Afrika Statistika

Afrika Statistika
Vol. 9, 2014, pages 685–694.
DOI: http://dx.doi.org/1016929.as.2014.685.62

ISSN 2316-090X

A Bayesian significance test of change in the presence of a single outlier

Abdeldjalil Slama[†,‡]

[†] Department of Mathematics and Computer Science, LDDI, University of Adrar, National Road No.06, Adrar, Algeria
[‡] Department of Probability and statistics, USTHB. PO Box 32 EL Alia 16111 Bab Ezzouar, Algiers, Algeria

Received 26 May 2014; Accepted 16 October 2014

Abstract. The Bayesian significance test for a change in independent gaussian samples in the presence of single outlier is considered. The impact of an outlier on the performance of the Bayesian significance test is studied.

Résumé. Dans ce travail, nous considérons un test de signification Bayésien pour la détection de rupture dans un échantillon gaussien en présence d'une observation aberrante. L'impact d'une contamonation sur la performance du test est étudié.

Key words: Gaussian models,; Change point; HPD region sets; p-value; Outlier.
AMS 2010 Mathematics Subject Classification : 62M10; 62F15; 62F03; 62F30.

1. Introduction

Suppose we have the observations $(x_1, \cdots, x_n)$ from the model of change point;

$$\begin{cases} X_i = \phi_0 + \varepsilon_i \text{ if } \quad i = 1, 2, \cdots, m \\ X_i = \phi_1 + \varepsilon_i \text{ if } i = m+1, \cdots, n \end{cases} \tag{1}$$

where the ε_i is normal random independent errors with mean zero and unknown constant variance σ^2, ϕ_0 and ϕ_1 are real unknown constants which represent the means of the variables X_i before and after the change-point m, n being the size of the sample.

A change point, which is generally the effect of an external event on the phenomenon of interest, may be represented by a change in the structure of the model or simply by a change of the value of some parameters. Since Page (1954, 1955) which developed a cumulative sum (Cusum) test to detect a location change, considerable attention has been given to

*Corresponding author Abdeldjalil Slama: slama_dj@yahoo.fr

this problem in a variety of settings. Hinkley (1970), Sen and Sen and Srivastava (1975), Siegmund (1986, 1988), Worsley (1983, 1986) and Kim (1996), who used likelihood ratio approaches. Worsley (1983, 1986) proposed a numerical method for computing the p-value of the generalized likelihood ratio test to detect a change in binomial probability and in location of an exponential family of distributions.

In a Bayesian context, the problem of detection of change was studied by many authors. We can cite Chernoff and Zacks (1964), Kander and Zacks (1966), Sen and Srivastava (1975) where the aim is to detect the change in the mean for normal random variables. Kim (1991), proposed a Bayesian significance test for stationarity of a regression equation using the highest posterior density credible set.

From a Monte Carlo simulation study, it has shown that the Bayesian significance test has stronger power than the Cusum and the Cusum of squares tests suggested by Brown *et al.* (1975). Ghorbanzadeh and Lounes (2001) proposed a Bayesian analysis of detection of a change of parameters in a sequence of independent random variables from an exponential family.

However, the observations can be contaminated by outiers. And it is natural to seek means of interpreting or categorizing outliers, of sometimes rejecting them to restore the propriety of the data, or at least of taking their presence properly into account in any statistical analysis Barnett and Lewis (1978). Verdinelli and Wasserman (1991) consider the Bayesian analysis of outlier models. They showed that the Gibbs sampler brings considerable conceptual and computational simplicity to the problem of calculating posterior marginals. Recentely, Belkacem and Fellag (2012) study the impact of an outlier on the performance of the Bayesian estimation of the change point in independent gaussian samples.

In this work, we propose a Bayesian significance test based on the HPD credible regions in independent gaussian samples in the presence of a single outlier, and our aim is to study the impact of a single outlier on the performance of the Bayesian significance test of change parameters model. The rest of paper is organized as follows. Section **??** presents the Bayesian analysis and the Bayesian significance test for change. Simulations results are given in In section **??**. Section **??** is our conclusion.

Assume that there exists a position k , $k \in \{1, 2, \cdots, n\}$, such that $(y_1, \cdots, y_n)$ are possible observations from the model,

$$\begin{cases} Y_k = X_k + \xi \\ Y_i = X_i \quad \forall i \in \{1, 2, \cdots, n\} \quad \text{with} \quad i \neq k \end{cases} \tag{2}$$

where the constant ξ is the magnitude of the contamination which occurs at a specified time, say k. Since the outlier can occur before or after the change-point, we will consider two cases where we derive the posterior density of the change point when an outlier occurs.

2. Bayesian analysis

We consider the contamination occurs before the change point m, i.e, $k \in \{1, \cdots, m\}$. Then the model (2) is written as follows :

$$\begin{cases} Y_i = X_i = \phi_0 + \varepsilon_i & i = 1, \cdots, k-1 \quad \text{and} \quad i = k+1, \cdots, m \\ Y_k = X_k + \xi = \phi_0 + \varepsilon_k + \xi & \\ Y_i = X_i = \phi_1 + \varepsilon_i & i = m+1, \cdots, n\,, \end{cases} \tag{3}$$

where $m \in \{1, \cdots, n-1\}$, $\phi_0, \phi_1 \in \mathbb{R}$, ($\phi_0 \neq \phi_1$), $\xi \in \mathbb{R}$, and $\varepsilon_i \sim N(0, \sigma^2)$ for $i = 1, \cdots, n$, $(\sigma > 0)$, with m, ϕ_0, ϕ_1 and σ are unknown parameters.

One has a parameter set $\theta = (m, \phi_0, \phi_1, r)$ where $r = 1/\sigma^2$. Since prior knowledge of $\theta' = (\phi_0, \phi_1, r)$ is often vague or diffuse, we employ a diffuse prior for θ'. The parameters m, (ϕ_0, ϕ_1) and r are assumed independent. The prior distribution of θ is, therefore

$$\pi(\theta) \propto \frac{1}{r}\,. \tag{4}$$

Note that the functional forms $\pi(.)$ and $\pi(./.)$ represent a prior and a posterior distribution, respectively.

The likelihood function based on the observations $y = (y_1, y_2, \cdots, y_n)$ is then

$$\begin{aligned} l(y/\theta) \propto & r^{-\frac{n}{2}} \exp\left\{ -\frac{r}{2} \left[\sum_{i=1}^{k-1} (y_i - \phi_0)^2 + (y_k - (\phi_0 + \xi))^2 \right.\right. \\ & \left.\left. + \sum_{i=k+1}^{m} (y_i - \phi_0)^2 + \sum_{i=m+1}^{n} (y_i - \phi_1)^2 \right] \right\}. \end{aligned} \tag{5}$$

The posterior distribution of θ, obtained by combination of (4) and (5) is

$$\begin{aligned} \pi(\theta/y) \propto & r^{-\frac{n}{2}-1} \exp\left\{ -\frac{r}{2} \left[\sum_{i=1}^{k-1} (y_i - \phi_0)^2 + (y_k - (\phi_0 + \xi))^2 \right.\right. \\ & \left.\left. + \sum_{i=k+1}^{m} (y_i - \phi_0)^2 + \sum_{i=m+1}^{n} (y_i - \phi_1)^2 \right] \right\}. \end{aligned} \tag{6}$$

The null hypothesis H_0, that there is no change in the parameters of model (1), is

$$H_0 : \delta = \phi_1 - \phi_0 = 0$$

For the Bayesian significance test, therefore, the posterior distributions of δ is needed to obtain the confidence region, i.e, the highest posterior density credible set of δ.

The followin theorem gives the posterior distribution of δ,

Theorem 1. *1. Given m and ϕ_0 the conditional posterior distribution of δ is:*

$$\pi(\delta|m,\phi_0,y) \propto \left\{1+\frac{(n-m)\left(\delta-\widehat{\delta}(m,\phi_0)\right)^2}{(n-1)S_1^2(m,\phi_0)}\right\}^{-\frac{n}{2}}, \tag{7}$$

where

$$\widehat{\delta}(m,\phi_0)=\frac{\sum_{i=m+1}^{n}(y_i-\phi_0)}{n-m},$$

$$S_1^2(m,\phi_0)=\frac{SS(m,\phi_0)}{(n-1)}$$

and

$$\begin{aligned} SS(m,\phi_0)=&\sum_{i=1}^{k-1}(y_i-\phi_0)^2+(y_k-(\phi_0+\xi))^2+\sum_{i=k+1}^{m}(y_i-\phi_0)^2\\ &+\sum_{i=m+1}^{n}(y_i-\phi_0)^2-\frac{\left[\sum_{i=m+1}^{n}(y_i-\phi_0)\right]^2}{n-m}, \end{aligned} \tag{8}$$

which is the Student t distribution with location parameter $\widehat{\delta}(m,\phi_0)$, precision $\dfrac{n-m}{S_1^2(m,\phi_0)}$, and $(n-1)$ degrees of freedom. Equivalently, the quantity

$$t(\delta)=\frac{(n-m)^{\frac{1}{2}}\left(\delta-\widehat{\delta}(m,\phi_0)\right)}{S_1(m,\phi_0)} \tag{9}$$

is distributed a posteriori as a conditional standard Student t distribution with $(n-1)$ degrees of freedom given m and ϕ_0.

2. Given ϕ_0, the conditional posterior distribution of m is:

$$\pi(m/\phi_0,y)\propto(n-m)^{-\frac{1}{2}}SS(m,\phi_0)^{-\frac{n-1}{2}}, \tag{10}$$

where $SS(m,\phi_0)$ is given in (8).

Proof. See Appendix A.

The unconditional posterior distributions of $t(\delta)$ is,

$$\pi(t(\delta)|y)=\sum_m\int_{\phi_0}\pi(t(\delta)|m,\phi_0,y)\pi(\phi_0|m,y)\pi(m|y) \tag{11}$$

One defines the highest posterior density credible sets of $t(\delta)$. The credible set will be used to define the unconditional p-value and therefore an unconditional test.

Given m, ϕ_0, the $(1-\alpha)$-credible set for $t(\delta)$ is defined as:

$$C_\delta = \left\{ t(\delta)/|t\left(\delta\right)| < t_{\alpha|2}\left(n-1\right) \right\},$$

where $t_{\alpha|2}\left(n-1\right))$ is the $(1-\alpha/2)$th quantile of an t-distribution with $(n-1)$ degrees of freedom. Hence, given m, ϕ_0, the decision rule for H_0, is to reject if $t(0) \in \overline{C_\delta}$, where $\overline{C_\delta}$ is the complement of C_δ.

The unconditional p-value of H_0, therefore, is calculated from (11) to yield:

$$\begin{aligned} P_{\delta=0/y} = & \sum_m \left(\int_{\mathbb{R}} \left[1 - \mathcal{T}_{n-1}\left(|\, t\left(0\right) |\right) \right] \pi(\phi_0|m,y) d\phi_0 \right) \pi(m|y) \\ & = 2 E_m E_{\phi_0} \left[1 - \mathcal{T}_{n-1}\left(|\, t\left(0\right) |\right) \right], \end{aligned} \tag{12}$$

where $\mathcal{T}_{n-1}$ is the cumulative density function of the standard Student t distribution with $(n-1)$ degrees of freedom, and the expectations E_m and E_{ϕ_0} are taken with respect to m and ϕ_0. The quantity given in 12 will be evaluated numerically.

Our test, therefore, rejects H_0, if $P_{\delta=0/Y}$ falls below α.

The quantity (12) will be evaluated numerically by Gibbs Sampler algorithm using the conditional posterior distributions given in Theorem 1 and Lemma 1.

Remark 1. The contamination can occur after the change-point, in this case, the same methodology than above can be adapted for determine the unconditional p-value of H_0. The model is written as follows:

$$\begin{cases} Y_i = X_i = \phi_0 + \varepsilon_i & i = 1, \cdots, m \\ Y_k = X_k + \xi = \phi_1 + \xi + \varepsilon_k & \\ Y_i = X_i = \phi_1 + \varepsilon_i & i = m+1, \cdots, k-1 \\ & i = k+1, \cdots, n\,. \end{cases} \tag{13}$$

where $m \in \{1, \cdots, n-1\}$, $\phi_0, \phi_1 \in \mathbb{R}$, ($\phi_0 \neq \phi_1$), $\xi \in \mathbb{R}$, and $\varepsilon_i \sim N(0, \sigma^2)$ for $i = 1, \cdots, n$, $(\sigma > 0)$, with m, ϕ_0, ϕ_1 and σ are unknown parameters.

The posterior distribution for the parameter $\theta = (m, \phi_0, \phi_1, r)$ $(r = 1/\sigma^2)$ is

$$\begin{aligned} \pi(\theta/y) \propto & r^{-\frac{n}{2}-1} \exp\left\{ -\frac{r}{2} \left[\sum_{i=1}^{m+1} (y_i - \phi_0)^2 + (y_k - (\phi_1 + \xi))^2 \right.\right. \\ & \left.\left. + \sum_{i=m+1}^{k-1} (y_i - \phi_0)^2 + \sum_{i=k+1}^{n} (y_i - \phi_1)^2 \right] \right\}. \end{aligned} \tag{14}$$

3. Simulation Study

Simulation has been used to study the effect of a single outlier on the Bayesian significance test based on the highest posterior density credible set (Kim, 1991; Ghorbanzadeh and Lounes, 2001).

We simulated a sample from the model (2) with $n = 70$, $m = 34$, $\sigma = 1$, $k = 20$ and for different values of ϕ_0, ϕ_1 and ξ. From these observations, by the application of the Gibbs sampler algorithm with 2000 repetitions, we approximate the unconditional p-values for the hypothesis $H_0 : \delta = 0$. The results are given in tables 1, 2 and 3.

$\delta = \phi_1 - \phi_0 = 1,5 - 0,5 = 1$						
ξ	0	5	10	20	40	80
$P_{\delta=0/y}$	$9,9.10^{-6}$	$4,7.10^{-5}$	$2,3.1^{-6}$	$4,0.10^{-5}$	$1,0.10^{-4}$	$8,6.10^{-5}$

Table 1. The unconditional p-values of H_0 for different values of ξ estimated by a Gibbs sampler algrorithm with 2000 repetitions.

$\delta = \phi_1 - \phi_0 = 1,0 - 0,5 = 0,75$						
ξ	0	5	10	20	40	80
$P_{\delta=0/y}$	0,008	0,005	0,001	0,003	0,005	0,002

Table 2. The unconditional p-values of H_0 for different values of ξ estimated by a Gibbs sampler algrorithm with 2000 repetitions.

$\delta = \phi_1 - \phi_0 = 1,0 - 0,5 = 0,5$						
ξ	0	5	10	20	40	80
$P_{\delta=0/y}$	0,037	0,049	0,038	0,069	0,068	0,029

Table 3. The unconditional p-values of H_0 for different values of ξ estimated by a Gibbs sampler algrorithm with 2000 repetitions.

The values of $P_{\delta=0/y}$ in Tables 1, 2 and 3 show that, for $\delta = 1$ and $\delta = 0,75$, the test rejected H_0 at significance level $\alpha = 0,01$, for all values of ξ, and for $\delta = 0,5$ the test rejects H_0 at significance level $\alpha = 0,10$ for all values of ξ.

To illustrate the impact of outliers on the Bayesian significance test for change, we simulated 100 samples from the contaminated model (2) where $n = 70$, $m = 34$, $\sigma = 1$, $k = 20$ and for different values of ϕ_0, ϕ_1 and ξ. And we computed the rejection rates for the hypothesis H_0 at different significance levels α. The results are given in the Tables 4, 5 and 6 as follows Tables 4, 5 and 6 showed that the rejection rates for H_0, for different significance levels α, is substantially the same for all values of the contamination ξ. For $\alpha = 0,10$ the test rejects H_0 with a rate of 98% for $\delta = 1$ and with a rate of more than 80% for all different values of contamination ξ.

This shows that the Bayesian significance test is insensitive to presence of a single outlier i.e, the test is stable under the presence of the single outlier in the data. But the test is sensitive

$\delta = \phi_1 - \phi_0 = 1,5 - 0,5 = 1$						
ξ	0	5	10	20	40	80
$\alpha = 0.01$	0,76	0,77	0,80	0,80	0,78	0,78
$\alpha = 0.05$	0,91	0,94	0,94	0,92	0,92	0,94
$\alpha = 0.10$	0,98	0,98	0,98	0,98	0,98	0,98
$\alpha > 0.10$	0,02	0,02	0,02	0,02	0,02	0,02

Table 4. The rejection rates of H_0 for different values of ξ for 100 samples, evaluated by a Gibbs sampler algrorithm with 2000 repetitions.

$\delta = \phi_1 - \phi_0 = 1,25 - 0,5 = 0,75$						
ξ	0	5	10	20	40	80
$\alpha = 0.01$	0,45	0,39	0,45	0,43	0,45	0,42
$\alpha = 0.05$	0,72	0,71	0,71	0,71	0,71	0,70
$\alpha = 0.10$	0,81	0,80	0,81	0,81	0,81	0,82
$\alpha > 0.10$	0,19	0,20	0,19	0,19	0,19	0,18

Table 5. The rejection rates of H_0 for different values of ξ for 100 samples, evaluated by a Gibbs sampler algrorithm with 2000 repetitions.

$\delta = \phi_1 - \phi_0 = 1,0 - 0,5 = 0,5$						
ξ	0	5	10	20	40	80
$\alpha = 0.01$	0,25	0,23	0,22	0,24	0,23	0,21
$\alpha = 0.05$	0,40	0,40	0,39	0,39	0,36	0,39
$\alpha = 0.10$	0,50	0,46	0,52	0,52	0,50	0,49
$\alpha > 0.10$	0,50	0,54	0,48	0,48	0,50	0,51

Table 6. The rejection rates of H_0 for different values of ξ for 100 samples, evaluated by a Gibbs sampler algrorithm with 2000 repetitions.

to the change of the magnitude of the shift in the mean. For $\alpha = 0,05$, the rejection rate of H_0 is more than 90% for $\delta = 1$ and it is only more than 70% for $\delta = 0,5$.

Remark 2. It is noticed that the position of contamination and of change point has no effect on the results of simulation.

4. Conclusion

In this paper, we presented a Bayesian significance test of change in mean in independent gaussian samples in the presence of single outlier. By numerical studies, we have showed that the bayesian significance test based on the HPD region is insensitive to the presence of outlier in the data. The cases where the position and for the magnitude of the contamination are unknown will be considered in an upcoming paper.

Appendix A: Proof of theorem

Derivation of the posterior distribution of δ and m :
By transforming the parameter set $\Theta = (m, \phi_0, \phi_1, r)$ into $\Phi = (m, \phi_0, \delta)$, we can form the posterior distribution of Φ; that is,

$$\pi(\Phi/y) = \int_r \pi(m, \phi_0, \delta + \phi_0, r/y)dr$$

$$\propto \left[\sum_{i=1}^{k-1}(y_i - \phi_0)^2 + (y_k - (\phi_0 + \xi))^2 + \sum_{i=k+1}^{m}(y_i - \phi_0)^2 + \sum_{i=m+1}^{n}(y_i - \delta - \phi_0)^2\right]^{-\frac{n}{2}} \tag{A1}$$

$$\propto \left[\sum_{i=1}^{k-1}(y_i - \phi_0)^2 + (y_k - (\phi_0 + \xi))^2 + \sum_{i=k+1}^{m}(y_i - \phi_0)^2 + \sum_{i=m+1}^{n}(y_i - \phi_0)^2 - \frac{\left[\sum_{i=m+1}^{n}(y_i - \phi_0)\right]^2}{n-m} + (n-m)\left(\delta - \widehat{\delta}(m, \phi_0)\right)^2\right]^{-\frac{n}{2}}. \tag{A2}$$

i) By application of Bayes theorem, the posterior conditional distribution of δ is obtained as given in (7).

ii) By integration with respect of δ, we obtained the joint posterior distribution of m and ϕ_0:

$$\pi(m, \phi_0/y) \propto (n-m)^{-\frac{1}{2}} SS(m, \phi_0)^{-\frac{n-1}{2}}.$$

and by application of Bayes theorem, the posterior conditional distribution of m is given as in (10).

Appendix B: Conditional posterior distribution of ϕ_0

Lemma 1. *Given m and δ, the conditional posterior distribution of ϕ_0 is:*

$$\pi\left(\phi_0|m, \delta, y\right) \propto \left\{1 + \frac{n\left(\phi_0 - \widehat{\phi_0}(m, \delta)\right)^2}{(n-1)S_2^2(m, \delta)}\right\}^{-\frac{n}{2}}, \tag{B1}$$

where,

$$\widehat{\phi_0}(m,\delta) = \frac{b(m,\delta)}{n}\,,$$

and

$$S_2^2(m,\delta) = \frac{1}{(n-1)}\left[a(m,\delta) - \frac{b^2(m,\delta)}{n}\right].$$

with,

$$a(m,\delta) = \sum_{i=1}^{k-1} y_i^2 + (y_k-\xi)^2 + \sum_{i=k+1}^{m} y_i^2 + \sum_{i=m+1}^{n} (y_i-\delta)^2\,,$$

and

$$b(m,\delta) = \sum_{i=1}^{k-1} y_i + (y_k-\xi) + \sum_{i=k+1}^{m} y_i + \sum_{i=m+1}^{n} (y_i-\delta)\,.$$

Thus is the Student t distribution with location parameter $\widehat{\phi_0}(m,\delta)$*, with precision* $\dfrac{n}{S_2^2(m,\delta)}$*, and* $(n-1)$ *degrees of freedom.*

Acknowledgements

The author would like to thank the editor and the anonymous referees for several helpful corrections, their careful comments and valuable suggestions that led to many improvements in the paper.

References

Belkacem, C. and Fellag, H., 2012. Bayesian change-point estimation in the presence of a single outlier, *Journal Afrika Statistika* Vol. **7**, 381-390.

Barnett, V. and Lewis, T., 1978. *Outliers in Statistical Data*, John Wiley & Sons, New York.

Brown, R.L., Durbin, J. and Evans, J.M. 1975. Techniques for testing the constancy of regression relationships over time (with discussion). *J. R. Statist. Soc. A* . **138**, 149-63.

Chernoff, H. and Zacks, S., 1964. Estimating the current mean distribution which is subjected to change in time, *Ann. Math. Statist.* **35**, 999-1018.

Ghorbanzadeh, D. and Lounes, R., 2001. Bayesian analysis for detecting a change in exponential family, *Applied Mathematics and Computation.* **124**, 1-15.

Kim, H.-J., 1996. Change-point detection for correlated observations, *Statistica Sinica* . **6**, 275-287.

Kander, A. and Zacks. S., 1966. Test procedure for possible change in parameters of statistical distributions occuring at unknown time point, *Ann. Math. Statist*, **37**, 1196-1210.

Kim, D., 1991. A Bayesian significance test of the stationarity of regression parametres, *Biometrika.* **78**, 667–675, 1991

Hinkley, D.V., 1970. Inference about the change-point in a sequence of random variables. *Biometrika.* **57**, 1-17.

Page, E.S., 1954. Continuous inspection schemes. *Biometrika.* **41**, 100-115.

Page, E.S., 1955. A test for change in a parameter occurring at an unknown point. *Biometrika*. **42**, 523-527.

Sen, A. and Srivastava, M.S., 1975. Some one-sided tests for change in level. *Technometrics*. **17**, 61-64.

Siegmund, D., 1986. Boundary crossing probabilities and statistical applications. *Ann. Statist*. **14**, 361-404.

Siegmund, D., 1988. Confidence sets in change point problem. *Int. Statist. Rev*. **56**, 31-48.

Verdinelli I. and Wasserman, L., 1991. Bayesian analysis of outlier problems using the Gibbs sampler. *Statistics and Computing*. **1**, 105-117.

Worsley, K.J., 1983. The power of likelihood ratio and cumulative sum tests for a change in a binomial probability. *Biometrika*. **70**, 455-464.

Worsley, K.J., 1986. Confidence regions and tests for a change-point in a sequence of exponential family random variables. *Biometrika*. **73**, 91-104.