Nonparametric method for change-point detection based on empirical likelihood

Dan-Ah Kim[1)], Sang-Un Park[2)] (Department of Applied Statistics, Yonsei University)

1) Graduate Student, Department of Applied Statistics, Yonsei University
2) Professor, Department of Applied Statistics, Yonsei University

Abstract

Summary: The change-point problem is detecting a location of change-point in data depending on time. In this paper, the nonparametric method based on empirical likelihood using quantile is proposed for detecting a change-point in distributions for independent random variables. We expand the constraints of empirical likelihood to the double quantiles. The proposed method demonstrates the better performance in comparison to the traditional nonparametric tests and recent methods through simulation. The proposed method is robust regardless of whether the observations are from the same distributions.

Keywords: Change-point, Empirical likelihood, Likelihood ratio test, Nonparametric test, Quantile

Description

Backgrounds of Change-point Problem

- In a time series, a change-point is the point in time when the statistical properties of the underlying process change.
- In finance and manufacturing, detecting the number of change-points and their locations are important issue.
- W. A. Shewhart (1931) first invented Control Chart.
- E. S. Page (1954) suggested CUSUM to monitor change detection.

Change-point Problem

- Consider a sequence of observations $x_1, x_2, \dots, x_n$ drawn from independent random variables $X_1, X_2, \dots, X_n$
- Multiple m change points $\tau_1, \tau_2, \dots, \tau_m$ exist in the data ⇒ *(m+1)* segments
- Then the distribution of the sequence can be written as

$$X_i \sim \begin{cases} F_1 & if\ i \le \tau_1 \\ F_2 & if\ \tau_1 < i \le \tau_2 \\ & \dots \\ F_{m+1} & if\ \tau_m < i \end{cases}$$

Change-point Model

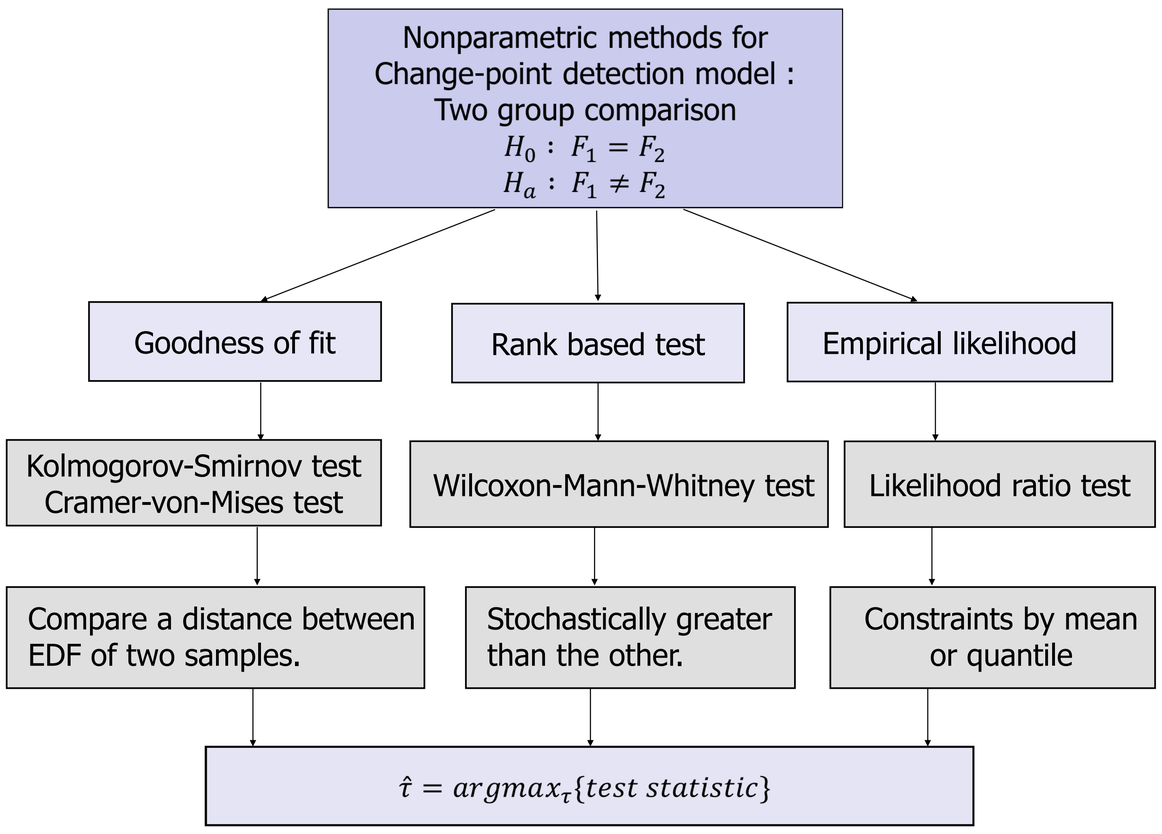
- Consider independent random variables $X_1 \sim G_1, \dots, X_n \sim G_n$.
- Assume that there is at most one change τ in the sequence of distributions above. We want to test the null hypothesis of no change

$$H_0: G_1 = G_2 = \dots = G_n = F,$$

against the following alternative of one change

$$H_a: F_1 = G_1 = G_2 = \dots = G_\tau \neq G_{\tau+1} = \dots = G_n = F_2.$$

- Using binary segmentation, it suffices to test and estimate the position of a single change point at each stage sequentially.
- Zou, C., Liu, Y., Qin, P., and Wang, Z. (2007): Suggest Empirical likelihood ratio test for the change-point problem.
- Zhou, Y., Fu, L., and Zhang, B. (2017): Based on two sample quantile empirical likelihood. (written as follows 'QLR')
- Hence, this paper is motivated to expand empirical likelihood to **double quantile** for the both extreme side.



Empirical Likelihood for Change-point

- Empirical likelihood is a nonparametric method first introduced by Owen(1988).
- Empirical likelihood function of $\{p_i\}_{i=1}^n$ is defined as L(F) = $\prod_{i=1}^n p_i$ and $L(F)$ is maximized at $p_i = 1/n$ and attains maximum n^{-n} under the full nonparametric model.
- When a population parameter θ identified by $E[m(X;\theta)] = 0$ is of interest, the empirical likelihood maximum when θ has the true value θ_0 is obtained subject to the additional constraint $\sum_{i=1}^n p_i m(x_i, \theta_0) = 0$.
- Two samples: $X_1, X_2, \dots, X_n \sim F_1$ and $Y_1, Y_2, \dots, Y_m \sim F_2$ and let $p_i = P(X = x_i)$ and $q_j = P(Y = y_j)$.
- Empirical likelihood function of $\{p_i\}_{i=1}^n$, $\{q_j\}_{j=1}^m$ is defined as $L(F) = \prod_{i=1}^n p_i \prod_{j=1}^m q_j$.

Double Quantile Likelihood Ratio for Two Sample (DLR)

- Proposed methodology: Double quantile likelihood by expanding expanding constraints to two quantiles for the both extreme side. (written as follows 'DLR')
- Let $p = F(\xi_p)$ and $1 - q = F(\xi_{1-q})$; hence, ξ_p is the p quantile of F and ξ_{1-q} is the *1-q* quantile of F. This satisfies

- We can construct the following quantile empirical likelihood test statistic under restriction, $E[I(X_i \le \xi_p) - p] = 0, \quad E[I(X_i \ge \xi_{1-q}) - q)] = 0$

$$\mathbf{R}(\xi_\mathbf{p}, \xi_{\mathbf{1-q}}) = \max \left\{ \prod_{i=1}^n np_i \prod_{j=1}^m mq_j \middle| \sum_{i=1}^n p_i I(X_i \le \xi_p)) = p, \right.$$
$$\left. \sum_{j=1}^m q_j I(Y_j \le \xi_p)) = p, \sum_{i=1}^m q_i I(X_i \le \xi_{1-q})) = 1 - q, \right.$$
$$\left. \sum_{j=1}^m q_j I(Y_j \le \xi_{1-q})) = 1 - q, p_i, q_j \ge 0, \sum_{i=1}^n p_i = \sum_{j=1}^m q_j = 1 \right\}$$
$$\therefore D_n = \sup_{\xi_p < \xi_{1-q}} \{-2 \log \mathbf{R}(\xi_\mathbf{p}, \xi_{\mathbf{1-q}})\}$$

- Using Lagrange multipliers to solve $\mathbf{R}(\xi_\mathbf{p}, \xi_{\mathbf{1-q}})$, we can get following unique $\lambda's$ and $\{p_i\}_{i=1}^n$, $\{q_j\}_{j=1}^m$.(Proof in appendix)
- When n or m is too small, the empirical likelihood estimators of $\lambda's$ may not exist. Therefore, use a trimmed statistic (Zou, C. (2007))

$$D_n^* = \sup_{c(n+m)^{-1/9} < \xi_p < \xi_{1-q} < 1 - c(n+m)^{-1/9}} \{-2 \log \mathbf{R}(\xi_\mathbf{p}, \xi_{\mathbf{1-q}})\} \text{ where } c > 0.$$

- Large values of D_n^* indicate that there is at least one change-point.
- The location τ can be estimated by $\hat{\tau} = \arg_\tau \max\{D_n^*\}$

Data Application

- Data application using Double quantile likelihood with a real data on Nile dataset.

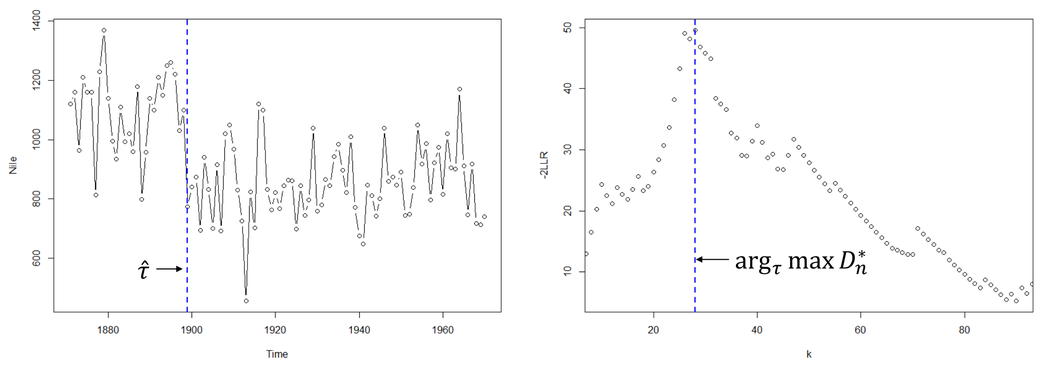


Figure 1. Double quantile empirical likelihood ratio test with Nile dataset: Nile dataset (right panel) and D_n^* (left panel). Change-point $\hat{\tau}$ (blue line) correspond to the maximum DLR(D_n^*).

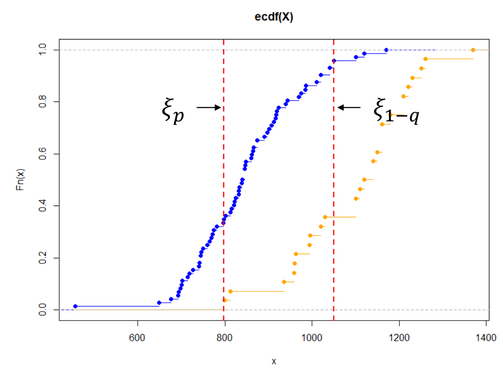


Table 1. Description of data application result

$\hat{\tau}$	-2LLR	ξ_p	ξ_{1-q}
	49.616	797	1050
28	p-value	$\hat{p}$	$\hat{q}$
	1.87e-12	0.25	0.21

Figure 2. EDF of F_1 (orange line) and F_2 (blue line) (left panel) and estimated value (right panel). ξ_p and ξ_{1-q} are red dotted line.

Simulation Results

- The proposed method demonstrates the better performance.
- The proposed method is robust regardless of whether the observations are from the same distributions. (Simulation results of test power under significance level using bootstrap 100,000 times.)

Table 2. Accuracy rate from two samples from different distribution of mean difference=2

$F_1 \sim Exp(3),\ F_2 \sim \chi^2(5)$

cpt	DLR	QLR	MWM	KS	CvM
25%	0.87	0.88	0.77	0.25	0.73
50%	0.89	0.86	0.94	0.01	0.87
75%	0.88	0.86	0.72	0.19	0.81
95%	0.83	0.9	0.59	0.89	0.13

$F_1 \sim Exp(3),\ F_2 \sim poisson\ (5)$

cpt	DLR	QLR	MWM	KS	CvM
25%	0.96	0.91	0.82	0.96	0.88
50%	0.98	0.98	0.91	0	0.88
75%	0.91	0.94	0.75	0.88	0.85
95%	0.86	0.91	0.62	0.83	0.07

Table 3. Power at a test level $\alpha = 0.05$ of two samples from different distribution of mean difference=2

$F_1 \sim Exp(3),\ F_2 \sim \chi^2(5)$

cpt	DLR	QLR	MWM	KS	CvM
25%	0.214	0.213	0.164	0.743	0
50%	0.216	0.197	0.178	0.753	0
75%	0.181	0.196	0.184	0.744	0
95%	0.2	0.186	0.17	0.745	0

$F_1 \sim Exp(3),\ F_2 \sim poisson\ (5)$

cpt	DLR	QLR	MWM	KS	CvM
25%	0.311	0.278	0.24	0.766	0.003
50%	0.307	0.277	0.243	0.745	0.001
75%	0.302	0.274	0.237	0.76	0.006
95%	0.311	0.269	0.224	0.746	0.002