Advances in Applied Mathematics and Mechanics
Adv. Appl. Math. Mech., Vol. **15**, No. 2, pp. 300-321

DOI: 10.4208/aamm.OA-2021-0209
April 2023

A Total Variation Based Method for Multivariate Time Series Segmentation

Min Li[1], Yumei Huang[1,*] and Youwei Wen[2]

[1] *School of Mathematics and Statistics, Center for Data Science, Lanzhou University, Lanzhou, Gansu 730000, China*
[2] *School of Mathematics and Statistics, Hunan Normal University, Changsha, Hunan 410081, China and Key Laboratory of Computing and Stochastic Mathematics (LCSM), Ministry of Education of China*

Received 21 June 2021; Accepted (in revised version) 29 November 2021

Abstract. Multivariate time series segmentation is an important problem in data mining and it has arisen in more and more practical applications in recent years. The task of time series segmentation is to partition a time series into segments by detecting the abrupt changes or anomalies in the time series. Multivariate time series segmentation can provide meaningful information for further data analysis, prediction and policy decision. A time series can be considered as a piecewise continuous function, it is natural to take its total variation norm as a prior information of this time series. In this paper, by minimizing the negative log-likelihood function of a time series, we propose a total variation based model for multivariate time series segmentation. An iterative process is applied to solve the proposed model and a search combined the dynamic programming method is designed to determine the breakpoints. The experimental results show that the proposed method is efficient for multivariate time series segmentation and it is competitive to the existing methods for multivariate time series segmentation.

AMS subject classifications: 65C20, 62M10, 91B84

Key words: Multivariate time series, segmentation, total variation, dynamic programming.

1 Introduction

Multivariate time series analysis plays an important role in the area of data mining [2, 3, 13, 17]. Abrupt changes or anomalies often appear in time series in practical applications [14, 21, 25–27, 32, 33]. For example, in hydrological and environmental time series, human activities or sudden changes of environment often cause abrupt changes in the

*Corresponding author.
Email: huangym@lzu.edu.cn (Y. Huang)

data [14,21]; in network traffic data, attacks against web service providers usually cause data anomalies [25,27]. The aim of time series segmentation is to partition the time series into several internally homogeneous segments by detecting the change points or anomalies in the time series [4,6,15,18–20,35–37]. Time series segmentation is a preprocessing step in data mining and it can provide a lot of meaningful information for further data analysis, prediction and policy decision etc. For example, in [22], after the time series is segmented, the fast pattern matching of the time series is completed; in [23], the fast, accurate classification and clustering of the time series are performed based on the segmentation results; in [31], an approach based on time series segmentation is proposed for subsequence search in sequence databases.

Time series segmentation has been actively studied in the literature. Because of the simplicity, univariate time series segmentation was mainly considered in the early study. For example, in [10], a sliding window segmentation algorithm was designed to segment an univariate autoregressive time series; in [12], Duncan and Bryant partitioned the univariate time series based on the idea that each segment can be fitted by an independent linear model; by utilizing the regression model, Kehagias et al. [21] constructed a segmentation method and applied the dynamic programming (DP) algorithm for univariate hydrological and environmental time series segmentation.

In recent years, the applications of multivariate time series have extensively arisen in data processing. The multivariate time series analysis has been becoming more and more important. Specifically, the segmentation of multivariate time series has drawn much attention and many segmentation methods have been proposed in the literature. For example, [3] proposed a modified fuzzy Gath-Geva clustering method for multivariate time series segmentation, where fuzzy sets were used to represent the segments and the local probabilistic principal component analysis model was applied to measure the homogeneity of the segments; in [11], Desobry et al. proposed a model free approach to determine the change points in the multivariate time series signal that could not be explained by any accurate and tractable model; [5] proposed a method that detected the change points of a multidimensional signal through approximating the original signal by a piecewise constant signal. The approximation was the solution of a group least absolute shrinkage and selection operator (Lasso) regression problem which was solved by a group least angle regression (LARS) procedure; by assuming that the multivariate time series in different segments are independent, and the multivariate time series in a same segment follow a same Gaussian distribution, [16] proposed the covariance-regularized likelihood maximization model for segmentation and designed a greedy Gaussian segmentation (GGS) algorithm to solve it. By taking advantage of the relationship formulas between the mean and the breakpoints, the covariance and the breakpoints in each segment of the time series, the proposed model was reformulated as a maximization problem with respect to the breakpoints and it became a combinatorial problem. An extremely scalable greedy algorithm was designed to solve this combinatorial problem.

In the GGS method for multivariate time series segmentation proposed by Hallac et al. [16], it assumes that both the means and the covariances of a same segment in the time