Change Detection and Data Segmentation Methods

THEODOR D. POPESCU
National Institute for R & D in Informatics
Research Department
8-10 Averescu Avenue, 011455 Bucharest
ROMANIA

MARIANE MANOLESCU
National Institute for R & D in Informatics
Research Department
8-10 Averescu Avenue, 011455 Bucharest
ROMANIA

Abstract: The problem of change detection and data segmentation has received considerable attention in a research context and appears to be the central issue in various application areas. The change detection and segmentation model used in this paper is the simplest extension of the linear regression models to data with abruptly changing properties. Usually, a change detection algorithm consists in two stages: residual generation and decision making. The residuals are analytical redundancy generated data representing the difference between the observed and expected system behavior. In the stage of decision making, the residuals are processed and analyzed under certain decision rules to determine the system change status. The following techniques are investigated: filtering techniques with a whiteness test, techniques based on sliding windows and distance measures, and maximum likelihood techniques for change point estimation. The results of some Monte-Carlo simulations for change detection and segmentation in signals with changes in the mean value and in the AR model parameters are presented.

Key–Words: Change detection, Data segmentation, Regression models, Decision making, Filtering techniques, Distance measure, Maximum likelihood.

1 Introduction

The problem of change detection or segmentation of data has received considerable attention during the last two decade in a research context and appears to be the central issue in various application areas.

The analysis of the behavior of such real data reveals the most of the changes that occur are either changes in the mean level, or changes in spectral characteristics. In this framework, the problem of segmentation between "homogenous" parts of the data (or detection of changes in the data) arises more or less explicitly. Actually, two main types of problems can be distinguished:

1. Segmentation of the data, the true model of which is not known, and where the model used for change or jump detection is simply a tool to locate the boundaries.

2. Segmentation of the data which are approximately represented by a large amount of models: the analysis is then of an artificial intelligence type, the changes may be not really abrupt.

The proposed problem formulation assumes off-line or batch-wise data processing, although the solution is sequential in data and an on-line data processing can be used. The change detection and segmentation model is the simplest possible extension of linear regression models to data with abruptly changing properties. It is assumed that the data can be described by one linear regression model within each segment with distinct parameter vector and noise variance.

2 Problem Formulation

The following problem is addressed: Let $\{Y_1\}$ and $\{Y_2\}$ two sets of stationary data, and one want to test the null hypothesis:

H_0 : $\{Y_1\}$ and $\{Y_2\}$ are generated by the same rule.
H_1 : $\{Y_1\}$ and $\{Y_2\}$ are generated by different rules.

Concerning the data generating mechanism, it is assumed that under H_0, data sets $\{Y_1\}$ and $\{Y_2\}$ are generated by an autoregressive AR(p) process, whose parameters may jump at some unknown time, i.e.,

$$y_t = \sum_{k=1}^{p} a_k^{(t)} y_{t-k} + \epsilon_t, \quad \text{var}(\epsilon_t) = \sigma_t \qquad (1)$$

where

$$a_k^{(t)} = a_k^{(1)}, \quad 1 \le k \le p, \quad \text{for} \quad t < \tau$$

$$\sigma_t = \sigma_1, \qquad \text{for} \quad t < \tau$$

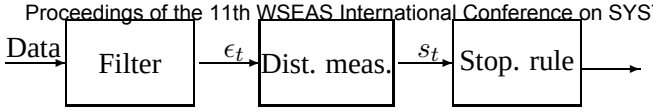

Figure 1: Change detection based on a whiteness test for filter residuals

$$a_k^{(t)} = a_k^{(2)}, \quad 1 \leq k \leq p, \quad \text{for} \quad t \geq \tau$$

$$\sigma_t = \sigma_2, \qquad\qquad \text{for} \quad t \geq \tau$$

and ϵ_t is a white noise sequence.

This assumption is not too restrictive since many stationary processes encountered in practice can be closely approximated by AR models. The advantage of this assumption consists of the computational simplicity of the resulted test procedures.

The change detection problem consists in the sequential detection of the change, and the estimation of the change time, τ, with few false alarms, short delay for detection and symmetrical detection (comparable performances when detecting a change from model (1) to model (2), or inverse).

There are different approaches to detect the changes in non-stationary signals. In this paper will be given the conceptual description of some methods for sequential detection of changes in non-stationary data, based on filtering and a whiteness test, sliding windows and distance measures and a MAP technique for segmentation.

3 Change Detection Based on Filtering

One useful approach for change detection consists in filtering of the observed data through a known or identified AR filter, and in looking for changes in the residual signal of innovations, $\{\epsilon_t\}$. Actually, the use of cusum techniques based upon the innovations (one-step prediction errors) ,$\{\epsilon_t\}$, or the squared innovations, $\{\epsilon_t^2\}$, is a standard approach for change detection in AR models. Such a technique, using $\{\epsilon_t^2\}$ is based upon the fact that, before the change $E(\epsilon_t^2) = \sigma_1$ and thus: $E(\epsilon_t^2/\sigma_1 - 1) = 0$.

To conclude, statistical whiteness tests can be used to test if the residuals are white noise as they should be if there is no change. Figure 1 shows the basic structure, where the filter residuals are transformed to a *distance measure*, that measures the deviation from the no-change hypothesis. The *stopping rule* decides whether the deviation is significant or not. The most natural distances are listed below, [1]:

- Change in the mean. The residual itself is used in the stopping rule and $s_t = \epsilon_t$.
- Change in variance. The squared residual substracted by a known residual variance λ is used and $s_t = \epsilon_t^2 - \lambda$.
- Change in correlation. The correlation between the residual and past outputs and/or inputs are used and $s_t = \epsilon_t y_{t-k}$ or $s_t = \epsilon_t u_{t-k}$ for some k.
- Change in sign correlation. For instance, one can use the fact that the white residuals should change sign every second sample in the average and use $s_t = \text{sign}(\epsilon_t \epsilon_{t-1})$.

The main problem in statistical change detection is now to decide what "large" are these distances. Many change detection algorithms can be recast into the problem of deciding on the following two hypotheses:

$$\begin{aligned} H_0 : E(s_t) &= 0, \\ H_1 : E(s_t) &> 0, \end{aligned}$$

where s_t is a *distance measure*. A stopping rule is essentially achieved by low-pass filtering s_t and comparing this value to a threshold. Below, two such low-pass filters are given:

- The CUmulative SUM (CUSUM) test of Page, [2]:

  $$g_t = \max(g_{t-1} + s_t - \nu, 0), \quad \text{change if} \quad g_t > h$$

  The *drift parameter* ν influences the low-pass effect, and the *threshold h* (and also ν) influences the performance of the detector.

- The Geometric Moving Average (GMA) test in Roberts, [3].

  $$g_t = \lambda g_{t-1} + (1 - \lambda) s_t, \quad \text{change if} \quad g_t > h.$$

  Here, the forgetting factor λ is used to tune the low-pass effect, and the threshold h is used to tune the performance of the detector. Using no forgetting at all (λ =0), corresponds to directly thresholding, which is one option.

It seems that classical approach consisting in testing how much the sequence of innovations ,$\{\epsilon_t\}$ is far from hypothesis "zero-mean white noise" is not sufficient for change detection in practice.

4 Change Detection Based on Sliding Windows

The main idea underlying this approach consists of comparison of two models: a model (M_2), based on data from a sliding window of size L ($y_{t-L+1}, \ldots, y_t$) is compared to a model (M_1) based on all data or a substantially larger sliding window ($y_1, y_2, \ldots, y_t$), [4]. If the model based on the larger data window gives larger residuals

$$\|\epsilon_t^1\| > \|\epsilon_t^2\|,$$

then a change is detected. The problem here is to choose a norm that corresponds to a relevant statistical measure. Some norms that have been proposed are:

- The Generalized Likelihood Ratio (GLR).
- The divergence test.
- Change in spectral distance. There are many methods to measure the distance between two spectra. One approach would be to compare the spectral distance of two models.

These criteria provide an s_t to be put into a stopping rule for instance, the CUSUM test. The choice of window size L is very critical here. On the one hand, a large value is need to get an accurate model in the sliding window and, on the other hand, a small value is needed to get quick detection.

Concerning the distance functions presented above, we will give in the following their expressions. In a linear regression model, AR(p), y_t can be written:

$$y_t = \phi_t^T \theta + \epsilon_t \tag{2}$$

with

$$\begin{aligned} \phi_t &= (y_{t-1}, y_{t-2} \ldots, y_{t-p})^T \\ \theta &= (a_1, a_2, \ldots, a_p)^T \end{aligned}$$

In [5], two different test statistics for the case of two different models are given. A straightforward extension of the generalized likelihood ratio test leads to:

$$d_{GLR} = L \log \frac{\sigma_1}{\sigma_2} + \frac{(y_t - \phi_t^T \theta_1)^2}{\sigma_1} - \frac{(y_t - \phi_t^T \theta_2)^2}{\sigma_2} \tag{3}$$

This test statistic was as the same time proposed in Appel and Brandt, [6] and will be referred as Brandt's GLR method.

To measure the distance between two models, any norm can be used. So, the Kullback discrimination information, [7] between two probability density functions p_1 and p_2 is defined as:

$$I(1,2) = \int p_1(x) \log \frac{p_1(x)}{p_2(x)} dx \geq 0 \tag{4}$$

In the special case of Gaussian distribution, we get

$$\begin{aligned} p_i(x) &= N(\hat{\theta}_i, P_i) \\ I(1,2) &= \frac{1}{2}\text{tr}(P_2^{-1} P_1 - I) + \\ &+ \frac{1}{2}(\hat{\theta}_1 - \hat{\theta}_2)^T P_2^{-1} (\hat{\theta}_1 - \hat{\theta}_2) - \\ &- \frac{1}{2} \log \left(\frac{\det P_1}{\det P_2} \right) \end{aligned}$$

The Kullback information is not a norm (it is not symmetric) and is not suitable as a distance measure. Instead, Kullback divergence is used:

$$V(1,2) = I(1,2) + I(2,1) \geq 0 \tag{5}$$

From Kullback divergence, the divergence test can be derived and it equals:

$$\begin{aligned} d_{DIV} &= L \left(\frac{\sigma_1}{\sigma_2} - 1 \right) + \left(1 + \frac{\sigma_1}{\sigma_2} \right) \frac{(y_t - \phi_t^T \theta_1)^2}{\sigma_1} - \\ &- 2 \frac{(y_t - \phi_t^T \theta_1)(y_t - \phi_t^T \theta_2)}{\sigma_2} \end{aligned} \tag{6}$$

The corresponding algorithm will be called the divergence test. d_{GLR} and d_{DIV} start to grow when a jump produced, and again the task of the stopping rule is to decide whether the growth is significant.

Concerning the parameter estimation of the models can be used the lattice implementation of the approximate least squares method, [8], for the long-term filter M_1, and the covariance method, [9], for the current filter M_2.

5 Change Detection Based on Segmentation

In segmentation, the goal is to find a sequence $k^n = (k_1, k_2, \ldots, k_n)$ of time indices, where both the n and the locations k_i are unknown, such that the signal can be accurately described as piecewise constant, i.e.

$$y_t = \theta_i + \epsilon_t, \quad \text{when} \quad k_{i-1} < t < k_t \tag{7}$$

is a good description of the observed signal y_t. The noise variance will be noted $E(\epsilon_t^2) = \sigma$.

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

$$\hat{k^n} = \arg \min_{0 < k_1 < \ldots < k_n = N} V(k^n) \tag{8}$$

where $n \geq 1$.

The procedure is illustrated below:

$$\begin{array}{ccc} y_1, y_2, \ldots, y_{k_1} & \ldots & y_{k_{n-1}+1}, \ldots, y_{k_n} \\ \text{Segment 1} & \ldots & \text{Segment n} \\ \hat{\theta}_1, \hat{\sigma}_1 & \ldots & \hat{\theta}_n, \hat{\sigma}_n \end{array}$$

Note that the segmentation k^n has $n-1$ degrees of freedom. Two types of optimality criteria have been proposed:

- Statistical criteria: The maximum likelihood or maximum a posteriori estimate (MAP) of k^n
- Information based criteria: The information of data in each segment is V_i (the sum of squared residuals) and the total information is the sum of these. Since the total information is minimized for the degenerated solution $k^n = 1, 2, \ldots, N$, giving $V_i = 0$, a penalty term is needed.

The main problem in segmentation is the dimensionality. The number of segmentation k^n is 2^N (can be a change or no change at each time instant). Several strategies have been proposed:

- Numerical searches based on dynamic programming or Markov chain Monte Carlo (MCMC) techniques.
- Recursive local searches schemes.

5.1 ML Change Time Sequence Estimation

Consider first an off-line problem, where the sequence of change times $k^n = k_1, k_2, \ldots, k_n$ is estimated from the data sequence y^t. We will use the likelihood for the data, given that the vector of change points is $p(y^t|k^n)$.

$$\begin{array}{ccc} y_1, y_2, \ldots, y_{k_1} & \ldots & y_{k_{n-1}+1}, \ldots, y_{k_n} = y_N \\ p\left(y_1^{k_1}\right) & \ldots & p\left(y_{k_{n-1}+1}^{k_n}\right) \end{array}$$

Repeatedly using independence of θ in different segments gives:

$$p(y^t|k^n) = \begin{cases} p(y^t) \\ p(y_1^{k_1}) \prod_{i=1}^{n=1} p\left(y_{k_i+1}^{k_{i+1}}\right) p\left(y_{k_n+1}^{t}\right) \end{cases} \tag{9}$$

for $n = 0$ and $n > 0$, respectively.

The *maximum likelihood (ML) estimate* is

$$(\widehat{n, k^n}) = \arg \max_{(n, k^n)} p(y^t|k^n) \tag{10}$$

In the Bayesian case, the change time has to be interpreted as a random variable, and the idea is to assign a probability q for a change at each time instant, and assuming independence:

$$P(\text{change at time } i) = q, \;\; 0 < q < 1. \tag{11}$$

Bayes'rule gives

$$p(k^n|y^t) = \frac{p(k^n)}{p(y^t)} p(y^t|k^n) \tag{12}$$

The maximing argument is called the *maximum a posteriory (MAP) estimate,* which is not influenced by the scaling factor $p(y^t)$

$$\begin{aligned} (\widehat{n, k^n}) &= \arg \max_{(n, k^n)} p(y^t|k^n) p(k^n) \\ &= \arg \max_{(n, k^n)} p(y^t|k^n) q^n (1-q)^{t-n} \end{aligned} \tag{13}$$

Note that with $q = 0.5$ the MAP and ML estimates coincide.

5.2 Information Based Segmentation

A natural estimation approach to segmentation would be to form a loss function. An off-line formulation for N observations is:

$$V_N(\theta^{n+1}, k^n) = \sum_{i=0}^{n} V_i(\theta_i) \tag{14}$$

$$V_i(\theta_i) = \sum_{t=k_i+1}^{k_{i+1}} (y_t - \theta_i)^2 \tag{15}$$

where $k_0 = 0$ and $k_{n+1} = N$ are used to define the first and the last segments. Straightforward minimization of $V_i(\theta_i)$ gives:

$$\begin{aligned}(\widehat{n,k^n}) &= \arg \min_{(n,k^n)} V_N(k^n) \\ &= \arg \min_{(n,k^n)} \sum_{i=0}^{n}(k_{i+1}-k_i)\hat{\sigma}_i\end{aligned} \tag{16}$$

It can be noted that the loss function is monotonously decreasing in n for all segmentation and this motivated the use of a penalty term for the number of change points. Penalty term occuring in model order selection problems can be used in this case:

- Akaike′s AIC, [10], with penalty term $2n(p+1)$
- The asymptotic equivalent criteria: Akaike′s BIC, [11], Rissanen's Maximum Description Length (MDL) approach, [12], and Schwartz criterion, [13]. The penalty term is $n(p+1)\log N$.

where p refers the number of parameters in the model. Both AIC and BIC are based on an assumption on a large number of data and tend to over segment the data.

6 Experimental Results

In the next subsections we present two case studies for change detection and segmentation in simulation, for changes in the mean of a signal and in the parameters of an AR model.

6.1 Change detection in the mean of a signal

The results obtained by Monte-Carlo simulation in the case of a change in the signal mean for 1000 noise realizations are presented. The signal contains a jump from 1 to 2 value at the instant 100. The experiments were performed for different values of the noise variance.

The results obtained for $\sigma = 0.01$ are presented in Fig. 2, Fig. 3 and Fig. 4 under the form of the histogram of the change detection instants, for the methods and stopping rules used.

6.2 Change detection in parameters of an AR model

The AR model used is a piecewise constant model of order 1:

$$y_t = \phi_1 y_{t-1} + \epsilon_t \tag{17}$$

with the values of the ϕ_1 parameter given in Table 1, when the first 300 samples of y_t were used.

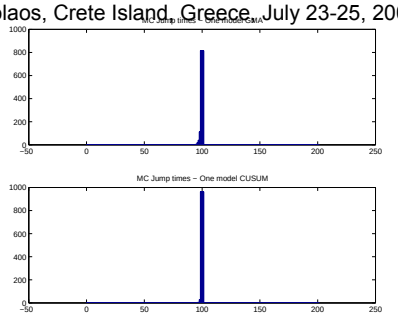

Figure 2: Histogram of the change instants for the filtering approach, with one model, GMA and CUSUM

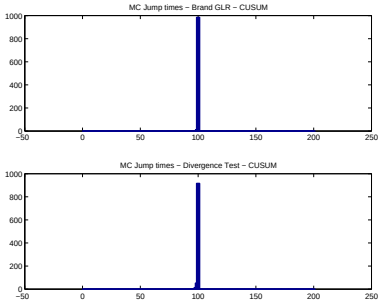

Figure 3: Histogram of the change instants for the sliding windows approach, with two models, GLR-CUSUM and DIV-CUSUM

The change detection results obtained by Monte-Carlo simulation, for sliding windows approach and MAP segmentation, and 0 delay in detection, are given in Fig. 5, Fig. 6 and Fig. 7 under the form of histogram of the change detection instants for a noise level of $\sigma = 0.01$.

Concerning the computation effort it is significant for segmentation MAP and reduced for the sliding windows and filtering approaches.

7 Conclusions

The paper gives the conceptual description of some change detection and data segmentation methods based on filtering, sliding windows and likelihood techniques and presents some Monte-Carlo simulation in two cases: change in the mean of a signal and change in the parameters of an AR model. Based on the obtained results it can be noted that the perfor-

t	[1-100)	[100-200)	[200-300]
ϕ_1	-0.4	0.8	-0.5

Table 1: Values of ϕ_1 parameter

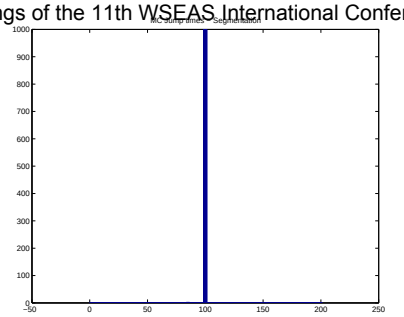

Figure 4: Histogram of the change instants for the segmentation MAP approach

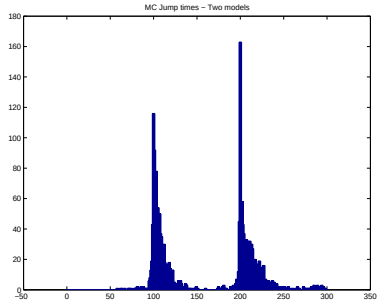

Figure 5: Histogram of the change instants for sliding windows approach and Brand GLR stopping test

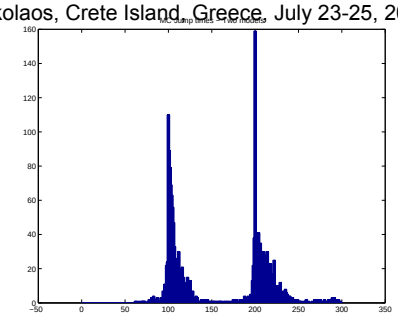

Figure 6: Histogram of the change instants for the sliding windows approach and divergence stopping test

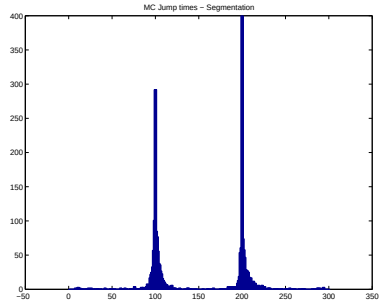

Figure 7: Histogram of the change instants for the segmentation MAP approach

mances of the MAP technique are superior to the other methods investigated, but with the price of the computation effort. The performances of the first approaches depend to a great extend of the choosing of the design parameters ν and h. The further evaluation,, in Monte Carlo simulation, of the methods, as well as their application in practice with real data and possibilities to validate the results are necessary.

References:

[1] F. Gustafsson, *Adaptive Filtering and Change Detection*, Wiley, NJ, 2001.

[2] E.S. Page, Continuous inspection schemes, *Biometrica*, 41, 1954, pp. 100–115.

[3] W. Roberts, Control charts based on geo metric moving averages, *Technometrics*, 8, 1959,pp. 411–430.

[4] M. Basseville, and I.V. Nikiforov, *Detection of abrupt changes: theory and applications*, Information and system science series, Prentice Hall, Englewood Cliffs, NJ, 1993.

[5] M. Basseville, and A. Benveniste, Sequential detection of abrupt changes in spectral characteristics of digital signals, *IEEE Trans. on Inf. Th.*, 29, 1983, pp. 709–724.

[6] U. Appel, and A.V. Brandt, Adaptive sequential segmentation of piecewise stationary time series, *Information Sciences*, 29, 1983, pp. 27–56.

[7] K.S. Kumamaru, K.S. Sagara and T. Söderstrom, Some statistical methods for fault diagnosis in dynamical systems, In R. Patton, P. Frank and R. Clark, editors, *Fault diagnosis in dynamic systems - Theory and applications*, pp. 439–476, Prentice Hall International, London, UK, 1989.

[8] J. Makhoul, Stable and efficient lattice methods for linear prediction, *IEEE Trans. on ASSP*, 25, 1977, pp. 423–428.

[9] J.D. Markel and A. H. Gray: *Linear Prediction of Speech*, Springer Verlag, 1976.

[10] H. Akaike, Fitting autoregressive models for prediction, *Annals of Institute for Statistical Mathematics*, 21, 1969, pp. 243–247.

[11] H. Akaike, On entropy maximization principle, *Proc. Symp. on Applications of Statistics*, 1977.

[12] J. Rissanen, *Stochastic complexity and modeling*, World Scientific, Singapore, 1989.

[13] G. Schwartz, Estimating the dimension of a model, *Annals of Statistics*, 6, 1978, pp. 461–464.