A Model for Non-stationary Time Series Analysis with Clustering Methods

Shai Policker **Amir B. Geva**

Electrical and Computer Engineering Department

Ben-Guriun University of the Negev

P.O.B. 653 Beer-Sheva 84105, Israel

email: geva@ee.bgu.ac.il

Abstract

The object of this paper is to present a model of non-stationary time series generated by switching between a finite number of random processes and to apply clustering algorithms to the task of estimating the model's parameters. We will also analyze the parameters which govern the algorithm's behavior to infer a novel cluster validity criterion for fuzzy clustering algorithms of temporal patterns.

The model defines a non-stationary composite source generated by randomly switching between elements of a finite number of random processes. The probability distribution which underlies the behavior of the switch is controlled by a temporal parameter vector process which is used to determine a different switching probability in each time instant. This definition allows us to analyze a drift between disjoint states of the composite model.

Key words: Time series analysis, fuzzy clustering, temporal pattern recognition

Introduction

Many physical and biological phenomena, which are observed in terms of a non-stationary time series, can be modeled by a set of stable or semi-stable states that the system traverses. In each of these states the system outputs a stationary distributed segment of observations. The task of segmenting the series to states and of estimating the distribution of each state can be handled by several algorithms of which the Hidden Markov Model (HMM) with the Baum-Welch algorithm [1] is probably the most popular.

In some of these cases it is of great importance to analyze the drift between any two states and to detect that such a drift is underway. An example for such an application is the early prediction of an epileptic seizure from a patient's EEG. The epileptic seizure itself can be detected in the sampled EEG signal and easily separated from the normal wake state [2], but the early prediction of the seizure, when the EEG diverges only slightly and infrequently from the normal wake signal, is a difficult task.

Applications utilizing the HMM can only partially support such a demand by assigning stages of the drift process to specific intermediate states. Describing such a drift process using a continuous amplitude, time dependent, signal which governs the temporal distribution of the series is more natural and can offer more flexibility and accuracy in extracting information about the drift. We suggest estimating such a signal by means of unsupervised fuzzy clustering algorithms. Furthermore, situations in which the time series distribution can be presented as a composite source of an unknown number of sub-sources raise another problem regarding standard versions of the HMM, which can only approximate the series by a constant, pre-determined number of Gaussian sub-sources.

In this contribution we propose a model for the generation of a non-stationary time series by a composite source such that its temporal distribution is controlled by an unknown temporal signal. First we will describe the model and provide an algorithm for estimating the process which underlies the time series distribution drift. Next we will analyze the algorithm's parameters and demonstrate it's performance with a simulation example.

Model Description

The output time series is generated from an unknown number, N, of sub-sources (see Fig. 1) by a *random switching function* $F(\theta(t), \mathbf{X}(t))$, where $\theta(t)\in\Re^{N}$ is a parameter vector process and $\mathbf{X}(t)\in\Re^{N\text{x}D}$ is a matrix with each of the N columns containing a random vector process with dimension D originated in a different sub-source. $\theta(t)$ is used as a parameter vector for determining the probabilities of selecting a single sub-source, in each time sample t, to be transmitted to the output.

$$y(t) = F(X(t),\theta(t)) = x_i(t) \text{ with probability } P(X(t), \theta(t)) \qquad (1)$$

The random function $F(\theta(t), \mathbf{X}(t))$ acts as a switch which takes a new position each time according to the probability vector inspired by $\theta(t)$ and outputs the vector $\mathbf{y}(t)\in\Re^D$ which equals one of the columns of $\mathbf{X}(t)$. The behavior of the model can be illuminated by comparing it to the Continuous Hidden Markov Model (CHMM) as defined in [1]. CHMMs of different kinds are widely used today in a variety of applications such as speech recognition [1]. If we are to define $\mathbf{X}(t)$ to be composed of Gaussian processes and F(•) as a random Markov process which can only receive a finite set of probability vectors for the switch and have a constant matrix governing the transition probabilities between these constant probability vectors then this model would have realized the common CHMM.

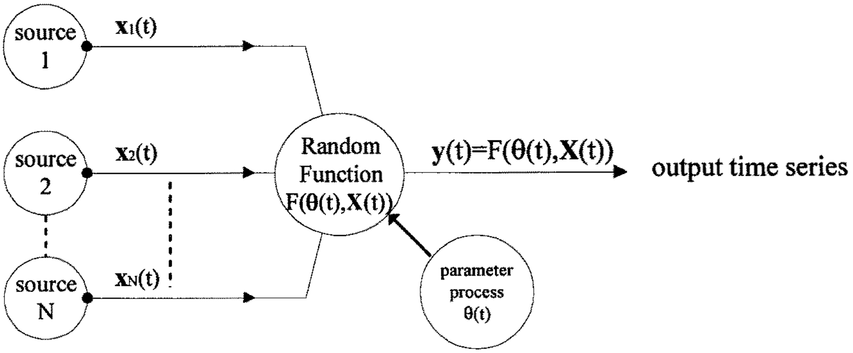

Figure 1
Model description

Examples for the random function $F(\theta(t),\mathbf{X}(t))$

In general, no limitations are imposed on F(•) and although other types of F(•) can be used we preferred, throughout this paper, to use (1) with a probability distribution function (pdf) which takes the value of $\theta(t)$ as a parameter.

In this case we must demand that $P(\theta(t))$ will realize a pdf over the possible values of $\mathbf{X}(t)$. This means that it should be non-negative and integrate to one. The simplest example for P(•) is:

$$P(X(t), \theta(t)) = \theta(t) \qquad (2)$$

i.e. the value of $\theta_i(t)$ itself will be used as the probability of selecting $x_i(t)$. In that case every element of $\theta(t)$ must be non-negative and follow:

$$\sum_{i=1}^{N} \theta_i = 1 \qquad (3)$$

We will use this simple possibility in the algorithm for estimating $\theta(t)$ which follows.

Another example for P(•) is the Gaussian distribution:

$$P(\mathbf{X}(t),\theta(t)) = N(\mu(\theta(t),\mathbf{X}(t)),\sigma_\theta^2) \tag{4}$$

where σ_θ^2 is a constant variance and $\mu(\theta(t),\mathbf{X}(t))$ is the Gaussian mean which is given by

$$\mu(\theta(t),\mathbf{X}(t)) = \sum_{i=1}^{N}\theta_i(t)\mathbf{x}_i(t) \tag{5}$$

This definition of F(•) can represent a tough but quite practical situation when both the sources internal distribution and the switching distribution is Gaussian.

An algorithm for estimating θ(t)

We now present an algorithm for the estimation of θ(t) in the case where all sub-sources are Gaussian with different means, drawn from a finite, D dimensional space. We assume that the output time series distribution $\{y(n)\in\Re^D \mid -\infty \le n \le \infty\}$ is governed by (1). The main idea is to use a sampled period of the time series as an input to a clustering algorithm and to estimate θ(t) using the clustering results.

The clustering space is composed of L sampled points from the time series: $\{y(n)\in \Re^D \mid 1\le n \le L\}$. We repeat the clustering for every possible number of clusters, N, in the range $[N_{min}, N_{max}]$ to receive a set of cluster means $\{\hat{\mu}_i^N \in \Re^D \mid 1 \le i \le N;\ N_{min} \le N \le N_{max}\}$ and membership matrices $\{\mathbf{U}^N\in[0,1]^{L\bullet N} \mid N_{min}\le N \le N_{max}\}$. We are currently using two alternatives for clustering: the unsupervised optimal fuzzy clustering algorithm (UOFC) defined in [3] and the deterministic annealing approach from [4]. After performing the clustering we divide the sampled points into a set of L/K segments, each including K samples, and use the membership matrices to produce N possible estimations of the series θ(t). Each possible estimation, $\theta^N(t)$, $N_{min}\le N \le N_{max,}$ assumes a different number of sub-sources. The estimation result is given by:

$$\left\{\hat{\theta}_i^N(\tau) = \frac{1}{K}\sum_{j=K(\tau-1)+1}^{K\cdot\tau} u_{i,j}^N \;\middle|\; 1\le\tau\le\frac{L}{K};\ 1\le i\le N;\ N_{min}\le N\le N_{max}\right\} \tag{6}$$

Where N is the number of clusters and $u^N_{i,j}$ is an element in the membership matrix, $\mathbf{U}^N$, which resulted from clustering to N clusters and represent the membership of the j-th sample in cluster i. The result for each value of N is a sampled version of the estimated θ(τ) in a sampling rate of f_s/K where f_s is the sampling rate of the given time series. The algorithm output is the series $\{\hat{\theta}_i^{N_c}(\tau) \mid 1 \le \tau \le L/K\}$ where N_c is the number of clusters which is selected by a special cluster validity criterion that will be detailed. The general algorithm will now be presented followed by a discussion on the selection of its parameters. Given a finite sample of the time series output of length L, perform the following steps to receive an estimate of $\theta(t)$:

```
1.  for every value of N in the range $N_{min}$ to $N_{max}$ :
    a)  use a fuzzy clustering method to calculate cluster means using all
        given data periods
    b)  set Nc=0 and segment the sampled series to consecutive samples of K
        elements each.
    c)  for every segment of length K do:
        i)   classify every point to a single sub-source.
        ii)  estimate the temporal value of $\theta^N(t)$ by (6).
    d)  if elements of the series $\theta_i^N(t)$ do not agree with the stopping condition
        then set $N_c$=N, else stop and return $N_c$.
2.  if $N_c$=0 return failure else return $\theta^{Nc}(t)$ .
```

Stopping condition for the algorithm

As a stopping condition we use the fact that when the number of estimated clusters exceeds the original number of clusters, then at least two estimated clusters contain elements which originated from the same sub-source. We deal with the common case where a true cluster splits into two or more pseudo-clusters with elements which originated in the same source. If two pseudo-cluster relate in that way then their relative temporal probability becomes:

$$P(c_1)/P(c_2)=\frac{P\{c_1 \mid P(y)=N(\mu_i,\sigma_i)\}+\varepsilon_1}{P\{c_2 \mid P(y)=N(\mu_i,\sigma_i)\}+\varepsilon_2}\approx \text{const} \tag{7}$$

where $P\{c_1|P(y)=N(\mu_i,\sigma_i)\}$ is the probability of selecting pseudo-cluster 1 given that the sample point y originated in sub-source i. $\varepsilon_{1,2}$ are the probability of error classification. If we ignore $\varepsilon_1,\varepsilon_2$ then we have that $P(c_1)/P(c_2)$ becomes the relative probability of selecting each pseudo-cluster given this sub-source which is constant. For a sub-source which results in two pseudo-clusters i,j we expect that the value of $\frac{\hat{\theta}_i}{\hat{\theta}_j}$ will be close to constant. We regard the case of two original sub-sources with constant relative probabilities as a single source regarding the temporal behavior of the series values distribution.

We will also select N_{max} large enough for the stopping condition to limit the number of clusters to a value smaller than $N_{max.}$.

The number of samples for each estimation of θ(t)

The number of elements in each segment used for estimating a single value of θ(t) can be determined by several constraints. We will examine two important examples. First, an error bound can be related to the number of samples in each segment by the maximal variance of the sub-sources σ^2_{max} [5]:

$$\varepsilon_{max} = \sigma^2_{max}/K \tag{8}$$

Another option is to use a band limit on θ(t) which is available in some applications, when all processes $\theta_i(t)$ are band limited by a mutual constant bound B which reflects the assumption of relatively slow transients between semi-static states of the system. The band limitation is given for each element $\theta_i(t)$ in θ(t) by:

$$\theta_i(\omega) = \frac{1}{T}\int_{-\infty}^{\infty}\theta_i(t)e^{-\omega t}dt = 0 \quad \forall\, N \geq i \geq 1\,,\ |\omega|>B \tag{9}$$

This limitation, together with the Nyquist sampling condition suggests an upper bound on the number of samples that can be used for each estimation of θ(t) in order to avoid aliasing. If the band limit can be expressed relative to the time series sampling rate:

$$R_B = \frac{f_s}{B} \tag{10}$$

where f_s is the time series sampling rate and B is defined in (9) then the maximal segment length should not exceed $R_B/2$. The two constraints can be combined by the demand:

$$\frac{f_s}{B} \geq K \geq \frac{\sigma^2_{max}}{\varepsilon_{max}} \tag{11}$$

Simulation example

As an example we present an estimation result of a simulated scalar composite source (D=1). The source distribution is given by: $P(t) = \theta_1(t)\mathcal{N}(10,4) + \theta_2(t)\mathcal{N}(40,4) + \theta_3(t)\mathcal{N}(60,4)$ where:

$\theta_1(t)=|0.5\sin(2\pi\omega_1 t)|$, $\theta_2(t) = |0.5\sin(2\pi\omega_2 t)|$, $\theta_3(t) = 1 - \theta_2(t) - \theta_1(t)$, $\omega_1=10^4$, $\omega_2=5\bullet10^3$. A plot of 1000 samples of the series is drawn in Figure 2. 5000 samples were used as the input for the algorithm and the segment length was chosen to be K=50 with 50% overlapping between the segments which yielded 200 temporal estimations of θ(t). The deterministic annealing algorithm [4] was used as the clustering algorithm and the stopping condition for N was the existence of any two estimated cluster with an averaged relation of C±0.1 where C is any constant. The algorithm correctly detected three clusters and the estimation of $\theta_1(t)$, $\theta_2(t)$ and $\theta_3(t)$ is given in Figure 3 together with the original temporal probability signal that was used in the generation model. The relative probabilities of the estimations for the case N=3 is given in Figure 4. The relation is presented with a logarithmic axis. The relative probability between two observed pseudo-clusters is presented in figure 5 for the case N=4 and it is demonstrates that the points from a single sub-source were split between two estimated pseudo-clusters. A threshold on the variance of the relation between clusters was used to stop the algorithm and to detect that N_c=3.

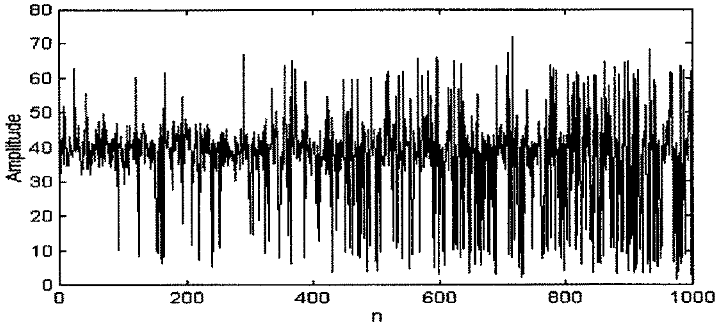

Figure 2

A sample of the time series output

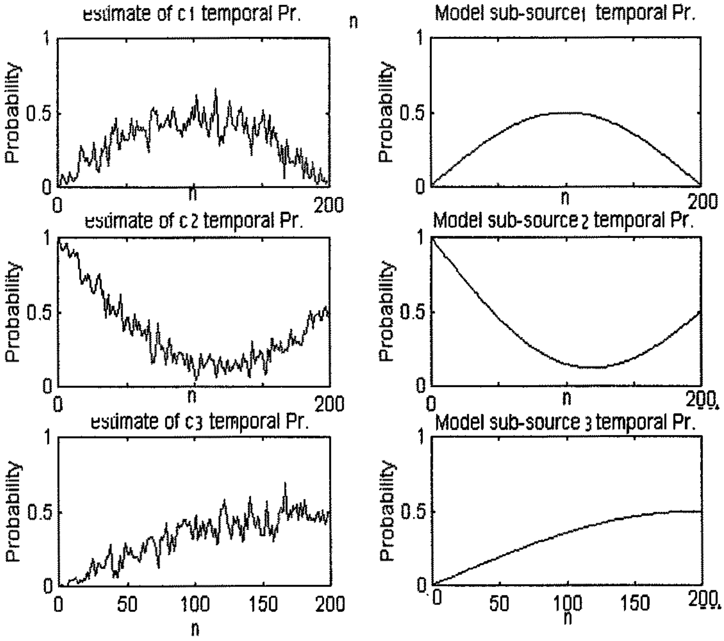

Figure 3

Estimation results for 3 sub-source example

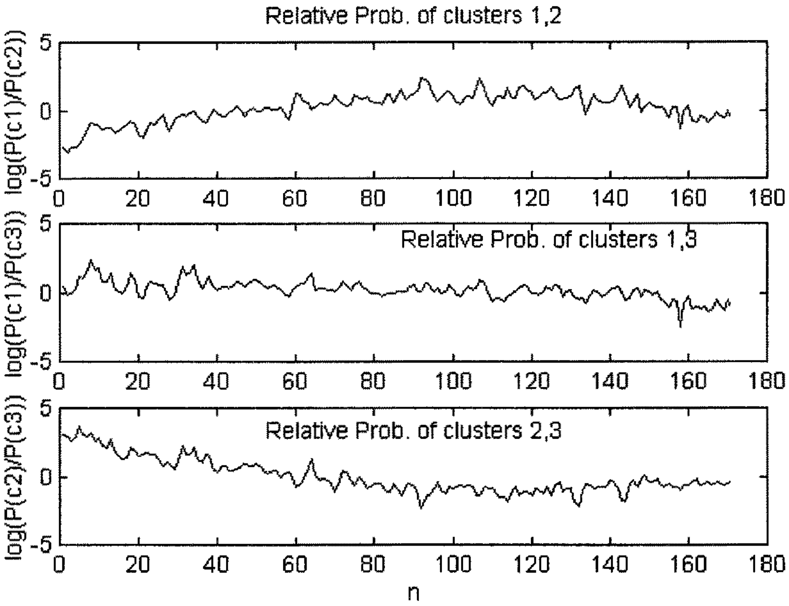

Figure 4

Relative probability of clusters

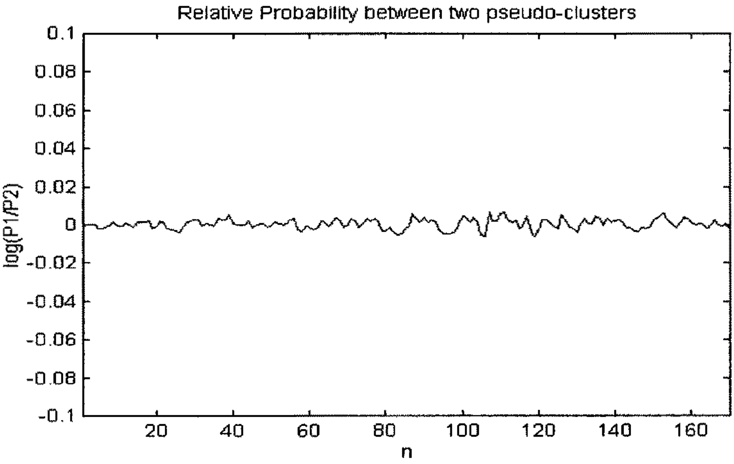

Figure 5

Relative probability of two pseudo-clusters originated in the same sub-source

Summary and open problems

A new model for the generation of a temporal composite source was presented together with a method for estimation of the statistical properties of the model. A stopping condition for an unsupervised fuzzy clustering algorithm was derived and serves as an example for a clustering validity criterion for time series analysis algorithms. Current work is focused on using the model in practical applications and deriving bounds on the quality of estimating θ(t).

References

[1] J. R. Deller, J. G. Proakis, J. H. L. Hansen *Discrete-time processing of speech signals*. Prentice-Hall, 1987

[2] E. R. Kandel, J. H. Schwartz and T. M. Jessell, *Principles of neural science*. Prentice-Hall, 1991

[3] I. Gath and A.B. Geva, “Unsupervised optimal fuzzy clustering”. IEEE Trans. On Pattern Anal. Machine Intell., Vol 7, pp. 773-781, 1989

[4] K. Rose, E. Gurewitz and G. Fox “A deterministic annealing approach to clustering”. IEEE Pattern Recognition Letters 11, 1990 589-594

[5] B. Porat, *Digital Processing of Random Signals*. Prentice-Hall, 1994.