JMLR: Workshop and Conference Proceedings vol (2012) ??–??

Multiscale parametric approach for change point detection

Nazar Buzun ABC@SAMPLE.COM
Address 1

Vladimir Spokoiny XYZ@SAMPLE.COM
Address 2

Alexandra Suvorikova XYZ@SAMPLE.COM
Address 3

Abstract

This work presents a novel algorithm for change point detection, that can be applied for analysis of data of unknown nature. It is based on likelihood-ratio test statistics, as its behaviour can be described in terms of χ^2-distribution even in case of model misspecification. To discover change point in the quickest way, statistics is calculated in a set of running windows of different scales. Algorithm is self-tuned: critical values are justified by data and calculated with multiplier bootstrap procedure. To make the method more robust for outliers, the concept of change-point patterns is presented.

Keywords: change point detection, multiscale inference

1. Introduction

The problem of change point detection has a wide range of applications, that varies from life-critical to pure scientific ones. It appears each time one needs to explore a set of random data and make a decision about homogeneity of its structure. In other words, the problem can be stated as two following questions: were there any structural changes in the nature of observed data? At which moments, if so? These and similar questions arise in many areas of theoretical and empirical research. For example, algorithms of change point detection are used in identification and elimination of faults of aeroplane's navigation system, so as to perform better geolocation **?**. There are many other examples of real-world applications, like analysis of stock markets **?** or anomaly detection in computer traffic **?**, **?**. The general problem can be stated as follows. Let $\mathbb{Y} = (Y_1, Y_2, ..., Y_N)$ denote the data, observed till the current moment N. Underlying assumption is that the structure of $\mathbb{Y}$ is homogeneous. In other words, each Y_i has similar statistical properties and $\mathbb{Y}$ is governed by some (presumably unknown) probability law $I\!\!P_1$. One says that there is a change point if the hypothesis of homogeneity is wrong, in other words, there exists a moment τ, $1 \leq \tau \leq N$, s.t. $(Y_1, ..., Y_{\tau-1}) \backsim I\!\!P_1$ and $(Y_\tau, ..., Y_N) \backsim I\!\!P_2$, where $I\!\!P_2$ is not known as well. The moment τ is called a change-point. The goal is to find τ as precise as possible and minimise the number of false alarms and missed cases at the same time. The problem of change point detection can be easily reduced to the problem of hypothesis testing:

$$H_0 : \{Y_i\}_{i=1}^{t-1} \backsim \ I\!\!P_1, \ \{Y_i\}_{i=t}^{N} \backsim \ I\!\!P_2$$

against its "homogeneous" alternative:

$$H_1 : \{Y_i\}_{i=1}^{N} \backsim I\!\!P_1,$$

Candidate t for a change point can be selected using simple enumerative technique. Many algorithms solve the problem introducing *parametric assumption* about observed data: $I\!\!P_1, I\!\!P_2 \in (I\!\!P(\theta), \theta \in \varTheta \subseteq I\!\!R^p)$.

In this work we propose a novel parametric approach that is robust for model misspecification: it performs well even if the parametric assumption is wrong. For automatic tuning of the critical values we use multipliers bootstrap **?**, **?**. Multiscale approach allows to balance between delay in detection and false-alarm rate. Combination of these techniques allows to detect relatively small changes. To make the method more robust for outliers we introduce the concept of a change point pattern, that is described in section **??**.

The paper is organised as follows: the section **??** contains a survey of existing methods, the idea and detailed description of the method is provided in Section **??**, theoretical results are presented in Section **??** and the results of simulations and comparison with some existing models are reported in Section **??**.

2. Related work

Methods that are used for change point detection can be classified in many different ways. Below we provide several standard ones.
Retrospective and sequential methods.
This approach divides methods into two groups not by their properties, but by the area of

their application. Under *retrospective* or *offline* setting observed data set is fixed and the goal is to extract homogeneous regions. These methods are widely applicable for analysis of data that is not changing over time, e.g. images or DNA **?**. A very detailed survey of existing methods can be found in **?**. *Sequential* or *online* methods solve the problem of the *quickest* change point detection. It is assumed, that the data is aggregated from running random process. The goal is to find changes in the nature of process as soon as possible. This problem arises across many scientific areas: quality control **?**, cybersecurity **?**, **?**, econometrics **?**, **?**, geodesy e.t.c. Overview of the state-of-art methods for quickest change point detection are described in **?** or **?**.

Frequentist and Bayesian methods.
Frequentist approaches do not make any preliminary *a prior* assumption about the stochastic nature of target parameter, i.d. it is supposed to be fixed value, not a random variable, e.g. **?**, **?**. *Bayesian* change point models, on the contrary, treat parameter as random variable, e.g. **?**, **?**. These methods are quite common in bio-statistics.

Parametric and non-parametric.
All algorithms that assume observed data to obey some unknown stochastic law $\mathbb{P}_\theta$, that belongs to some known parametric family $(\mathbb{P}_\theta, \theta \in \varTheta \subseteq I\!\!R^p)$ are called *parametric*, e.g. **?**, **?**. Up-to-date survey of exiting methods and its applications can be found also in **?**. *Non-parametric* methods have more wide range of application, as they do not use any assumptions of this type, e.g.**?**, **?**. Many non-parametric methods can be found in **?**.

The concept of multiscality is, for example, exploited in **?**, **?** and **?**. It means, that observed data is analysed on different scales simultaneously. In this work we broaden the idea of multiscale change point detection proposed in **?**.

As this realm of research is developing rapidly, more and more methods combine several of described techniques, e.g. *bayesian*, *parametric* or *non-parametric* sequential change point detection **?**, **?**. There is a significant cohort of free soft-ware for researchers written in R and MatLab **?**, **?**, **?**.

3. Algorithm

Let $(\mathbb{P}(\theta), \theta \in \varTheta \subseteq \mathbb{R}^p)$ be a parametric assumption about observed data $\mathbb{Y} = (Y_1, ..., Y_N)$. For structural break detection the algorithm uses the likelihood-ratio test statistic (LRT) in a rolling window. Let $2h$ be a size of the rolling window and t be a candidate for a change point, $1 + h \leq t \leq N - h + 1$, then the hypothesis testing problem can be stated as follows:

$$H_0 : \{Y_i\}_{i=t-h}^{t-1} \backsim \mathbb{P}_1, \; \{Y_i\}_{i=t}^{t+h-1} \backsim \mathbb{P}_2$$
$$H_1 : \{Y_i\}_{i=t-h}^{t+h-1} \backsim \mathbb{P}_1,$$

where $\mathbb{P}_1 = \mathbb{P}(\theta_1^*)$, $\mathbb{P}_2 = \mathbb{P}(\theta_2^*)$, and $\theta_1^*, \theta_2^* \in \varTheta$. A possible solution is likelihood-ratio test **?**, **?**, **?**:

$$\begin{aligned} T_h(t) = & \sup_{\theta \in \varTheta} L(\theta; Y_{t-h}, ..., Y_{t-1}) + \sup_{\theta \in \varTheta} L(\theta; Y_t, ..., Y_{t+h-1}) \\ & - \sup_{\theta \in \varTheta} \{L(\theta; Y_{t-h}, ..., Y_{t-1}) + L(\theta; Y_t, ..., Y_{t+h-1})\}, \end{aligned}$$

$L(\theta; \cdot)$ is a log-likelihood function.

The statistic $T_h(\cdot)$ behaves as the one having a non-central χ_p^2 distribution if the hypothesis H_0 is not rejected. Otherwise, $T_h(\cdot)$ is distributed as χ_p^2, where p is a dimension of the parametric space.

Proposition 1 *Under conditions* **??**, **??**, *it is hold with probability* $1 - 4e^{-x}$, *that*

$$|\sqrt{2T_h(t)} - \|\xi_t + b_h(t)\|| \leq 10\Diamond(r, x),$$

where $\xi_t \backsim \mathcal{N}(0, I_p)$, I_p *is an identity matrix,*

$$b_h(t) = \begin{cases} \varSigma_h |\theta_1^* - \theta_2^*|, & \textit{if } H_0, \\ \textit{0}, & \textit{if } H_1; \end{cases}$$

$b_h(t)$ *is a systematic drift component;* $\varSigma$ *is a positive matrix,* θ_1^* *and* θ_2^* *are target parameters before and after the structural change respectively.*

Proposition **??** is proved in Section **??**.

Instead of analysing a single value of the statistic $T_h(t)$ at a single point of interest t, we do so for all points inside of a running window: $\mathbb{T}_h(t) = (T_h(t-h), ..., T_h(t+h-1))$. The vector $\mathbb{T}_h(t)$ has the same geometric properties as a vector ${}_h(t) = (b_h(t-h), ..., b_h(t+h-1))$ has. This vector forms a type of a pattern depending on a type of a change. It is referred to as a *change-point-pattern*. Several illustrations of the concept are presented in Fig.**??** - **??**.

The Fig.**??** shows a triangle pattern. It is produced by an abrupt change in a parameter of an observed signal. The jump takes place at the moment $t = 200$. Until that moment, the signal is distributed according to $\mathbb{P}(\theta_1^*) = N(30, 5)$ and after the jump it is obeyed to $\mathbb{P}(\theta_2^*) = N(35, 5)$, $\theta_1^* = 30$, $\theta_2^* = 35$. A smooth transition from θ_1^* to θ_2^* leads to the

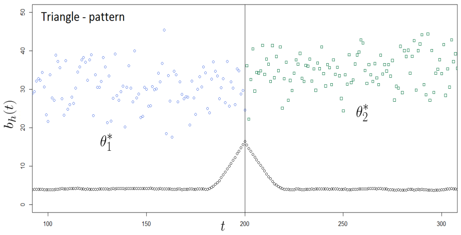


Figure 1: Behaviour of $b_h(t)$ in case of abrupt change point

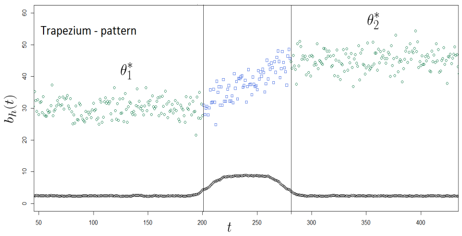


Figure 2: Behaviour of $b_h(t)$ in case of smooth transition

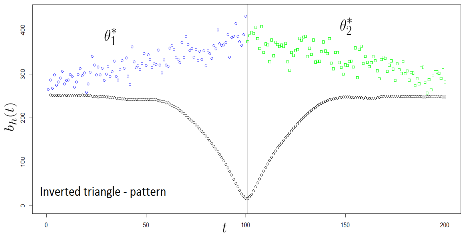


Figure 3: Behaviour of $b_h(t)$ in case of break in trend

formation of trapezium pattern. In Fig.**??** is presented the transition between two Gaussian random processes: from $N(30, 5)$ to $N(45, 5)$, $\theta_1^* = 30$, $\theta_2^* = 45$. The last pattern (Fig.**??**) is an inverted triangle. It appears because of a change point in a piece-wise linear regression model. Namely, $Y_i = i + \varepsilon_i$ before the change point and $Y_i = (-i + 100) + \varepsilon_i$ after the change point, where $\varepsilon_i \backsim N(0, 25)$ is normally distributed homogeneous noise.

The distinguishability of the change point pattern depends on the window size $2h$ and the size of a change point $|\theta_1^* - \theta_2^*|$. In other words, for each size of a change point exists the optimal scale h_{opt}, s.t. for each $h < h_{opt}$ the pattern is poorly visualised. The detailed description of the h_{opt}-concept can be found in Section **??**. An example is presented in Fig.**??**. Gaussian data with abrupt change in mean at the moment $\tau = 342$ is depicted at the first line. The next two panels show test statistics $T_{2h=4}(t)$ and $T_{2h=10}(t)$ respectively. In both cases the change point pattern is still uncertain. The last test statistic $T_{2h=20}(t)$ forms a distinguishable triangle. For this example $5 < h_{opt} \leq 10$. The procedure monitors not the changes in statistic $T_h(t)$ itself, but in its convolution with a target *change-point*

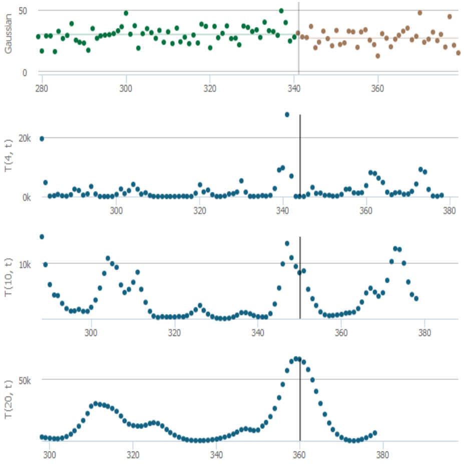


Figure 4: Pattern on different scales

pattern P_h. This makes algorithm less sensitive for outliers.

$$\widehat{T}_h(t) = \langle \mathbb{T}_h(t), P_h \rangle$$

where

$$\mathbb{T}_h(t) = (T_h(t-h), ..., T_h(t+h-1)).$$

Abrupt change in an observed parameter entails triangle-like behaviour of test statistics $T_h(\cdot)$. Therefore, in this case the target pattern P_h is an isoscales triangle with a unit height and a base equal to $2h$.

Naturally, the most desirable way is to use only the smallest window $2h_{opt}$. In practice, it is not possible, as it was mentioned before, $2h_{opt}$ depends on parameters before and after structural break: $h_{opt} = h_{opt}(\theta_1^*, \theta_2^*)$. Here we propose a *multiscale* approach. It implies that the statistic $\widehat{T}_h(t)$ is computed simultaneously in running windows of different sizes $H = (h_1, ..., h_n)$. This allows algorithm to estimate the smallest window size $2h_{opt}$ automatically: exists $h_k \in H$, s.t. $h_k < h_{opt} \leq h_{k+1}$. Critical values $\{z(h)\}_{h \in H}$ for $\{\widehat{T}_h(\cdot)\}$ are computed using the multiplier bootstrap. Under this approach, introduction of multiscality entails the problem of multiplicity correction. To overcome it we use synchronisation technique. The detailed description of the multiplier bootstrap procedure and the synchronisation are presented in **?**.

The presented algorithm can be applied for both sequential and retrospective settings. Under *online* framework, it marks a time moment τ as a change point, if test statistics $\widehat{T}_h(\tau + h)$ exceeds critical value $z(h)$ at the moment $\tau + h$:

$$\{\tau : \widehat{T}_h(\tau - h) > z(h)\}.$$

This means, that the smallest delay of detection is h_{min}.

Under *offline* setting, τ is a change point if

$$\{\tau = \operatorname*{argmax}_{t \in \{1,...,N\}} \widehat{T}_h(t)\}.$$

The greater number k of such scales $h'_{i_1}, ..., h'_{i_k}$ where τ is marked as change point, the more sure algorithm is, that τ the *true* change point is.

4. Theoretical results

4.1. LRT statistic

This section presents main results that describe theoretical properties of the likelihood-ratio statistics (LRT). They are essential for the proposed algorithm of change point detection. Further assume that log-likelihood function $L(\theta) = L(Y, \theta)$ is well approximated by its quadratic part in local region $\varTheta_0(r)$ of θ^*, $\varTheta_0(r) \subseteq I\!R^p$, where

$$\theta^* = \operatorname*{argmax}_{\theta} I\!E L(\theta), \quad \widehat{\theta} = \operatorname*{argmax}_{\theta} L(\theta)$$

and $\varTheta_0(r) = \{\|D(\theta - \theta^*)\| < r\}$. Conditions for justified quadratic approximation and parameter concentration in the local region are presented in **?**.

Define variables that describe the approximation error:

$$\alpha(\theta, \theta_0) = L(\theta) - L(\theta_0) - (\theta - \theta_0)^T \nabla L(\theta_0) + \frac{1}{2}\|D(\theta - \theta_0)\|^2,$$

$$\chi(\theta, \theta_0) = D^{-1}\nabla\alpha(\theta, \theta_0) = D^{-1}(\nabla L(\theta) - \nabla L(\theta_0)) + D(\theta - \theta_0).$$

Assume that two following inequalities are fulfilled in region $\varTheta_0(r)$ with probability $1 - e^{-x}$:

$$\frac{|\alpha(\theta, \theta^*)|}{\|D(\theta - \theta^*)\|} \le \diamondsuit(r, x), \quad \|\chi(\theta, \theta^*)\| \le \diamondsuit(r, x), \tag{A}$$

where $\diamondsuit(r, x) = (\delta(r) + 6 v_0 z_H(x)\omega) r$,

$$D^2(\theta) = -\nabla^2 I\!E L(\theta), \quad D = D(\theta^*), \tag{D}$$

$$\|D^{-1} D^2(\theta) D^{-1} - I_p\| \le \delta(r), \tag{L}$$

$$\forall \lambda \le g,\ \gamma_1 \gamma_2 \in I\!R^p : \quad \log I\!E \exp\left\{ \frac{\lambda}{\omega} \frac{\gamma_1^T \nabla^2 \overset{o}{L}(\theta) \gamma_2}{\|D\gamma_2\| \|D\gamma_2\|} \right\} \le \frac{\nu_0^2 \lambda^2}{2}, \tag{ED2}$$

$$z_H(x) = \sqrt{H} + \sqrt{2x} + \frac{g^{-2} x + 1}{g} H, \quad H = 6p.$$

Condition **??** ensures quadratic approximation of $I\!E L(\theta)$ and **??** ensures linear approximation of centered likelihood $\overset{o}{L}(\theta) = L(\theta) - I\!E L(\theta)$.

Firstly provide simple non-strict explanation of what kind of distribution the main statistic T_h is supposed to have. Review T_h as

$$T_h = L(\widehat{\theta}) - L(\widehat{\theta}_{H_0}), \quad L(\theta) = L_1(\theta_1) + L_2(\theta_2), \quad L_1 = L(Y[1:h]),\ L_2 = L(Y[h:2h]),$$

where $\widehat{\theta}_{H_0}$ is argmax of L under condition $H_0 : \theta_1 = \theta_2$. Then under quadratic approximation assumption T_h could be presented in Tailor equation with point $\widehat{\theta}$:

$$T_h \approx \frac{1}{2}\|D(\widehat{\theta} - \widehat{\theta}_{H_0})\|^2.$$

If $\widehat{\theta}$ and $\widehat{\theta}_{H_0}$ tend to be Normal and H_0 is true then their difference are close to a centered Normal variable. If H_0 is false – the Normal variable will have mean that is equal to $\theta^* - \theta^*_{H_0}$.

Obtain strict equation for LRT statistic distribution in quadratic model case.

$$L(\theta) = L_1(\theta) + L_2(\theta) = L_1(\widehat{\theta}_1) + L_2(\widehat{\theta}_2) - \frac{1}{2}(\theta - \widehat{\theta}_1)^T D_1^2 (\theta - \widehat{\theta}_1) - \frac{1}{2}(\theta - \widehat{\theta}_2)^T D_2^2 (\theta - \widehat{\theta}_2) =$$

$$= L(\widehat{\theta}) - \frac{1}{2}(\theta - \widehat{\theta})^T D^2 (\theta - \widehat{\theta}),$$

$$\widehat{\theta} = D^{-2}(D_1^2 \widehat{\theta}_1 + D_2^2 \widehat{\theta}_2), \quad D^2 = D_1^2 + D_2^2.$$

$$T_h = L_1(\widehat{\theta}_1) + L_2(\widehat{\theta}_2) - L(\widehat{\theta}) =$$

$$= \frac{1}{2}(\widehat{\theta} - \widehat{\theta}_1)^T D_1^2 (\widehat{\theta} - \widehat{\theta}_1) + \frac{1}{2}(\widehat{\theta} - \widehat{\theta}_2)^T D_2^2 (\widehat{\theta} - \widehat{\theta}_2).$$

$$\widehat{\theta} - \widehat{\theta}_1 = D^{-2}(D_1^2 \widehat{\theta}_1 + D_2^2 \widehat{\theta}_2) - \widehat{\theta}_1 = D^{-2} D_2^2 (\widehat{\theta}_2 - \widehat{\theta}_1),$$

$$\widehat{\theta} - \widehat{\theta}_2 = D^{-2}(D_1^2 \widehat{\theta}_1 + D_2^2 \widehat{\theta}_2) - \widehat{\theta}_2 = D^{-2} D_1^2 (\widehat{\theta}_1 - \widehat{\theta}_2).$$

$$2T_h = (\widehat{\theta}_2 - \widehat{\theta}_1)^T \varSigma^2 (\widehat{\theta}_2 - \widehat{\theta}_1),$$

where

$$\varSigma^2 = D_2^2 D^{-2} D_1^2 D^{-2} D_2^2 + D_1^2 D^{-2} D_2^2 D^{-2} D_1^2 = D_1^2 D^{-2} D_2^2 \approx \frac{1}{4} D^2. \tag{S}$$

Make replacement of $\widehat{\theta}_2$, $\widehat{\theta}_1$ in the equation for T_h with regard to condition $\chi(\theta, \theta^*) = 0$ in quadratic model using following equations:

$$D_1(\widehat{\theta}_1 - \theta_1^*) = \xi_1 = D_1^{-1} \nabla L(\theta_1^*), \quad D_2(\widehat{\theta}_2 - \theta_2^*) = \xi_2 = D_2^{-1} \nabla L(\theta_2^*).$$

Present generalized result for non-quadratic model.

Theorem 2 *Assume condition (**??**) and quadratic Laplace approximation (**??**) of L_1 and L_2 are fulfilled with probability $1 - 2e^{-x}$, additionally with probability $1 - 2e^{-x}$*

$$\|\xi_i\| \le z(x), \quad z^2(x) = \max_i p_{B_i} + 6\lambda_{B_i} x,$$

$$B_i = D_i^{-1} \operatorname{Var}(\nabla L_i(\theta^*)) D_i^{-1}, \quad p_B = \operatorname{tr}(B), \quad \lambda_B = \lambda_{\max}(B). \tag{B}$$

Then in the local region with probability $1 - 8e^{-x}$

$$2T_h = \|\xi_{12} + \theta_{12}^*\|^2 + O(\{r + z(x)\}\diamondsuit(r, x)),$$

where

$$\xi_{12} = \varSigma(D_2^{-1}\xi_2 - D_1^{-1}\xi_1), \quad \theta_{12}^* = \varSigma(\theta_2^* - \theta_1^*).$$

Remark 3 *In increasing sample size $n \to \infty$ the stochastic component tends to Normal distribution:*

$$\xi_{12} \to \mathcal{N}(0, B_1 + B_2).$$

Remark 4 *For the condition $\widehat{\theta} \in \Theta_1(r) \cap \Theta_2(r)$ the restriction of the parameter variability θ^* is required*

$$\|D(\theta_1^* - \theta_2^*)\| \le r. \tag{L*}$$

Prove a similar statement (theorem **??**) for statistic $\sqrt{2T_h}$. From condition (**??**) one get with probability $1 - 2e^{-x}$

$$\left| T_h(\widehat{\theta}_1, \widehat{\theta}_2) - \frac{1}{2}\|\Sigma(\widehat{\theta}_2 - \widehat{\theta}_1)\|^2 \right| \le$$

$$\le 2\|D_1(\widehat{\theta}_1 - \widehat{\theta})\|\diamondsuit(r,x) + 2\|D_2(\widehat{\theta}_2 - \widehat{\theta})\|\diamondsuit(r,x) \le 4\|\Sigma(\widehat{\theta}_2 - \widehat{\theta}_1)\|\diamondsuit(r,x).$$

Use inequality $|a-b| \le |a^2 - b^2|/b, \ b > 0$:

$$\left| \sqrt{2T_h(\widehat{\theta}_1, \widehat{\theta}_2)} - \|\Sigma(\widehat{\theta}_2 - \widehat{\theta}_1)\| \right| \le 8\diamondsuit(r,x).$$

Replace $(\widehat{\theta}_1, \widehat{\theta}_2)$ with $(D_1^{-1}\xi_1 + \theta_1^*, \ D_2^{-1}\xi_2 + \theta_2^*)$:

$$\left| \|\Sigma(\widehat{\theta}_2 - \widehat{\theta}_1)\| - \|\xi_{12} + \theta_{12}^*\| \right| \le$$

$$\le \|\Sigma(\widehat{\theta}_1 - \theta_1^*) - \Sigma D_1^{-1}\xi_1\| + \|\Sigma(\widehat{\theta}_2 - \theta_2^*) - \Sigma D_2^{-1}\xi_2\| \le 2\diamondsuit(r,x).$$

Summarize:

Theorem 5 *Assume condition (**??**) and quadratic Laplace approximation (**??**) with probability $1 - 2e^{-x}$ are fulfilled. Then with probability $1 - 4e^{-x}$ in the local region $\Theta_1(r) \cap \Theta_2(r)$ took place*

$$\left| \sqrt{2T_h} - \|\xi_{12} + \theta_{12}^*\| \right| \le 10\diamondsuit(r,x).$$

*where ξ_{12} and θ_{12}^* are defined in theorem **??**.*

Remark 6 *The constant near $\diamondsuit(r,x)$ could be decreased, expanding series of $L_1(\theta)$, $L_2(\theta)$ and $L(\theta)$ in the local regions around θ_1^*, θ_2^* and θ^* instead of MLE values:*

$$2T_h = -\|\xi\|^2 + \|\xi_1\|^2 + \|\xi_2\|^2 - 2\xi_1^T D_1 D^{-2} D_2^2(\theta_2^* - \theta_1^*) + 2\xi_2^T D_2 D^{-2} D_1^2(\theta_2^* - \theta_1^*) +$$

$$+\|D_1 D^{-2} D_2^2(\theta_2^* - \theta_1^*)\|^2 + \|D_2 D^{-2} D_1^2(\theta_2^* - \theta_1^*)\|^2 \pm (2\diamondsuit(r,x)r + 2\delta(r)r^2) =$$

$$= -\|\xi\|^2 + \|\xi_1\|^2 + \|\xi_2\|^2 + 2(D_2^{-1}\xi_2 - D_1^{-1}\xi_1)^T \Sigma^2(\theta_2^* - \theta_1^*) + +\|\Sigma(\theta_2^* - \theta_1^*)\|^2$$

$$\pm(2\diamondsuit(r,x)r + 2\delta(r)r^2).$$

*Replace $\|\xi\|^2$ with $\|D^{-1}(D_1\xi_1 + D_2\xi_2)\|^2 \pm 2\diamondsuit(r,x)z(x)$, referring to condition **??**.*

$$-\|\xi\|^2 + \|\xi_1\|^2 + \|\xi_2\|^2 = \|\Sigma(D_2^{-1}\xi_2 - D_1^{-1}\xi_1)\|^2 \pm 2\diamondsuit(r,x)z(x).$$

That leads to result

$$\left| 2T_h - \|\xi_{12} + \theta_{12}^*\|^2 \right| \le (4\diamondsuit(r,x)r + 2\delta(r)r^2).$$

4.2. Optimal window size

The CP detection algorithm described above has rather meaningful parameter window size (h) that determines sample sizes on which MLE ($\widehat{\theta}_1$, $\widehat{\theta}_2$) will be compared. Find out the minimal required sample size from the condition

$$h\mathcal{KL}(\theta_1^*, \theta_2^*) > h\mathcal{KL}(\widehat{\theta}_1, \theta_1^*) + h\mathcal{KL}(\widehat{\theta}_2, \theta_2^*).$$

From Wilks theorem (reg. **?**) one could obtain approximation with probability $1 - 10e^{-x}$

$$h\mathcal{KL}(\widehat{\theta}_1, \theta_1^*) + h\mathcal{KL}(\widehat{\theta}_2, \theta_2^*) \leq 2r\diamondsuit(r,x) + \frac{\|\xi_1\|^2}{2} + \frac{\|\xi_2\|^2}{2},$$

where with probability $1 - 4e^{-x}$

$$\frac{\|\xi_1\|^2}{2} + \frac{\|\xi_2\|^2}{2} \leq z^2(x) = p_B + 6\lambda_B x,$$

Consider the case with

$$r\diamondsuit(r,x) = \sqrt{\frac{C(p_B + x)^3}{h}}, \quad h > C(p_B + x),$$

that leads to lower bound estimation

$$h\mathcal{KL}(\theta_1^*, \theta_2^*) > 3p_B + (6\lambda_B + 2)x.$$

Why the optimal h should be limited? Increasing a sample size one decreases an impact of stochastic part of $\|\xi_{12} + \theta_{12}^*\|$ since $\|\theta_{12}^*\|$ grows. But at the same time $\|\theta_{12}^*\|$ will not be changed with window replacement when $h \to \infty$. Note also that angle of $\|\theta_{12}^*\|$ growth decreases with h, so the optimal window size is the smallest one that is sufficient to overcome random fluctuations in convolution of $\|\xi_{12}(i) + \theta_{12}^*(i)\|$ with linear function $f(i) = i$. Define new variables

$$b = \|\theta_{12}^*\| = \sqrt{h} b_0, \quad b_i = \frac{i}{h} b,\ i > 0, \quad \xi_i = \xi_{12}(i).$$

Optimal window size for online CP detection is to be derived from the following inequality.

$$\sum_{i=1}^{h} i\|\xi_i + b_i\| \geq \sum_{i=1}^{h} i\Big(\|\xi_i\| + 10\diamondsuit(r,x)\Big).$$

Use theorem 4.1 from paper **?** that ensures following inequality with probability $1 - 2e^{-x}$

$$\|\xi_i + b_i\| \geq \sqrt{\|\xi_i\|^2 + \|b_i\|^2 - 2\|b_i\| - 2\delta_1(x)} \geq$$

$$\geq \|b_i\| - 2 - \sqrt{4 + 2\delta_1(x)}.$$

With probability $1 - 4e^{-x}$ under condition that statement from theorem **??** is true come to a final estimation of the minimal sufficient window size:

$$h \geq \frac{9(2 + \sqrt{4 + 2\delta_1(x)} + z(x) + 10\diamondsuit(r,x))^2}{4b_0^2} \sim \frac{c_1 + c_2 p}{b_0^2}.$$

5. Experiments

5.1. Experiments with synthetic data

This section presents results of the comparison of the proposed algorithm of change point detection (referred as *LRTOnline* or *LRTOffline*) with two other methods: *Bayesian online changepoint detection (BOCPD)* **?** and *cpt.meanvar(PELT,...)* (RMeanVar) from **?**. The first method is constructed for online inference, but so far as it returns CP location with each CP signal, it is also applicable for offline testing scenario. It is based on idea of predictive filtering: the method forecasts a new data point using only the information have been observed already, where the distribution family is fixed (Normal for the tests in this paper). The length of the observed data (from the last CP) is calculated by Bayesian inference. The second algorithm uses the model that is preliminary specified by user. It is designed for finding multiple changes in mean and variance in Normally (another distributions also supported) distributed data. The returned set of change points is the result of sequential testing H_0 (existing number of change points) against H_1 (one extra change point) applying the likelihood ratio statistic of the whole data coupled with the penalty for CP count. Originally it is applied for offline change point detection, but one could adapt such method for online case by buffering incoming data elements and clearing the buffer when at least one CP have been observed in the buffered data. In total each of these two algorithms was modified in the way that allows one to use it in both online and offline testing mode.

LRTOffline configuration:
window sizes $(h_1, \dots, h_W) = (10, 20, 40, 70)$; confidence for the upper bound of convolution with pattern = 0.1; window weights $(u_1, \dots, u_W) = (1.0, 2.0, 0.5, 0.2)$.

LRTOnline configuration:
window sizes $(h_1, \dots, h_W) = (30, 50, 70)$; confidence = 0.1.

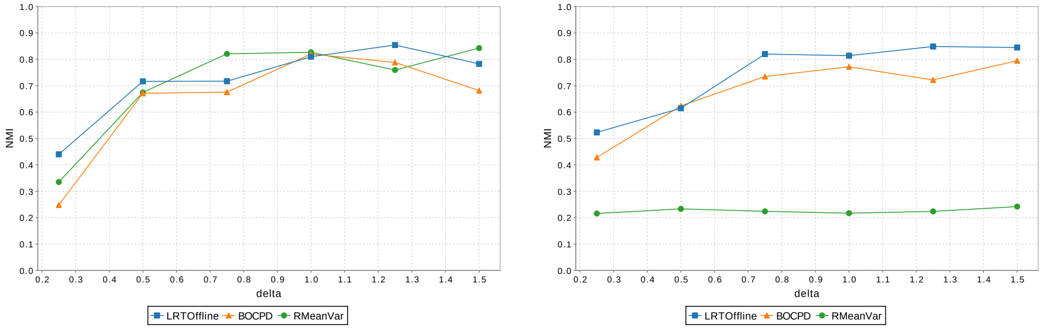
To measure the quality use three following metrics: Normalised Mutual Information (NMI), Delay (average time interval in which CP have been detected after it had taken place), Precision and Recall. NMI measure of two partitions (X, Y) of time range by change points is defined as

$$\mathrm{NMI}(X, Y) = 2\frac{H(X) + H(Y) - H(X, Y)}{H(X) + H(Y)}.$$

Higher NMI values (they are in $[0, 1]$) correspond to better quality. NMI measure is applied for quality comparison in offline case, for online – Delay, Precision and Recall.

Synthetic test data have been generated for different values of difference norm of the data distribution parameter. Such values are denoted as *delta*. For each delta 10 data sequences have been sampled over which one compute measure average. In online mode each data sequence could have one or none change points, in offline mode – two, one or none change points.

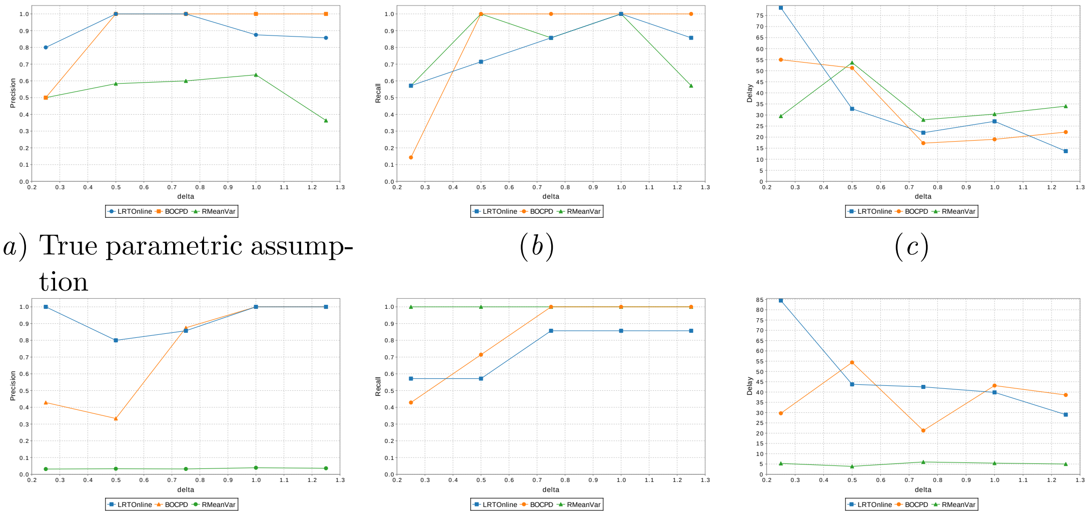
In the offline tests with Normal data all the methods achieves similar NMI scores, nonetheless LRTOffline is more stable for decreasing strength of CP (delta). In the tests with Poisson data (misspecification) RMeanVar has relatively low quality. In the the online tests the proposed method (LRTOnline) shows more conspicuous stability along different delta values what is accomplished by multiscale heuristic.



(a) True parametric assumption

(b) Wrong parametric assumption

Figure 5: Offline mode. First data: $\mathcal{N}(\theta(1), \theta(2))$, second data: $\mathrm{Po}(\theta)$, delta $= \|\theta^*_{12}\|$, data size $= 340$, PA for all methods is $\mathcal{N}(\theta(1), \theta(2))$, NMI – Normalized Mutual Information between predicted and reference partitions of time interval with change points, tests per delta $= 10$, change point per test $= \{0, 1, 2\}$.



(a) True parametric assumption

(b)

(c)

(d) Wrong parametric assumption

(e)

(f)

Figure 6: Online mode. First row data: $\mathcal{N}(\theta(1), \theta(2))$, second row data: $\mathrm{Po}(\theta)$, delta $= \|\theta^*_{12}\|$, data size $= 340$, PA for all methods is $\mathcal{N}(\theta(1), \theta(2))$, tests per delta $= 10$, change point per test $= \{0, 1\}$.

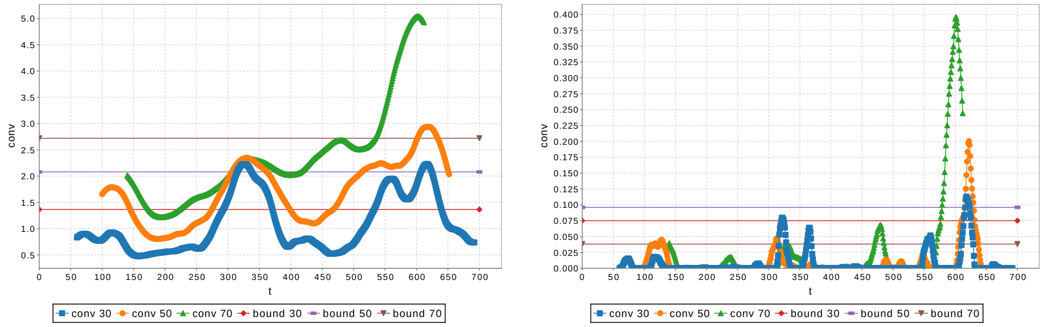
During the experiments following meaningful properties of the proposed method configuration were noted:

1. Quality is sensitive to selection of interval for upper bound calibration of convolution in offline mode. For example in data $\mathcal{N}(0,1)$.sample(100) $\cup$ $\mathcal{N}(1,2)$.sample(100) is preferable to use only slice of 0 to 100 elements for calibration, because of lower Var ξ_{12}. Generally one should find change points in $tr(B_1 + B_2)$ according to remark **??** from section **??** and run calibration in the range with the lowest values of $tr(B_1 + B_2)$. This improvement additional requires approximation of the convolution maximum in larger data ranges.

2. It is influenced to find out the minimal h sufficient for bootstrap usage. Small h improves Delay but makes unable to keep high level of Precision and Recall in online mode.

5.2. Experiments with real data

Here data from 1972-75 Dow Jones Returns is taken **?** that describes several major events with potential macroeconomic effects (the most significant among them are the Watergate affair and the OPEC oil embargo). Convolutions plot with its upper bounds on this dataset appeared to be a nice illustration of multiscale search importance: CP near $t = 325$ is better perceptible when window size is equal to 30 and CP near $t = 600$ has more perceptible convolution when window size is equal to 70. Two plots presented below includes convolutions with Static and Fitted Patterns, where one could remark better separability of convolution peaks in fitted case.



(a) Static pattern (b) Fitted pattern

Figure 7: Data: daily returns of the Dow Jones Industrial Average from July 3, 1972 to June 30, 1975. Left plot – convolution with static triangle pattern; right plot – convolution with fitted triangle pattern. The time axis is in business days, conv 30 (50, 70) corresponds to pattern with window size 30 (50, 70), bound 30 (50, 70) corresponds to convolution upper bound. Three reference CP are presented: the conviction of G. Gordon Liddy and James W. McCord, Jr. on January 30, 1973 ($t = 142$); the beginning of the OPEC embargo against the United States on October 19, 1973 ($t = 325$); the resignation of President Nixon on August 9, 1974 ($t = 548$).

5.3. Sources

Demo of the LRTOnline method is available by link: demos.wias.de/cpd.

Scala project with LRTOffline and LRTOnline methods could be cloned from https://github.com/nazarblch/cpd, which also includes testing system for binary CP detection applications and generated data used in the experiments.

6. Conclusion

The conclusion goes here.