Communications in Statistics—Theory and Methods, 34: 1935–1938, 2005
Copyright © Taylor & Francis, Inc.
ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610920500201145

Inference

Change Point Detection in a General Class of Distributions

R. HABIBI, S. M. SADOOGHI-ALVANDI, AND A. R. NEMATOLLAHI

Department of Statistics, Shiraz University, Shiraz, Iran

We consider the change point problem in a general class of distributions, and derive a test statistic T_n which reduces to the statistic obtained by Kander and Zacks (1966) for the exponential family. Properties of the test, including its asymptotic distribution, are discussed.

Keywords Change point; Likelihood ratio; Marginal likelihood.

Mathematics Subject Classification 62G20; 62E20.

1. Introduction

Change point problems are concerned with inferences about an unknown parameter m, such that given a sequence of independent random variables $X_1, X_2, \ldots, X_n$, which are distributed according to $\{F_\theta, \theta \in \Theta\}$, the first m observations come from a distribution F_{θ_0} and the remaining $n - m$ observations come from another distribution F_{θ_1} with $F_{\theta_0} \neq F_{\theta_1}$. The change point problem arises in various areas, such as industrial quality control, medicine, epidemiology, and economics under other names such as Surveillance, Statistical Process Control, and Monitoring; see Wessman (1998).

There are various techniques for making inferences about m: see, e.g., Chernoff and Zacks (1964), Kander and Zacks (1966), Hinkely (1970), Broemeling (1974), Smith (1975), Henderson (1986), Brodsky and Darkhovsky (1993), Broemeling and Gregurich (1996), and Csorgo and Horvath (1997).

In this article, we derive a test statistic T_n for the presence of a change point in a general class of distributions, which reduces to the test statistic obtained by Kander and Zacks (1966) in the case of one parameter exponential family.

Received March 10, 2004; Accepted April 15, 2005
Address correspondence to A. R. Nematollahi, Department of Statistics, Shiraz University, Shiraz, Iran; E-mail: nematollahi@susc.ac.ir

2. Main Results

Suppose that $X_1, X_2, \ldots, X_n$ are independent random variables from a one parameter family with density function $f_\theta(\cdot)$ with $\theta \in \Theta$. We are interested in testing the null hypothesis that $X_1, X_2, \ldots, X_n$ are *iid* from f_{θ_0} (with θ_0 known), against the alternative hypothesis that at some unknown point m a change occurs in parameter θ, i.e., for some $m \in \{1, 2, \ldots, n-1\}$, $X_1, X_2, \ldots, X_m$ are *iid* from f_{θ_0} and $X_{m+1}, \ldots, X_n$ are *iid* from f_{θ_1} (with $\theta_1 = \theta_0 + \delta$, where δ is an unknown positive real number).

To derive the test statistic and study its properties, the following mild regularity conditions are assumed:

(i) For every value x, the derivative $\frac{\partial}{\partial\theta} \log f_\theta(x)$ exists for every $\theta \in \Theta$, and $E_{\theta_0}\left|\frac{\partial}{\partial\theta} \log f_\theta(X)\right|^3 < \infty$, so that $I(\theta_0) = Var_{\theta_0}\left[\frac{\partial}{\partial\theta} \log f_\theta(x)\right]$ exists.

(ii) For any function $h(\cdot)$ with $E_\theta|h(X)| < \infty$, $\frac{d}{d\theta} \int h(x) f_\theta(x) dx = \int h(x) \frac{\partial}{\partial\theta} f_\theta(x) dx$.

Following Kander and Zacks (1966), we derive a test statistic using a quasi-Bayesian approach by considering the point of change m as a realization of a random variable M, with a uniform prior density $\pi(m) = \frac{1}{n-1}$; $m = 1, 2, \ldots, n-1$.

The marginal likelihood of the sample under H_1 is

$$\frac{1}{n-1} \sum_{m=1}^{n-1} \exp\left\{ \sum_{i=1}^{m} \log f_{\theta_0}(x_i) + \sum_{i=m+1}^{n} \log f_{\theta_0+\delta}(x_i) \right\}.$$

Since $\frac{\partial}{\partial\theta} \log f_\theta(x)$ exists for every x and θ, we have

$$\log f_{\theta_0+\delta}(x_i) = \log f_{\theta_0}(x_i) + \delta \frac{\partial}{\partial\theta} \log f_\theta(x_i)|_{\theta=\theta_0} + o(\delta), \quad \text{as } \delta \to 0.$$

So the marginal likelihood under H_1 is

$$\left\{ \prod_{i=1}^{n} f_{\theta_0}(x_i) \right\} \left\{ 1 + \frac{\delta}{n-1} \sum_{i=1}^{n} (i-1) \frac{\partial}{\partial\theta} \log f_\theta(x_i)|_{\theta=\theta_0} + o(\delta) \right\}, \quad \text{as } \delta \to 0.$$

Since the likelihood under H_0 is $\prod_{i=1}^{n} f_{\theta_0}(x_i)$, the likelihood ratio (LR) is

$$\text{LR} = 1 + \frac{\partial}{n-1} \sum_{i=1}^{n} (i-1) \frac{\partial}{\partial\theta} \log f_\theta(x_i)|_{\theta=\theta_0} + o(\delta), \quad \text{as } \delta \to 0,$$

and the null hypothesis H_0 is rejected whenever the likelihood ratio is large. So as $\delta \to 0$, H_0 is rejected whenever

$$T_n = \sum_{i=1}^{n} (i-1) \frac{\partial}{\partial\theta} \log f_\theta(x_i)|_{\theta=\theta_0},$$

is larger than a suitably chosen value. We propose to use T_n as our test statistic. It is easily verified that, in the special case of exponential family, T_n reduces to the statistic proposed by Kander and Zacks (1966).

It can be shown, as in Kander and Zacks (1966), that the test procedure based on T_n for testing H_0 is locally most powerful (see Habibi et al., 2004).

The exact distribution of T_n can be found in some special cases. Using the results of Chernoff and Zacks (1964) and Kander and Zacks (1966), the exact distribution of T_n is easily obtained in the case of normal, exponential, and binomial distributions. The exact distribution of T_n can also be found in the logistic case (see Habibi et al., 2004). However, since the exact distribution of T_n is very complicated in many cases, we derive its asymptotic distribution.

Proposition 2.1. *Assuming regularity conditions* (i) *and* (ii), *under the null hypothesis, we have*

$$n^{\frac{-3}{2}} T_n \xrightarrow{d} N\left(0, \frac{I(\theta_0)}{3}\right),$$

where $I(\theta_0)$ is the Fisher information at θ_0.

Proof. Under the regularity conditions we have $E_{\theta_0}(T_n) = 0$ and $Var_{\theta_0}(T_n) = \frac{(2n-1)(n-1)(n)}{6} I(\theta_0)$. Now write $T_n = \sum_{i=1}^n (i-1)Z_i$, where $Z_i = \frac{\partial}{\partial\theta} \log f_\theta(X_i)|_{\theta=\theta_0}$, $i = 1, 2, \ldots, n$, which are *iid* under the null hypothesis. Since $\frac{\max_{j\le n}(j-1)^2}{\sum_{j=1}^n (j-1)^2} \to 0$ as $n \to \infty$ and $E_{\theta_0}(T_n) = 0$, we conclude that under H_0, $\frac{T_n}{\sqrt{Var_{\theta_0}(T_n)}} \xrightarrow{d} N(0, 1)$ (see Ferguson, 1996, p. 34), and using the Slutsky theorem, we have $n^{\frac{-3}{2}} T_n \xrightarrow{d} N\left(0, \frac{I(\theta_0)}{3}\right)$.

The rate of convergence is considered in the following proposition; cf. Kander and Zacks (1966).

Proposition 2.2. *Under conditions of Proposition* 2.1, *we have, for every x*

$$|F_n(x) - \Phi(x)| = O(n^{\frac{-1}{2}}) \quad \text{as } n \to \infty,$$

where $F_n(\cdot)$ is the distribution function of standardized variable $\frac{T_n}{\sqrt{Var_{\theta_0}(T_n)}}$ and $\Phi(\cdot)$ is the distribution function of the standard normal.

Proof. As in the proof of Proposition 2.1, $T_n = \sum_{i=1}^n Y_i$, where $Y_i = (i-1)Z_i$. It can be shown (see Feller, 1965, p. 544) that

$$|F_n(x) - \Phi(x)| < 6 \frac{\sum_{i=1}^n E|Y_i|^3}{\left(\sum_{i=1}^n \sigma_i^2\right)^{\frac{3}{2}}}, \quad -\infty < x < +\infty,$$

where σ_i^2 denotes the variance of Y_i and F_n is the distribution function of $\sum_{i=1}^n Y_i / \left(\sum_{i=1}^n \sigma_i^2\right)^{\frac{1}{2}}$. But

$$\frac{\sum_{i=1}^n E|Y_i|^3}{\left(\sum_{i=1}^n \sigma_i^2\right)^{\frac{3}{2}}} = \frac{\sum_{i=1}^n (i-1)^3 E_{\theta_0}|Z_1|^3}{\left[\sum_{i=1}^n (i-1)^2 Var_{\theta_0}(Z_1)\right]^{\frac{3}{2}}} = O(n^{\frac{-1}{2}}), \quad \text{as } n \to \infty,$$

(since Z_is are *iid*), and the result follows.

Simulation results (see Habibi et al., 2004) confirm that the normal approximation is satisfactory.

It can be shown that under the additional regularity condition:

$$E_{\theta_1}\left|\frac{\partial}{\partial\theta}\log f_\theta(X_{m+1})\bigg|_{\theta=\theta_0}\right|^3 < \infty,$$

the asymptotic distribution of T_n under the alternative hypothesis is given by

$$\frac{T_n - \sum_{i=m+1}^{n} E(Y_i)}{\sqrt{\sum_{i=1}^{n} Var(Y_i)}} \xrightarrow{d} N(0,1),$$

where $Y_i = (i-1)Z_i$ with $Z_i = \frac{\partial}{\partial\theta}\log f_\theta(X_i)|_{\theta=\theta_0}$, and the rate of convergence is $O(n^{\frac{-1}{2}})$ (see Habibi et al., 2004).

References

Brodsky, B. E., Darkhovsky, B. S. (1993). *Nonparametric Methods in Change Point Problems*. Dordrecht, The Netherlands: Kluwer Academic Press.

Broemeling, L. D. (1974). Bayesian inferences about a changing sequence of random variables. *Commun. Statist. Theor. Meth.* 3:243–255.

Broemeling, L. D., Gregurich, M. A. (1996). On a Bayesian approach for the shiftpoint problem. *Commun. Statist. Theor. Meth.* 25:2267–2279.

Chernoff, H., Zacks, S. (1964). Estimating the current mean of a normal distribution which is subjected to changes in time. *Ann. Math. Statist.* 35:999–1018.

Csorgo, M., Horvath, L. (1997). *Limit Theorems in Change-Point Analysis*. New York: Wiley & Sons.

Feller, W. (1965). *An Introduction to Probability and Its Applications*. Vol. 2. New York: Wiley & Sons.

Ferguson, S. T. (1996). *A Course in Large Sample Theory*. London: Chapman & Hall.

Habibi, R., Sadooghi-Alvandi, S. M., Nematollahi, A. R. (2004). Change point detection in a general class of distributions. *Technical Report*. Department of Statistics, Shiraz University.

Henderson, R. (1986). Change point problem with correlated observations, with an application in material accountancy. *Technometrics* 28:381–389.

Hinkely, D. V. (1970). Inference about the change point in a sequence of random variables. *Biometrika* 57:1–16.

Kander, Z., Zacks, S. (1966). Test procedure for possible changes in parameter of statistical distributions occurring at unknown time points. *Ann. Math. Statist.* 37:1196–1210.

Smith, A. F. M. (1975). A Bayesian approach to inference about change point in a sequence random variable. *Biometrika* 62:407–416.

Wessman, P. (1998). Some principles for surveillance adopted for multivariate processes with a common change point. *Commun. Statist. Theor. Meth.* 27(5):1143–1161.

Copyright of Communications in Statistics: Theory & Methods is the property of Marcel Dekker Inc.. The copyright in an individual article may be maintained by the author in certain cases. Content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.