East Asian Journal on Applied Mathematics
doi: 10.4208/eajam.180921.050122

Vol. **12**, No. 2, pp. 353-366
May 2022

An L_0-Norm Regularized Method for Multivariate Time Series Segmentation

Min Li and Yu-Mei Huang*

School of Mathematics and Statistics, Center for Data Science, Lanzhou University, Lanzhou 730000, China.

Received 18 September 2021; Accepted (in revised version) 5 January 2022.

Abstract. A multivariate time series segmentation model based on the minimization of the negative log-likelihood function of the series is proposed. The model is regularized by the L_0-norm of the time series mean change and solved by an alternating process. We use a dynamic programming algorithm in order to determine the breakpoints and the cross-validation method to find the parameters of the model. Experiments show the efficiency of the method for segmenting both synthetic and real multivariate time series.

AMS subject classifications: 37M10, 62M10, 91B84

Key words: Multivariate time series, segmentation, L_0-norm, dynamic programming.

1. Introduction

Multivariate time series arise in various areas, including meteorology [6, 10, 12, 21], medicine [23,32,37,40], finance [14,28] and speech processing [2,7,16,33]. Anomalies or abrupt changes often appear in the time series at the breakpoints [12,21,25,26,31,32]. For example, unusual changes of the environment may cause abnormal changes in the observed hydrological and the corresponding environmental time series data [12, 21] and attacks against web service providers cause the anomalies in network traffic data [25, 26]. Time series segmentation aims to detect such breakpoints of abrupt changes/anomalies in time series, thus dividing them into homogeneous segments [2, 6–9, 12–14, 17–20, 24, 34–36]. Time series segmentation is an important problem in the analysis of time series and it can provide meaningful information for further data analysis, prediction and policy decision.

The earliest work on breakpoints detection can be traced back to the Page's works [29, 30]. Since then, the time series segmentation problem has been actively studied. Yao [39] introduced an estimator of the number of breakpoints in a univariate independent normal sequence based on Schwarz' criterion, Rabiner [33] used hidden Markov models (HMM) to detect the changes in speech signal series, Duncan and Bryant [8] proposed a univariate time series segmentation method derived from the idea that each segment can be fitted by

*Corresponding author. *Email address:* `huangym@lzu.edu.cn` (Y.-M. Huang)

an independent linear model. We also note that Kehagias *et al.* [21] applied a regression model to construct a univariate hydrological and environmental time series segmentation and used a dynamic programming (DP) algorithm to detect the breakpoints. A dynamic programming algorithm was also used by Jackson *et al.* [19]. Besides, in order to determine the location of breakpoints in the one-dimensional piecewise constant signal corrupted by white noise, Harchaoui and Lévy-Leduc [15] used a regularized least-square criterion with a total variation (TV) regularization.

The segmentation of multivariate time series has been also studied in literature. Abonyi *et al.* [1] proposed a modified fuzzy Gath-Geva clustering method. The method uses fuzzy sets to represent the segments and local probabilistic principal component analysis model to measure the homogeneity of the segments. Vert and Bleakley [37] constructed a segmentation model for a multidimensional signal. The method involves the L_1-norm regularization and approximates the signal by piecewise constant aggregates. Using the algorithm in [19], Killick *et al.* [22] developed pruned exact linear time method (PELT), which allows to find the exact solution of the segmentation model by the L_0-norm regularization. The PELT algorithm adds a pruning step to the DP algorithm, thus reducing the computational cost. Hallac *et al.* [14] proposed a covariance-regularized likelihood maximization model for multivariate time series segmentation and designed a greedy Gaussian segmentation (GGS) algorithm to solve it. More exactly, the model is reformulated as a combinatorial optimization problem on the breakpoints and solved by an extremely scalable greedy algorithm.

Breakpoints are the points in the time series where an abrupt data change occurs. We note that breakpoints are usually scattered. Therefore, it is natural to assume that the derivative of the time series mean is sparse and does not vanish at the breakpoints, so that the L_0-norm prior of the derivative can be naturally adopted in the model. On account of this, here we apply the maximum likelihood principle and construct an L_0-norm regularized segmentation model for multivariate time series. Thus we expect that such an approach allows to obtain better results in terms of the number of the breakpoints and the time series segmentation.

The rest of the paper is organized as follows. In Section 2, we introduce an L_0-norm regularized model for multivariate time series segmentation. In Section 3, we solve the model by an alternating process and utilize a dynamic programming algorithm to determine the breakpoints. In order to establish the parameters of the model under consideration, the cross-validation (CV) method is employed in Section 4. Section 5 demonstrates the segmentation for synthetic and real time series. Finally, our conclusion is given in Section 6.

2. Segmentation Model

Considering a multivariate time series $x_1, x_2, \cdots, x_T \in \mathbb{R}^n$, we assume that x_t are independent random variables such that

$$x_t \sim f(\cdot|\mu_t, \Sigma_t), \tag{2.1}$$