Journal of the Korean Data & Information Science Society
2015, **26**(6), 1565–1572

http://dx.doi.org/10.7465/jkdi.2015.26.6.1565
한국데이터정보과학회지

Robust CUSUM test for time series of counts and its application to analyzing the polio incidence data[†]

Jiwon Kang[1]

[1]Research Institute for Basic Sciences, Jeju National University

Received 7 September 2015, revised 23 November 2015, accepted 24 November 2015

Abstract

In this paper, we analyze the polio incidence data based on the Poisson autoregressive models, focusing particularly on change-point detection. Since the data include some strongly deviating observations, we employ the robust cumulative sum (CUSUM) test proposed by Kang and Song (2015) to perform the test for parameter change. Contrary to the result of Kang and Lee (2014), our data analysis indicates that there is no significant change in the case of the CUSUM test with strong robustness and the same result is obtained after ridding the polio data of outliers. We additionally consider the comparison of the forecasting performance. All the results demonstrate that the robust CUSUM test performs adequately in the presence of seemingly outliers.

Keywords: Minimum density power divergence estimator, Poisson autoregressive model, robust parameter change test, the polio incidence data.

1. Introduction

In recent years, time series of counts are commonly observed in real-world applications. The fields include insurance industry (e.g., the number of claim counts), economics (e.g., the number of transactions of some stock), engineering (e.g., the number of traffic accidents) and epidemiology (e.g., the number of people with a certain disease). For a general review of time series of counts, we refer to Jung *et al.* (2006) and Weiß (2008). It is widely recognized that time series models applied to various fields such as finance often undergo parameter changes. Since such changes can lead to an invalid inference, the change point problem has attracted much attention from many authors such as Lee and Lee (2007), Aue and Horváth (2013) and Kitabo and Kim (2015). For historical background and general review, Csörgő and Horváth (1997) and Chan and Gupta (2000), and the articles cited therein. As Kang and Lee (2014) addressed, integer-valued time series (particularly in the field of epidemiology) can also undergo a significant change in parameters due to, for example, changes in the quality of health care and the state of patients' health. However, only recently have researchers begun to study on the detection of parameter changes for integer-valued time series, Kang and Lee (2014) and Doukhan and Kengne (2015).

[†] This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (2015R1A6A3A01060651).

[1] Special researcher, Research Institute for Basic Sciences, Jeju National University, Jeju 690-756, Korea. E-mail: jwkang.stats@gmail.com

It is also well known that outliers affect parameter estimations and test procedures. In time series models, this problem has been investigated by many authors. See, for example, Tsay (1988) and Lee and Park (2001). However, to the best of our knowledge, few works are devoted to integer-valued time series models. Recently, Kang and Song (2015) proposed the robust cumulative sum (CUSUM) test for detecting change points in Poisson autoregressive models and demonstrated that the proposed test can be a useful tool in testing for parameter change when outliers are suspected to contaminate data.

The objective here is to employ the robust CUSUM test to analyze the monthly number of polio cases in the US from January 1970 to December 1983. The polio data is one of the most famous data sets in the context of time series of counts. The data are published by the United States (US) Centers for Disease Control and consist of 168 observations with some strongly deviating data points. This data set has been previously studied by many researchers, including Zeger (1988), Davis *et al.* (2000) and Jung and Tremayne (2011). Recently, Kang and Lee (2014) performed the CUSUM test to detect change points in this data based on Poisson autoregressive models. They showed that there exist a significant change in parameter, which may be because the outliers affect the test procedure. In this study, we employ the robust CUSUM test of Kang and Song (2015) to analyze the polio data, and compare with the CUSUM test of Kang and Lee (2014).

This paper is organized as follows. In Section 2, we review the robust CUSUM test for Poisson autoregressive models introduced by Kang and Song (2015) and its asymptotic properties. In Section 3, we apply the proposed test for analyzing the polio data, and compare the one-step-ahead forecasting performance of the several models. Section 4 concludes the paper.

2. Robust CUSUM test for Poisson autoregressive model

The Poisson autoregressive model is defined by

$$X_t|\mathcal{F}_{t-1} \sim Poisson(\lambda_t), \quad \lambda_t = f_\theta(\lambda_{t-1}, X_{t-1}) \;\text{ for all } t \in \mathbb{Z}, \tag{2.1}$$

where f_θ is a known positive function on $[0,\infty) \times \mathbb{N}_0, \mathbb{N}_0 = \mathbb{N} \cup \{0\}$, depending on unknown parameter $\theta \in \Theta \subset \mathbb{R}^d$, and $\mathcal{F}_{t-1}$ is the σ-field generated by $\{X_{t-1}, X_{t-2}, \ldots\}$. Denote the true value of θ by θ_0. In what follows, it is assumed that the function f_θ satisfies the following condition:

> For all $\theta \in \Theta$, $|f_\theta(\lambda, x) - f_\theta(\lambda', x')| \leq \kappa_1|\lambda - \lambda'| + \kappa_2|x - x'|$ for all $\lambda, \lambda' \geq 0$ and $x, x' \in \mathbb{N}_0$, where κ_1 and κ_2 are nonnegative real numbers with $\kappa := \kappa_1 + \kappa_2 < 1$.

Under the above condition, there is a strictly stationary and ergodic solution for model (2.1) and any order moments of X_t and λ_t are finite (Neumann, 2011; Doukhan *et al.*, 2012).

In particular, model (2.1) with a linear specification is referred to as the INGARCH(1,1) model, such that,

$$X_t|\mathcal{F}_{t-1} \sim Poisson(\lambda_t), \quad \lambda_t = w + a\lambda_{t-1} + bX_{t-1} \;\text{ for all } t \in \mathbb{Z},$$

where $\theta = (w, a, b)^T$ and $w > 0, a \geq 0, b \geq 0$. Ferland *et al.* (2006) showed that a strictly stationary ergodic solution exists and all moments of X_t and λ_t are finite if $a + b < 1$. Count time series data are often overdispersed, that is, the variance is bigger than the mean of data.

This model is broadly used to analyze dependent count data with overdispersion because of the fact that $E(X_t) < Var(X_t)$.

In this paper, we consider model (2.1) with θ replaced by θ_t and set up the null and alternative hypotheses as follows:

$$\begin{aligned} &H_0 : \theta_1 = \cdots = \theta_n = \theta \quad vs. \\ &H_1 : \text{not } H_0. \end{aligned}$$

To perform the above test, Kang and Lee (2014) proposed the CUSUM test based on maximum likelihood estimator (MLE):

$$T_n^{mle} = \max_{1 \le k \le n} \frac{k^2}{n} (\hat{\theta}_k - \hat{\theta}_n)^T \hat{I}_n (\hat{\theta}_k - \hat{\theta}_n),$$

where $\hat{\theta}_k$ is the MLE of θ_0 based on $X_1, \cdots X_k$ for $k = 1, \cdots, n$ and $\hat{I}_n$ is a consistent estimator of the Fisher information matrix. However, it is well known that the MLE and the test procedure based on MLE are strongly influenced when the data are contaminated by outliers.

As a robust estimator, Basu *et al.* (1998) proposed the minimum density power divergence estimator (MDPDE) and demonstrated that the MDPDE has strong robustness properties, with low losses in asymptotic efficiency relative to the MLE. Consequently, for model (2.1) in the presence of outliers, we consider the robust CUSUM test, which is introduced by Kang and Song (2015):

$$T_n^{mdpde} = \max_{1 \le k \le n} \frac{k^2}{n} (\hat{\theta}_{\alpha,k} - \hat{\theta}_{\alpha,n})^T \hat{J}_\alpha \hat{K}_\alpha^{-1} \hat{J}_\alpha (\hat{\theta}_{\alpha,k} - \hat{\theta}_{\alpha,n}),$$

where $\hat{\theta}_{\alpha,k}$ is the MDPDE of θ_0 based on $X_1, \cdots X_k$ for $k = 1, \cdots, n$. Here, $\hat{K}_\alpha$ and $\hat{J}_\alpha$ are consistent estimators of K_α and J_α, which are those defined in (2.3), respectively. More specifically, the MDPDE is obtained by

$$\hat{\theta}_{\alpha,n} = \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{t=1}^{n} \tilde{l}_{\alpha,t}(\theta),$$

where

$$\tilde{l}_{\alpha,t}(\theta) := \begin{cases} \displaystyle\sum_{y=0}^{\infty} \left(\frac{e^{-\tilde{\lambda}_t} \tilde{\lambda}_t^y}{y!} \right)^{1+\alpha} - \left(1 + \frac{1}{\alpha}\right) \left(\frac{e^{-\tilde{\lambda}_t} \tilde{\lambda}_t^{X_t}}{X_t!} \right)^{\alpha} & , \alpha > 0\,, \\ \tilde{\lambda}_t - X_t \log \tilde{\lambda}_t + \log(X_t!) & , \alpha = 0, \end{cases} \tag{2.2}$$

and $\tilde{\lambda}_t$ are defined recursively by $\tilde{\lambda}_t = f_\theta(\tilde{\lambda}_{t-1}, X_{t-1})$ with arbitrarily chosen $\tilde{\lambda}_1$. Note that

$$K_\alpha := \frac{1}{(1+\alpha)^2} E\left(\frac{\partial l_{\alpha,t}(\theta_0)}{\partial \theta} \frac{\partial l_{\alpha,t}(\theta_0)}{\partial \theta^T} \right) \quad \text{and} \quad J_\alpha := -\frac{1}{1+\alpha} E\left(\frac{\partial^2 l_{\alpha,t}(\theta_0)}{\partial \theta \partial \theta^T} \right), \tag{2.3}$$

where $l_{\alpha,t}(\theta)$ is defined by substituting λ_t for $\tilde{\lambda}_t$ in (2.2). As consistent estimators of K_α and J_α, we consider to use

$$\hat{K}_\alpha = \frac{1}{(1+\alpha)^2} \frac{1}{n} \sum_{t=1}^{n} \frac{\partial \tilde{l}_{\alpha,t}(\hat{\theta}_{\alpha,n})}{\partial \theta} \frac{\partial \tilde{l}_{\alpha,t}(\hat{\theta}_{\alpha,n})}{\partial \theta^T} \quad \text{and} \quad \hat{J}_\alpha = -\frac{1}{(1+\alpha)} \frac{1}{n} \sum_{t=1}^{n} \frac{\partial^2 \tilde{l}_{\alpha,t}(\hat{\theta}_{\alpha,n})}{\partial \theta \partial \theta^T}.$$

Assuming suitable regularity conditions, Kang and Song (2015) showed that under H_0,

$$T_n^{mdpde} \xrightarrow{d} \sup_{0\le s\le 1} \|\mathbf{B}_d^{\circ}(s)\|^2,$$

where $\mathbf{B}_d^{\circ}$ is a standard d-dimensional Brownian bridge, and H_0 is rejected if T_n^{mdpde} is large. They demonstrated, via a simulation study, that T_n^{mdpde} is valid when outliers are involved in observations.

Remark. The tuning parameter α controls the trade-off between the efficiency and robustness of the MDPDE. Note that the MDPDE with $\alpha = 0$ is exactly the same as the MLE, which is fully efficient. For $\alpha > 0$, the MDPDE becomes more robust against outliers but less efficient as α increases (Basu *et al.*, 1998).

3. Real data analysis

In this section, we apply the robust CUSUM test in Section 2 to analyze the polio data. The top row of Figure 1 provides the time series plot of the polio data, which shows that there are some strongly deviating data points. The empirical mean and variance of the data are 1.33 and 3.5, respectively. This suggests that there is some overdispersion in the data.

3.1. Robust CUSUM test

We focus on investigating whether a parameter change exists in this time series. First, we fit an INGARCH(1,1) model to the polio data. In order to test for a parameter change, Kang and Lee (2014) performed T_n^{mle} and obtained $T_n^{mle} = 5.859$, which indicates rejection of the null hypothesis at the nominal level 0.05. T_n^{mle} has a maximum at $t = 35$, which corresponds to November 1972. In this study, T_n^{mdpde} is performed at the nominal level 0.05. The result is presented in “the polio data” of Table 3.1. It can be seen that the null hypothesis is not rejected in the case of T_n^{mdpde} with large α. This means that the test procedure based on MDPDE with large α seems to be not affected by suspected outliers because of the strong robustness of the MDPDE.

To remove the effect of outliers, we clean the polio data by using the approach introduced by Fokianos and Fried (2010), which is designed for the stepwise detection, classification, and elimination of multiple intervention effects. The plot of the clean polio data is at the bottom of Figure 3.1. After completing the data cleaning procedure, we also fit an INGARCH(1,1) model to the clean polio data and perform T_n^{mle} and T_n^{mdpde} at the nominal level 0.05. From “the clean polio data” of Table 3.1, there is no significant change in all the cases, which is in agreement with that of T_n^{mdpde} with large α. All these results demonstrate that outliers apparently seem to affect test procedures and thus the robust CUSUM test performs adequately in the presence of seemingly outliers.

Table 3.1 T_n^{mle} and T_n^{mdpde} after fitting the INGARCH model for the polio data and for the clean polio data

	T_n^{mle}	T_n^{mdpde} with α				
		0.1	0.2	0.3	0.5	1.0
the polio data	5.859*	6.546*	4.167*	3.840*	2.909	2.781
the clean polio data	2.016	2.019	1.891	1.727	1.428	1.341

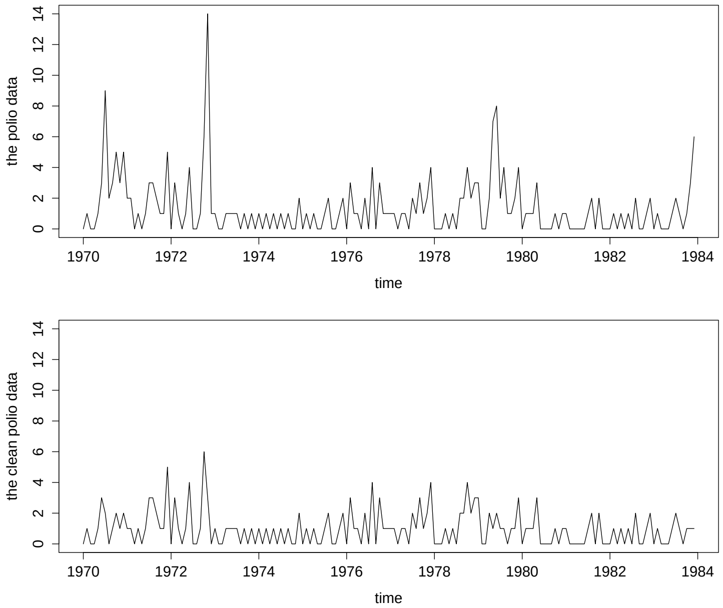

Figure 3.1 Plots of the polio data (top) and the clean polio data (bottom)

3.2. Comparison of forecasting performance

In Section 3.1, the robust CUSUM test has provided the different results according to the α, for the polio data. In order to study whether there is a significant change in parameters or not, we additionally consider the comparison of the forecast performance of the ordinary INGARCH model and the INGARCH model with a break. The estimations of the models can be carried out in MLE and MDPDE. The data set is divided into an in-sample period comprising the first 118 observations and an out-of-sample period composed by the remaining 50 observations, which will be used to assess the forecasting abilities.

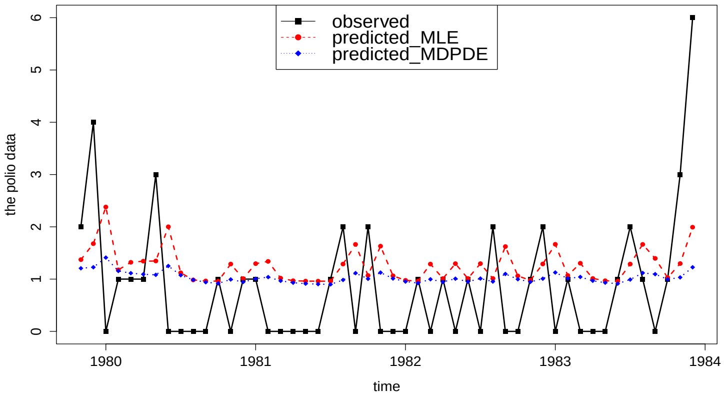

Figure 3.2 The observed and predicted values for the ordinary INGARCH model based on MLE and MDPDE

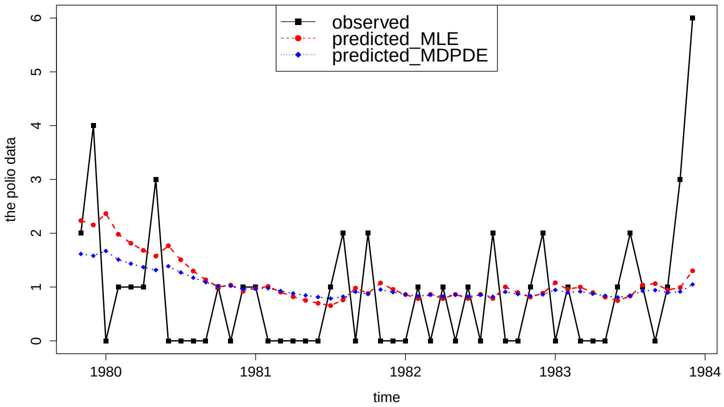

Figure 3.3 The observed and predicted values for the INGARCH model with a break based on MLE and MDPDE

First, we consider the ordinary INGARCH(1,1) model. Let $(\hat{w}, \hat{a}, \hat{b})$ is the MLE of the INGARCH model by using the in-sample data. For forecasting purposes, we calculate the one-step-ahead forecasts error which is obtained from the estimated model, as follows

$$e_t = X_t - \hat{\lambda}_t$$

where $\hat{\lambda}_t$ are defined recursively by $\hat{\lambda}_t = \hat{w} + \hat{a}\hat{\lambda}_{t-1} + \hat{b}X_{t-1}$ and $\hat{\lambda}_1 = 0$. Comparison of the predictive performance among the models is implemented by calculating the forecasting accuracy measures. There are two measures of mean square error (MSE) and mean absolute error (MAE), which are defined by

$$\text{MSE} = \frac{\sum_{t=119}^{168} e_t^2}{50} \quad \text{and} \quad \text{MAE} = \frac{\sum_{t=119}^{168} |e_t|}{50}.$$

Whereas MSE can lead to spurious inference in the presence of outliers, MAE is inherently robust to outliers (Park, 2002). Thus, in this paper, we compute not only MSE but also MAE to evaluate the predictive performances. In the same way, we calculate two measures by using MDPDE with $\alpha = 0.5$. The result is presented in “ordinary INGARCH model” of Table 3.2.

We now consider the INGARCH model with a break at $t = 35$. We obtain the MLE and the MDPDE of the INGARCH model by using the subseries $36 \leq t \leq 118$, and then compute MSEs and MAEs respectively. The result is presented in “INGARCH model with a break” of Table 3.2. Figure 3.2 and Figure 3.3 are the observed versus predicted plots in out-of-sample period for two models. In terms of MLE, the INGARCH model with a break is superior to the ordinary INGARCH model on both two measures. Furthermore, all the models based on MDPDE outperform the models based on MLE, and perform similarly in forecasting whether a parameter change is considered or not. These may be because outliers damage the MLE and the forecasting performance of the model based on MLE, while the MDPDE is robust to outliers. It is noteworthy that in all of forecasting measures, the ordinary INGARCH model based on MDPDE is superior to the INGARCH model with a break based on MLE. Although not reported here, we could see that the results obtained by other α except 0.1 are similar to the case of $\alpha = 0.5$.

Overall, our findings suggest that when outliers are involved in observations, the ordinary INGARCH model based on MDPDE has a tendency to improve substantially in forecasting over the INGARCH models considering a parameter change.

Table 3.2 Out-of-sample forecasting results after fitting the INGARCH model with and without a break based on MLE and MDPDE

	ordinary INGARCH Model		INGARCH Model with a break	
	MLE	MDPDE	MLE	MDPDE
MSE	1.528	1.420	1.442	1.406
MAE	0.974	0.885	0.934	0.893

4. Conclusion

The parameter estimations and the test statistics can be affected in the presence of extreme values. This motivation allowed the paper to use the robust CUSUM test, which can reduce the impact of outliers. In real data analysis, we employed the robust CUSUM test after fitting Poisson autoregressive models to the polio data. The result of the CUSUM test based on MLE indicated a significant change, while there is no significant change in the case of the CUSUM test based on MDPDE with large α. Furthermore, there is also no significant change in the clean polio data. Thus, the robust CUSUM test performs adequately in the presence of outliers. Additionally, through the comparison of forecasting performance, we demonstrated that substantial improvement of forecasts may indeed be achieved by the approach of the robust estimation instead of considering a parameter change. In fact, there are no universal rules for the selection of the α even though it is needed in practice. We leave this as a task for our future study.

References

Aue, A. and Horváth, L. (2013). Structural breaks in time series. *Journal of Time Series Analysis*, **34**, 1-16.

Basu, A., Harris, I. R., Hjort, N. L. and Jones, M. C. (1998). Robust and efficient estimation by minimizing a density power divergence. *Biometrika*, **85**, 549-559.

Chan, J. and Gupta, A. K. (2000). *Parametric Statistical Change Point Analysis*, Birkhauser, Boston.

Csörgő, M. and Horváth, L. (1997). *Limit theorems in change-point analysis*, Wiley, New York.

Davis, R. A., Dunsmuir, W. and Wang, Y. (2000). On autocorrelation in a Poisson regression model. *Biometrika*, **87**, 491-505.

Doukhan, P., Fokianos, K. and Tjøstheim, D. (2012). On weak dependence conditions for Poisson autoregressions. *Statistics and Probability Letters*, **82**, 942-948.

Doukhan, P. and Kengne, W. (2015). Inference and testing for structural change in general Poisson autoregressive models. *Electronic Journal of Statistics*, **9**, 1267-1314.

Ferland, R., Latour, A. and Oraichi, D. (2006). Integer-valued GARCH process. *Journal of Time Series Analysis*, **27**, 923-942.

Fokianos, K. and Fried, R. (2010). Interventions in INGARCH processes. *Journal of Time Series Analysis*, **31**, 210-225.

Jung, R. C., Kukuk, M. and Liesenfeld, R. (2006). Time series of count data: Modeling, estimation and diagnostics. *Computational Statistics and Data Analysis*, **51**, 2350-2364.

Jung, R. C. and Tremayne, A. R. (2011). Useful models for time series of counts or simply wrong ones. *Advances in Statistical Analysis*, **95**, 59-91.

Kang, J. and Lee, S. (2014). Parameter change test for Poisson autoregressive models. *Scandinavian Journal of Statistics*, **41**, 1136-1152.

Kang, J. and Song, J. (2015). Robust parameter change test for Poisson autoregressive models. *Statistics and Probability Letters*, **104**, 14-21.

Kitabo, C. A. and Kim, J. (2015). Comparative analysis of Bayesian and maximum likelihood estimators in change point problems with Poisson process. *Journal of the Korean Data & Information Science Society*, **26**, 261-269.

Lee, J. and Lee, H. (2007). Change point estimators in monitoring the parameters of an AR(1) plus an additional random error model. *Journal of the Korean Data & Information Science Society*, **18**, 963-972.

Lee, S. and Park, S. (2001). The cusum of squares test for scale changes in infinite order moving average processes. *Scandinavian Journal of Statistics*, **28**, 625-644.

Neumann, M. (2011). Absolute regularity and ergodicity of Poisson count processes. *Bernoulli*, **17**, 1268-1284.

Park, B. (2002). An outlier robust GARCH model and forecasting volatility of exchange rate returns. *Journal of Forecasting*, **21**, 381-393.

Tsay, R. S. (1988). Outliers, level shifts, and variance changes in time series. *Journal of Forecasting*, **7**, 1-20.

Weiß, C. H. (2008). Thinning operations for modeling time series of counts-a survey. *Advanced in Statistical Analysis*, **92**, 319-341.

Zeger, S. L. (1988). A regression model for time series of counts. *Biometrika*, **75**, 621-629.