Journal of Statistical Planning and Inference 145 (2014) 102–112

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

# Retrospective change detection for binary time series models

Konstantinos Fokianos [a], Edit Gombay [b,*], Abdulkadir Hussein [c]

[a] University of Cyprus, Nicosia, Cyprus
[b] University of Alberta, Edmonton, AB, Canada
[c] University of Windsor, Windsor, ON, Canada

ARTICLE INFO

*Article history:*
Received 26 February 2013
Received in revised form
7 August 2013
Accepted 23 August 2013
Available online 7 September 2013

*Keywords:*
Binary time series
Logistic regression
Maximum partial likelihood estimator
Weak convergence

ABSTRACT

Detection of changes in health care performance, financial markets, and industrial processes have recently gained momentum due to the increased availability of complex data in real-time. As a consequence, there has been a growing demand in developing statistically rigorous methodologies for change-point detection in various types of data. In many practical situations, the data being monitored for the purpose of detecting changes are autocorrelated binary time series. We propose a new statistical procedure based on the partial likelihood score process for the retrospective detection of change in the coefficients of a logistic regression model with $AR(p)$-type autocorrelations. We carry out some Monte Carlo experiments to evaluate the power of the detection procedure as well as its probability of false alarm (type I error). We illustrate the utility using data on 30-day mortality rates after cardiac surgery and to data on IBM share transactions.


## 1. Introduction

Advances in information technology have made complex data available and therefore stimulated the need for detection of regime changes in many scientific disciplines. Real-time monitoring, also known as surveillance, has been used in detecting changes in financial markets, health care and industrial process performances as well as in public health disease surveillance; see Steiner et al. (2000), Gandy et al. (2010), Frisen (2003) and Gombay et al. (2011), among other references.

In general, the assumption of independent observations over time is not realistic. For instance, consider monitoring surgeon performance via Risk-adjusted CUSUM charts (RACUSUM). In this case, the assumption that patients operated by the same surgeon are independent over time is questionable. In Hussein et al. (in preparation), the authors studied the effect of autocorrelations on the performance of the RACUSUM chart and pointed out that the average run length (ARL) of the chart becomes extremely misleading if autocorrelations are not properly accommodated.

Recently, Gombay (2008) and Gombay and Serban (2009) studied score tests for prospective and retrospective detection of changes in the parameters of an $AR(p)$ time series model, respectively. These test statistics perform change point detection and are usually computationally less demanding than likelihood ratio test statistic processes. Gombay (2010) proposed a retrospective score test for detection of changes in the coefficients of linear regression model with time series errors; Berkes et al. (2009) studied the change detection in the covariance structure of linear processes. Related work can be found in Antoch et al. (2004), Hušková et al. (2007, 2008) and in Horváth (1993). Recently, Kengne (2012) proposed a change detection procedure based on quasi-likelihood estimators for the parameters of a class of continuous time series models including $AR(\infty)$. Also, Kirch and Tadjuidje Kamgaing (2012) extended the usual CUSUM charts to change detection in continuous nonlinear time series process by approximating the nonlinear part via neural network.

* Corresponding author. Tel.: +1 780 492 2337.
*E-mail address:* egombay@ualberta.ca (E. Gombay).

0378-3758/$ - see front matter 
http://dx.doi.org/10.1016/j.jspi.2013.08.017

There are several situations of practical interest in which the data follow binary regression models. For instance, monitoring the 30-day mortality rate of patients after cardiac surgery calls for a logistic regression model with possible changes in regression parameters over time. Similarly, in environmental studies, monitoring daily precipitation in a given geographic region calls for change detection methods in binary regression models. The authors are not aware of any treatment of retrospective change detection in the coefficients of a logistic regression model in which the binary outcomes have some sort of serial dependence and possibly depend upon covariate information. Lai (1995) and Vexler (2008) studied change-point detection in general dependent sequences of observations without adjusting for covariate information. Höhle (2010) proposed a CUSUM procedure based on the generalized likelihood ratio statistic for change-point detection in regression models for categorical time series. His use of CUSUM is not in the retrospective but in the sequential framework, hence they are not comparable to out methodology.

The objective of this paper is to propose a retrospective change point detection procedure for the coefficients of a logistic regression model for binary time series with an autoregressive-type dependence. Such models have been extensively studied in Kedem and Fokianos (2002).

In Section 2 we describe the proposed method based on the standardized score statistic as in Gombay (2008). In Section 3 we report the results of Monte Carlo simulations assessing the type I error, power and average run lengths of the proposed method. In Section 4 the methodology will be illustrated by using data on 30-day mortality rates after cardiac surgery as well as data on IBM share transactions. All proofs are given in the Appendix.

## 2. Model and proposed change detection procedure

### 2.1. The model and hypotheses of change

Consider a binary time series, $\{Y_t\}$, with probability of success $\pi_t(\beta)$ which depends on an unknown parameter vector $\beta$ and an accompanying $p$-dimensional vector of covariates $\{Z_t\}$. Following Kedem and Fokianos (2002), let us denote the history of the binary process and past covariate information by $\{\mathfrak{F}_t\}$, that is a filtration generated by $\{Y_{t-1}, Y_{t-2}, \dots ; Z_{t-1}, Z_{t-2}, \dots\}$. The vector of covariates $Z_t$ is allowed, in general, to contain lagged values of the binary response itself, thus permitting an $AR(p)$-type of serial dependence over time. The conditional probability mass function of the series $\{Y_t\}$ is given by

$$f(y_t;\beta|\mathfrak{F}_{t-1}) = \exp\left\{ y_t \log\left(\frac{\pi_t(\beta)}{1-\pi_t(\beta)}\right) + \log(1-\pi_t(\beta))\right\},$$

and the dependence on the covariate vector $\{Z_t\}$ can be modeled through the logit link function as

$$g(\pi_t(\beta)) = \eta_t = \log\left(\frac{\pi_t(\beta)}{1-\pi_t(\beta)}\right) = \beta' Z_{t-1}, \tag{1}$$

where $\beta \in \Re^p$, and $Z_t$ is assumed to be $\mathfrak{F}_t$-measurable. Although we focus on the logistic link function, the results can be generalized to other link functions, including the probit model. The logistic link function is the most natural choice for the binary time series. In fact, it can be shown that regardless of the degree and type of dependence between $Y_t$ and its past values, the logistic regression model occurs naturally; for more see Kedem and Fokianos (2002, Chapter 2).

Retrospective change-point detection assumes that a series of observations $y_1, \dots, y_n$ generated by the logistic model (1) is available to the investigator and that detection of the presence of break points in the coefficients of the model, $\beta = (\beta_1, \dots, \beta_p)$ is desired. In other words, retrospective change-point hypotheses can be formulated as follows:

$$H_0 : \eta_t = \beta_0' Z_{t-1}, \quad \text{for all } 1 \le t \le n, \ \beta_0 \text{ unknown},$$
$$H_A : \eta_t = \beta_0' Z_{t-1} \quad \text{for } 1 \le t \le \tau, \quad \text{and} \quad \eta_t = \beta_A' Z_{t-1} \text{ for } \tau < t \le n,$$

where $\beta_A \ne \beta_0$ is also unknown, and $1 < \tau < n$ is the unknown time when a change occurs in some of the $p$ regression coefficients. In this paper we explore a test statistic based on a standardized score obtained via a partial likelihood function to test hypothesis $H_0$.

In general, inferences concerning the binary time series model (1) are based on the so-called partial likelihood function defined in Kedem and Fokianos (2002) as

$$\prod_{t=1}^{n} f(y_t;\beta|\mathfrak{F}_{t-1}) = \prod_{t=1}^{n} \pi_t^{y_t}(\beta)(1-\pi_t(\beta))^{(1-y_t)},$$

or equivalently on the log-partial likelihood function,

$$l(\beta) = \sum_t l_t(\beta) = \sum_{t=1}^{n}\left[ y_t \log\frac{\pi_t(\beta)}{1-\pi_t(\beta)} + \log(1-\pi_t(\beta))\right].$$

The score vector of this log-partial likelihood is given by

$$S_n(\beta) = \sum_t \nabla_\beta l(\beta) = \sum_{t=1}^{n} Z_{t-1}(Y_t - \pi_t(\beta)) = \sum_{t=1}^{n} Z_{t-1}\left(Y_t - \frac{\exp(\beta' Z_{t-1})}{1+\exp(\beta' Z_{t-1})}\right). \tag{2}$$