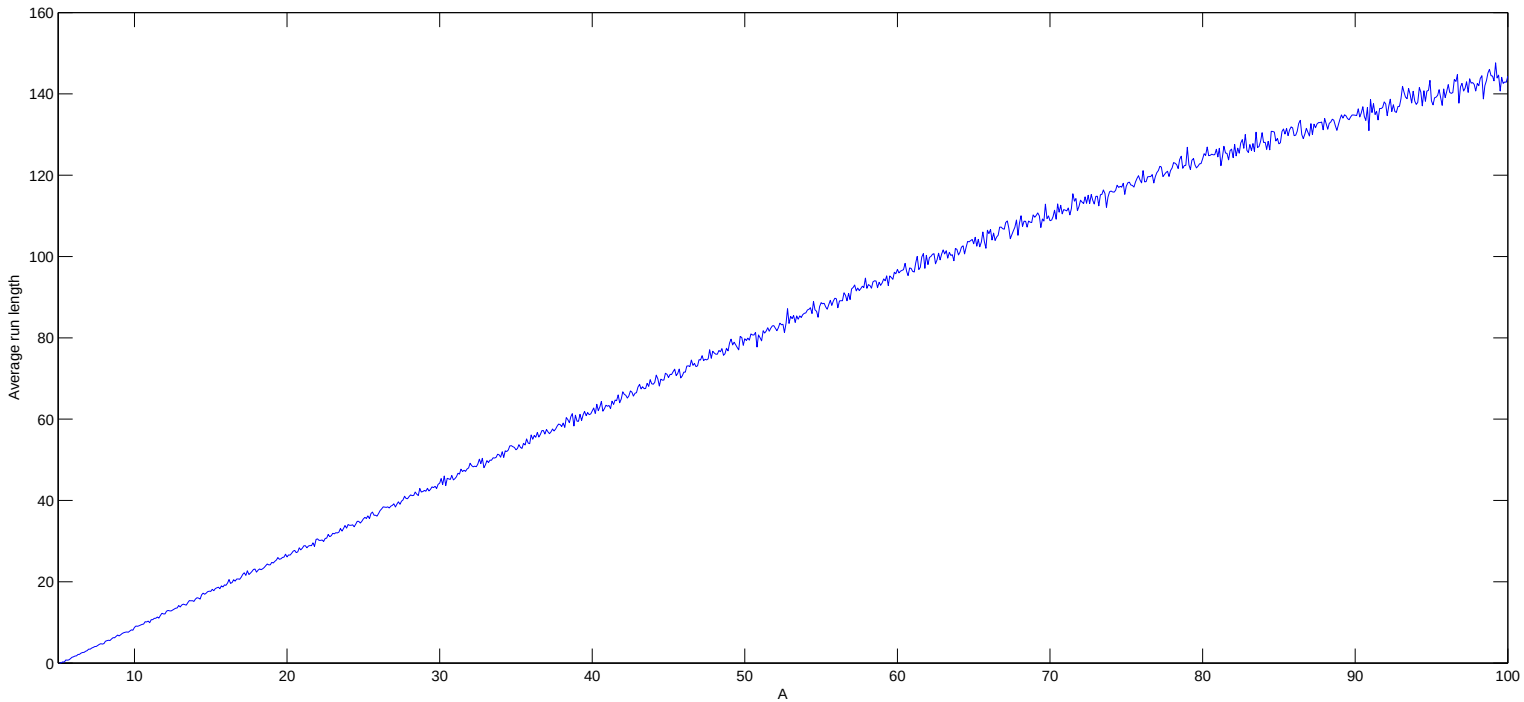
Table : CE-method

N_1	10000	5000	2000	1000	500
$N_2 = 100$	472.873	372.746	88.735	36.493	30.031
$N_2 = 200$	963.380	568.024	156.675	78.884	45.487
$N_2 = 300$	1432.528	642.752	204.038	104.317	67.349

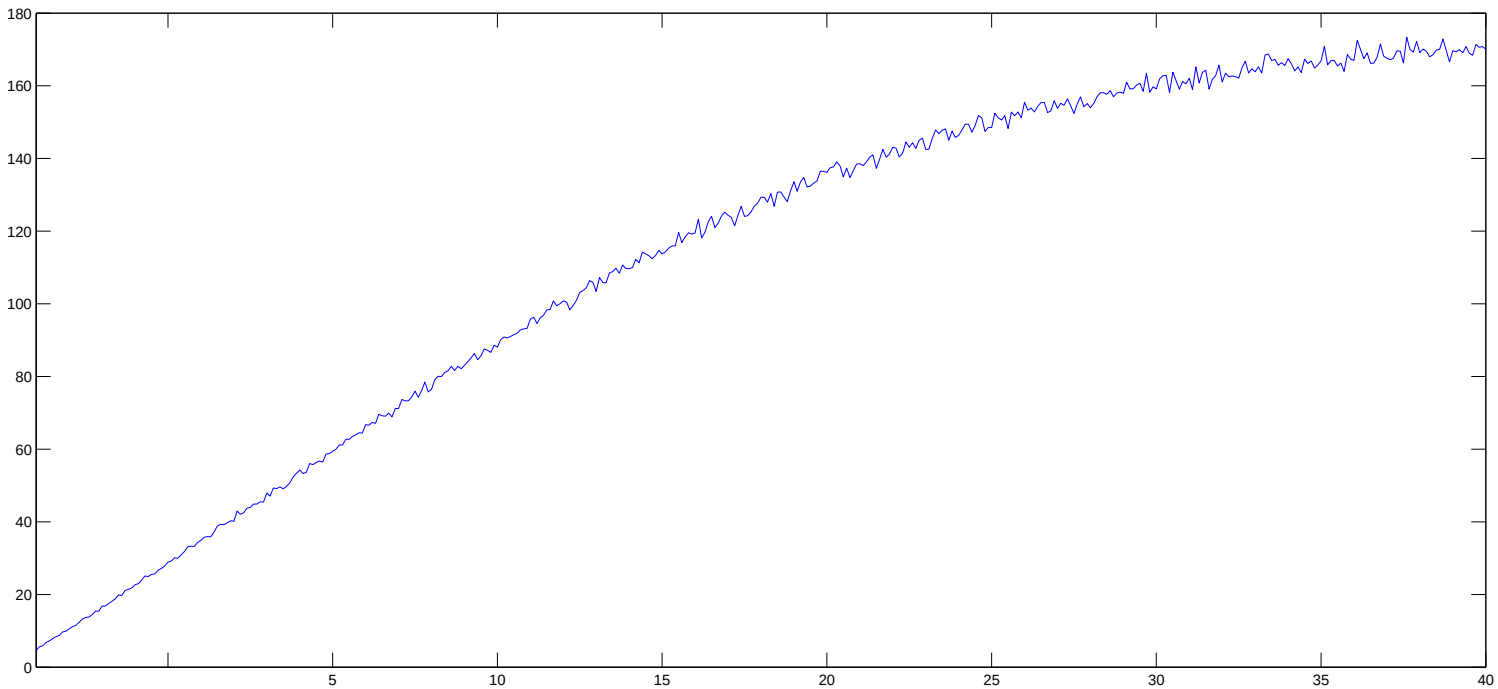
Table : CE-method + bisection

N_1	10000	5000	2000	1000	500
$N_2 = 100$	102.478	52.061	28.347	13.012	8.752
$N_2 = 200$	179.394	91.495	38.126	22.070	13.739
$N_2 = 300$	257.246	128.650	54.112	29.643	17.688

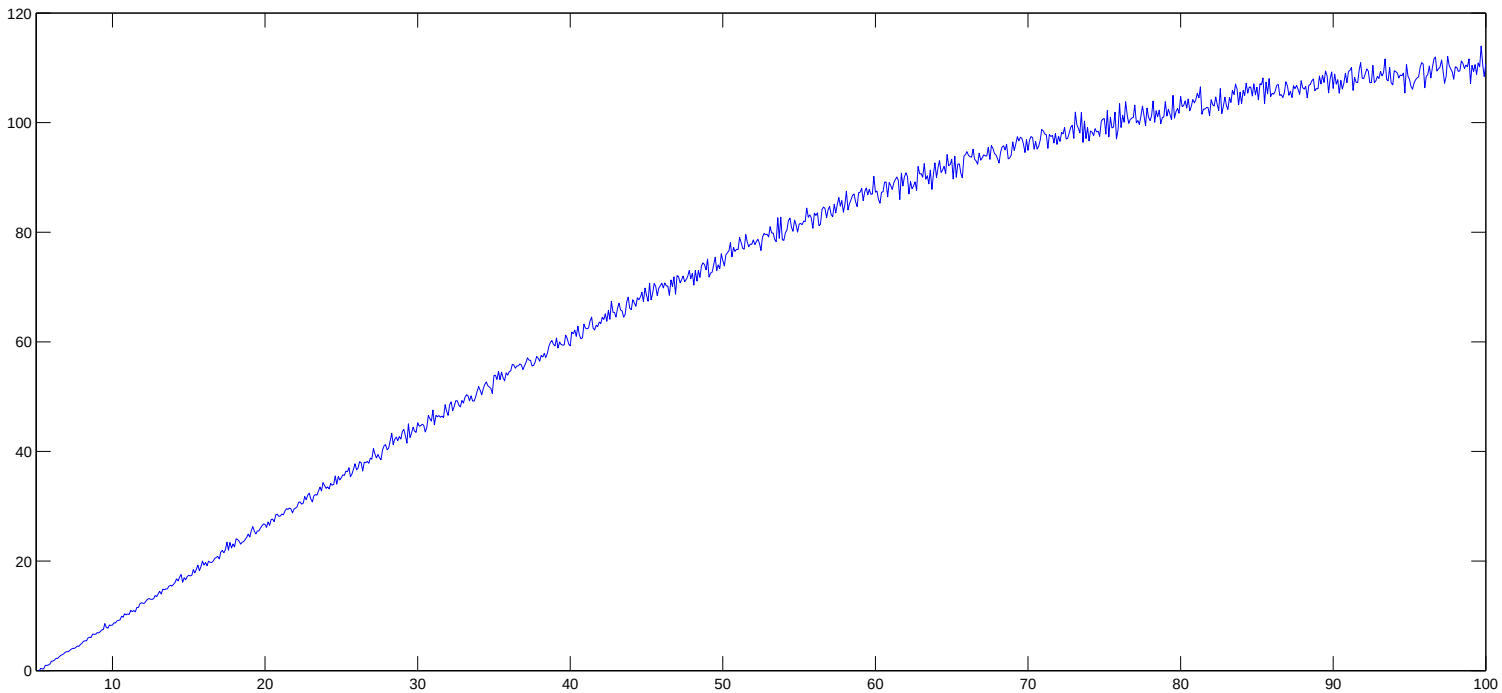
Average run length as a function of A



Example 2. CUSUM procedure



Example 2. Shiryaev-Roberts procedure



Concluding remarks

- Modified cross-entropy method gives a gain in CPU time
- The number of steps N_2 should be large
- We will consider the quickest change-point problem with a few change-points
- Application for different biological data

  Polushina T., Sofronov G. Change-point detection in biological sequences via genetic algorithm. In: CEC 2011 IEEE Congress on Evolutionary computation), 1966–1971 (2011)

References

- Andersson E. A monitoring system for detecting starts and declines of influenza epidemics. Research report. Göteborg University. 12 (2003)
- Boys R. J., Henderson D.A. A Bayesian approach to DNA sequence segmentation. Biometrics 60, 573–588 (2004)
- Frisén M. Optimal Sequential Surveillance for Finance, Public Health, and Other Areas, Sequential Analysis: Design Methods and Appl., 28:3, 310–337 (2009)
- Page E. S. Continuous Inspection Scheme. Biometrika, 41. 100–115 (1954)
- Pettersson M. Monitoring a freshwater fish population.Statistical surveillance of biodiversity. Environmetrics. 9, 139–150 (1998)
- Roberts S.W. A comparison of some control chart procedures. Technometrics, 8. 411–430 (1966)
- Pollak M., Tartakovsky A. Optimality properties of the Shiryaev-Roberts procedure. Statistica Sinica, 1729–1739 (2009)
- Rodionov S. A sequential algorithm for testing climate regime shifts. Geophysical research letters. V. 31. L90204 (2004)
- Rubinstein R., Kroese D. The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning. Springer-Verlag, New York (2004)
- Shiryaev A.N. The problem of the most rapid detection of a disturbance in a stationary process. Dokl. Math., 2. 795–799 (1961)
- Shiryaev A.N. Optimal stopping rules. Springer, New York (1978)

Thank you!