A TEST FOR SELF-EXCITING CLUSTERING MECHANISM

YUCHEN FAMA AND VLADIMIR POZDNYAKOV

DEPARTMENT OF STATISTICS, UNIVERSITY OF CONNECTICUT

Abstract. In this paper, we propose a test to distinguish between data with cluster pattern where the variables are dependent in a self-exciting fashion versus independently identically distributed random variables. We also developed asymptotic distribution of the test statistic with closed-form covariance structure. Comparisons with scan statistics are discussed in the context of simulated earthquake data. Applications to two data sets are discussed.

1. Introduction

As it happens in quality control framework one often needs to test the uniformity of discrete data against a clustering alternative, see Bowman (1999), Yu and Zelterman (2002), and Glaz and Zhang (2006). A type of cluster pattern that we have in mind here is the one that is caused by a "self-exciting" mechanism. For example, when an excess of failures in the recent history increase the chance of failure in the future. Starting from a somewhat artificial switching model first we introduce a simple statistic that can track the clustering of this nature. Then some asymptotic results are presented for a sequence of non-negative integer-valued random variables. Via simulations we can show that in a certain situation the new

Key words and phrases. Asymptotic normality, Exponential family, Self-exciting, Cluster detection, Scan statistics, Moving sum.

Address. 745 Merrow Road Unit 180, Coventry, CT 06238, USA. *E-mail.* yfama@travelers.com.

test outperforms a popular approach based on scan statistics. Finally, application of the test to two data sets are discussed.

2. Motivation

Let us first consider a hypothesis testing procedure that tries to discriminate between a sequence of independent identically distributed (iid) random variables and k-order ($k \geq 1$) Markov chain with state space $Z_+ = \{0, 1, 2, ...\}$. More specifically, we assume that under both the null hypothesis and the alternative the discrete random variables $\{X_i\}_{1 \leq i \leq k}$ (with values from Z_+) has the same (arbitrary) initial distribution. But under the null $\{X_i\}_{i>k}$ are iid with a discrete density

$$f_{\theta_0}(x) = h(x) \exp[\theta_0 t(x) - \psi(\theta_0)]$$

from the exponential family.

Under the alternative hypothesis, the random variables $\{X_i\}_{i \geq 1}$ forms a kth-order Markov chain. The transition probabilities are fully determined by a binary function that we call switch:

$$s : Z_+^k \mapsto \{0, 1\}.$$

Let $s_i = s(X_{i-k}, X_{i-k+1}, ..., X_{i-1})$ for $i > k$. If the switch $s_i = 0$ then X_i has density $f_{\theta_0}(\cdot)$, if $s_i = 1$, then X_i has density $f_{\theta_1}(\cdot)$ from the same exponential density but with $\theta_1 > \theta_0$. In other words, the conditional density of X_i is given by

$$\begin{aligned} f\left(x_i | \bigcap_{i-k \leq j \leq i-1} X_j = x_j\right) &= \begin{cases} f_{\theta_0}(x_i), & \text{if } s_i = 0, \\ f_{\theta_1}(x_i), & \text{if } s_i = 1 \end{cases} \\ &= f_{\theta_0}(x_i)(1 - s_i) + f_{\theta_1}(x_i) s_i. \end{aligned}$$

Suppose that the switch function s is given and one wants to test an iid behavior versus Markov chain based on the data set $\{X_i\}_{1\le i\le N}$. Since when $\theta_0 = \theta_1$, the Markov chain described above is, in fact, an iid sequence, we can formally write the testing as

$$H_0 : \theta_1 = \theta_0 \quad vs \quad H_1 : \theta_1 > \theta_0.$$

Now, one can show that the loglikelihood ratio for testing H_0 vs H_1 is given by

$$\begin{aligned}\log R(x) &= \log\left(\prod_{k+1\le i\le N}\left[\frac{f_{\theta_1}(x_i)}{f_{\theta_0}(x_i)}\right]^{s_i}\right)\\ &= (\theta_1-\theta_0)\sum_{i=k+1}^{N} s_i t(x_i) - (\psi(\theta_1)-\psi(\theta_0))\sum_{i=k+1}^{N} s_i\end{aligned}$$

Thus we can see two statistics $\sum_{i=k+1}^N s_i$ and $\sum_{i=k+1}^N s_i t(x_i)$ must play an important role. Note that both statistics have very simple meanings. Random variable $\sum_{i=k+1}^N s_i$ just counts how often the switch is activated, and the statistic $\sum_{i=k+1}^N s_i t(x_i)$ adds up $t(x_i)$ for which the preceding observations in the window of size k trigger the switch.

Admittedly, it is very unrealistic to hope that any real world data would follow this simple model, and, moreover, the assumption that the switch function is known is too strong. But the expression of the loglikelihood ratio suggests that the large values of $\sum_{i=k+1}^N s_i t(x_i)$ will indicate in favor of the alternative. So, we propose a cluster detection test in the following form:

$$\Psi(x) = \begin{cases} 1, & \text{if } \sum_{i=k+1}^N s_i t(X_i) > C,\\ 0, & \text{otherwise,}\end{cases}$$

for some natural choice of the switch function s. For example, one can consider a moving average/threshold type switch:

$$s(x_1, ..., x_k) = \mathbf{1}_{x_1+\cdots+x_k \geq L}. \tag{2.1}$$

The next question is how to choose the constant C.

Remark 1. If the switch is unknown, one should be cautious using the primary asymptotic distribution theory for the likelihood ratio test. Because it requires the assumption that the model is quadratic mean differentiable. However, this assumption is violated when the switching parameters are not identified under the null hypothesis, thus the likelihood function is *flat* (with respect to the unidentified parameters) at the optimum.

One solution is to bound the asymptotic distribution of standardized likelihood ratio statistics, as suggested by Hansen (1992). Practically, we can form a grid over the parameters defining the switch and calculate the likelihood ratio statistic at each cross point. The supremum of these likelihood ratio statistics is used as test statistic, and the asymptotic distribution can be generated using normal random variables with empirical covariance function.

3. Asymptotic Theory for the Null Case

If the sample size N is large then $S_N = \sum_{i=k+1}^{N} s_i t(X_i)$ is asymptotically normal. More specifically, we have the following result.

Proposition 1. *Let $\{X_i\}_{1\leq i\leq N}$ be iid random variables with $E[X_i^{12}] < \infty$.Then as $N \to \infty$*

$$\frac{S_N - N\vartheta}{\sigma\sqrt{N}} \to \mathcal{N}(0,1)$$

in distribution, where

$$\vartheta = E(s_{k+1}t(X_{k+1})) = P(X_1 + \cdots + X_k \geq L)E(t(X_1))$$

and

$$\sigma^2 = Var(s_{k+1}t(X_{k+1})) + 2\sum_{c=1}^{k} Cov\left[s_{k+1}t(X_{k+1}), s_{k+1+c}t(X_{k+1+c})\right]. \tag{3.1}$$

Proof. First, note that $\{s_i t(X_i)\}_{i\geq k+1}$ is a stationary sequence. Second, since $s_i t(X_i)$ is a function of $X_{i-k}, X_{i-k+1}, ..., X_{i-1}, X_i$, $\{s_i t(X_i)\}_{i\geq k+1}$ is a stationary sequence of k-dependent random variables. Therefore, by the central limit theorem for this dependent sequence (for instance, Billingsley (1995, p. 364)) we immediately obtain the asymptotic normality of S_N. □

3.1. **Calculation of the covariance.** So, the main question now is how to compute σ in Proposition 1. For common discrete distributions and moving average/threshhold type switch as in (2.1) this task is quite straightforward and one can easily derive the exact result for the covariance structure. More specifically, let us consider the case when density f_{θ_0} is such that $t(x) = x$. Define

$$\delta = s_{k+1}X_{k+1},$$

$$\eta = s_{c+k+1}X_{c+k+1}, \text{ where } c \geq 1,$$

$$P(X_1 + ... + X_t = T) = f(t, T),$$

$$P(X_1 + ... + X_t \geq T) = F(t, T),$$

$$P(X_1 = T) = f(1, T) = f(T) = f_{\theta_0}(T).$$

Proposition 2. *If $EX_i^2 < \infty$, then*

$$E\delta = E\eta = F(k, L)EX_1,$$

for $c = 1$

$$E(\delta\eta) = EX_1 \sum_{i=1}^{\infty} if(i) \sum_{m=L-i}^{\infty} F(k-1, m)F(1, L-m),$$

for $2 \leq c \leq k-1$

$$E(\delta\eta) = EX_1 \sum_{i=1}^{\infty} if(i) \sum_{m=0}^{\infty} f(k-c, m)F(c, L-m)F(c-1, L-m-i),$$

and for $c = k$

$$E(\delta\eta) = EX_1 \sum_{i=1}^{\infty} if(i)F(k, L)F(k-1, L-i).$$

Proof. Because of independence of $\{X_i\}$ and stationarity of $\{s_iX_i\}$ we get

$$\begin{aligned} E\delta = E\eta &= Es_{k+1}X_{k+1} \\ &= E(\mathbf{1}_{X_1+...+X_k \geq L} \cdot X_{k+1}) \\ &= P(X_1 + ... + X_k \geq L)EX_{k+1} \\ &= F(k, L)EX_1. \end{aligned}$$

Now, it is obvious that

$$\begin{aligned} E(\delta\eta) &= E\left(\mathbf{1}_{X_1+...+X_k \geq L} \cdot X_{k+1} \cdot \mathbf{1}_{X_{c+1}+...+X_{c+k} \geq L} \cdot X_{c+k+1}\right) \\ &= E\left(\mathbf{1}_{X_1+...+X_k \geq L} \cdot X_{k+1} \cdot \mathbf{1}_{X_{c+1}+...+X_{c+k} \geq L}\right) EX_1 \\ &= E^* EX_1. \end{aligned}$$

To deal with the first term E^* we subset the sequences $\{X_1,...,X_k\}$, $\{X_{k+1}\}$ and $\{X_{c+1},...,X_{c+k}\}$ into non-overlapping subsequences. For $2 \leq c \leq k-1$ we have

$$\begin{aligned}E^* =&E\left(\mathbf{1}_{X_1+...+X_k\geq L}\cdot X_{k+1}\cdot \mathbf{1}_{X_{c+1}+...+X_{c+k}\geq L}\right)\\ =&\sum_{i=0}^{\infty} iP\left(X_1+...+X_k \geq L \cap X_{k+1}=i\cap X_{c+1}+...+X_{c+k}\geq L\right)\\ =&\sum_{i=0}^{\infty} i\sum_{m=0}^{\infty} P(X_1+...+X_c\geq L-m\cap X_{c+1}+...+X_k=m\\ &\qquad\qquad \cap X_{k+1}=i\cap X_{k+2}+...+X_{c+k}\geq L-m-i).\end{aligned}$$

Since vectors $\{X_1,...,X_c\}$, $\{X_{c+1},...,X_k\}$, $\{X_{k+1}\}$, and $\{X_{k+2},...,X_{c+k}\}$ are independent, we obtain that

$$\begin{aligned}E^* =&\sum_{i=1}^{\infty}\sum_{m=0}^{\infty} iF(c,L-m)f(k-c,m)f(i)F(c-1,L-m-i)\\ =&\sum_{i=1}^{\infty} if(i)\sum_{m=0}^{\infty} f(k-c,m)F(c,L-m)F(c-1,L-m-i).\end{aligned}$$

If $c=1$ then

$$\begin{aligned}E^* =&E\left(\mathbf{1}_{X_1+...+X_k\geq L}\cdot X_{k+1}\cdot \mathbf{1}_{X_2+...+X_{k+1}\geq L}\right)\\ =&\sum_{i=0}^{\infty} iP\left(X_1+...+X_k\geq L\cap X_{k+1}=i\cap X_2+...+X_{k+1}\geq L\right)\\ =&\sum_{i=0}^{\infty} iP(X_{k+1}=i\cap X_1+...+X_k\geq L\cap X_2+...+X_k\geq L-i)\\ =&\sum_{i=0}^{\infty} if(i)\sum_{m=L-i}^{\infty} P(X_2+...+X_k=m\cap X_1\geq L-m)\\ =&\sum_{i=0}^{\infty} if(i)\sum_{m=L-i}^{\infty} F(k-1,m)F(1,L-m).\end{aligned}$$

Finally, when $c = k$ we get

$$\begin{aligned} E^* =& E\left(\mathbf{1}_{X_1+\ldots+X_k\geq L}\cdot X_{k+1}\cdot \mathbf{1}_{X_{k+1}+\ldots+X_{2k}\geq L}\right) \\ =& \sum_{i=0}^{\infty} iP\left(X_1+\ldots+X_k\geq L\cap X_{k+1}=i\cap X_{k+1}+\ldots+X_{2k}\geq L\right) \\ =& \sum_{i=0}^{\infty} if(i)P(X_1+\ldots+X_k\geq L\cap X_{k+2}+\ldots+X_{2k}\geq L-i) \\ =& \sum_{i=0}^{\infty} if(i)F(k,L)F(k-1,L-i). \end{aligned}$$

□

Note that the formulas in Proposition 2 are easy to compute in case of common distributions. For example, if $\{X_i\}$ are iid Bernoulli with probability of success p, then $\delta\eta$ has Bernoulli distribution as well. As a result, for instance, in the case when $2 \leq c \leq k-1$ the formula in Proposition 2 collapses to a much simpler sum

$$E(\delta\eta) = \sum_{m=0}^{k-c} p^2 f(k-c,m)F(c,L-m)F(c-1,L-m-1),$$

where $f(n,i) = C_i^n p^i (1-p)^{n-i}$ is just the binomial density, and $F(n,j) = \sum_{i=j}^{n} f(n,i)$. In fact, the distribution functions of sum of independent random variables from common discrete distributions, such as Bernoulli, Binomial, Geometric, Negative Binomial and Poisson distributions often are “named” distributions (see Table 1).

TABLE 1. Distributions of the sum of t discrete independent random variables

$f(\cdot)$	$Bern(p)$	$Bin(N,p)$	$Geom(p)$	$NegBin(r,p)$	$Pois(\lambda)$
$f(t,\cdot)$	$Bin(t,p)$	$Bin(Nt,p)$	$NegBin(t,p)$	$NegBin(rt,p)$	$Pois(t\lambda)$

3.2. **The bottom line.** Given the distribution of X_i, the window width k and threshold level L, the variance can be easily calculated. And if the sample size N is large enough normality of the test statistic can be employed to obtain the critical value C. In practice, k and L can be pre-specified based on the historical studies and general knowledge about the nature of the data. Preliminary simulated studies have shown that for a chosen k (14 or 20 days, for example), a threshold level L between $k \times \bar{X}$ and $1.5 \times k \times \bar{X}$ can be used. However, this is an ad-hoc "rule" for detecting possible cluster pattern, more detailed simulation studies should be conducted to determine the optimal choice (perhaps something along lines of Kulldorff (1997)). We also ran extensive simulation studies to determine when N is "large enough". Basically, similar to normality in a classical case of binomial distribution, one needs to make sure that we have a relatively large number of non-zero terms in $\sum_{i=k+1}^{N} s_i t(X_i)$.

4. Comparison with Scan Statistics: Simulated Studies

In this section we compare the power of our test with the popular scan approach using simulated earthquake data generated with help of the Epidemic Type Aftershock-Sequence (ETAS) model (see Ogata (1998)). The ETAS model is one of the most popular approach to assess the probabilistic risk of earthquake occurrence. It assumes that the risk of future earthquakes is based on the past seismicity in the region. The model includes background activity of constant occurrence rate μ (i.e., a stationary Poisson process) and also includes aftershocks sequences of magnitude M_0 and larger. The rate of occurrence of the whole earthquake at time t based on the history of occurrence $\mathcal{H}_t = \{(t_i, M_i) : t_i < t\}$ is given by:

$$\lambda(t|\mathcal{H}_t) = \mu + A \sum_{i:t_i<t} e^{\alpha(M_i - M_0)} \Big/ \left[1 + \frac{t - t_i}{c}\right]^p \tag{4.1}$$

where t_i denotes the event times and the summation is taken over i such that $t_i < t$. The parameters A and c are geographic characteristic of the region, and p and α characterize the temporal pattern of seismicity. In our simulations, we will set $\mu = 0.05$, $c = 1$, $p = 2$ and $\alpha = 0$ (the magnitude is not taken into consideration).

Scan statistics are commonly used to detect cluster of events versus a homogeneous process (see Glaz et al. (2001), and some recent developments in Glaz et al. (2009)). There are many variants of the scan statistics, and we will use a simple fixed window and one dimensional scan statistics for discrete data considered by Naus (1965). The principle idea is moving a fixed size window continuously from start to end of a discrete process, then the scan statistics is defined as the maximum number of events among all windows. If the probability of observing at least this maximum number of events is considerably small, given that the null hypothesis is true, there is a high chance of clustering. Although the formulation of the statistic is simple, the probabilistic nature of scan statistics is very complex due to the dependent nature of the events and large number of overlapping window locations. Here we will use Monte-Carlo simulations to approximate the significance level of the scan test.

Under the null hypothesis the data follows an iid Poisson process with a sample size of 5000 and $\lambda = 0.05$ (this is equivalent to $A = 0$ in (4.1)). In case of the scan test, we reject the null hypothesis if we observe at least 11 events out of a window of size 50. This corresponds to a type I error of $\alpha = 0.036$ by the simulations.

Table 2. Comparisons of the Power, Number of Simulations = 1000, N=5000, $\lambda = 0.05$

A	Scan statistic	SCD statistic	SCD statistic	SCD statistic
		$(k = 50, L = 3)$	$(k = 70, L = 5)$	$(k = 30, L = 2)$
0	0.036	0.038	0.033	0.035
0.05	0.055	0.079	0.073	0.066
0.5	0.399	0.627	0.658	0.659
1	0.898	0.986	0.995	0.985
2	1.000	1.000	1.000	1.000

If the window of $k = 50$ and the threshold level $L = 3$ are chosen for the cluster test, then Proposition 1 and 2 give us critical value $C = 146.87$ that corresponds to the standardized z-value 1.8 and significance level $\approx .036$. With type I error of both tests controlled approximately at the same level, the power of the tests are computed on the simulated earthquake data generated under (4.1). The result of Table 2 shows that the power of cluster detection test dominates over scan test. Let us stress an important point: the exact clustering mechanism here is very different from the one in Section 2. Also we ran simulations with windows that are about 50% different from $k = 50$, and the empirical power is almost the same. It shows that the test is relatively stable with respect to a choice of the window size k and the threshold level L.

Remark 2. Note that in the case of the Poisson distribution the covariance calculation requires finding a sum of double series. Since all the summands involve Poisson distribution probabilities the convergence is reasonably fast. Because of that we did not address the issue of the numerical calculations in any intelligent way, instead a brutal force was employed. More specifically, we approximated $\sum_{i=1}^{\infty}\sum_{m=0}^{\infty} p_{im}$ by $\sum_{i=1}^{n}\sum_{m=0}^{m_i} p_{im}$. For every i the limit m_i was chosen by the rule $p_{im_i} < 10^{-12}$.

The outer limit n corresponds to $\sum_{m=0}^{m_n} p_{nm} < 10^{-9}$. Then simulations were used to verify the result.

5. Applications: Coalition Casualties and Residential Burglary Data

In this section we present cluster testing results for coalition casualty data in Iraq. In the present study, we use daily coalition casualty data for time period from January 2004 to March 2006, when the time process is relatively stationary. The data was collected with help of a publicly available internet source: http://icasualties.org/.

The goodness-of-fit test indicates that the geometric distribution with the average 1.93 provides a good fit for the overall histogram. But the question is whether observations are iid. For the cluster detection test, we choose a window of 14 days (two weeks) and a threshold level of 40 (the 1.5*average rule). The p-value of the test is 0.040. It indicates that the data are not iid. After converting the data to "0-1" observations, Kulldorff's scan test (the online test for the geometric distribution is not available) is applied and the p-value is 0.014. These results can be explained by the self-exciting nature of the terrorism phenomena: violence typically prompts more violence.

The self-exciting pattern of certain types of crimes has been observed by criminologists and research suggests that victims of personal or property crimes are more likely to be victimized in the near future (see Farrell and Pease (2001), Short et al. (2008)). If there exists a significant temporal clustering, the municipal police would benefit greatly. First, it will help identify two levels of crime activities: high and low. Therefore, an advanced, synchronized alarm system can be developed to warn the citizens. It also helps target intense patrols during high-activity period and schedule training and vacation during low-activity period. The question is, does a pure temporal clustering exist?

The temporal cluster pattern of burglary data is examined by both the cluster detection test and the scan test. The data set is the residential burglary count in the southeast region of Los Angeles area collected by Los Angeles Police Department between January, 2004 and December, 2005. Residential burglary is defined as the number of recorded offences of housebreaking at night. Based on the preliminary analysis the geometric distribution with average 0.96 is chosen as the underlying distribution. We use the cluster detection test with window size 14 and threshold level 20. The resulting p-value is 0.429. Kulldorff's scan test gives p-value 0.937. We also consider the central west valley region in our analysis, and again both cluster detection test and scan test indicate no significant clustered pattern. We should note, however, that the data is sampled during a two-year span, so perhaps a longer time series should be studied before generalizing the conclusions.

Acknowledgement

The authors thank the Los Angeles Police Department and Mr. Ong for providing the data on residential burglary.

References

[1] Billingsley, P., 1995. Probability and measure, Wiley-Interscience, New York, 3rd Edition.

[2] Bowman, D., 1999. A parametric independence test for clustered binary data, Statistics & Probability Letters 41, 1-7.

[3] Glaz, J., Naus, J., Wallenstein, S. 2001. Scan Statistics, Springer, New York.

[4] Glaz, J., Pozdnyakov, V., Wallenstein, S., eds., 2009. Scan Statistics: Methods and Applications, Birkhauser, Boston.

[5] Glaz, J., Zhang, Z. 2006. Maximum scan score-type statistics, Statistics & Probability Letters 76, 1316-1322.

[6] Farrel, G, Pease, K., 2001. Repeated Victimization, Criminal Justice Press, New York.

[7] Hansen, B., 1992. The likelihood Ratio Test under Nonstandard Conditions: Testing the Markov Switching Model of GNP, Journal of Applied Econometrics, 7, 61-82.

[8] Kulldorff, M., 1997. A spatial scan statistic, Communications in Statistics: Theory and Methods, 26, 1481-1496.

[9] Naus, J. 1965. The distribution of the size of the maximum cluster of points on a line, Journal of the American Statistical Association 60, 532-538.

[10] Ogata, Y., 1998. Space-time point-process models for earthquake occurrences, Annals of the Institute of Statistical Mathematics 50, 379-402.

[11] Short, B., D'Orsogna, R., Brantingham, J., Tita, E. 2008. Measuring repeat and near-repeat burglary efects, preprint.

[12] Yu, C. and Zelterman, D., 2002. Sums of dependent Bernoulli random variables and disease clustering, Statistics & Probability Letters 57, 363-373.