Currency Exchange Rate Segmentation and Modelling

Costas Panagiotakis*, Georgios Tziritas* and Athanasios Papadopoulos+

*Department of Computer Science, University of Crete, Heraklion, Greece
+Department of Economics, University of Crete, Rethymno, Greece

Abstract

In this paper, we propose a method for investigating the dynamics of markets. The goal is to segment/classify the time into intervals, where the market characteristics are homogeneous and at the same time to provide a description of each segment of the signal. We have combined a linear model and Fourier series to describe the market changes in long and short term, respectively. The method solves the time segmentation problem based on the Iso Level Algorithm (ILA), solving the EquiPartition problem (EP). Several market characteristics related to the used modelling efficiently summarize the behavior of the market in each segment. The whole methodology provides the critical times where the market behavior changes and a description of each estimated segment. We have successfully applied the proposed method on EURO-USD exchange rate of last ten years.

Key Words: segmentation, classification, Fourier transform, spectral analysis, euro-usd.

1 Introduction

The market time series analysis [1–5], the segmentation/classification and the description of the market dynamics [6] is a challenging problem. In [1,2] nonlinear time series analysis to high-frequency currency exchange data is performed and the correlations of foreign exchange markets are studied. In [4], the return time series of the rates of exchange of various currencies versus the U.S. dollar are modelled as a Markov process providing a signal reconstruction and prediction method as well as drift and diffusion coefficients. In [6], a segmentation method of financial time series is proposed. The method is based on a genetic approach solving the optimization problem of segmentation. There are three major categories of approaches to time series clustering [7]: 1) raw-data-based approaches, in which the series compared are considered as normally sampled at the same interval; 2) features-based approaches, in which the series are compared using some selected features; 3) model-based methods, where the time series are considered similar when the models characterizing them are similar. The approach proposed in this work belongs to the third category. Moreover, in [7], a model-based method is proposed following the tradition of AR processes to capture the similarity among time series. In [5] a model-based classification method of time series is proposed. Time series are considered in the same class when the models characterizing them are similar. In [3], a new measure of distance between time series based on the normalized periodogram is proposed. In [8] signal segmentation is done under fixed length segments for subsequent pattern discovery. In [9], an iterative algorithm is proposed stopping when a likelihood criterion is satisfied. More recently, signal processing techniques (power spectral analysis and filtering) were applied to characterize change in the dynamics of foreign exchange markets [10].

The traditional representation of time series data as a sequence of numerous consecutive time windows, each of which corresponds to a constant time interval, while being adequate for viewing a file, presents a number of limitations for the new emerging signal services such as content-based search, classification, understanding, retrieval, and browsing. Most of the above mentioned approaches address the signal summarization problem focusing either on a restricted signal content, minimizing metric criteria on feature domain, using constant time windows or applying simple clustering-based techniques. To overcome the aforementioned difficulties, a non-sequential (non-linear) content representation has to be provided, by extracting a small but meaningful information of the signal content. Therefore, it is important to segment the signals into homogenous segments in content and to describe each segment by a small and sufficient number of content descriptors providing a symbolic representation [8]. In this paper, signal summarization is performed by the use of an innovative computational geometry algorithm [11], which equally partitions the *signal* resulting time segments that are *equivalent* in the content domain under any type of signal content description. Moreover a signal description per segment is provided using the proposed modelling, yielding an efficient signal summarization. We have applied equipartition problem (EP) to segment and classify financial time series (currency exchange). We have used the EURO-USD exchange rate signal of last ten years.

More formally, let us consider a curve $C(t),\ t \in [0,1]$, describing an "event" in the n-dimensional feature space. How can we give a discrete summary of this event? Naturally, we can simply select a finite sample $\{C(t_i) : i \in \{0,\cdots,N\},\ t_i \in [0,1]\}$, but based on which criteria should we select the sampling time-moments t_i? A constant sampling frequency is not the best solution, since the evolution of the system need not be uniform at all. In many of the above applications it could be interesting to have a uniform representation according to an appropriate quality measure. The objective is the partition of the feature sequence into "homogenous" segments with uniform characteristics according to a predefined criterion. Let us consider a simple case of EP problem under Euclidean distance metric, where $N = 2$. Then we have to locate a curve point P, so that $|AP| = |PB|$. This point can be given as the intersection of the curve with the AB segment bisector. It holds that when $N > 2$, there is not a trivial method to solve EP [11].

The rest of the paper is organized as follows: Section 2 presents the proposed method. First, the problem is defined and then its reduction to EP and the algorithm description are given. Section 3 describes the proposed market characteristics that are used for the summarization of the market behavior. The experimental results are given in Section 4. Finally, conclusions and discussion are provided in Section 5. In Appendix Section mathematical formulations and an analysis of the proposed algorithm is provided.

2 Financial Time Series Analysis

2.1 Modelling

We have used as input financial time series of the EURO-USD exchange rate signal. The goal of the work is to automatically segment and to describe each segment of the given signal. The proposed method selects for each detected segment a compact representation, so that the signal is segmented into segments with equal reconstruction errors and number of coefficients per segment, solving simultaneously the signal segmentation and modelling. The model coefficients in each segment are used for signal summarization.

In bibliography, spectral analysis have been successfully used to model financial time series [10]. In this work, a time segment of the given signal $f(t)$, $t \in \{1,2,\cdots,T\}$ has been modelled by $g(t)$, a sum of linear model $l(t)$

and S most important Fourier coefficients $c(t)$.

$$\begin{aligned} l(t) &= a \cdot t + b & (1) \\ c(t) &= \frac{1}{T}\sum_{k=1}^{S} w_k \cdot e^{2\pi i \cdot f_k \cdot t} & (2) \\ g(t) &= l(t) + c(t) & (3) \end{aligned}$$

The parameters of the two models are computed independently in two steps. First, a and b are estimated by solving a linear system of T equations.

$$\begin{aligned} a + b &= f(1) \\ 2 \cdot a + b &= f(2) \\ 3 \cdot a + b &= f(3) \\ &\ldots \\ T \cdot a + b &= f(T) \end{aligned}$$

Next, w_k, f_k, $k \in 1, 2, ..., S$ are computed by the Fourier Transform of the signal $h(t) = f(t) - l(t)$. w_k and f_k, denote the weight and the frequency of k Fourier coefficient, respectively. We have used the S highest in energy coefficients of Fourier series [12] of the signal $h(t)$. These coefficients correspond to a robust reconstruction of the signal reducing noise and providing at the same time meaningful information of the market behavior. S is a parameter of the proposed model that can be defined by the user (e.g. $S = 5$).

2.2 Segmentation

The goal of the method is to provide a signal segmentation into homogenous time segments. Let N be the number of the time segments $[1, t_1] \cup [t_1, t_2] \cup \cdots \cup [t_{N-1}, T]$, where $t_i \leq t_{i+1} \in \{1, \cdots, T\}, i \in \{1, \cdots, N-1\}$ be the time that define the end time of $i-1$ segment and the start time of i segment $(t_0 = 1, t_N = T)$. Then, the approximation of $f(t)$, $g(t)$ is given by applying the above methodology for each time segment.

We consider that the approximation error $(E(f, g))$ between f and g is defined as the maximum error between the segments of f and their corresponding segments of g,

$$E(f, g) = \max_{i \in \{0,1,2,\cdots,N-1\}} d(t_i, t_{i+1}) \qquad (4)$$

This definition of error is used on polygonal approximation problem[1] [13]. Therefore, the goal of the method is to select the segments so that the approximation error is minimized. A near optimal solution of the segmentation

[1] Given an $N-$vertex polygonal curve P in the n-dimensional space $\Re^n$, the curve

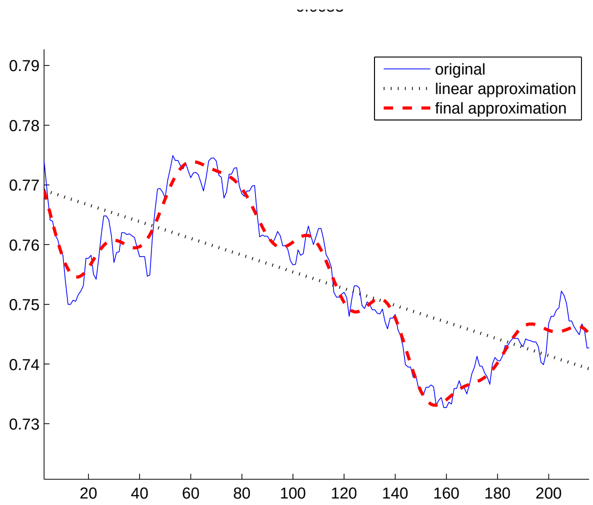


Fig. 1: The 220 days of EURO-USD (from 11/22/2006 to 06/30/2007) exchange rate signal (blue line) is approximated by a linear model (dotted line) and by the linear-plus-Fourier model (dashed line).

problem is achieved when the approximation errors per segment are equal, as the error is shared between all the segments,

$$\epsilon = d(1, t_1) = d(t_1, t_2) = \cdots = d(t_{N-1}, T) \tag{5}$$

This is the equal errors (EE) criterion. A detailed analysis of EE criterion is given in [13]. According to the EE criterion the segments are *equivalent* in the content domain yielding segments of the same number of degrees of freedom and equal errors in reconstruction. Therefore, the segmentation and the signal modelling are given at the same time by the reduction to the EP problem. The straightforward implementation of the EP algorithm provides directly N segments. The definition of the distance function $d(.,.)$ and the proposed algorithm analysis are given on Appendix Section 6. Hereafter, the proposed market characteristics are described that are estimated for each detected segment.

3 Market Characteristics

3.1 Linear model features

According to the used linear model, the parameter a corresponds to the mean market slope (increasing or decreasing rate) in long term analysis. If a is close to zero, then the market is almost stable, in long term. However, in most of the cases there is an increasing or decreasing rate, in long term analysis, corresponding to positive or negative a, respectively.

Fig. 1 illustrates the approximation of a financial time series using the proposed model. In this example $a = -0.000143$ meaning that the signal slightly decreases in long term analysis.

3.2 Fourier features

We have used the mean ν and standard deviation σ of Fourier Coefficients as features to describe locally the market behavior.

$$\nu = \frac{\sum_{k=1}^{S} |w_k| \cdot f_k}{\sum_{k=1}^{S} |w_k|} \tag{6}$$

$$\sigma = \sqrt{\frac{\sum_{k=1}^{S} |w_k| \cdot (f_k - \nu)^2}{\sum_{k=1}^{S} |w_k|}} \tag{7}$$

ν is an approximation of the mean frequency of the signal that corresponds to how often the market rate changes in short term analysis. σ is a measurement of how bandlimited is the signal. If σ is low (tends to zero), it means that the signal includes a major frequency and it is mainly periodic and predictable. Otherwise, as σ increases the financial signal changes are more chaotic and unpredictable and the market is unstable. Therefore, ν and σ measure how often the market changes and if the change follows a main frequency, respectively.

Fig. 1 illustrates the approximation of a financial time series using the proposed model. In this example $\nu = 0.0265$. ν corresponds to the mean frequency of the signal, showing that the signal period in short term analysis is $\frac{1}{0.0265} = 37.5$ days. This means that the signal period is about 37 days. In this case $\sigma = 0.0227$. As our experiments show, if $\sigma < 0.03$ means that the market is almost stable in short term analysis. More detailed analysis of experimental results is given in next section.

approximation P consists in computation of another $M-$vertex polygonal curve in the n-dimensional space $\Re^n$ that approximates the original curve, according to a predefined error criterion.

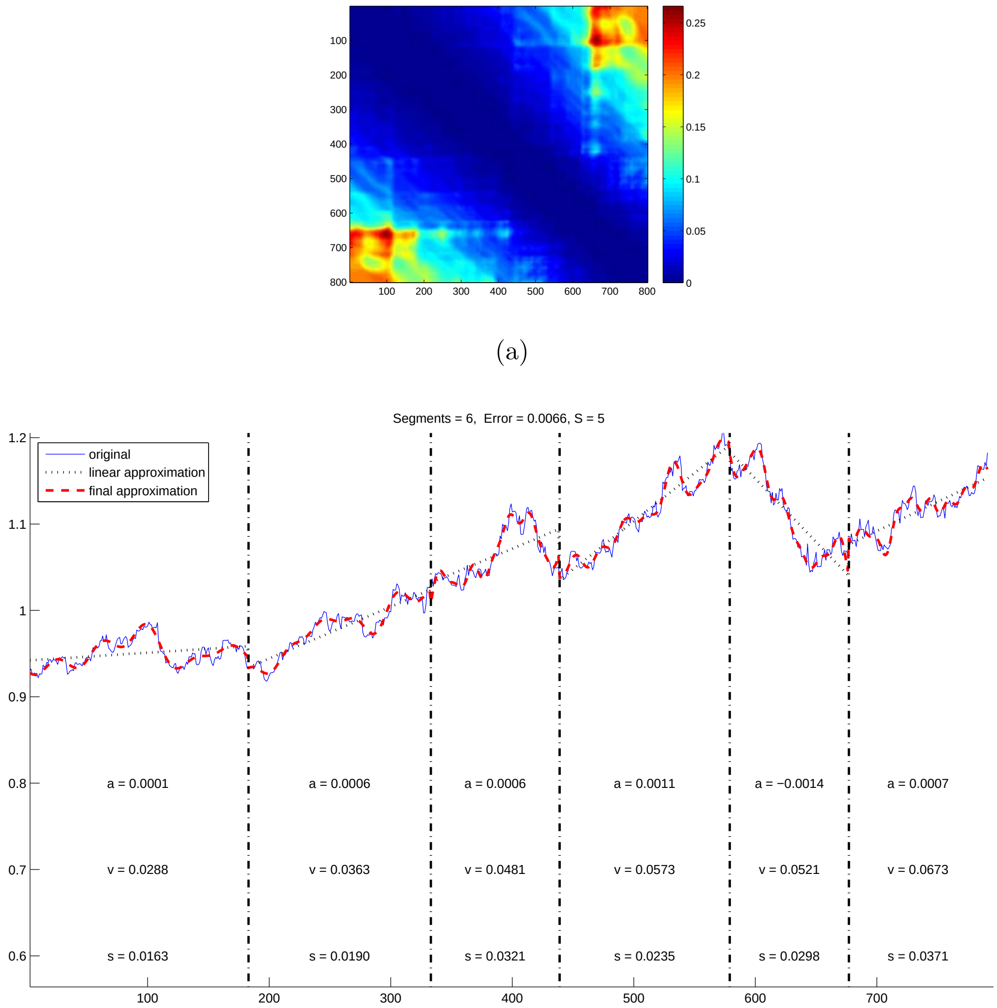
4 Experimental Results

In this section, the experimental results of the proposed method are presented. We have tested the proposed algorithm on a data set consisting the EURO-USD exchange rate of the last ten years. Figs. 2, 3 and 4 illustrate the results of the proposed algorithm for various EURO-USD time series of the last ten years, for $S = 5$. In each case, N was automatically estimated by the method of Section 6.2. Figs. 2(a), 3(a) and 4(a) illustrate the error function $d(u, v)$ for each case. $d(u, v)$ is minimized on its diagonal getting zero values. The highest values of $d(u, v)$ correspond to non homogeneous segments (high error on approximation). For example, in Fig. 4(a) $d(u, v)$ is maximized for segment (120,600) that is the most non homogeneous segment of the signal.

More specifically, the blue and the red dashed curves of Fig. 2 correspond to the EURO-USD exchange rate (from 24/03/1999 to 01/06/2001) and its reconstruction using the proposed modelling, respectively. The dotted curve corresponds to the linear approximation. The proposed method gives 6 segments. The estimated model parameters per segment are:

- $a = 0.0001$, $\nu = 0.00288$ and $\sigma = 0.0163$ (first segment)
- $a = 0.0006$, $\nu = 0.00363$ and $\sigma = 0.0190$ (second segment)
- $a = 0.0006$, $\nu = 0.00481$ and $\sigma = 0.0321$ (third segment)
- $a = 0.0011$, $\nu = 0.00573$ and $\sigma = 0.0235$ (fourth segment)
- $a = -0.0014$, $\nu = 0.00521$ and $\sigma = 0.0298$ (fifth segment)
- $a = 0.0007$, $\nu = 0.00673$ and $\sigma = 0.0371$ (sixth segment)

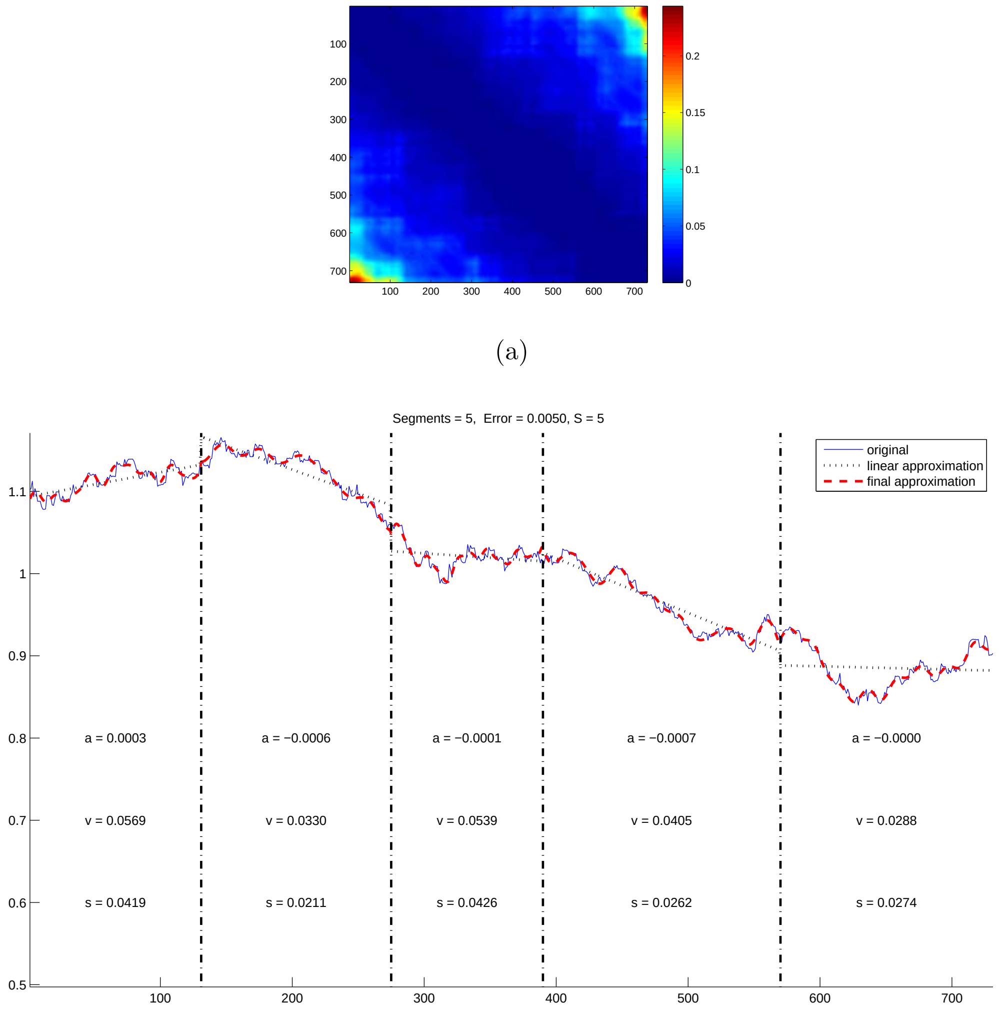
During first segment (24/03/1999 - 28/9/1999) the financial signal is almost stable ($a \simeq 0$)in long term analysis. The mean frequency as well σ are the lowest of all segments. This means that during this segment, the EURO-USD exchange rate is characterized by periodic behavior and local stability. During the second segment, the market has almost the same behavior, but the parameters ν and σ start to increase. During the next segments, ν and σ continue to increase driving the market to be unstable (see last two segments). Especially, during the last segment (02/02/2001 - 01/06/2001), where the exchange rate gradually increases, the mean frequency and σ are the highest of all segments, while during the previous one (26/10/2000 - 01/02/2001) the financial time series gradually decreases with the highest rate of all the segments.



(a)

(b)

Fig. 2: (a) The function $d(u, v)$ and **(b)** the segmentation results of the proposed method for 800 days of EURO-USD (from 24/03/1999 to 01/06/2001) exchange rate.



(a)

(b)

Fig. 3: **(a)** The function $d(u, v)$ and **(b)** the segmentation results of the proposed method for 2 years of EURO-USD (from 09/09/2001 to 09/09/2003) exchange rate.

Fig. 3 illustrates the results of the proposed algorithm for EURO-USD time series starting from 09/09/2001 and ending to 09/09/2003. The proposed method gives 5 segments. The estimated model parameters per segment are:

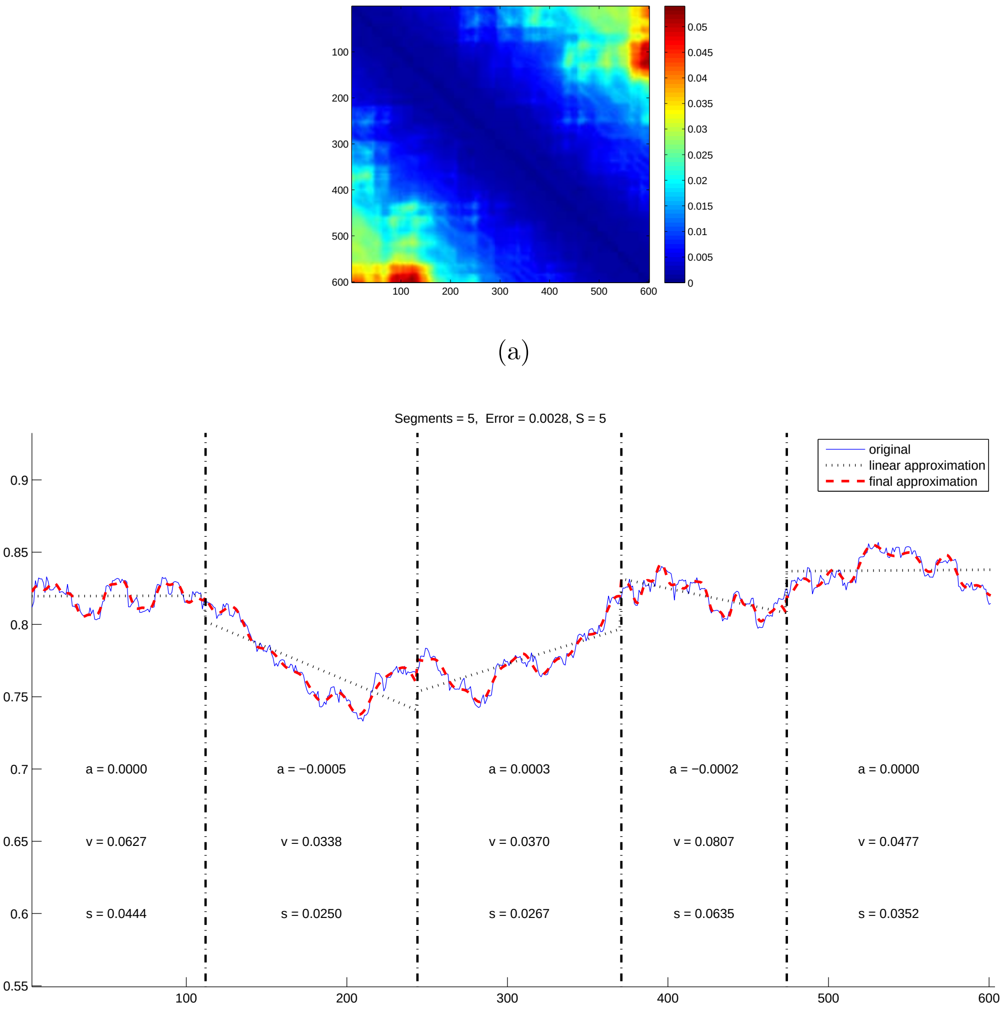
- $a = 0.0003$, $\nu = 0.0330$ and $\sigma = 0.0211$ (first segment)
- $a = -0.0006$, $\nu = 0.0363$ and $\sigma = 0.0190$ (second segment)
- $a = -0.0001$, $\nu = 0.0539$ and $\sigma = 0.0426$ (third segment)
- $a = -0.0007$, $\nu = 0.0405$ and $\sigma = 0.0262$ (fourth segment)
- $a = 0.0000$, $\nu = 0.0288$ and $\sigma = 0.0274$ (fifth segment)

During the first segment (09/09/2001 - 12/01/2002) and the third segment (16/06/2002 - 04/10/2002) signal is almost stable ($a \simeq 0$) having similar high values for ν and σ. Therefore, the financial signal has similar behavior in short and long term analysis, characterizing by fast and unpredictable alternations due to high frequency and high standard deviation, respectively. During the rest segments, ν and σ are low, and the financial signal is almost periodic, with slow alternations in short term analysis.

Fig. 4 illustrates the results of the proposed algorithm for EURO-USD time series starting from 05/06/2004 and ending to 26/01/2006. The proposed method gives 5 segments. The estimated model parameters per segment are:

- $a = 0.0000$, $\nu = 0.0627$ and $\sigma = 0.0444$ (first segment)
- $a = -0.0005$, $\nu = 0.0338$ and $\sigma = 0.0250$ (second segment)
- $a = 0.0003$, $\nu = 0.0370$ and $\sigma = 0.0267$ (third segment)
- $a = -0.0002$, $\nu = 0.0807$ and $\sigma = 0.0635$ (fourth segment)
- $a = 0.0000$, $\nu = 0.0477$ and $\sigma = 0.0352$ (fifth segment)

During the first segment (from 05/06/2004 to 22/09/2004) and the last two segments (from 15/06/2005 to 23/09/2005 and from 24/09/2005 to 26/01/2006) the financial signal is almost stable ($a \simeq 0$) in long term analysis having similar high values for ν and σ. Therefore, the financial time series has similar behavior in short and long term analysis, characterizing by fast and unpredictable, non periodic alternations due to high frequency and high standard deviation, respectively. During the rest segments, ν and σ are low, and the financial signal is almost periodic, with slow alternations



(a)

(b)

Fig. 4: **(a)** The function $d(u, v)$ and **(b)** the segmentation results of the proposed method for 600 days of EURO-USD (from 05/06/2004 to 26/01/2006) exchange rate.

in short term analysis. The main difference between these two successive segments (second and third segment) is appeared in decreasing/increasing market rate in long term analysis, due to strong negative and positive value of a, respectively.

Apart from the segmentation the proposed methodology can be applied for financial time series noise reduction, since we have used the highest in energy Fourier series coefficients. The red dotted curves of Figs. 2, 3 and 4 correspond to robust reconstruction of the financial time series using the proposed modelling.

5 Conclusions

In this paper, we have discussed a currency exchange segmentation method based on curve equipartition principle. The segments are homogeneous in content and can be described by a small number of coefficients that are provided by the proposed method. More specifically, a description of each time segment of the signal is done using a linear model and the highest in energy Fourier series coefficients providing short and long term analysis.

Moreover, a robust reconstruction (with noise reduction) of the signal is provided. The main advantage of the method is that the equal approximation error is the same per segment, so the reconstructed homogenous in content segments are equivalent on the approximation. The method is flexible on changes of signal modelling and signal dimension. Instead of Fourier series modelling, splines curves or wavelets can be used without any significant change on methodology. Apart from noise reduction, the method can be used for signal browsing and short/long term prediction of market behavior.

6 Appendix

6.1 Distance function

In this section the distance function is defined. Let $u, v,\ u \leq v$ be the start and the end time of a segment of the given signal $f(t)$, $t \in [u, v]$ and $g(t)$ be the corresponding approximation function using the proposed model. Let $d(u, v)$ be the error between the two functions. According to the EP problem [13], $d(u, v)$ should be continuous satisfying the following properties:

1. $d(u, v) = 0 \Leftrightarrow u = v$ (isolation).

2. $d(u, v) = d(v, u)$ (symmetry).

These properties are satisfied by the difference in energy of signal $h(t)$ and $c(t)$ that is used as error function.

$$d(u,v) = \frac{1}{v-u+1}(\sum_{t=u}^{v} h^2(t) - \sum_{t=u}^{v} c^2(t)) = (\frac{1}{v-u+1})(\sum_{t=u}^{v} h^2(t) - \frac{\sum_{k=1}^{S} |w_k|^2}{T}) \tag{8}$$

The error of the approximation is a good metric to show if the proposed model fits well to the data. It holds that $\frac{\sum_{k=1}^{S}|w_k|^2}{T} le \sum_{t=u}^{v} h^2(t)$ and $\frac{\sum_{k=1}^{S}|w_k|^2}{T} = \sum_{t=u}^{v} h^2(t) \Leftrightarrow d(u,v) = 0$, if and only if, $h = c$. If $d(u,v)$ is getting high, it means that the signal can not be described well by fourier basis. Therefore, if the error is low, it means that the segmentation is good and the content description of the signal by the proposed descriptors is valid. Moreover, it means that the segment is homogenous in content, since it can be described by a small number of coefficients that are related by the global signal content.

6.2 Iso-Level Algorithm (ILA)

The input of the proposed method is the number of segments N. In addition, it needs the values of symmetric matrix $g(t_k, t_l), k, l \in \{1, 2, \cdots, N\}$ of distortions. The detailed description of the algorithm used can be found in [13]. It is an iterative method. Thus, when it is executed for N segments, it uses the precomputed results for $N-1$ segments. In each iteration step l, the algorithm computes the zero level [13] L_l by the L_{l-1}. According to our analysis [11], the equipartition problem (EP) always admits at least one solution.

The number of key frames N can be given by the user or can be estimated automatically by terminating the EP algorithm when the estimated "distortion" exceeds a predefined error, similar with the problem of minimum number of segments $(min - \#)^2$ [13]. Both cases are solved in $O(N \cdot T^2)$ steps thanks to the property of the method that it solves the problem for values less than N without additional cost [13]. An important algorithm property is that the computation cost is independent of content curve dimension n, that coincides with the feature vector space dimension. This means that it is independent of content descriptors selection.

The selection of N could be done by the user to fit specific preferences and information needs. A 'semiautomatic' computation of N can be done by terminating the EP algorithm when the estimated "distortion" (error)

[2] $min - \#$ is used for the problem of finding the minimum number of segments that gives error lower than the given error (at polygonal approximation).

exceeds a predefined error. However, it is crucial to develop a mechanism able to automatically estimate the most appropriate number of segments N. N can be provided without any user interaction estimating if the segments at iteration l of the algorithm suffice to approximate the given signal.

Let Q_l (see Equation (9)) is a measurement of the distortion between the given signal and its approximation using the l segments. As Q_l we use the error of the approximation (r_l) multiplied by the number of segments (l).

$$Q_l = l \cdot r_l \tag{9}$$

This is done since the error under the proposed criterion the error is locally summed and the size of segments in time is inversely proportional to the number of segments or the level of EP algorithm.

Q_l is usually decreasing as l increasing, $Q_l \geq 0$. Q_l has characteristics of a convex function, that is, if we smooth it we will get a convex function. Therefore, we have to introduce a new criterion instead of minimum of this function. Thus, we propose to select the appropriate level l so that the numerical approximation of the second derivative of Q_l, $\ddot{Q}_l$, is maximized.

$$\ddot{Q}_l = Q_{l+1} + Q_{l-1} - 2 \cdot Q_l, \quad l \in \{2, 3, \cdots, N-2\} \tag{10}$$

This is due to the fact that the second derivative expresses a measure of the curvature of the content curve.

Acknowledgments

The author would like to thank Professors G. Georgakopoulos and K. Athanassopolos and for the fruitful collaboration on analysis of curve equipartition problem.

References

[1] F. Strozzia, J.-M. Zaldivarb, J. P. Zbilut, Application of nonlinear time series analysis techniques to high-frequency currency exchange data, Physica 312 (2002) 520–538.

[2] A.-H. Sato, Frequency analysis of tick quotes on the foreign exchange market and agent-based modeling: A spectral distance approach, Physica 382 (2007) 258–270.

[3] J. Caiado, N. Crato, D. Pena, A periodogram-based metric for time series classification, Computational Statistics & Data Analysis 50 (10) (2006) 2668–2684.

[4] F. Farahpour, Z. Eskandari, A. Bahraminasab, G. R. Jafari, F. Ghasemi, M. Sahimi, M. R. Rahimi Tabar, A langevin equation for the rates of currency exchange based on the markov analysis, Physica A Statistical Mechanics and its Applications 385 (2007) 601–608.

[5] M. Corduas, D. Piccolo, Time series clustering and classification by the autoregressive metric, Computational Statistics & Data Analysis 52 (4) (2008) 1860–1872.

[6] T. chung Fu, F. lai Chung, V. Ng, R. Luk, Evolutionary segmentation of financial time series into subsequences, in: Proc. of Evolutionary Computation, Vol. 1, 2001, pp. 426–430.

[7] E. Otranto, Clustering heteroskedastic time series by model-based procedures, Tech. rep. (2006).

[8] G. Das, K.-I. Lin, H. Mannila, Rule discovery from time series, in: Proc. of ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 1998, pp. 16–22.

[9] V. Gurslnik, J. Srivastava, Event detection from time series data, in: Proc. of ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 1999, pp. 33–42.

[10] T. Holden, K. Xu, S. L. Ho, Using signal processing techniques to characterise change in the dynamics of foreign exchange markets following political and economic events, in: Proc. of IEEE Int. Engineering Management Conf., 2004, pp. 1293–320.

[11] C. Panagiotakis, G. Georgakopoulos, G. Tziritas, On the curve equipartition problem: a brief exposition of basic issues, in: European Workshop on Computational Geometry, 2006.

[12] P. Duhamel, M. Vetterli, Fast fourier transforms: A tutorial review and a state of the art, Signal Processing 19 (1990) 259–299.

[13] C. Panagiotakis, G. Tziritas, Any dimension polygonal approximation based on equal errors principle, Pattern Recogn. Lett. 28 (5) (2007) 582–591.