Nonparametric method for changepoint detection using Empirical likelihood

Danah Kim, Sangun Park

Department of Applied Statistics
Yonsei University

October 22, 2019

Table of Contents

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 many practical situations, a statistician is faced with the problem of detecting the number of change-points and their locations.
- W. A. Shewhart (1931) fisrt invented Control Chart.
- E. S. Page (1954) suggested CUSUM to monitor change detection.

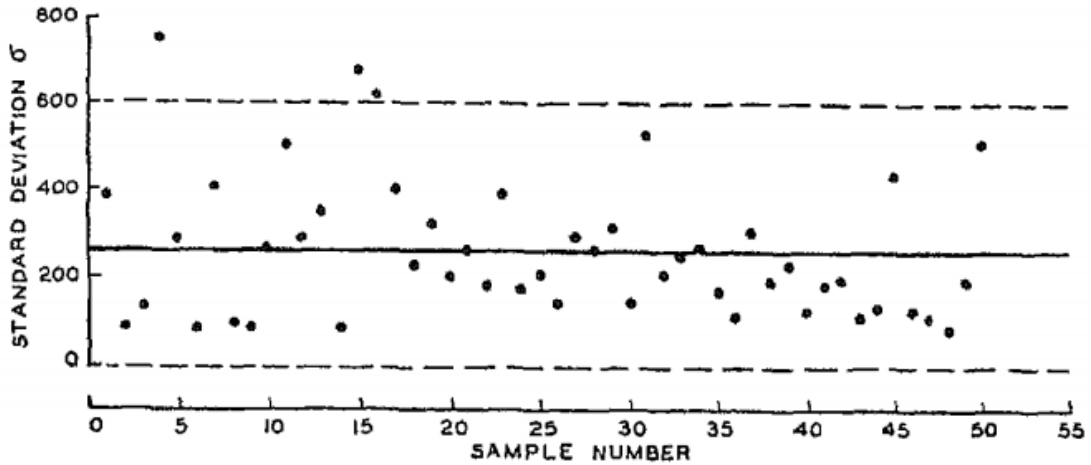


FIG. 111.—CONTROL CHART FOR STANDARD DEVIATIONS OF SAMPLES OF FOUR—DATA OF TABLE 2.

Figure: Example of Shewhart's control chart

Change-point Problem

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

$$X_i \sim \begin{cases} F_1 & \text{if } i \leq \tau_1 \\ F_2 & \text{if } \tau_1 < i \leq \tau_2 \\ \ldots & \\ F_{m+1} & \text{if } \tau_m < i \end{cases} \tag{1}$$

- Structural changes : Change in mean, Change in variance, Change in distribution

Change-point Model

- Consider independent random variables $X_1 \sim G_1, \ldots, X_n \sim G_n$.
- 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

$$\mathbf{H_0} : G_1 = G_2 = \ldots = G_n = F, \tag{2}$$

against the following alternative of one change

$$\mathbf{H_a} : F_1 = G_1 = G_2 = \ldots = G_\tau \neq G_{\tau+1} = \ldots = G_n = F_2. \tag{3}$$

where $1 \leq \tau < n$ and neither F nor G is degenerate.

- Using binary segmentation, it suffices to test and estimate the position of a single change point at each stage sequentially.

Parametric method on change-point analysis

- Numerous studies related to change-point analysis largely in parametric and nonparametric methods.
- The most investigated change point problem is that of testing a change in the mean of independent normal variables with a constant variance. (Chernoff, H. and Zacks, S. (1964))
- While testing a change in the variance of normal variables with a common mean has been explored. (Chen, J. and Gupta, A. K. (2011))
- However, existing limitation as follows:
  - Parametric assumptions may **not be satisfied in practice**.
  - Too sensitive to the effect of **outliers**.
  - The value at the extreme varies greatly **depending on the distribution**.

Nonparametric method on change-point analysis

- Nonparametric models work with less assumptions, more acceptable under various conditions.
- Traditional nonparametric methods include Kolmogorov-Smirnov Test, Cramer-Von-Mises Test, Wilcoxon-Mann-Whitney test.
- 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.
- Hence, this paper is motivated to expand empirical likelihood to **double quantile** for the both extreme side.

Empirical likelihood

- Empirical likelihood is a nonparametric method first introduced by Owen(1988). The main idea is to place an unknown probability mass at each observation.
- Assume that independently and identically distributed observation $x_1, ..., x_n$ are from an unknown population distribution F. Let $p_i = P(X = x_i)$.
- Empirical likelihood function of $\{p_i\}_{i=1}^n$ is defined as

$$L(F) = \prod_{i=1}^{n} p_i, \tag{4}$$

where p_i satisfy the constraints $p_i \geq 0$ and $\sum_{i=1}^n p_i = 1$

Empirical likelihood

- It is clear that $L(F)$ 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

$$\sum_{i=1}^{n} p_i m(x_i, \theta_0) = 0. \tag{5}$$

To find $\{p_i\}_{i=1}^n$ under the restrictions, solve the Laglange Multiplier

$$\sum_{i=1}^{n} \log p_i + \lambda_0 (\sum_{i=1}^{n} p_i - 1) + \lambda_1 (\sum_{i=1}^{n} p_i m(x_i, \theta_0)). \tag{6}$$

Empirical likelihood

- The ELR statistic to test $\theta = \theta_0$ is given by

$$\mathbf{R}(\theta_0) = \frac{L(F)}{L(F_n)} = \max \left\{ \prod_{i=1}^{n} np_i | \sum_{i=1}^{n} p_i m(x_i, \theta_0) = 0, p_i \geq 0, \sum_{i=1}^{n} p_i = 1 \right\} \tag{7}$$

- Under the null model $\theta = \theta_0$ with mild regular conditions, $-2 \log \mathbf{R}(\theta_0) \to \chi^2_r$ in distribution as $n \to \infty$, where r is dimension of $m(x, \theta)$ (Owen, 1988). The empirical likelihood method can be extended to other constraints using Lagrange multiplier method to find $\{p_i\}_{i=1}^n$.

Empirical likelihood for Two groups comparison

- Two samples: $X_1, X_2, \ldots, X_n \sim F_1$ and $Y_1, Y_2, \ldots, Y_m \sim F_2$ 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

$$L(F) = \prod_{i=1}^{n} p_i \prod_{j=1}^{m} q_j, \tag{8}$$

where p_i and q_j satisfy the constraints $p_i \geq 0, q_j \geq 0$ and $\sum_{i=1}^n p_i = 1$, $\sum_{j=1}^m q_j = 1$

- This hypothesis (2) and (3) is equivalent to

$$\mathbf{H_0} : F_1 = F_2, \tag{9}$$

against

$$\mathbf{H_a} : F_1 \neq F_2. \tag{10}$$

The hypothesis (9) and (10) becomes two sample test.

Quantile empirical likelihood ratio for two sample (QLR)

- Zhou, Y., Fu, L., and Zhang, B. (2017)
- Under the null hypothesis, for any given x, we have $F_1(x) = F_2(x) = F(x)$. Let $p = F(\xi_p)$; hence, ξ_p is the p quantile of F and ξ_p needs to satisfy

$$E[I(X_i \leq \xi_p) - p] = 0, \quad \text{for } 1 \leq i \leq n+m, \tag{11}$$

- We can construct the following quantile empirical likelihood test statistic under restriction,

$$\mathbf{R}(\xi_{\mathbf{p}}) = \max \left\{ \prod_{i=1}^{n} np_i \prod_{j=1}^{m} mq_j | \sum_{i=1}^{n} p_i I(X_i \leq \xi_p)) = p, \right. \tag{12}$$
$$\left. \sum_{j=1}^{m} q_j I(Y_j \leq \xi_p)) = p, p_i, q_j \geq 0, \sum_{i=1}^{n} p_i = \sum_{j=1}^{m} q_j = 1 \right\}$$

Double quantile likelihood ratio for two sample (DLR)

- Proposed methodology: Double quantile likelihood
- Expand (11) to **double quantile likelihood** for the both extreme side.
- 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

$$E[I(X_i \leq \xi_p) - p] = 0, \quad E[I(X_i \geq \xi_{1-q}) - q)] = 0 \tag{13}$$

  where $0 < p < 1 - q < 0$ for $1 \leq i \leq n + m$.
- Using (13), double quantile empirical likelihood test statistic under restriction is

$$\mathbf{R}(\xi_{\mathbf{p}}, \xi_{\mathbf{1-q}}) = \max \left\{ \prod_{i=1}^{n} np_i \prod_{j=1}^{m} mq_j | \sum_{i=1}^{n} p_i I(X_i \leq \xi_p)) = p, \right. \tag{14}$$

$$\sum_{j=1}^{m} q_j I(Y_j \leq \xi_p)) = p, \sum_{i=1}^{m} q_i I(X_i \leq \xi_{1-q})) = 1 - q,$$

$$\left. \sum_{j=1}^{m} q_j I(Y_j \leq \xi_{1-q})) = 1 - q, p_i, q_j \geq 0, \sum_{i=1}^{n} p_i = \sum_{j=1}^{m} q_j = 1 \right\}$$

Double quantile likelihood ratio for two sample (DLR)

- Using Lagrange multipliers to solve (14), we can get following unique $\lambda's$ and $\{p_i\}_{i=1}^{n}$, $\{q_j\}_{j=1}^{m}$. (Proof in Appendix.A)
- This leads to double quantile likelihood ratio(DLR) test statistic

$$\mathbf{R}(\xi_{\mathbf{p}}, \xi_{\mathbf{1-q}}) = \left(\frac{np}{n_1}\right)^{n_1} \left(\frac{nq}{n_2}\right)^{n_2} \left(\frac{n(1-p-q)}{n-n_1-n_2}\right)^{n-n_1-n_2} \left(\frac{mp}{m_1}\right)^{m_1} \left(\frac{mq}{m_2}\right)^{m_2} \left(\frac{m(1-p-q)}{m-m_1-m_2}\right)^{m-m_1-m_2} \tag{15}$$

where $\sum_{i=1}^{n} I(X_i \leq \xi_p) = n_1$, $\sum_{i=1}^{n} I(X_i > \xi_{1-q}) = n_2$ and $\sum_{j=1}^{m} I(Y_j \leq \xi_p) = m_1$, $\sum_{j=1}^{m} I(Y_j > \xi_{1-q}) = m_2$

$$\therefore D_n = \sup_{\xi_p < \xi_{1-q}} \{-2 \log \mathbf{R}(\xi_{\mathbf{p}}, \xi_{\mathbf{1-q}})\} \tag{16}$$

- Large values of D_n indicate that there is at least one change-point.

Algorithm for change-point detection

- Change-point detection problem is to detect τ where

$$F_1 = G_1 = G_2 = \ldots = G_\tau \neq G_{\tau+1} = \ldots = G_n = F_2 \tag{17}$$

- Two sample test: $X_1, \ldots, X_n \sim F_1$ and $Y_1, \ldots, Y_m \sim F_2$.
- 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))

$$D_n^* = \sup_{c(n+m)^{-1/9} < \xi_p < \xi_{1-q} < 1 - c(n+m)^{-1/9}} \{-2 \log \mathbf{R}(\xi_{\mathbf{p}}, \xi_{\mathbf{1-q}})\} \tag{18}$$

  where c is a positive constant.

- The location τ can be estimated by

$$\hat{\tau} = \arg_\tau \max\{D_n^*\} \tag{19}$$

Simulation Setting

- Simulation steps
  1. Assume that $X_1, ..., X_n$ from F_1, and $Y_1, ..., Y_m$ from F_2 with different distributions by setting δ satisfying $\delta = E_{F_1}(X) - E_{F_2}(X)$
  2. Change location m takes 25%, 50%, 75%, and 95% quantiles of the number of samples.
  3. ξ_p and ξ_{1-q} are the value of $x's$ satisfying the rank(ξ_p)-rank(ξ_{1-q}) $\geq 0.5(n+m)$ for computation.
  4. Calculate D_n^* and detect change-point $\hat{\tau}$.
  5. For each case, 100 simulations are carried out.
  6. Calculate the accuracy rate.

Simulation Results - Case 1

- In case 1, the two samples are from two different distribution families.

$F_1 \sim \exp(3)$, $F_2 \sim \chi^2(5)$					
cpt	DLR	QLR	WMW	KS	CvM
25%	0.89	0.88	0.77	0.25	0.73
50%	0.89	0.86	0.94	0.21	0.87
75%	0.88	0.86	0.72	0.19	0.81
95%	0.83	0.89	0.59	0.89	0.13
$F_1 \sim \exp(3)$, $F_2 \sim$ pois(5)					
cpt	DLR	QLR	WMW	KS	CvM
25%	0.96	0.91	0.82	0.96	0.88
50%	0.90	0.98	0.91	0.70	0.88
75%	0.91	0.94	0.75	0.88	0.85
95%	0.90	0.91	0.62	0.83	0.07

Table: Accuracy rate of δ =2

Simulation Results - Case 2

- In case 2, the samples are from the same distribution family but with different mean parameters.

$F_1 \sim$pois(1), $F_2 \sim$pois(3)					
cpt	DLR	QLR	WMW	KS	CvM
25%	0.34	0.35	0.44	0.4	0.09
50%	0.43	0.32	0.54	0.4	0.26
75%	0.47	0.31	0.45	0.32	0.11
95%	0.38	0.39	0.45	0.3	0.38
$F_1 \sim N(0,1)$, $F_2 \sim N(2,1)$					
cpt	DLR	QLR	WMW	KS	CvM
25%	0.24	0.24	0.35	0.29	0.05
50%	0.27	0.2	0.23	0.2	0.05
75%	0.29	0.25	0.4	0.31	0.07
95%	0.15	0.25	0.32	0.19	0.24

Table: Accuracy rate of δ =2

Application to real data: Nile River data

- We illustrate the proposed methods by analyzing a well-known real example, the Nile River data (Cobb, 1978), which has been studied by many authors in the area of change-point analysis. The change-point was 1898. (See Appendix.B)

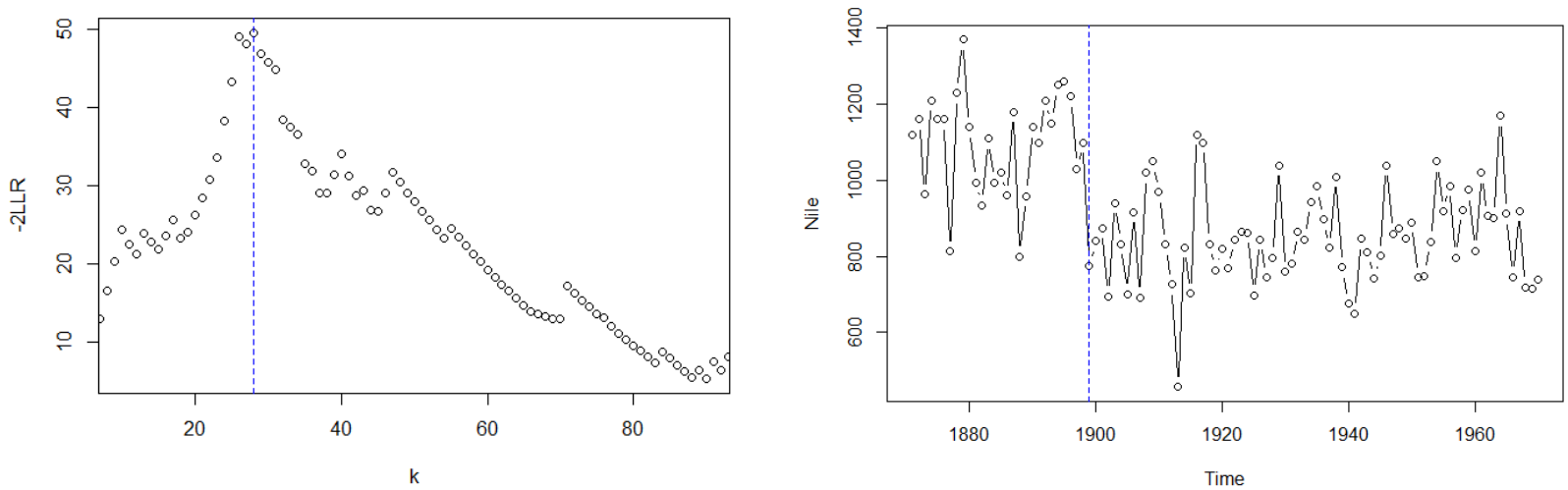


Figure: DLR result : $\hat{\tau} = 28$, change-point=1898.

Future Works

- Needs to evaluate the performance of this method with power or p-value.
- Needs to compare the performance of this method with that of parametric change-point method.
- Needs to find more efficient way to take ξ_p and ξ_{1-q}.
- Needs to extend the proposed method to detect multiple change-points.

Appendix.A - Proof of Double quantile likelihood

- To solve $\mathbf{R}(\xi_{\mathbf{p}}, \xi_{\mathbf{1-q}})$, Use Lagrange multipliers to find maximum value of $\{p_i\}_{i=1}^n$, $\{q_j\}_{j=1}^m$,
  The solution can be found to solve the following Lagrangian equation:

$$\sum_{i=1}^{n} \log p_i + \sum_{j=1}^{m} \log q_j + \lambda_0(\sum_{i=1}^{n} p_i - 1) + \lambda_1(\sum_{j=1}^{m} q_j - 1)$$
$$+\lambda_2(\sum_{i=1}^{n} p_i I(X_i \leq \xi_p)) - p) + \lambda_3(\sum_{j=1}^{m} q_j I(Y_j \leq \xi_p)) - p)$$
$$+\lambda_4(\sum_{i=1}^{m} q_i I(X_i \geq \xi_{1-q})) - q) + \lambda_5(\sum_{j=1}^{m} q_j I(Y_j \geq \xi_{1-q})) - q) = 0$$

- The the solution is

$$\hat{p}_i = -(n - \lambda_2(I(X_i \leq \xi_p) - p) - \lambda_4(I(X_i \geq \xi_{1-q}) - q))^{-1}$$

Appendix.A - Proof of Double quantile likelihood

- and the following equation gives a unique solution for $\lambda's$

$$\sum_{i=1}^{n} \frac{I(X_i \leq \xi_p)}{n - \lambda_2(I(X_i \leq \xi_p) - p) - \lambda_4(I(X_i > \xi_{1-q}) - q)} = p$$

$$\sum_{i=1}^{n} \frac{I(X_i > \xi_{1-q})}{n - \lambda_2(I(X_i \leq \xi_p) - p) - \lambda_4(I(X_i > \xi_{1-q}) - q)} = q$$

- $\hat{q}_i$ can be found in the same way.

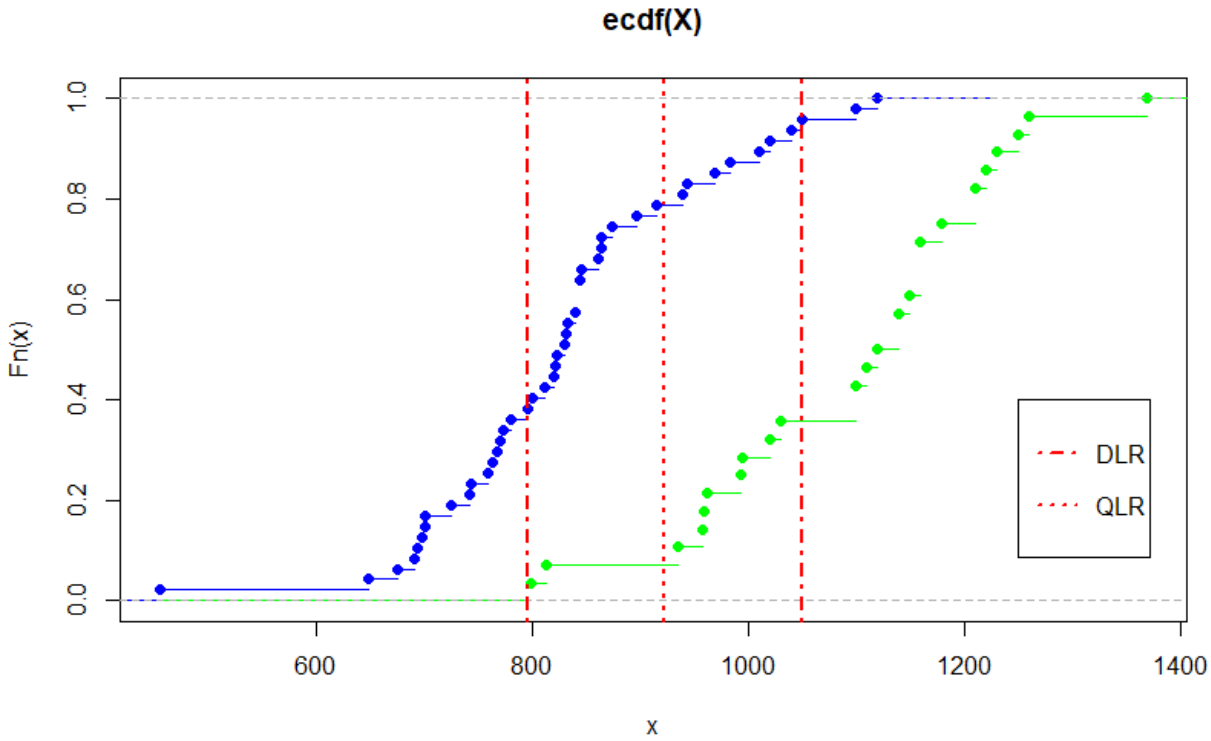
Appendix.A - Proof of Double quantile likelihood

$$\begin{aligned}\hat{p}_i &= \frac{p}{n_1} \text{ for } i = 1, \ldots, n_1 \\ &= \frac{1-p-q}{1-n_1-n_2} \text{ for } i = n_1+1, \ldots, n-n_2 \\ &= \frac{q}{n_2} \text{ for } i = n-n_2+1, \ldots, n\end{aligned}$$

$$\begin{aligned}\hat{q}_j &= \frac{p}{m_1} \text{ for } j = 1, \ldots, m_1 \\ &= \frac{1-p-q}{1-m_1-m_2} \text{ for } j = m_1+1, \ldots, m-m_2 \\ &= \frac{q}{m_2} \text{ for } j = m-m_2+1, \ldots, m\end{aligned}$$

where $\sum_{i=1}^{n} I(X_i \leq \xi_p) = n_1$, $\sum_{i=1}^{n} I(\leq X_i > \xi_{1-q}) = n_2$ and $\sum_{j=1}^{m} I(Y_j \leq \xi_p) = m_1$, $\sum_{j=1}^{m} I(Y_j > \xi_{1-q}) = m_2$

Appendix.B - Results of Data application



DLR	$\hat{\tau}$	-2LLR	p-value	$\hat{\xi}_p$	$\hat{\xi}_{1-q}$	$\hat{p}$	$\hat{q}$
	28	49.616	1.870e-12	797	1050	0.25	0.21
QLR	$\hat{\tau}$	-2LLR	p-value	$\hat{\xi}_p$		$\hat{p}$	
	28	45.371	1.630e-11	923		0.58	

Table: The results of DLR and QLR function

References

- Chen, J. and Gupta, A. K. (2011).Parametric statistical change point analysis: with applicationsto genetics, medicine, and finance. Springer Science Business Media.
- Jing, B.-Y. (1995). Two-sample empirical likelihood method.Statistics & probability letters, 24(4):315-319.
- Owen, A. B. (1988).Empirical likelihood ratio confidence intervals for a single functional.Biometrika, 75(2):237-249.
- Owen, A. B. (2001).Empirical likelihood. Chapman and Hall/CRC.
- Ross, G. J. and Adams, N. M. (2012). Two nonparametric control charts for detecting arbitrary distribution changes. Journal of Quality Technology, 44(2):102-116.

References

- Zhang, J. (2002). Powerful goodness-of-fit tests based on the likelihood ratio.Journal of the RoyalStatistical Society: Series B (Statistical Methodology), 64(2):281-294.
- Zhang, J. (2006). Powerful two-sample tests based on the likelihood ratio.Technometrics, 48(1):95-103.
- Zhou, Y., Fu, L., and Zhang, B. (2017). Two non parametric methods for change-point detectionin distribution.Communications in Statistics-Theory and Methods, 46(6):2801-2815.
- Zou, C., Liu, Y., Qin, P., and Wang, Z. (2007). Empirical likelihood ratio test for the change-pointproblem.Statistics probability letters, 77(4):374-382.