Resampling-based Change Point Estimation

Jelena Fiosina and Maksims Fiosins *

Clausthal University of Technology,
Instutute of Informatics,
Julius-Albert Str. 4,
D-38678, Clausthal-Zellerfeld, Germany
(Jelena.Fiosina, Maksims.Fiosins)@gmail.com

Abstract. Change point detecting problem is an important task in data mining applications. Standard statistical procedures for change point detection, based on maximum likelihood estimators, are complex and require building of parametric models of data. Instead, the methods of computational statistics replace complex statistical procedures with an amount of computation; so these methods become more popular in practical applications. The paper deals with two resampling tests for change point detection. In well known bootstrap-based CUSUM test we derive the formulas to estimate the efficiency of this procedure by taking an expectation and variance of the estimator into account. We propose also another simple pairwise resampling test and analyze its properties and efficiency too. We illustrate our approach by numerical example considering the problem of decision making of vehicles in city traffic.

Keywords: Change point, resampling, estimation, expectation, variance, CUSUM test

1 Introduction

Change point (CP) analysis is an important part of data mining, which purpose is to determine if and when a change in a data set has occurred. Online detection of CP is useful in modeling and prediction of data sequence in application areas such as finance, biometrics [9], robotics and traffic control [4].

CP analysis can be used: 1) for determining if changes in the process control led to changes in an output, 2) for solving a class of problems, such as control, forecasting etc., and 3) trend change analysis ([9]).

Traditional statistical approach to the problem of CP detection is maximum likelihood estimation (MLE). In this approach, a model of data is constructed, the likelihood function for CP is written and the estimator of CP is a result of the likelihood function minimization. Such an approach requires knowledge of exact data model and its parameters as well as complex analytical or numerical manipulations with likelihood function [6].

* We would like to thank Lower Saxony Technical University project "Planning and Decision Making for Autonomous Actors in Traffic" and European Commission FP7 Marie Curie IEF for Career Development programm grant for support.

In the case of small samples this approach does not allow us to choose the probability distributions correctly and properly estimate their parameters.

Alternatively, methods of computational statistics (CST) [10] are widely used, where complex statistical procedures are replaced with a big amount of computations. One technique for assessing if and when a CP (shift) has occurred is a cumulative sum chart (CUSUM chart). The form of a CUSUM chart allows to see visually if there is a CP. A confidence level may be assigned for each detected change. It can be constructed using bootstrapping approach[5], which belongs to a class of methods of computational statistics [11].

The paper firstly deals with bootstrap-based CUSUM CP test, slightly modified and described in terms of the resampling approach [1], [2], [3], [7], which allows its more accurate analysis by estimating its theoretical properties. We derive the analytical formulas to estimate the efficiency of this technique by taking expectation and variance as efficiency criteria. Secondly we propose another simple resampling-based test, based on pairwise comparisons of randomly selected data and estimate its efficiency too.

We illustrate our approach by numerical examples considering the problem of decision making of vehicles in city traffic [8].

The paper is organized as follows. In Section 2, we formulate CP problem formally and describe a standard CUSUM approach. In Section 3, we present the efficiency analysis of CUSUM -based approach in resampling terms. Section 4 proposes alternative simple pairwise resampling test and analyses its efficiency. Section 5 demonstrates a case study, with numerical examples in traffic applications. Section 6 contains final remarks and concludes the paper.

2 Problem Formulation

Let us formulate CP detection problem. Let $\{X_1, X_2, \ldots, X_n\}$ be a sequence of random variables. Let us divide the sample $\mathbf{X}$ as $\mathbf{X} = \{\mathbf{X}^B, \mathbf{X}^A\}$. We say that there is a CP at a position k, if the variables $\mathbf{X}^B = \{X_1, X_2, \ldots, X_k\}$ have a distribution $F_B(x, \Theta^B)$, but the variables $\mathbf{X}^A = \{X_{k+1}, X_{k+2}, \ldots, X_n\}$ have a distribution $F_A(x, \Theta^A)$, $\Theta^B \neq \Theta^A$. The aim of CP analysis is to estimate the value of k (clear that in the case of $k = n$ there is CP absence). We are interested in the case, when the distributions $F^B(\cdot)$ and $F^A(\cdot)$ differ with a mean value.

CP analysis using CUSUM charts in combination with Bootstrap approach [9] is the following: first, a CUSUM chart is constructed, which presents a difference between sample data and a mean. If there is no CPs in the mean of the data, the CUSUM chart will be relatively flat. From other side, in the case of CP existence, there will be obvious minimum or maximum in CUSUM chart.

The cumulative sum S_i at each data point i is calculated as follows: $S_i = \sum_{j=1}^{i}(X_i - \bar{X})$, where $i = 1, 2 \ldots n$, X_i is the current value, and $\bar{X}$ is the mean.

A CUSUM chart starts at zero ($S_0 = 0$) and always ends at zero ($S_n = 0$). Increasing (decreasing) of the CUSUM chart means that the data X_i are permanently greater (smaller) than the sample mean. A change in the direction in the CUSUM chart allows to spread about the CP in the mean.

Figure 1 presents an example of initial sample (left) and corresponding CUSUMs (right); a bold line represents CUSUMs calculated on initial sample, dotted lines - CUSUMs on bootstrapped data. An initial CUSUM chart well detects a change point at $k = 10$.

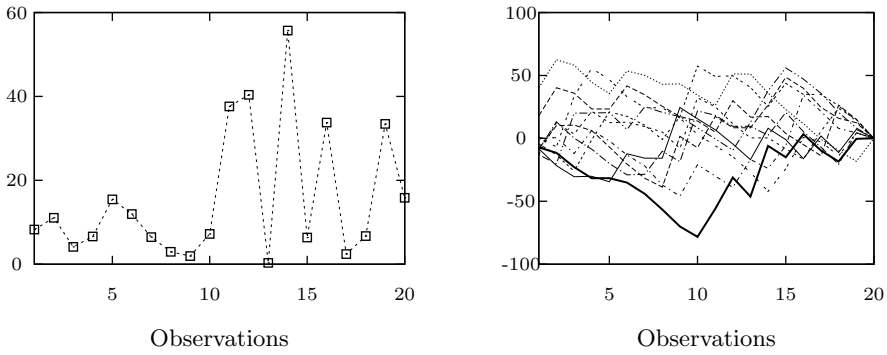

Fig. 1. Sample data (left) and sample CUSUM chart(right)

For each change it is possible to calculate a confidence level using bootstrapping of initial data. For this purpose, initial data are randomly permuted (bootstrapped). N bootstraps are produced with the sample data set. For each bootstrapped data set we construct CUSUMS $S^*(r)$ and estimate its range, so for the r-th bootstrap iteration $\Delta S^*(r) = \max_i\{S_i^*(r)\} - \min_i\{S_i^*(r)\}$, $i = 1, 2, \ldots, n$.

The final step in determining the confidence level is to calculate the percent of times that the range for the original CUSUM data $\Delta S = \max_i\{S_i\} - \min_i\{S_i\}$ exceeds the range for the bootstrapped CUSUM data $\Delta S^*(r)$, $r = 1, 2 \ldots N$. For this purpose we need to build an empirical distribution function for bootstrap ranges $\Delta S^*(r)$, $r = 1, 2, \ldots N$, as

$$\hat{F}_{\Delta S^*}(x) = \frac{\#\{\Delta S^*(r) \le x\}}{N} = \frac{1}{N}\sum_{r=1}^{N} 1_{\{\Delta S^*(r) \le x\}}. \tag{1}$$

Let us consider a hypotheses H^0, about CP absence in the data against the alternative H^1 that there is CP. It is appropriate to set a predetermined threshold confidence level γ beyond which a change is considered significant. Typically, $\gamma = 0.95$ or $\gamma = 0.99$ is selected.

Then using the cdf $\hat{F}_{S^*}(x)$ (1) we construct a bootstrap approximation of a confidence interval for ΔS: $[\hat{F}^{-1}_{\Delta S^*}(\frac{1-\gamma}{2}); \hat{F}^{-1}_{\Delta S^*}(\frac{1+\gamma}{2})]$, where $\hat{F}^{-1}_{\Delta S^*}(\gamma)$ is the quantile of the distribution $\hat{F}_{S^*}$ of the level γ. If the interval does not cover the value ΔS, then we can conclude, that initial and bootstrapped data significantly differ, and we reject H^0 (and this means CP existence).

3 Statistical properties of CUSUM Test

CUSUM test interpretation as a resampling [3] test allows to derive some its properties and estimate the efficiency on the base of expectation and variance of the estimator. Let us consider a test for a CP at the point k under H^0. Here, we do not deal with the ranges of CUSUMs, but with values itself.

We produce N iterations of resampling procedure. At r-th iteration, we extract, without replacement, k elements from the sample $\mathbf{X}$, forming the resample $\{X_1^{*r}, X_2^{*r}, \ldots, X_k^{*r}\}$ and construct the CUSUM estimator for a point k:

$$S_k^*(r) = \sum_{i=1}^{k}(X_i^{*r} - \bar{X}) = \sum_{i=1}^{k} X_i^{*r} - k\bar{X}, \tag{2}$$

where $\bar{X}$ is an average over the sample $\mathbf{X}$.

After N such realizations, obtaining a sequence $S_k^*(1)$, $S_k^*(2)$, …, $S_k^*(N)$.

Now we calculate the resampling estimator $F_k^*(x)$ of the distribution function of bootstrapped CUSUMS: $F_k^*(x) = \frac{1}{N}\sum_{r=1}^{N} 1_{\{S_k^*(r)\le x\}}$.

We are interested in the expectation and variance of the estimator $F_k^*(x)$. The expectation of $F_k^*(x)$ can be expressed as

$$\begin{aligned} E[F_k^*(x)] = E\left[\frac{1}{N}\sum_{r=1}^{N} 1_{\{S_k^*(r)\le x\}}\right] = P\{S_k^*(r)\le x\} = \\ = P\left\{\left(\sum_{i=1}^{k} X_i^{*r} - k\bar{X}\right) \le x\right\} = P\left\{\sum_{i=1}^{k} X_i^{*r} \le k\bar{X} + x\right\} = \\ = P\left\{\sum_{i=1}^{k} X_i^{*r} \le x'\right\} = P\left\{BA_k^*(r) \le x'\right\}. \end{aligned} \tag{3}$$

In order to find the distribution of the sum $BA_k^*(r)$, it is important to note that the resample $\{X_1^{*r}, X_2^{*r}, \ldots, X_k^{*r}\}$ contains elements from samples $\mathbf{X}^B$ and $\mathbf{X}^A$ with possibly different distributions $F^B(x)$ and $F^A(x)$. So let us split the sum $BA_k^*(r)$ into two sums: the sum $B_{y_r}^*(r)$ contains elements selected from $\mathbf{X}^B$ and the sum $A_{k-y_r}^*(r)$ contains elements selected from $\mathbf{X}^B$: $BA_k^*(r) = B_{y_r}^*(r) + A_{k-y_r}^*(r)$, where y_r is a number of elements extracted from $\mathbf{X}^B$ on r-th step.

Now we can write a formula for the distribution of $BA_k^*(r)$:

$$\begin{aligned} F_{BA}^k(x) = P\left\{BA_k^*(r)\le x\right\} = P\left\{\sum_{i=1}^{k} X_i^{*r} \le x\right\} = \\ = \sum_{y_r=0}^{k} q(y_r)\cdot P\left\{B_{y_r}^*(r) + A_{k-y_r}^*(r) \le x\right\} = \\ = \sum_{y_r=0}^{k} q(y_r)\cdot \int_{-\infty}^{\infty} F_B^{(y_r)}(x-u)dF_A^{(k-y_r)}(u), \end{aligned} \tag{4}$$

where $F^{(c)}(x)$ is c-th fold convolution of cdf $F(x)$ with itself, $F^{(0)}(x) \equiv 1$.

Note that the probability $q(y_r)$, that the sum $BA_k^*(r)$ contains a fixed number y_r of elements from $\mathbf{X}^B$, and the rest elements are from $\mathbf{X}^A$ is

$$q(y_r) = \binom{k}{y_r}\binom{n-k}{k-y_r}/\binom{n}{k}. \tag{5}$$

Taking into account the previous discussion (4)-(5), the formula (3) can be now rewritten as $E[F_k^*(x)] = P\{S_k^*(r) \le x\} = F_{BA}^k(x')$.

Now we derive an expression for the variance of the estimator (2)

$$\begin{aligned} &Var[F_k^*(x)] = \left(\frac{1}{N}\right)^2 Var\left[\sum_{r=1}^{N} 1_{\{S_k^*(r)\le x\}}\right] = \\ &= \frac{1}{N} Var\left[1_{\{S_k^*(r)\le x\}}\right] + \frac{(N-1)}{N} Cov\left[1_{\{S_k^*(r)\le x\}}, 1_{\{S_k^*(p)\le x\}}\right],\ r \ne p. \end{aligned} \tag{6}$$

The variance $Var\left[1_{\{S_k^*(r)\le x\}}\right]$ does not depend on the resampling procedure:

$$\begin{aligned} &Var\left[1_{\{S_k^*(r)\le x\}}\right] = E\left[(1_{\{S_k^*(r)\le x\}})^2\right] - \left(E\left[1_{\{S_k^*(r)\le x\}}\right]\right)^2 = \\ &= F_{BA}^k(x') - (F_{BA}^k(x'))^2. \end{aligned} \tag{7}$$

The term $Cov\left[1_{\{S_k^*(r)\le x\}}, 1_{\{S_k^*(p)\le x\}}\right]$ depends on the resampling procedure:

$$\begin{aligned} &Cov\left[1_{\{S_k^*(r)\le x\}}, 1_{\{S_k^*(p)\le x\}}\right] = \\ &= E\left[1_{\{S_k^*(r)\le x\}} \cdot 1_{\{S_k^*(p)\le x\}}\right] - E\left[1_{\{S_k^*(r)\le x\}}\right] E\left[1_{\{S_k^*(p)\le x\}}\right], \end{aligned} \tag{8}$$

which can be expressed using a mixed moment μ_{11} as

$$\mu_{11} = E\left[1_{\{S_k^*(r)\le x\}} \cdot 1_{\{S_k^*(p)\le x\}}\right] = P\left\{BA_k^*(r) \le x', BA_k^*(p) \le x'\right\}. \tag{9}$$

For the fixed value of $\mathbf{y} = (y_r, y_p)$ we have

$$\mu_{11}(\mathbf{y}) = P\left\{B_{y_r}^*(r) + A_{k-y_r}^*(r) \le x', B_{y_p}^*(p) + A_{k-y_p}^*(p) \le x'\right\}. \tag{10}$$

For $\mu_{11}(\mathbf{y})$ calculation we use the notation of $\boldsymbol{\alpha}$-pair ([1], [2], [7]). Let $\boldsymbol{\alpha} = (\alpha^B, \alpha^A)$, where α_B be the number of common elements in $B_{y_r}^*(r)$ and $B_{y_p}^*(p)$ and α_A be the number of common elements in $A_{k-y_r}^*(r)$ and $A_{k-y_p}^*(p)$.

Denote $W_{y_r}^B(r) \subset \mathbf{X}^B$ a set of elements which produce the sum $B_{y_r}^*(r)$ and $W_{k-y_r}^A(r) \subset \mathbf{X}^A$ a set of elements which produce the sum $A_{k-y_r}^*(r)$ given y_r, $\mathbf{W}_{k,y_r}(r) = \{W_{y_r}^B, W_{k-y_r}^A\}$. Let $M_k^B(y) = \{\max(0, y_r - y_p - k), \ldots, \min(y_r, y_p)\}$, $M_k^A(y) = \{\max(0, 3k - n - y_r - y_p), \ldots, \min(k - y_r, k - y_p)\}$, $M_k(y) = M_k^B(y) \times M_k^A(y)$. We say that $\mathbf{W}_{k,y_r}(r)$ and $\mathbf{W}_{k,y_p}(p)$ produce an $\boldsymbol{\alpha}$-pair, $\boldsymbol{\alpha} \in M_k(y)$, if $|W_{y_r}^B(r) \cap W_{y_p}^B(p)| = \alpha^B$ and $|W_{k-y_r}^A(r) \cap W_{k-y_p}^A(p)| = \alpha^A$.

Let $A_{r,p}(\boldsymbol{\alpha}, y)$ denote an event "$\mathbf{W}_{k,y_r}(r)$ and $\mathbf{W}_{k,y_p}(p)$ produce α-pair", $P_{r,p}(\boldsymbol{\alpha}, y) = P\{A_{r,p}(\boldsymbol{\alpha}, y)\}$ its probability. All realizations $r = 1 \ldots N$ are statistically equivalent and we can omit the lower indices r, p and write $P(\boldsymbol{\alpha}, y)$.

The probability $P(\boldsymbol{\alpha}, y)$ to obtain an $\boldsymbol{\alpha}$-pair given y, $\boldsymbol{\alpha} \in M_k(y)$ is

$$P(\boldsymbol{\alpha}, y) = \binom{y_r}{\alpha_B}\binom{k-y_r}{y_p-\alpha_B}/\binom{k}{y_p} \cdot \binom{k-y_r}{\alpha_A}\binom{n-2k+y_r}{k-y_p-\alpha_A}/\binom{n-k}{k-y_p}. \quad (11)$$

Let $\mu_{11}(\boldsymbol{\alpha}, y)$ be a conditional mixed moment μ_{11} given $\boldsymbol{\alpha}$ and y. Then μ_{11} (9) can be expressed in the following form:

$$\mu_{11} = \sum_{y} \sum_{\boldsymbol{\alpha} \in M_k(y)} \mu_{11}(\boldsymbol{\alpha}, y) P(\boldsymbol{\alpha}, y) q(y_r) q(y_p). \quad (12)$$

To obtain an expression for $\mu_{11}(\boldsymbol{\alpha}, y)$, we consider the sums $B^*_{y_r}(r)$ and $B^*_{y_p}(p)$ at two different realizations r and p, which produce an $\boldsymbol{\alpha}$-pair and y is given: they contain α^B common and $y_r - \alpha^B$ and $y_p - \alpha^B$ different elements. Let us split each of these sums into two parts, which contain common and different elements for realizations r and p: $B^*_{y_r}(r) = B^{com}_{\alpha_B}(r,p) + B^{dif}_{y_r-\alpha_B}(r,p), B^*_{y_p}(p) = B^{com}_{\alpha_B}(p,r) + B^{dif}_{y_p-\alpha_B}(p,r)$. The same can be made for the sums: $A^*_{k-y_r}(r) = A^{com}_{\alpha_A}(r,p) + A^{dif}_{k-y_r-\alpha_A}(r,p), A_{k-y_p}(p) = A^{com}_{\alpha_A}(p,r) + A^{dif}_{k-y_p-\alpha_A}(p,r)$.

Then the previous sums for two realizations r and p can be written as follows:

$$\begin{aligned} BA^*_k(r) &= B^{com}_{\alpha_B}(r,p) + B^{dif}_{y_r-\alpha_B}(r,p) + A^{com}_{\alpha_A}(r,p) + A^{dif}_{k-y_r-\alpha_A}(r,p), \\ BA^*_k(p) &= B^{com}_{\alpha_B}(p,r) + B^{dif}_{y_p-\alpha_B}(p,r) + A^{com}_{\alpha_A}(p,r) + A^{dif}_{k-y_p-\alpha_A}(p,r). \end{aligned} \quad (13)$$

Then

$$\begin{aligned} &\mu_{11}(\boldsymbol{\alpha}, y) = P\{BA^*_k(r) \le x', BA^*_k(p) \le x' | \boldsymbol{\alpha}, y\} = \\ &= P\left\{ B^{com}_{\alpha_B}(r,p) + A^{com}_{\alpha_A}(r,p) + B^{dif}_{y_r-\alpha_B}(r,p) + A^{dif}_{k-y_r-\alpha_A}(r,p) \le x', \right. \\ &\left. B^{com}_{\alpha_B}(p,r) + A^{com}_{\alpha_A}(p,r) + B^{dif}_{y_p-\alpha_B}(p,r) + A^{dif}_{k-y_p-\alpha_A}(p,r) \le x' | \boldsymbol{\alpha}, y \right\} = \\ &= \int_{-\infty}^{\infty} P\left\{ B^{dif}_{y_r-\alpha_B}(r,p) + A^{dif}_{k-y_r-\alpha_A}(r,p) \le x' - u | \boldsymbol{\alpha}, y \right\} \cdot \\ &P\left\{ B^{dif}_{y_p-\alpha_B}(p,r) + A^{dif}_{k-y_p-\alpha_A}(p,r) \le x' - u | \boldsymbol{\alpha}, y \right\} dF^{\alpha^B,\alpha^A}_{BA}(u) = \\ &= \int_{-\infty}^{\infty} F^{y_r-\alpha_B, k-y_r-\alpha_A}_{BA}(x'-u) \cdot F^{y_p-\alpha_B, k-y_p-\alpha_A}_{BA}(x'-u) dF^{\alpha^B,\alpha^A}_{BA}(u), \end{aligned} \quad (14)$$

where $F^{b,a}_{BA}(x)$ is cdf of the sum of b elements from $\mathbf{X}^B$ and a elements from $\mathbf{X}^A$.

For some distributions we can obtain explicit expressions.

Exponential case. Firstly need to find the distribution of the sum of a independent exponentially distributed r.v. $\{X_i\}$ with parameter λ and b independent exponentially distributed r.v. $\{Y_i\}$ with parameter ν. Following Afanasyeva [1], where a case of the difference of the sums of random variables was discussed

$$\begin{aligned} &F^{b,a}_{BA}(x) = P\left\{ \sum_{i=1}^{a} X_i + \sum_{i=1}^{b} Y_i \le x \right\} = \int_{-\infty}^{\infty} F^{(a)}(x-u) f^{(b)}(u) du = \\ &= F_{Er(\nu,a)}(x) - \\ &- \frac{e^{-\lambda x}\nu^a}{(a-1)!} \sum_{i=0}^{b-1} \frac{\lambda^i}{i!} \sum_{p=0}^{i} \binom{i}{p} (-1)^{i-p} x^p \cdot \frac{(i-p+a-1)!}{(\nu-\lambda)^{a-p+i}} F_{Er(\nu-\lambda, i-p+a)}(x), \end{aligned}$$

where $F_{Er(\nu,a)}(x)$ is cdf of Erlang distribution with parameters ν, a.

Now using (4) - (14) we obtain the properties of the resampling estimator.

Normal case. We consider a case of a sum of a independent normally distributed r.v. $\{X_i\}$ with parameters β_X and σ_X and b independent normally distributed r.v. $\{Y_i\}$ with parameters β_Y and σ_Y. As the sum of normally distributed r.v. is normally distributed:

$$F_{BA}^{b,a}(x) = P\left\{\sum_{i=1}^{a} X_i + \sum_{i=1}^{b} Y_i \le x\right\} = \Phi\left(\frac{x - (\beta_X \cdot a + \beta_Y \cdot b)}{\sqrt{a \cdot \sigma_X^2 + b \cdot \sigma_Y^2}}\right), \tag{15}$$

where $\Phi(x)$ is cdf of standard normal distribution $N(0,1)$.

Now we can use all formulas from (4) to (14) to obtain the properties of the resampling estimator.

4 Pairwise Resampling CP Test and its efficiency

We propose an alternative resampling CP test. Let us test a point k. The idea behind this method is based of the consideration of the probability $P\{X \le Y\}$, where r.v. X is taken randomly from the subsample $\mathbf{X}^B$ and r.v. Y from the subsample $\mathbf{X}^A$. If the samples $\mathbf{X}^B$ and $\mathbf{X}^A$ are from one distribution, this probability should be equal to 0.5. However, for our test we scale this value by multiplying to the difference $y - x$ in the case when $x \le y$.

So our characteristic of interest is: $\Psi(x,y) = I_{\{x<y\}} \cdot (y - x)$.

Now we produce N realizations of the following resampling procedure. On the realization r we extract one value X^{*r} from the sample $\mathbf{X}^B$ and one value Y^{*r} from the sample $\mathbf{X}^A$, compare them and calculate value of $\Psi(x,y)$.

The resampling estimator is an average over all realizations of $\Psi(x,y)$:

$$\Theta^* = \frac{1}{N}\sum_{r=1}^{N} \Psi(X^{*r}, Y^{*r}). \tag{16}$$

We are interested in such efficiency criteria as expectation and variance of (16). Firstly the expectation of (16) can be expressed as follows:

$$\begin{aligned} E[\Theta^*] &= \frac{1}{N}\sum_{r=1}^{N} E\left[I_{\{X^{*r}<Y^{*r}\}} \cdot (Y^{*r} - X^{*r})\right] = \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{y} (y-x) f_B(x) \cdot f_A(y) dx dy = \\ &= \int_{-\infty}^{\infty} y f_A(y) F_B(y) dy - \int_{-\infty}^{\infty} f_A(y) \int_{-\infty}^{y} x f_B(x) dx dy. \end{aligned} \tag{17}$$

Secondly the variance of (16) for $p \ne r$ can be expressed as follows:

$$Var(\Theta^*) = \frac{1}{N} Var(\Psi(X^{*r}, Y^{*r})) + \frac{N-1}{N} Cov(\Psi(X^{*r}, Y^{*r}), \Psi(X^{*p}, Y^{*p})). \tag{18}$$

Let us consider the variance $Var(\Psi(X^{*r}, Y^{*r}))$:

$$Var(\Psi(X^{*r}, Y^{*r})) = E\left(\Psi(X^{*r}, Y^{*r})^2\right) - \left(E\left(\Psi(X^{*r}, Y^{*r})\right)\right)^2. \tag{19}$$

Here we should find the expression for the second moment:

$$\begin{aligned} &E\left(\Psi(X^{*r}, Y^{*r})^2\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{y}(y-x)^2 f_B(x)\cdot f_A(y)dxdy = \\ &= \int_{-\infty}^{\infty} y^2 f_A(y)F_B(y)dy - 2\int_{-\infty}^{\infty} y f_A(y)\int_{-\infty}^{y} x f_B(x)dxdy + \\ &+ \int_{-\infty}^{\infty} f_A(y)\int_{-\infty}^{y} x^2 f_B(x)dxdy. \end{aligned} \tag{20}$$

Now let us consider the covariance term from (18): $Cov(\Psi(X^{*r},$ $Y^{*r}), \Psi(X^{*p}, Y^{*p})) = E(\Psi(X^{*r}, Y^{*r})\cdot\Psi(X^{*p}, Y^{*p})) - E(\Psi(X^{*r}, Y^{*r}))^2$.

The mixed moment $\mu_{11} = E(\Psi(X^{*r}, Y^{*r})\cdot\Psi(X^{*p}, Y^{*p})) = \sum_{i=1}^{4}\mu_{11}^i\cdot p_i$ can be calculated differently for 4 following cases: 1) both elements X^{*r}, Y^{*r} and X^{*p}, Y^{*p} are different for two realizations p and r; 2) X^{*r} and X^{*p} are common for two realizations p and r; 3) Y^{*r} and Y^{*r} are common for two realizations p and r; 4) both X^{*r}, Y^{*r} and X^{*p}, Y^{*p} are common for two realizations p and r, with probabilities

$$p_1 = \frac{(k-1)(n-k-1)}{k\cdot(n-k)},\ p_2 = \frac{(k-1)}{k\cdot(n-k)},\ p_3 = \frac{(n-k-1)}{k\cdot(n-k)},\ p_4 = \frac{1}{k\cdot(n-k)}.$$

The corresponding moments are:

$$\begin{aligned} &\mu_{11}^1 = E(\Psi(X^{*r}, Y^{*r})\cdot\Psi(X^{*p}, Y^{*p})) = E(\Psi(X^{*r}, Y^{*r}))^2, \\ &\mu_{11}^2 = E(\Psi(X^{*r}, Y^{*r})\cdot\Psi(X^{*p}, Y^{*p})) = \\ &= \int_{-\infty}^{\infty)}\int_{x}^{\infty}\int_{x}^{\infty}(y_1-x)(y_2-x)\cdot f_B(x)f_A(y_1)f_A(y_2)dy_1dy_2dx, \\ &\mu_{11}^3 = E(\Psi(X^{*r}, Y^{*r})\cdot\Psi(X^{*p}, Y^{*p})) = \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{y}\int_{-\infty}^{y}(y-x_1)(y-x_2)\cdot f_B(x_1)f_B(x_2)f_A(y)dx_1dx_2dy, \\ &\mu_{11}^4 = E(\Psi(X^{*r}, Y^{*r})\cdot\Psi(X^{*p}, Y^{*p})) = E\left(\Psi(X^{*r}, Y^{*r})^2\right). \end{aligned} \tag{21}$$

Exponential case. Let elements from the sample $\mathbf{X}^B$ be exponentially distributed with parameter ν and elements from the sample $\mathbf{X}^A$ be exponentially distributed with parameter λ. Then

$$\begin{aligned} &E(\Theta^*) = \frac{1}{\lambda} - \frac{1}{\nu} + \frac{\lambda}{\nu(\lambda+\nu)},\ \mu_{11}^2 = \frac{\nu}{\lambda^2(2\lambda+\nu)}, \\ &E\left(\Phi(X^{*r}, Y^{*r})^2\right) = \frac{2}{\lambda^2} - \frac{2}{\nu\cdot\lambda} + \frac{2}{\nu^2} - \frac{2\lambda}{\nu^2(\lambda+\nu)}, \\ &\mu_{11}^3 = \frac{2}{\lambda^2} - \frac{2}{\nu\lambda} + \frac{2\lambda}{\nu(\lambda+\nu)^2} + \frac{1}{\nu^2} - \frac{2\lambda}{\nu^2(\lambda+\nu)} + \frac{\lambda}{\nu^2(2\nu+\lambda)}. \end{aligned} \tag{22}$$

5 Numerical examples

Normal case. We consider a vehicle routing problem in a street network, where vehicles receive data about travel times and are applying the shortest path algorithm looking for a fastest path to their destination. As travel times are subject to change, that's why CP analysis is performed. If CP is detected, only the data part after the last CP is taken into account. Suppose, that travel times are normally distributed with initial parameters $\mu_B = 60$, $\sigma_B = 10$ before the CP and $\mu_A = 70$, $\sigma_A = 10$ after the CP, this means that CP occurs. We show the behavior of our tests' estimators depending on different parameters of initial data (Fig. 2 - 4). We can see (Fig. 2) how the probability of CP absence decreases depending on the increasing of the difference between the means before μ_B and after μ_A the CP. A big variance of the probability of interest (standard deviation = 0.29) (left), i.e. there exists a big risk of considering some point as a CP, if it is not one. In Fig. 4 (right) we see very good CP detection of CUSUM test. For the pairwise test (Figure 4)(left) in the case of CP absence we see smaller variance of the probability of interest (σ= 0.14). i.e. a risk of considering some point as a CP, if it is not one, is lower than for CUSUM test. Here we see larger variance of the estimator.

So we can conclude that CUSUM test detects CP very well; however, it often considers as classifies as CP some point, which are not. In opposite, pairwise test is more reliable in the case of CP absence; however, it can miss some CP. So for streets where CPs are rare, it is better to use the pairwise test; CUSUM test is better for streets with often CPs in travel times.

Exponential case. We consider a traffic control device (TCD) (traffic lights or matrix signs), which regulate the vehicle flows in the street network. TCD receives data about the intervals between vehicles intensity arriving to this TCD. The working regime of TCD depends on the state of traffic. When TCD noticed some CP in traffic flow characteristics it changes its working regime. That's why data mining using CP analysis is of great importance. Suppose for our investigation the initial parameters of intensity of traffic flow are $\lambda_B = 0.1$ before the CP and $\lambda_A = 0.05$ after the CP. Following the previous example we change the initial parameters to verify the correct behavior of our tests.

We show the behavior of our tests' estimators depending on different parameters of initial data (Fig. 5 - 7). For the exponential case, we can see (Fig. 5) a bit worse performance of CP tests; even for the difference $1/\nu - 1/\lambda = 10$ the CUSUM test demonstrates the probability of CP = 0.2; the variance of the pairwise test increases. In Fig. 6 and 7 we see the similar behavior of CUSUM and pairwise test: the first better detects existing CP, the second is more reliable in the case of CP absence.

So we can conclude that CUSUM test detects CP very well; however, it often considers as classifies as CP some point, which are not. In opposite, the pairwise test is more reliable in the case of CP absence; however, it can miss some CPs. So for streets, where CPs are rare, it is better to use the pairwise test; CUSUM test is better for streets with often CPs in travel times.

6 Conclusions

CP detecting problem was considered and the methods of computational statistics were implemented. In the paper, we consider two resampling tests for CP detection. In well known bootstrap-based CUSUM test we derive the formulas to estimate the efficiency of this procedure based on expectation and variance. We propose also another simple pairwise resampling test and analyze its properties and efficiency too. We illustrate the application of our approach by numerical example solving the problem of decision making of vehicles in city traffic. CUSUM test demonstrated good detection of CPs; however has a big variance in the case of CP absence. The pairwise test has smaller variance in the case of CP absence, however worse detects existing CPs. First experiments show the properties of demonstrated approach, depending on various parameters of initial data. In the future we are going to work with CP trend analysis and autoregression, applying bootstrapping and resampling technique.

References

1. Afanasyeva, H.: Resampling-approach to a task of comparison of two renewal processes. In: Proc. of the 12th International Conference on Analytical and Stochastic Modelling Techniques and Applications. pp. 94–100. Riga (2005)
2. Andronov, A., Fioshina, H., Fioshin, M.: Statistical estimation for a failure model with damage accumulation in a case of small samples. Journal of Statistical Planning and Inference 139(5), 1685 – 1692 (2009)
3. Andronov, A., Merkuryev, Y.: Optimization of statistical sample sizes in simulation. Journal of Statistical Planning and Inference 85(1-2), 93 – 102 (2000)
4. Carslaw, D.C., Ropkins, K., Bell, M.C.: Change-point detection of gaseous and particulate traffic-related pollutants at a roadside location. Environmental Science & Technology 40(22), 6912–6918 (2006)
5. Efron, B., Tibshirani, R.: An introduction to the Bootstrap. Chapman and Hall, New York (1993)
6. Ferger, D.: On the almost sure convergence of maximum likelihood-type estimators for a change-point. Theory Stoch. Processes 8(1–2), 81–87 (2002)
7. Fioshin, M.: Efficiency of resampling estimators of sequential-parallel systems reliability. In: Proc. of 2nd International Conference Simulation, Gaming, Training and Business Process Reengineering in Operations. pp. 112–117. Riga (2000)
8. Fiosins, M., Fiosina, J., Müller, J., Görmer, J.: Agent-based integrated decision making for autonomous vehicles in urban traffic. In: Demazeau, Y., Pechoucek, M., Corchado, J., Pérez, J. (eds.) Advances on Practical Applications of Agents and Multiagent Systems, Advances in Intelligent and Soft Computing, vol. 88, pp. 173–178. Springer, Berlin / Heidelberg (2011)
9. Gavit, P., Baddour, Y., Tholmer, R.: Use of change-point analysis for process monitoring and control. BioPharm International 22 (2009)
10. Gentle, J.E.: Elements of Computational Statistics. Springer (2002)
11. Hinkley, D.V.: Inference about the change-point from cumulative sum tests. Biometrika 58(3), 509–523 (1971)

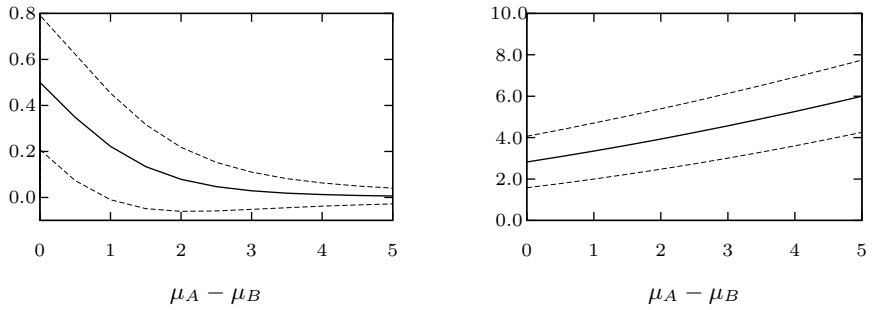

Fig. 2. Normal case, sample size = 20. CUSUM test (left), pairwise test (right); $E[F^*(x)]$ and $E[\Theta^*]$ - straight lines; $E[F^*(x)] \pm \sqrt{(Var[F^*(x)])}$ and $E[\Theta^*] \pm \sqrt{(Var[\Theta^*])}$ - dashed lines

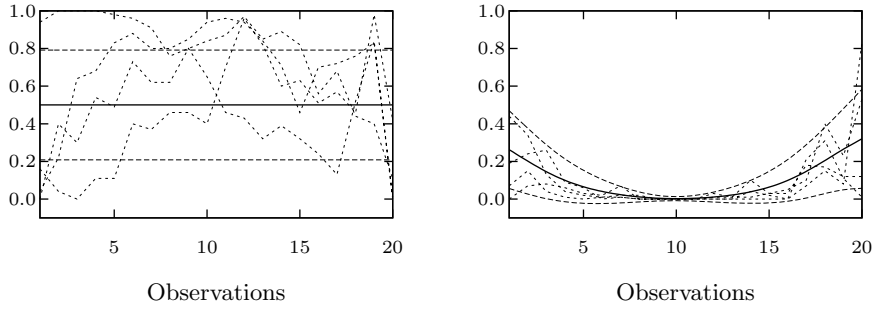

Fig. 3. Normal case. CUSUM test without CP (left) and with CP at $k = 10$ (right); $E[F^*(x)]$ - straight line; $E[F^*(x)] \pm \sqrt{(Var[F^*(x)])}$ - dashed line; several realizations of $F^*(x)$ - dotted lines

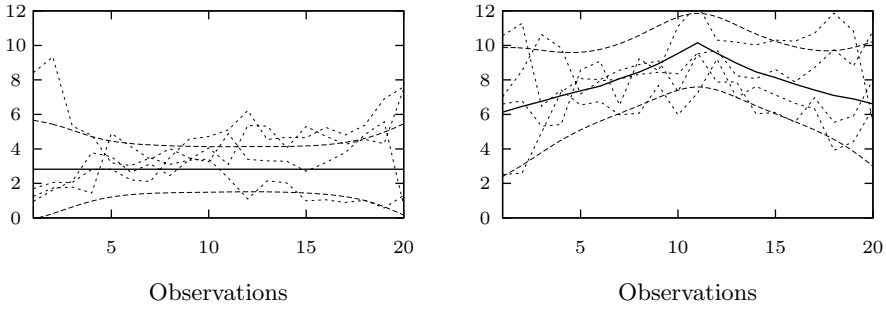

Fig. 4. Normal case. Pairwise test without CP (left) and with CP at $k = 10$ (right); $E[\Theta^*]$ - straight line; $E[\Theta^*] \pm \sqrt{(Var[\Theta^*])}$ - dashed lines; several realizations of Θ^* - dotted lines

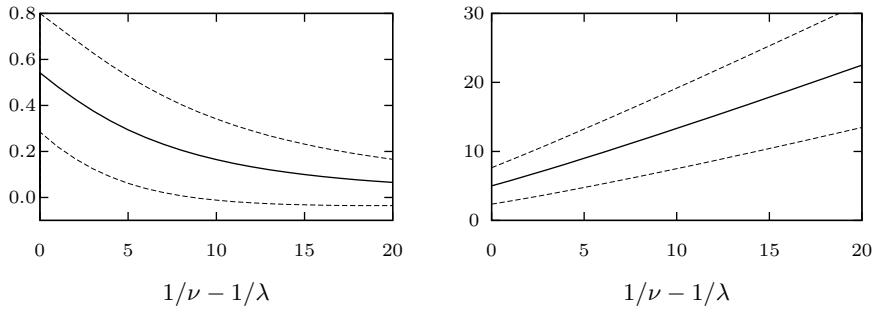

Fig. 5. Exponential case, sample size = 20. CUSUM test (left), pairwise test (right); $E[F^*(x)]$ and $E[\Theta^*]$ - straight lines; $E[F^*(x)] \pm \sqrt{(Var[F^*(x)])}$ and $E[\Theta^*] \pm \sqrt{(Var[\Theta^*])}$ - dashed lines

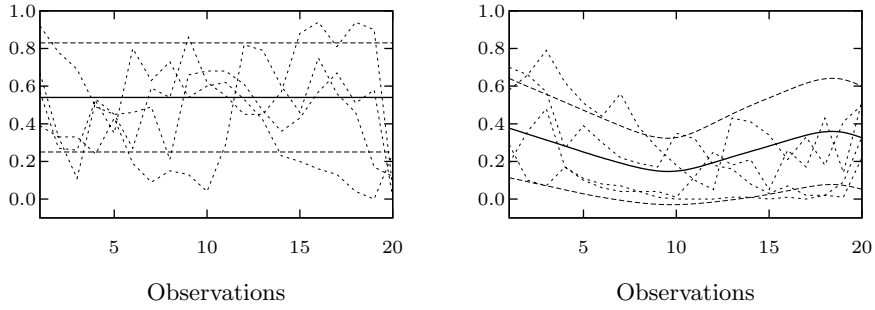

Fig. 6. Exponential case. CUSUM test without CP (left) and with CP at $k = 10$(right); $E[F^*(x)]$ - straight line; $E[F^*(x)] \pm \sqrt{(Var[F^*(x)])}$ - dashed lines; several realizations of $F^*(x)$ - dotted lines

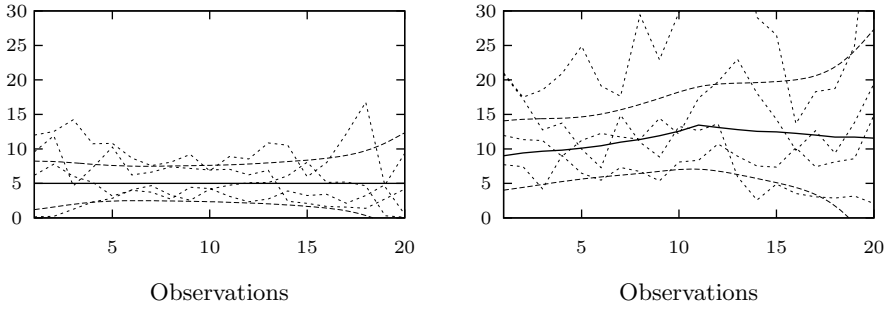

Fig. 7. Exponential case. Pairwise test without CP (left) and with CP at $k = 10$ (right); $E[\Theta^*]$ - straight line; $E[\Theta^*] \pm \sqrt{(Var[\Theta^*])}$ - dashed lines; several realizations of Θ^* - dotted lines