Data Science in Finance and Economics

http://www.aimspress.com/journal/dsfe

DSFE, 2(3): 237–253.
DOI:10.3934/DSFE.2022012
Received: 20 May 2022
Revised: 12 July 2022
Accepted: 18 July 2022
Published: 09 August 2022

Research article

Retrospective technology of segmentation and classification for GARCH models based on the concept of the ϵ-complexity of continuous functions

Alexandra Piryatinska[1,*], **Boris Darkhovsky**[2]

[1] Department of Mathematics, San Francisco State University, 1600 Holloway ave, San Francisco, CA 94132, USA

[2] FRC CSC RAS

* **Correspondence:** Email: alpiryat@sfsu.edu.

Abstract: We consider a retrospective segmentation and classification problem for GARCH models. Segmentation is the partition of a long time series into homogeneous fragments. A fragment is homogeneous if only one mechanism generates it. The points of "concatenation" of homogeneous segments we call (by analogy with the term used in the stochastic literature) points of disorder or change-points. We call classification the separation of two relatively short time series generated by different mechanisms. By classification, we mean the way in which two groups of time series with unknown generating mechanism (in particularly, generated by GARCH models) can be distinguished , and the new time series can be assigned to the class. Our model free technology is based on our concept of ϵ-complexity of individual continuous functions. This technology does not use information about the time series generation mechanism. We demonstrate our approach on time series generated by GARCH models. We present simulations and real data analysis results confirming the effectiveness of the methodology.

Keywords: ϵ-complexity; GARCH models; model-free segmentation; classification

JEL Codes: C14, C19, C22, C29

1. Introduction

The generalized autoregressive conditional heteroskedasticity models GARCH were developed to address the problem of forecasting volatility in asset prices (Bollerslev T (1986)). They are a generalization of the Autoregressive Conditional Heteroskedasticity (ARCH) models (see, Engle R (1982)). Models ARCH and GARCH are used for modeling of the volatility of returns on a wide variety of financial assets such as exchange rate, individual stock, and stock indices (see, e.g., Brooks C

(2014)).

GARCH-models are non-linear models with multiple parameters. The moments of parameter changes correspond to the moments of changes in the volatility of the process. Obtaining estimates for the parameters of such models due to their non-linearity is a complex problem. If the model's parameters changed during data collection, then an important task is to detect the moments of change *without using estimates of the model parameters*. Failure to take into account the change points in financial time series results in many undesirable consequences. For example, Lamoureux CG, Lastrapes WD (1990) found that the persistence of the change in the variance of stock-return data may be overstated if deterministic structural shifts are ignored in the model.

We consider the problem of retrospective detecting the moments of changes in the parameters of GARCH models and the problem of classification of GARCH models with different parameters. There is extensive literature on the detection of change points. The extensive review of change-point detection methods can be found in Truong C, Oudre L, et al. (2020).

In econometrics, the change points are are also called structural breaks. The review on this topic can be found in Andreou E and Ghysels E (2009) and Aue A, & Horváth L (2013). An early attempt to detect change points in GARCH was made in Chu C (1995) who studies the properties of supremum–type F and Lagrange multiplier (LM) tests. He finds that these tests generally have good power but are somewhat sensitive to the assumption of normality. For the comparison of the performance of the different change point detection procedures in detecting structural breaks in GARCH models, see Smith D (2008) and bibliography there. In Li Q, Wang L Qiu F (2015) the Bayesian methods have been used to detect the change points. In Berkes I, Horváth L, and Kokozka P (2004) and Song J, Kang J (2018) the quasi-likelihood score function was used to construct test statistics in GARCH models and ARMA-GARCH models, respectively.

Time series classification (TSC) is a growing research topic due to the substantial amount of time-series data being collected in various fields. This is important in many areas of research and applications where the analysis of time-dependent or financial data might need to be analyzed to support a business decision. TSC uses supervised machine learning to analyze multiple labeled classes of time series data and then predict or classify the class that a new data set belongs. There are many algorithms that are developed to perform time series classification. For the reviews of such algorithms see, e.g Abanda A, Mori U, Lozano JA (2019), Faouzi J (2022). For application of time series classification see, e.g. Bagnall A, Janacek G (2014), Susto GA, Cenedese A, and Terzi M (2018). The time series classification problem is meaningful in finance too. For the application of TSC in finance see, e.g. Chao L, Zhipeng J, Yuanjie Z (2019) and Majumdar S, & Laha A K (2020).

Time series classification methods can be divided into three main categories: feature-based, model-based, and distance-based methods (see, Xing Z, Pe J, and Keogh E (2010)). In feature based classification methods, the time series are transformed into feature vectors and then classified by a conventional classifier such as a neural network or a decision tree. For overview of other methods see Abanda A, Mori U, Lozano JA (2019).

In our paper, we consider the problem of binary classification of relatively short time series. Each class is generated by the same generating mechanism and different classes are trajectories of different mechanisms. As a generated mechanism the GARCH (1,1) models are used.

The main idea of our approach is to obtain a solution to segmentation and classification problems without using the knowledge about the model and its parameters. In other words, the methodology we propose is model-free. Even we demonstrate the performance for GARCH(1,1) models, the information about the model is not used in the technology. Our solution is based on the concept of the ϵ-complexity of continuous functions. Such an approach enables us to develop model-free methods for segmentation and classification of time series of arbitrary nature.

The paper is organized as follows. In Section 2, we present the basic concepts and results of the theory of ϵ-complexity at a meaningful level, referring the reader to the exact formulations in Darkhovsky B (2020). The main ideas of segmentation and classification of time series based on ϵ-complexity coefficients are also given in Section 2. In section 3, we show the results of simulations. We detect changes in the parameters of GARCH models. In section 4, we consider examples of segmentation and classification of stock market data. Section 5 provides conclusions.

2. Methodology: Brief description of the ϵ-complexity theory and its application to segmentation and classification problems

2.1. Basic principles of our approach

Our approach to the segmentation and time series classification problems is fundamentally different from those described in the literature. Mainly, we do not use any information about the mechanisms of generation of the time series to solve these problems. Thus, our technology is applicable to a series of arbitrary nature (stochastic, deterministic, or mixed).

Obtaining a priori information about the mechanisms of the generation of a time series is a difficult task. If the series is a concatenation of several different fragments, then in principle, such information cannot be extracted without preliminary segmentation. Only for homogeneous (i.e., generated by an invariable mechanism) fragments of time series one can perform statistical analysis based on the laws of large numbers. As far as we know, most segmentation algorithms in the literature rely either on knowledge of the series generation mechanism up to parameters or a priori known set of models or on knowledge of which probabilistic characteristic (if we are talking about a stochastic process) can change. Thus, a vicious circle is obtained: segmentation requires a priori information about the generation mechanism, and such information can be extracted only from homogeneous (i.e., obtained after segmentation) fragments. Therefore the development of model-free methods for detecting change points is an important problem in this area. Recently model-free approaches have been developed, see Darkhovsky B, Piryatinska A (2018), Truong C, Oudre L, and Vayatis N (2019). In Truong C, Oudre L, and Vayatis N (2019) authors consider the search for change-points in a stochastic process under the assumption that there exists a linear transformation of the process, in which the change-points of the process turn into changes in the mathematical expectation after the transformation. In such a search, the prior information about what type of changes took place in the process in order to be able to specify the corresponding Hilbert kernel should be used. They also consider only examples in which the number of change-point is known a priory.

We consider more general case. We provide characteristics, the ϵ-complexity coefficients, for which the change in generating mechanisms leads to the change in the mean values of the ϵ-complexity coefficient. This approach is applicable for all types of stochastic processes as well as for chaotic deterministic or mixed processes Darkhovsky B, Piryatinska A (2018).

The situation is similar to the classification problem. For this problem, one needs to know a feature space in which the series can be separated. The definition of this space must be based on a priori information about the mechanism of generation of the series.

Our model-free technology, based on the concept of ϵ-complexity of continuous functions, allows us to break this vicious circle. In the next subsection, we describe this concept's main definitions and results at a meaningful level. In Section 3 the ideas of model-free segmentation and classification are provided.

We apply our technology to GARCH models. However, we would like to emphasize that the technology works for any time series model. Moreover, in contrast to all other methods in the literature, it allows detecting changes not only in stochastic processes but in processes of arbitrary nature, e.g., in chaotic deterministic or mixed processes (see, e.g Darkhovsky B, Piryatinska A (2018)). It is essential for financial series since the set of models of such series is much richer than models of the GARCH type. A researcher does not know a priori what kind of model should be sought to describe a particular time series. Therefore, in the general case, the segmentation and classification of time series should be carried out using model-free technology.

2.2. The ϵ-complexity theory

This subsection presents some results of the theory of the ϵ-complexity of continuous finite-dimensional maps. We restrict ourselves to showing the results of the theory on a meaningful level as applied to scalar functions of one variable, referring the reader to general precise formulations to Darkhovsky B (2020). Firstly, the theory was developed and presented in Darkovsky B, Piryatinska A (2014). In Darkhovsky B (2020) the definitions are more precise and theory is refined. Therefore, we refer to this paper.

Consider a continuous function $x(t) : [0, 1] \to \mathbb{R}$. Denote $R = \max_{t\in[0,1]} |x(t)|$ and will assume that $R > 0$.

Let us denote by $\mathcal{F}$ a given at most countable collection of methods for recovering of a function $x(t)$ from a finite set of its values on some uniform grid of $[0, 1]$.

The set of continuous functions that cannot be accurately reconstructed by any fixed set of methods $\mathcal{F}$ is everywhere dense in the space of continuous functions on $[0, 1]$.

Informally, we can say that the ϵ-complexity (more precisely, $(\epsilon, \mathcal{F})$-complexity) of function $x(t)$ estimates the number of its uniform discrete samples (more precisely, the logarithm of this number), which are necessary to recover this function by a given set of *approximation methods* $\mathcal{F}$ with the given accuracy. In other words, this quantity estimates the minimum amount of information (in the language of approximation theory) required to describe the function.

In this respect, the concept of ϵ-complexity is consistent with the concept of "complexity" of objects proposed by A.N. Kolmogorov in the mid-60s. The main idea of the Kolmogorov approach (see, for example, Kolmogorov A (1983)) is that: "complex" object requires a lot of information for its description, and a "simple" object requires a small amount of information for this. It is advisable to estimate the complexity of an object through the minimum amount of information required to describe it.

Suppose a function is exactly reconstructed from *a finite set of its values* (e.g., such is a linear function, since any reasonable set of recovery methods includes a linear or piece-wise linear approximation method). In that case, we set its complexity to be zero. However, it follows from the above that "almost any" (and in applications - any, due to the inevitable measurement noise) continuous function can not be exactly reconstructed by any fixed set of approximation methods and, therefore, has a nonzero $(\epsilon, \mathcal{F})$ -

complexity for any fixed set of recovery methods $\mathcal{F}$.

In majority of applications, a researcher deals with time series given by a discrete set of their values on a uniform grid. Assuming that such a collection of values is the restriction of a continuous function on some uniform grid, we can extend the theory of ϵ-complexity to this case.

Let a number $0 < S < 1$ be chosen. We discard a part of the initial n values of the function so that after discarding $[Sn]$ values will retain. The sample points should be discarded so that the remaining sample points are approximately evenly spaced; here and below symbol $[a]$ means the integer part of a. Thus, S is the fraction (of the total n) of sample points that remain after discarding.

Denote by $\epsilon(n, \mathcal{F}, S) \overset{\text{def}}{=} \epsilon(\cdot)$ *minimal* recovery error for function $x(t)$ (now it is a time series) with the remaining $[Sn]$ time points. The recovery is performed by all methods of the list $\mathcal{F}$. The recovery error can be measured in any finite dimensional standard norm.

We set

$$\log\rho = \log\frac{\epsilon}{R} + \log\epsilon \tag{1}$$

Let us present the main result of the theory of ϵ-complexity for the case when a continuous function is given by its restriction on a fixed uniform grid:

For any Hölder function from an everywhere dense set, given by its restriction on a fixed uniform grid, the following basic relation holds

$$\log\rho \approx A(n) + B(n)\log S. \tag{2}$$

The richer the set of approximation methods $\mathcal{F}$ and the greater the number of sample points n on a fixed time interval, the more accurate recovery.

The coefficients A, B in (2) will be called the *ϵ-complexity coefficients*. The complexity coefficients have nothing to do with the time series generation mechanism (i.e., it's model). Therefore, the methods which employ the ϵ-complexity will be automatically model-free.

2.3. Segmentation

Let $X = \{x(t)\}_{t=1}^{N}$ be a time series with unknown change points $t_i, i = 2, \dots, k$ (it is not known whether there are such moments). The mechanisms of generation of a time series are also unknown and can be *stochastic, deterministic or mixed.*

Intervals $[t_i, t_{i+1}],\ t_1 = 1, \dots, t_{k+1} = N$, on which the generation mechanism does not change, we call *homogeneous*. We assume that the homogeneity intervals are sufficiently long.

The key assumption of the proposed segmentation methodology is the following *hypothesis*: on the *i*th homogeneity interval $[t_i, t_{i+1}]$ of the series X for $t_i \le t,\ (t+n) < t_{i+1}$ the two-dimensional vector of ϵ-complexity coefficients satisfies the relation

$$\mathbb{R}_t = \mathbb{R}_i + \xi_t^i,$$

where ${\xi_t}^i$ is a two-dimensional random process with zero expectation. Here $n \ll \min_i(t_{i+1} - t_i))$ is the size of the sliding window in which the ϵ-complexity coefficients of the series are calculated.

Thus, if the above hypothesis is true, the problem of time series segmentation is reduced to the problem of detecting the "disorder" in the mean value of the diagnostic vector sequence $\mathbb{R}_t$.

To solve this last problem, we use a family of statistics, the first version of which was proposed in 1979 in Brodsky BS, Darkhovsky BE (1979). In Brodsky B, Darkhovsky B (2013) it is shown that under broad assumptions about random sequences $\{\xi_t^i\}$, these statistics lead to asymptotically (for $N \to \infty$) minimax estimates of the moments of change in the generation mechanism.

A more detailed study of the procedure with a description of computational experiments can be found in Darkhovsky B and Piryatinska A (2019).

2.3.1. Classification

We consider the classification problem of relatively short time series into two groups. In practice, the following problem arises. Assume you observe two time series with unknown generating mechanisms . We would like to identify whether these series are generated by the same mechanism or not. At the same time, it is not known a priori that the series are described by GARCH models, their mechanisms can be much more complicated. To solve such a problem, it is not necessary to look for series models. It can be done using our model-free technology. Our technology makes it possible not to look for any models, and this is essential for practical work.

Here we propose to use the ϵ-complexity coefficients of the time series itself and its finite differences (the latter are analogs of derivatives) as a feature space. After that the standard classification algorithms such as Random forest (Breiman L (2001)) or support vector machine (SVM), Cortes C, Vapnik V (1995) can be applied to get a classification. Such a feature space does not use any series models. For detail description see Darkhovsky B, Piryatinska A (2017) . In the above work we were able to classify, with an accuracy of the order of 85%, two groups of subjects - healthy and with signs of schizophrenia - according to the 16-channel EEG records, in a feature space of dimension 4, composed of complexity coefficients.

3. Simulations

An autoregressive moving average model (ARMA) is a model for forecasting of stationary time series. It assumes that the time series is generated by the white noise by passing white noise through a recursive and through a nonrecursive linear filter, consecutively. The ARMA model is a combination of an autoregressive (AR) model and a moving average (MA) model-stealth. Autoregressive Conditional Heteroskedasticity (ARCH) is a method that explicitly models the change in variance over time in a time series.

In this paper, we consider Generalized Conditional Heteroscedasticity (GARCH) models. The generalized autoregressive conditional heteroskedasticity (GARCH) process is an approach to estimating the volatility of financial markets. In particular, we focus on GARCH(1,1) models.

A general GARCH(p,q) model is defined in Bollerslev T (1986).

Let us specify a GARCH(1,1) model:

$$u_t = \sigma_t z_t$$

$$\sigma_t^2 = \omega + \alpha u_{t-1}^2 + \beta \sigma_{t-1}^2$$

here $z_t \backsim N(0,1)$ are independent identically distributed random variables.

Let us notice that GARCH models have observable component u_t and unobservable component σ_t , and we will work only with the observable component.

3.1. *Demonstration of basic relation*

Firstly let us demonstrate the basic relation (2) on two examples of GARCH(1,1) models.
Example 1: GARCH(1,1) with coefficients $\omega = 0.0002, \beta = 0.15$, $\alpha = 0.3$.
Example 2: GARCH(1,1) with coefficients $\omega = 0.1, \beta = 0.7$, $\alpha = 0.25$.

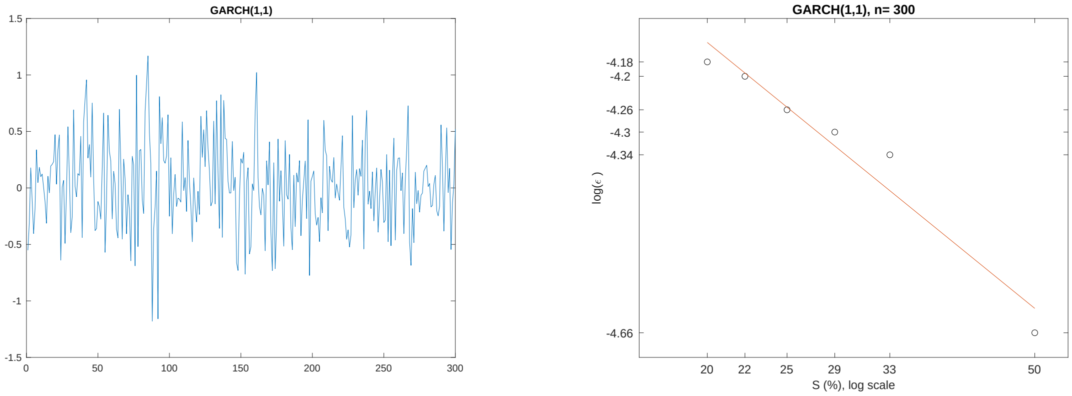

Figure 1. Example 1. Example, GARCH (1,1) model, coefficients: $\omega = 0.0002, \beta = 0.15$, $\alpha = 0.3$. Left: simulated time series, Right: Linear dependence, equation (2) . Circles correspond to the points $(\log(S_i), \log(\epsilon_i))$, and the line is the linear regression line

Figure 2 left provides the original time series of the GARCH(1,1) with coefficients $\omega = 0.1, \beta = 0.7$, $\alpha = 0.25$. as well as demonstration of the relationship (2) (right plot). In the right plot Circles correspond to the points $(\log(S_i), \log(\epsilon_i))$, and the linear regression line.

3.2. *Demonstration of our change-points algorithm*

We consider the time series obtained by simulation of the non-homogeneous GARCH(1,1) model trajectories. In particular, we simulated five homogeneous segments of the GARCH(1,1) of length 30000. The coefficients are presented in the following Table 1. We concatenated these segments. After concatenation, we separate each time series into non-overlapping segments of length 300. For each segment, the ϵ-complexity coefficients are calculated. As a result, we generate two-dimensional diagnostic sequences. For each component of a diagnostic sequence, we will apply the 3-step non-parametric change-point detection procedure of Brodsky and Darkhovsky (see, Darkhovsky B and Piryatinska A (2019)). If we observe a change in at least one component of the diagnostic sequence, we will assume that the change occurred. To ensure the stability of the results, we perform 1000 replications of each numerical experiment.

Figure 3 provides an example of such simulations. The left plot shows the example of the simulated time series (the blue line). The right plot provides the diagnostic sequence, solid black line, and blue line corresponds to the local means for homogeneous increments. The vertical red lines correspond to the true change points.

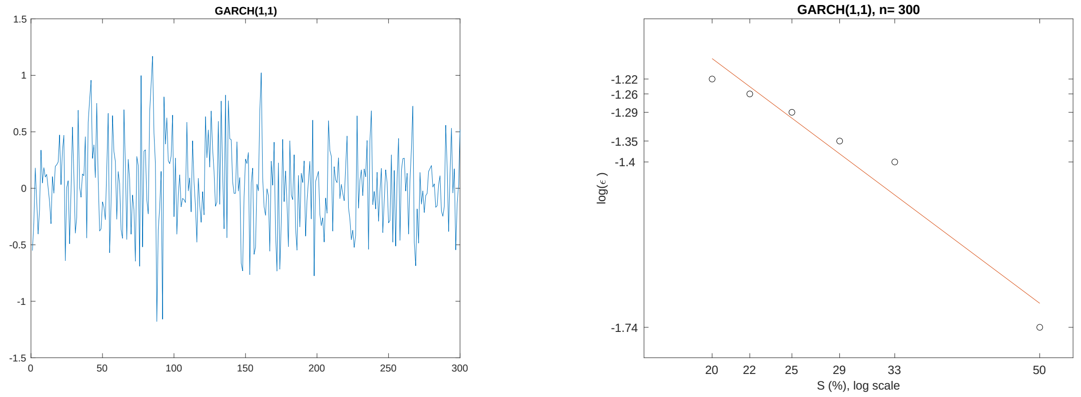

Figure 2. Example 1. Example, GARCH (1,1) model, coefficients: $\omega = 0.1, \beta = 0.7, \alpha = 0.25$. Left: simulated time series, Right: Linear dependence, equation (2) . Circles correspond to the points $(\log(S_i), \log(\epsilon_i))$, and the line is the linear regression line.

The percentage of the number of detected points is presented in Table 2.

Table 1. Coefficients of the GARCH(1,1) for the increments of the non-homogeneous time series

	ω	α	β
Segment 1	0.0002	0.36	0.15
Segment 2	0.0002	0.6	0.31
Segment 3	0.0002	0.6	0.1
Segment 4	0.001	0.6	0.1
Segment 5	0.001	0.01	0.1

Table 2. The percentage of the number of detected points

♯ of detected points	coeff $A(t)$	coeff $B(t)$
1	0%	6%
2	0%	65.9%
3	0%	1.9%
4	78.1%%	0.4%
5	18%	0%
6	3.5%	0%

One can see that only coefficient A is useful to perform a segmentation of non-homogeneous time

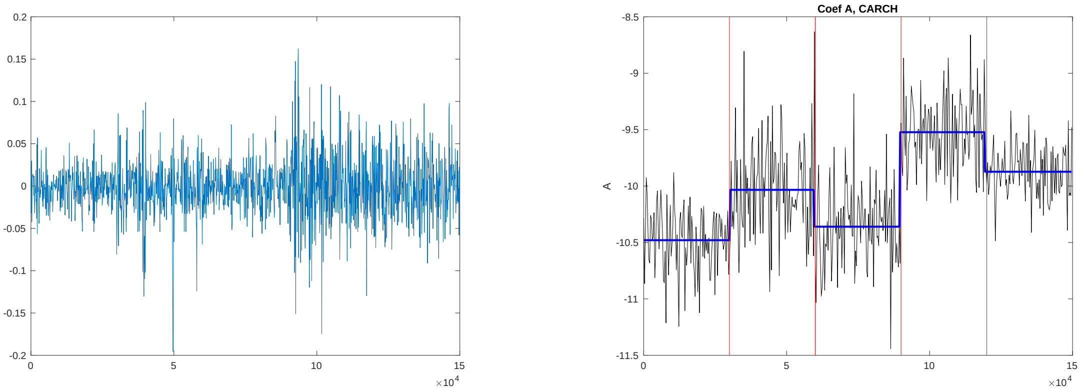

Figure 3. Left: simulated time series, Right: Diagnostic sequence $A(t)$

series with segments from GARCH (1,1) models. Below we present results only for the coefficient A.

The true positive rate and bootstrap confidence intervals are given in the Table 3.

Table 3. The true positive rate and bootstrap confidence intervals

change-point	true positive A	CI, coeff A
1	97%	(29100,32100)
2	91.1%	(55582, 63900
3	96%	(89250, 90600)
4	74.8 %	(115800, 122400)

Let us notice that the first change point corresponds to the change in two parameters of the GARCH (1,1) model, α and β. Our detection rate is 97%. The second point corresponds to the change in the parameter β, and we detected this change 91 % of the time. The third point corresponds to the mean value ω change, and the detection rate was 96%. The fourth point corresponds to the change in the parameter α, and the detection rate is 74.8%. We obtained the best detection rate when both coefficients α and β had been changed. We observe that only ϵ-complexity coefficient A helped to detect changes in the coefficients of GARCH (1,1) model. Let us emphasize that we used the same approach to detect these disorders and did not use any knowledge about the model.

3.3. Demonstration of our classification approach

We create two groups from GARCH(1,1) processes with the following coefficients.

Example 3.

Group 1: $\omega_1 = 0.001, \alpha = 0.4, \beta = 0.15$

Group 2 a): $\omega_2 = 0.003$ $\alpha = 0.4, \beta = 0.15$

Group 2 b): $\omega_2 = 0.002$ $\alpha = 0.4, \beta = 0.15$

Example 4.
Group 1: $\omega = 0.001, \alpha_1 = 0.015, \beta = 0.15$
Group 2 (a): $\omega = 0.001\ \alpha_2 = 0.5, \beta = 0.15$
Group 2 (b): $\omega = 0.001\ \alpha_2 = 0.3, \beta = 0.15$
Example 5.
Group 1: $\omega = 0.001, \alpha = 0.3, \beta_1 = 0.1$
Group 2 (a): $\omega = 0.001, \alpha = 0.3, \beta_2 = 0.43$
Group 2 (b): $\omega = 0.001, \alpha = 0.3, \beta_2 = 0.3$

Then we simulate 1000 realizations of length 300 for each process. For each realization, we estimated complexity coefficients A and B. Then we plot them on the coordinate plane (A, B). In Figure 4, the left plots corresponds to example 3 (the top one is (a) the bottom one is (b)), the middle plots to example 4, and the right plots corresponds to example 5. Red circles represent the points of the first group. Black points present the points in the second group.

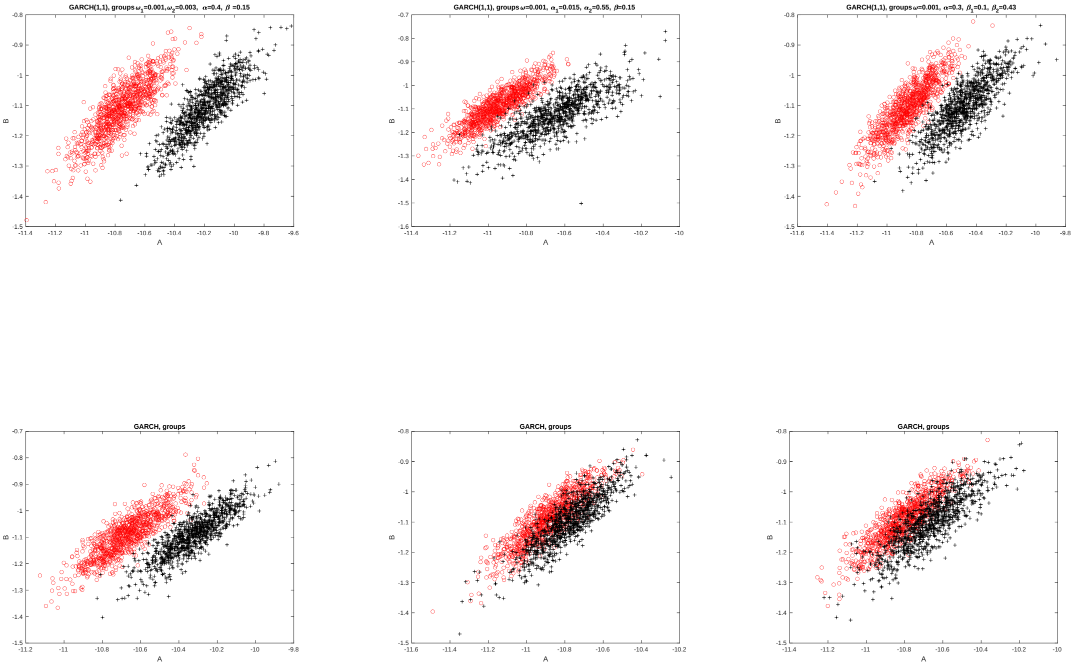

Figure 4. The points of the first group are represented by red circles. The points in the second group are presented by black points. Top left: Example 3 (a), . Bottom left: Example 3 (b). Top middle: Example 4 (a), Bottom middle: Example 4 (b). Top right: Example 5 (a), Bottom right: Example 5 (b)

We applied support vector machine classifier to each example. The classification errors for the 10-fold cross-validation are presented in the Table 4, column 2. We also found ϵ–complexity coefficients for the first order difference of the time series and employed SVM to 4-dimensional feature space of the ϵ-complexity coefficients of time series and its first order differences. The classification errors for the 10-fold cross-validation are presented in Table 4.

Table 4. The classification errors for the 10-fold cross-validation

Examples	Classification error	Classification error
	2-D features space	4-D features space
3 a)	0%	0%
3 b)	1.35%	0.35%
4 a)	1.2%	1.1%
4 b)	20.3.67%	16.7%
5 a)	0.85%	0.8%
5 b)	17.4%	12.2%

One can see, that in Example 3 we have a perfect separation. In the Example 4 we got 98.8% correct classification if we use only ϵ-complexity coefficients A and B. Let us notice that in the Example 3 we changed the coefficient ω in the GARCH model.

The adding of the ϵ-complexity coefficients of first differences improve the classification accuracy only on 0.1%. In the Example 4 we got 99.15% accuracy for the SVM classifier if we use only ϵ-complexity coefficients. It is improved by 0.05% if the ϵ-complexity coefficients of first differences are added to the feature space.

4. Applications to real data

4.1. Segmentation of Stock price data

We tested our approach on the high frequency Microsoft (MSFT) stock price data during the Flash crash. The flash crash occurred on May 6, 2010. The flash crash started at 2:36 pm and lasted approximately 36 minutes. The stock prices data were provided by nanex.net. In this example, we consider data for five days from May 3, 2010 until May 7, 2010. This recordings are are collected every five seconds. In this analysis, the median between bid, ask, last trade price values at each time point was used. Figure 5 (left) presents MSFT stock price from May 3, 2010 until May 10 of 2010. The vertical red line separates days. Figure 5 (right) presents corresponding log-returns data (difference of the log of the original data).

We considered the disjoint intervals of length 120 (it corresponds to 10 minutes intervals). For each interval we calculated the ϵ-complexity coefficients and used them as diagnostic sequences for the non-parametric change-point detection procedure discussed above. The results are presented in Figure 6. The left plot corresponds to the diagnostic sequence of the coefficients A_t and the right plot corresponds to the diagnostic sequences of the coefficients B_t. We observe the diagnostic sequence of parameters A_t detecting three changes in generating mechanisms.

We assume that that the intervals between points of disorders are homogeneous. Let us notice that the third change point in the diagnostic sequence of the coefficient A_t is very closed to the change point

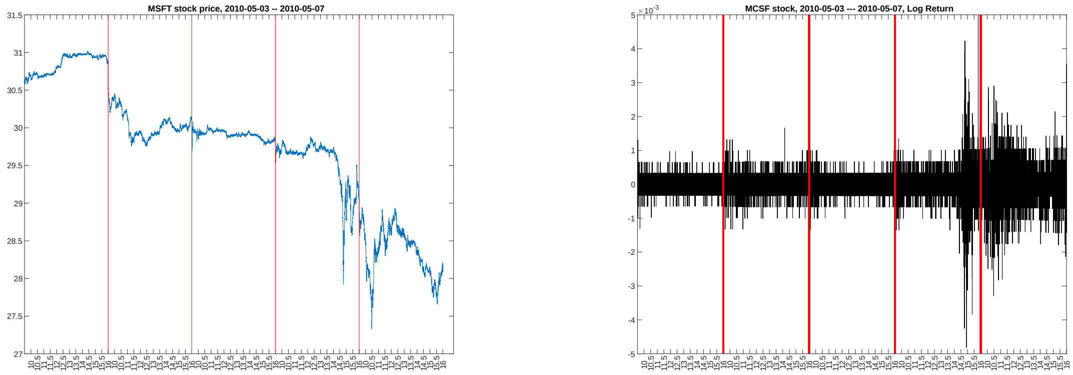

Figure 5. Left: MSFT stock price, Right Log returns of MSFT stock price, May 3, 2010–May 7, 2010. The vertical red line separates days.

of the coefficients B_t diagnostic sequence. Therefore we used the moments of changes in the generated mechanism of the coefficients A_t as point of the separation into homogeneous intervals.

For each homogeneous interval we model data using GARCH models for the log return data. The MATLAB build in function was used for the parameter estimation. We observe that for the first and the last intervals GARCH models are not applicable. The p-values for coefficients α, β equal one. For the second interval the ARCH(1) model is applicable. The estimated coefficient $\alpha = 0.01$ and the corresponding p-value is $2.8 \cdot 10^{-5}$. For the third interval GARCH(2,1) model is applicable. The corresponding estimated coefficients are $\alpha_1 = 0.245$ (p-val$\approx$ 0), $\alpha_2 = 0.135$ (p-val$= 2.5 \cdot 10^{-10}$), $\beta = 0.232$ (p-val$\approx$ 0).

4.2. *Application to classification problem*

In this subsection, we use daily (at the end of the day) stock market prices for two companies NVIDIA and AMD, during five years period (from 02/08/2013 until 02/07/18). This data can be found at `chttps://github.com/CNuge/kaggle-code/tree/master/stock_data`. Firstly we found the log returns of these data.

After that, we model the data of log returns using GARCH models. The GARCH(1,1) model turns out to be the best model for these two data sets. We estimated coefficients of GARCH(1,1) using MATLAB function and received the following result:

NVIDIA: $\omega_1 = 0.0003, \alpha_1 = 0.2872, \beta_1 = 0.1533$

AMD: $\omega_2 = 0.0004, \alpha_2 = 0.3836, \beta_2 = 0.4455$

After that, we simulated 1000 replications of GARCH(1,1) models with these coefficients, calculated complexity coefficients A and B for each process replication, and plotted them on (A, B) coordinate plane. We used SVM to classify the data. The classification error for the 10-fold cross-validation is 0.1%.

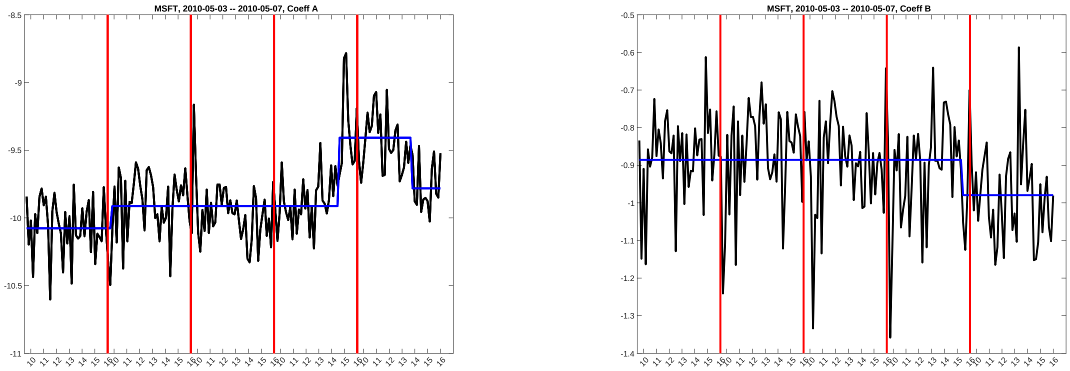

Figure 6. Left: Coefficient A, Right: Coefficient B. Red vertical line separate days, horizontal blue lines correspond to the local means between detected disorders

Figure 7 (left) provides plot of the stock price data of NVIDIA (solid blue line) and AMD (dashed black line). The middle plot of Figure 7 shows the corresponding log-returns data for both companies. Figure 7 (right) provides values of the ϵ-complexity coefficients for both groups of time series The blue points corresponds to the coefficients (A, B) of time series simulated from the GARCH model which fits NVIDIA data, and the black points correspond to the coefficients (A, B) of time series simulated from the GARCH model which fits AMD data.

We observe the perfect separation between these two groups. We employed the support vector machine classifier from a MATLAB. Classification error for the 10-fold cross validation is 0.25%. We also calculated first order differences of the time series and found its ϵ-complexity coefficients. If we use support vector machine classifier than the classification error for the 10-fold cross-validation is 0.1%.

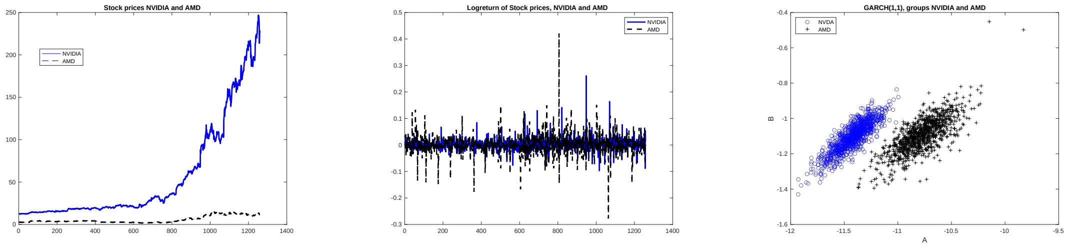

Figure 7. Left: Stock price data: NVIDIA (solid blue line) and AMD (dashed black line). Middle: Log-returns data for NVIDIA (blue line) and AMD (black line) stock prices. Right: The ϵ-complexity coefficients (A,B): for both groups of time series. The blue points correspond to the coefficients (A,B) of time series simulated from the GARCH model which fits NVIDIA data, and the black points correspond to the coefficients (A, B) of time series simulated from the GARCH model which fits AMD data.

5. Conclusions

In this paper, we propose model-free technologies for a retrospective segmentation of time series into the homogeneous increments and binary classification of the short time series. This approach was demonstrated on GARCH(1,1) models.

To perform a segmentation of long time series we employed the ϵ-complexity coefficients calculated for the disjoint segments (the sliding window also can be used) as a diagnostic sequence. The change in the mean of the ϵ-complexity coefficients reflects the change in the generating mechanism of time series. Subsequently, the change point detection algorithm of Brodsky B, Darkhovsky B (2013) is employed to detect change points.

We apply this approach to GARCH (1,1) models and detect changes in coefficients of the model without any knowledge about the model.

In our simulation study, we demonstrated the applicability of our approach for the data simulated using GARCH(1,1) models with different coefficients and then concatenated. We apply this approach to the high frequency stock market data.

Our approach was also applied to classify the short time series simulated by two GARCH(1,1) models with different coefficients. We applied it to the real stock price data.

A key feature of our approach is independence from the model of the observed process. Let us emphasize that most of the methods used in the literature (maximum likelihood, estimation of parameters in a window with subsequent monitoring of estimates, etc.) cannot be used in principle without knowledge of the model. The independence from the process model is achieved by utilizing our theory of the ϵ-complexity of continuous vector functions, which is consistent with the general idea of A.N. Kolmogorov on how it is expedient to evaluate the complexity of an object.

However, the limitation of our method is that it requires a relatively long sequence of time series. To calculate the complexity coefficient, we used segments of length 300. In the example with stock market data the length of 100-120 were sufficient. Our approach is beneficial for the high-frequency stock market data, when the time series are very long.

Our numerous experiments with different data types show that it is sufficient to use 100 points in most cases. However, for GARCH(1,1), model 300 provides more stable results and helps us detect better change points. To ensure that the limiting distribution for statistics from our 3-step algorithm will start to work, the diagnostic sequence for each homogeneous increment should be several dozen. In our examples, we used non-overlapping windows.

Note that when a non-overlapping window intersects any change-point, the mathematical expectation of the sequence of complexity coefficient varies according to a particular transient process from one constant to another. However, when the window size is much less than the length of any homogeneity interval, such a transient process does not significantly affect the change point estimates.

Acknowledgment

We would like to express our gratitude to the anonymous referees, who read the paper very carefully and whose comments were especially important to improve the paper. Boris Darkhovsky's research was funded by RFFI No. 20-07-00221

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

Abanda A, Mori U, Lozano JA (2019) A review on distance based time series classification. *Data Min Knowl Disc* 33: 378-412. https://doi.org/10.1007/s10618-018-0596-4

Andreou E, Ghysels E (2009) Structural breaks in financial time series. *Hand financ time ser* 839-870. https://doi.org/10.1007/978-3-540-71297-8_37

Aue A, Horváth L (2013). Structural breaks in time series. *J Time Ser Anal* 34: 1–16. https://doi.org/10.1111/j.1467-9892.2012.00819.x

Bagnall A, Janacek G (2014) A run length transformation for discriminating between autoregressive time series. *J Classif* 31: 274–295. https://doi.org/10.1007/s00357-013-9135-6

Berkes I, Horváth L, Kokoszka P (2004) Testing for parameter constancy in GARCH (p, q) models. *Stat probabil lett* 70: 263–273. https://doi.org/10.1016/j.spl.2004.10.010

Bollerslev T (1986) Generalized Autoregressive Conditional Heteroskedasticity. *J Econometrics* 31: 307–327. https://doi.org/10.1016/0304-4076(86)90063-1

Breiman L (2001) Random Forests. *Mach Learn* 45: 5–32. https://doi.org/10.1023/A:1010933404324

Brodsky BS, Darkhovsky BE (1979) Identification of the change time in the random sequence.

Brodsky E, Darkhovsky B (2013) *Nonparametric methods in change point problems.* Springer Science Business Medi, New York.

Brooks C (2014) *Introductory Econometrics for Finance (3rd ed.)* Cambridge: Cambridge University Press.

Chao L, Zhipeng J, Yuanjie Z (2019) A novel reconstructed training-set SVM with roulette cooperative coevolution for financial time series classification. *Expert Syst Appl* 123: 283–298. https://doi.org/10.1016/j.eswa.2019.01.022

Chu C (1995) Detecting parameter shift in GARCH models. *Economet Rev* 14: 241–66. https://doi.org/10.1080/07474939508800318

Cortes C, Vapnik V (1995) Support-vector networks. *Mach learn* 20: 273-297. https://doi.org/10.1007/BF00994018

Darkhovsky B (2020) On the complexity and dimension of continuous finite-dimensional maps. *Theory of Probab Its Appl* 65: 375–387. https://doi.org/10.1137/S0040585X97T990010

Darkhovsky B, Piryatinska A (2014) New approach to the segmentation problem for time series of arbitrary nature. *P Steklov I Math* 287: 54–67. https://doi.org/10.1134/S0081543814080045

Darkhovsky B, Piryatinska A, Kaplan A (2017) Binary classification of multichannel-eeg records based on the ϵ-complexity of continuous vector functions. *Comput meth prog bio* 152: 131–139. https://doi.org/10.1016/j.cmpb.2017.09.001

Darkhovsky B, Piryatinska A (2018) Model-free offline change-point detection in multidimensional time series of arbitrary nature via ϵ-complexity: Simulations and applications. *Appl Stoch Model Bus* 34: 633–644. https://doi.org/10.1002/asmb.2303

Darkhovsky B, Piryatinska A (2019) Detection of Changes in Binary Sequences. In Workshop on Stochastic Models, Statistics and their Application 157–176. Springer, Cham. https://doi.org/10.1007/978-3-030-28665-1_12

Engle R (1982) Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica* 50: 987–1007. https://doi.org/10.2307/1912773.

Faouzi J (2022) Time Series Classification: A review of Algorithms and Implementations. *Mach Learn* https://hal.inria.fr/hal-03558165

Kokoszka P, Leipus R (2000) Change-point estimation in ARCH models. *Bernoulli* 6: 513–539. https://doi.org/10.2307/3318673

Kolmogorov AN (1983) Combinatorial foundations of information theory and the calculus of probabilities. *Russ math surv* 38: 29–40.

Lamoureux CG, Lastrapes WD (1990) Persistence in variance, structural change, and the GARCH model. *J Bus Econ Stat* 8: 225–234. https://doi.org/10.1080/07350015.1990.10509794

Li Q, Wang L, Qiu F (2015) Detecting the structural breaks in GARCH models based on Bayesian method: The case of China share index rate of return. *J Syst Sci Inf* 3: 321–333. https://doi.org/10.1515/JSSI-2015-0321

Majumdar S, Laha AK (2020) Clustering and classification of time series using topological data analysis with applications to finance. *Expert Syst Appl* 162, 113868. https://doi.org/10.1016/j.eswa.2020.113868

Smith D (2008) Testing for structural breaks in GARCH models. *Appl Financ Econ* 18: 845–862, https://doi.org/10.1080/09603100701262800.

Song J, Kang J (2018) Parameter change tests for ARMA-GARCH models. *Comput Stat Data Anal* 121: 41–56. https://doi.org/10.1016/j.csda.2017.12.002

Susto GA, Cenedese A, Terzi M (2018) Time-series classification methods: Review and applications to power systems data. *Big data appl power syst* 179–220. https://doi.org/10.1016/B978-0-12-811968-6.00009-7

Truong C, Oudre L, Vayatis N (2019) Greedy Kernel Change-Point Detection. *IEEE T Signal Proces* 67: 6204–6214. https://doi.org/10.1109/TSP.2019.2953670

Truong C, Oudre L, Vayatis N (2020) Selective review of offline change point detection methods. *Signal Process* 167: 107299. https://doi.org/10.1016/j.sigpro.2019.107299

Xing Z, Pe J, Keogh E (2010) A brief survey on sequence classification. *ACM Sigkdd Explor Newsl* 12: 40–48. https://doi.org/10.1145/1882471.1882478

AIMS Press