Signal Processing 138 (2017) 307–312

Contents lists available at ScienceDirect

Signal Processing

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

Short communication

Nonparametric modeling and break point detection for time series signal of counts

Qi Gao[a], Thomas C.M. Lee[a,*], Chun Yip Yau[b]

[a] Department of Statistics, University of California, Davis, One Shields Avenue, Davis, CA 95616, USA
[b] Department of Statistics, Chinese University of Hong Kong, Shatin, N.T., Hong Kong

ARTICLE INFO

Article history:
Received 2 December 2016
Accepted 23 March 2017
Available online 4 April 2017

Keywords:
Astronomical signals
Change-point
Genetic algorithms
Minimum description length principle
Structural break detection

ABSTRACT

This paper considers the problem of flexible modeling as well as break point detection for time series signal of counts. In particular, the Poisson Generalized Autoregressive Moving Average (GARMA) models paired with radial basis expansions are used to fit such signals. A genetic algorithm is developed to find the possible breaks and the best fitting model derived from the minimum description length principle. The empirical performance of the proposed methodology is illustrated via a simulation study and a practical analysis of the bursts in the BATSE gamma ray data. Lastly, the consistency of the estimated break points and the model parameters is established under some regularity conditions.

1. Introduction

Time series of counts arise in a wide range of scientific research such as signal processing, astrophysics and epidemiology. For example, the number of photons emitted by astronomical objects over time is important in studying their activities. Different models have been proposed to take into account the dependency structure and other complications introduced by count data such as discreteness. These models can be classified into two major categories: parameter-driven models and observation-driven models where different strategies are used to model the dependency structure. Parameter-driven models (see e.g., [1]) introduce autocorrelation through a latent process, while the conditional distribution of the current observation is specified as explicit functions of its lagged values in observation-driven models (see e.g., [2]). Please refer to Camreon and Trivedi [3] for a review of the subject.

Moreover, structural breaks are commonly observed in such data (e.g., sudden increase in photon counts due to gamma ray bursts), which draws attention to detecting and characterizing these deviations from stationarity. This problem of break detection, also known as time series segmentation, has been widely studied in signal processing and other fields, see e.g., [4–7]. However, there is less work on segmenting time series of counts. A notable exception is a nonparametric algorithm called *Bayesian Blocks* [8] which aims to find the optimal segmentation of astronomical time series of count data that separates local light-curve features from the random observational errors. Piecewise constant models are fitted between change points in *Bayesian Blocks* which works similarly as a histogram with unequal bin width. See also [9] where a hierarchical Bayesian approach is used to segment two or more related time series of counts. There are also hypothesis test type methods (e.g., cumulative sum based tests) for detecting change points in time series of counts, see e.g., [10,11]. Nevertheless, for all these methods, little emphasis is placed on flexible model fitting between change points.

In this paper, a new method is proposed which achieves both flexible model fitting and consistent break point detection in time series of counts. We choose a class of observation-driven models called the generalized autoregressive moving average (GARMA) model proposed by Benjamin et al. [12] and specifically the Poisson-GARMA model to fit the count data, considering its resemblance to the classic ARMA models and the relative simplicity in model estimation. A nonparametric modeling method called radial basis expansion is also used within the GARMA models to improve flexibility in model fitting. Break point detection is considered as a model selection problem solved by using genetic algorithms based on the minimum description length (MDL) principle, in view of their success in tackling break point detection problems in similar context.

The rest of this paper is organized as follows. First we introduce our modeling strategy with Poisson GARMA model and radial basis expansion in Section 2. In Section 3 we apply the MDL principle to our break point detection problem and develop a genetic algorithm for solving the optimization involved. Following this, we state and

* Corresponding author.
E-mail addresses: qigao@ucdavis.edu (Q. Gao), tcmlee@ucdavis.edu (T.C.M. Lee), cyyau@sta.cuhk.edu.hk (C.Y. Yau).

http://dx.doi.org/10.1016/j.sigpro.2017.03.034
0165-1684/

prove the weak consistency of the MDL solution in Section 4. In Section 5, we study the performance of our method via simulations, and we apply it to several gamma ray bursts datasets in BATSE catalog in Section 6. Lastly, we conclude our paper with a summary and a discussion of possible generalizations in Section 7.

2. Model formulation

The GARMA model proposed by Benjamin et al. [12] is an observation-driven model for non-Gaussian time series data $y_1, \ldots, y_n$. Similar to the generalized linear model (GLM), the conditional density of each observation y_t given the previous information set $\mathbf{H}_t = (\mathbf{x}_t, \ldots, \mathbf{x}_1, y_{t-1}, \ldots, y_1, \mu_{t-1}, \ldots \mu_1)$ comes from the same exponential family; i.e.,

$$f(y_t|\mathbf{H}_t) = \exp\left\{\frac{y_t\theta_t - b(\theta_t)}{\phi} + d(y_t, \phi)\right\}, \tag{1}$$

where θ_t and ϕ are canonical and scale parameters, respectively. The functions $b(\cdot)$ and $d(\cdot)$ together define the particular exponential family, where $\mathrm{E}(y_t|\mathbf{H}_t) = \mu_t = b'(\theta_t)$ and $\mathrm{var}(y_t|\mathbf{H}_t) = \phi b''(\theta_t)$. The $\mathbf{x}_t$s are design points and in the current context, they are radial basis expansion values of time index which will be explained in more details later. Unlike the classical GLMs, here the predictor $\eta = g(\mu_t) = \mathbf{x}_t'\boldsymbol{\beta} + \tau_t$ has an additional component τ_t which allows the inclusion of autoregressive moving average terms, resulting in

$$\begin{aligned} g(\mu_t) = \eta_t &= \mathbf{x}_t'\boldsymbol{\beta} + \sum_{j=1}^{p}\phi_j\{g(y_{t-j}) - \mathbf{x}_{t-j}'\boldsymbol{\beta}\} \\ &\quad + \sum_{j=1}^{q}\psi_j\{g(y_{t-j}) - \eta_{t-j}\}, \end{aligned} \tag{2}$$

and GARMA(p, q) models are defined by (1) and (2).

For time series of counts, Poisson GARMA(p, q) models are used where the conditional distribution for y_t is Poisson. The log function is chosen as the link and any zero values of y_{t-j} in (2) are replaced by a threshold parameter c such that $0 < c < 1$.

We consider the problem of segmenting time series of counts data into stationary pieces where each piece follows a Poisson GARMA model. Similar to the settings in Davis et al. [13], for an observed time series of length n, let τ_j, $j = 1, \ldots, B$ be the break points between the jth and $(j+1)$th pieces. All the pieces are assumed to be independent and the jth piece of the time series W_t can be modeled by a stationary Poisson GARMA(p, q) model $Y_{t,j}$; i.e.

$$W_t = Y_{t+1-\tau_{j-1},j}, \quad \tau_{j-1} \le t < \tau_j, \tag{3}$$

and $\boldsymbol{\theta}_j = (\boldsymbol{\beta}_j, \boldsymbol{\phi}_j, \boldsymbol{\psi}_j)$ is the corresponding parameter vector. Here B and τ_js are unknown while the order p and q are assumed to be known for simplicity. The methodology can be straightforwardly extended to the cases where p and q are unknown.

We also consider fitting each piece using radial basis expansion which is a nonparametric regression method to improve flexibility and accuracy (see [14] for more details). Radial basis function is a real-valued function whose value depends only on the distance from some point k, known as knot; i.e., $\delta(t, k) = \delta(\|t-k\|)$. Our choice of δ for fitting is $\delta(t) = (|t-k|)^3$ which is used to generate design points $\mathbf{x}$ in (2). To be more specific,

$$h(t) = \mathbf{x}_t'\boldsymbol{\beta} = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + \sum_{i=1}^{K} b_i(|t-k_i|)^3,$$

where K is the total number of knots and $\{k_i\}_{i=1}^K$ are locations of the knots. Here K and k_is are also assumed to be unknown.

Therefore, when comparing to existing methods for modeling time series of count data, a novelty of the current work is that, in addition to break point detection, it also models the mean $h(t)$ of each piece nonparametrically.

3. Model selection and practical fitting

3.1. Model selection with MDL

The two-part MDL developed by Rissanen [15] is used here to derive an objective function upon which a "best" segmented model is selected, including the knots' locations in each piece. The MDL principle defines the best fitting model as the one which allows the greatest compression of the data, as reflected by achieving the minimum total code length. More detailed introductions of MDL can be found for examples in Hansen and Yu [16] and Lee [17].

Let $\mathcal{F}$ be the class of piecewise Poisson GARMA processes defined in (3). The total code length of the data $\mathbf{Y} = (y_1, y_2, \ldots, y_n)$ denoted by $\mathrm{CL}(\mathbf{Y})$ for any model $\mathcal{M} \in \mathcal{F}$ can be decomposed into two parts: the code length of the fitted model $\mathrm{CL}(\hat{\mathcal{M}})$ and that of the corresponding residuals $\mathrm{CL}(\hat{\boldsymbol{\epsilon}}|\hat{\mathcal{M}})$. In other words, $\mathrm{CL}(\mathbf{Y}) = \mathrm{CL}(\hat{\mathcal{M}}) + \mathrm{CL}(\hat{\boldsymbol{\epsilon}}|\hat{\mathcal{M}})$, and the "best" fitting model is the one that minimizes $\mathrm{CL}(\mathbf{Y})$. Next we need to obtain expressions for $\mathrm{CL}(\hat{\mathcal{M}})$ and $\mathrm{CL}(\hat{\boldsymbol{\epsilon}}|\hat{\mathcal{M}})$, and we begin with $\mathrm{CL}(\hat{\mathcal{M}})$.

For brevity, for the rest of this paper we set $p = 1$ and $q = 0$ in (2); other values of p and q can be handled in a straightforward manner. That is, we are considering GARMA(1, 0) models, which can also be denoted as GAR(1) models. Now to encode any fitted model $\hat{\mathcal{M}}$, we need to take into account the total number and locations of break points as well as the model fitted in each segment. To encode each segment, we need to encode the number and locations of the knots and the estimated coefficients $\hat{a}_i$s and $\hat{b}_i$s from radial basis expansion. To be more specific,

$$\begin{aligned} \mathrm{CL}(\hat{\mathcal{M}}) &= \mathrm{CL}(B) + \mathrm{CL}(\tau_1, \tau_2, \ldots \tau_B) + \sum_{i=1}^{B+1}\mathrm{CL}(\{W_t\}_{\tau_{i-1}\le t<\tau_i}) \\ &= \mathrm{CL}(B) + \mathrm{CL}(\tau_1, \tau_2, \ldots \tau_B) + \sum_{i=1}^{B+1}[\mathrm{CL}(K_i) + \mathrm{CL}(\{k_{ij}\}) \\ &\quad + \mathrm{CL}(\{a_{ij}\}) + \mathrm{CL}(\{b_{ij}\})] \end{aligned}$$

where K_i is the number of knots in ith segment and $\{k_{ij}\}$ are the locations of knots. Based on the results from Rissanen [15], it requires approximately $\log_2(I)$ bits to encode an integer I and we can use this rule to encode numbers and locations of break points and knots. Moreover, the code length required to encode a maximum likelihood estimate computed from n data points is $\frac{1}{2}\log_2(n)$. Suppose that there are n_i data points in the ith segment, $\{a_{ij}\}$ and $\{b_{ij}\}$ are maximum likelihood estimates based on the n_i data points in this piece. Therefore, $\mathrm{CL}(\{a_{ij}\}) + \mathrm{CL}(\{b_{ij}\}) = \frac{(K_i+4)}{2}\log_2(n_i)$. Putting these together,

$$\begin{aligned} \mathrm{CL}(\hat{\mathcal{M}}) &= \log_2(B) + \sum_{i=1}^{B}\log_2(\tau_i) + \sum_{i=1}^{B+1}\log_2(K_i) + \sum_{i=1}^{B+1}\sum_{j=1}^{K_i}\log_2(p_{ij}) \\ &\quad + \frac{1}{2}\sum_{i=1}^{B+1}(K_i+4)\log_2(n_i). \end{aligned} \tag{4}$$

Next we need an expression for $\mathrm{CL}(\hat{\boldsymbol{\epsilon}}|\hat{\mathcal{M}})$, which can be well approximated by the negative log-likelihood of the fitted model $\hat{\mathcal{M}}$, as shown by Rissanen [15]. A such expression for the Poisson GAR(p) model log-likelihood is given in (A.5) below. As this expression is rather lengthy, we shall denote it as $\log L$ here. Consequently, the MDL criterion for the proposed piecewise Poisson GAR model with radial basis expansion is: