Digital Signal Processing 24 (2014) 14–26

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Signal segmentation using changing regression models with application in seismic engineering

Theodor D. Popescu

National Institute for Research and Development in Informatics, 8-10 Averescu Avenue, 011455 Bucharest, Romania

ARTICLE INFO

Article history:
Available online 17 September 2013

Keywords:
Change detection
Data segmentation
MAP estimator
Monte Carlo simulation
Seismic signal processing

ABSTRACT

The change detection and segmentation methods have gained considerable attention in scientific research and appear to be the central issue in various application areas. The objective of the paper is to present a segmentation method, based on maximum a posteriori probability (MAP) estimator, with application in seismic signal processing; some interpretations and connections with other approaches in change detection and segmentation, as well as computational aspects in this field are also discussed. The experimental results obtained by Monte Carlo simulations for signal segmentation using different signal models, including models with changes in the mean, in FIR, AR and ARX model parameters, as well as comparisons with other methods, are presented and the effectiveness of the proposed approach is proved. Finally, we discuss an application of segmentation in the analysis of the earthquake records during the Kocaeli seism, Turkey, August 1999, Arcelik station (ARC). The optimal segmentation results are compared with time–frequency analysis, for the reduced interference distribution (RID). The analysis results confirm the efficiency of the segmentation approach used, the change instants resulted by MAP appearing clear in energy and frequency contents of time–frequency distribution.

1. Introduction

The problem of change detection and diagnosis has gained considerable attention during the last three decades in the research context and appears to be the central issue in various application areas. From a statistical point of view, change detection tries to identify changes in the probability distribution of a stochastic process. In general, the problem involves both detecting whether or not a change has occurred, or whether several changes might have occurred, and identifying the times of any such changes.

In the off-line applications, it is available a batch of data, and the goal is to find the time instants for system changes as accurately as possible. This is usually called segmentation. Deeper interactions between the control, signal processing, and statistical communities have recently contributed to the creation of new insights into the change detection problem in a significant way.

The analysis of the behavior of real data reveals that most of the changes that occur are either changes in the mean level, variance, or changes in spectral characteristics. In this framework, the problem of segmentation between “homogeneous” parts of the data (or detection of changes in the data) arises more or less explicitly.

E-mail address: pope@ici.ro.
URL: http://www.ici.ro/ici/homepage/thpopescu.html.

1051-2004/$ – see front matter © 2013 Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.dsp.2013.09.003

In our opinion, a coherent methodology is now available to the designer, together with the corresponding set of tools, which enables him to solve a large variety of change detection problems in dynamical systems. It is interesting to note that the theory has been used in many successful applications [1,2]. Also, many books, journals and conference publications are concerned with these applications. Among them can be mentioned applications in mechanical engineering [2–4], industrial process monitoring [5–7], civil infrastructure [8–10], medical diagnosis and treatment [11–13], speech segmentation [14,15], underwater sensing [16,17], video surveillance [18,19], and driver assistance systems [20–22].

The detection of events in seismic signals has been a subject of great interest during the last thirty years. Most of the methods in this area have been based on detecting special patterns or clusters in seismic data [23–28]. Other approaches make use of AR and ARMA models, used in conjunction with the Akaike information criterion (AIC) method for change detection and isolation, as well as to detect the primary (P-waves) and secondary (S-waves) waves [8,29–34]. These techniques currently employed for event detection in seismic waves use single- or three-component recordings.

Another class of methods makes use of time–frequency analysis. Many earthquake engineering processes are characterized by non-stationary and nonlinear features that are often obscured in the traditional Fourier-based analysis schemes. As these representations provide an averaged-sense of frequency content, they do not distinguish noteworthily certain frequency components that are of short duration and high frequency as well as those arising from persistent, low-amplitude features. The ability to differentiate these contributions is critical in earthquake engineering. The time–frequency methods are capable of resolving energy content in such signals with both frequency and time. As a consequence, it was required to develop and use some non-stationary spectral analysis techniques. In this sense, significant efforts have been made in order to represent the temporal evolution of non-stationary spectral characteristics, assured by time–frequency analysis [35–37] among others. A new method based on a time–frequency analysis through the Wigner Distribution (WD) is presented and applied in [38] and [39]. The method consists on defining an appropriate entropic measure through a suitable time–frequency distribution, acting as probability distribution function. The information entailed by WD is explored by means of Rényi entropy. The method is based on identification of the events as those temporal clusters having the highest amount of information (entropy). In [40], the seismic signal analysis is performed by Smoothed Pseudo Wigner–Ville Distribution (SPWVD) in order to obtain instantaneous frequency (IF) information. Based on the time–frequency behavior, a pattern to characterize the seismic signal is estimated. In the same time it is analyzed the energy signal envelope, which is the derivative of the filtered cumulative energy, serving to estimate the different transitions along the seismic signal. Other applications of time–frequency analysis in earthquake engineering are presented in [41, 42] and [43], among others.

The outline of this paper is as follows. In Section 2, it is presented the segmentation problem formulation from a statistical perspective. In Section 3 the conceptual description of the statistical criteria for segmentation: the Maximum Likelihood (ML) and Maximum A posteriori Probability (MAP) estimate, some interpretations and connections, as well as computational aspects in this field are discussed. In Section 4, some experimental results obtained by Monte Carlo simulations for signal segmentation using different signal models, referred in literature, are presented and the effectiveness of the proposed approach is proved. Section 5 presents some comparisons of MAP approach with other change detection and segmentation methods, in the same simulation framework. Finally, Section 6 presents the experimental results obtained in segmentation of the two-component recordings (NS and WE) of the Kocaeli, Turkey, August 1999, earthquake, Arcelik station (ARC). The efficiency of the segmentation approach, in this case, is proved by time–frequency analysis, when segmentation results are evaluated using the reduced interference distribution (RID).

2. Problem formulation

We consider the following linear regression model with piecewise constant parameters,

$$y_t = \phi_t^T \theta_t + e_t, \qquad E\left(e_t^2\right) = R_t, \tag{1}$$

as a good description of the observed signal y_t. Here θ_t is a d-dimensional parameter vector, ϕ_t is the regressor, and the measurement vector is assumed to have dimension p. The noise e_t is assumed to be Gaussian with a known time-varying noise variance R_t, for generality.

The task of determining θ_t from y_t is referred as estimation, and change detection is the task of finding abrupt, or rapid, changes in θ_t, which is assumed to start at time k, referred as the change time. The basic assumptions on model (1) in change detection are the following:

- The component θ_t undergoes an abrupt change at time $t = k$. Once this change has been detected, the procedure will start all over again to detect the next change. The alternative is to consider θ_t as piecewise constant and focus on a sequence of change times $k_1, k_2, \ldots, k_n$. This sequence is denoted k^n, where both k_i and n are free parameters. The segmentation problem is to find both the number and locations of change times in k^n.
- In statistical approaches, the noise will be assumed to be white and Gaussian $e_t \in N(0, R_t)$.

We introduce now the general segmentation problem for linear regression model with piecewise constant parameters. As we mentioned above, in segmentation the goal is to find a sequence of time indices $k^n = k_1, k_2, \ldots, k_n$, where both the number n and the locations k_i are unknown, such that a linear regression model with piecewise constant parameters,

$$y_t = \phi_t^T \theta(i) + e_t, \qquad E\left(e_t^2\right) = \lambda(i) R_t \tag{2}$$

when $k_{i-1} < t \leqslant k_i$ is a good description of the observed signal y_t. Here $\theta(i)$ is the d-dimensional parameter vector in segment i, ϕ_t is the regressor and k_i denotes the change times. The noise e_t is assumed to be Gaussian with variance $\lambda(i) R_t$, where $\lambda(i)$ is a possibly segment dependent scaling of the noise and R_t is the nominal covariance matrix of the noise; the model (2) represents an extension of model (1). We can think of λ either as a scaling of the noise variance or variance itself ($R_t = 1$). Neither $\theta(i)$ or $\lambda(i)$ are known. The Gaussian assumption on the noise is a standard one, partly because it gives analytical expressions and partly because it has proven to work well in practice. We will assume R_t to be known and the scaling as a possibly unknown parameter. The model (2) is referred to as changing regression, because it changes between regression models. Its important feature is that the jumps divide the measurements into a number of independent segments, since the parameter vectors in different segments are independent. Some important cases of the model (2) are the changing mean model, the autoregressive (AR) model, the autoregressive model with exogenous variable (ARX) and finite impulse response (FIR) model, etc., where ϕ_t has different expressions.

The assumption on the regression models in (2) is not too restrictive since many stationary processes encountered in practice can be closely approximated by such models. The identification and parameters estimation methods represent only tools to perform change detection and segmentation. Good and precise models offers high performance in these schemes, but also biased parametric models can be used for change detection and segmentation. This bias decreases, but does not annihilate the performance of the detection and segmentation procedures.

One way to guarantee that the best possible solution is found, is to consider all possible segmentation k^n, estimate one linear regression model in each segment, and then choose the particular k^n that minimizes an optimality criteria:

$$\widehat{k^n} = \arg \min_{n \geqslant 1, 0 < k_1 < \cdots < k_n = N} V\left(k^n\right). \tag{3}$$

For the measurements in the ith segment, that is $y_{k_{i-1}+1}, \ldots, y_{k_i} = y_{k_{i-1}+1}^{k_i}$, the least square estimate and its covariance matrix are denoted:

$$\hat{\theta}(i) = P(i) \sum_{t=k_{i-1}+1}^{k_i} \phi_t R_t^{-1} y_t, \tag{4}$$

$$P(i) = \left(\sum_{t=k_{i-1}+1}^{k_i} \phi_t R_t^{-1} \phi_t^T \right)^{-1}. \tag{5}$$

The following quantities, V – the sum of squared residuals, D – $-\log\det$ of the covariance matrix P and N – the number of data in each segment, are given by