

Article

Retrospective Change-Points Detection for Multidimensional Time Series of Arbitrary Nature: Model-Free Technology Based on the ϵ -Complexity Theory

Alexandra Piryatinska ^{1,*},† and Boris Darkhovsky ^{2,†} 

¹ Department of Mathematics, San Francisco State University, 1600 Holloway Ave., San Francisco, CA 94132, USA

² Institute for Systems Analysis, FRC CSC RAS 9 Pr. 60-Letiya Oktyabrya, 117312 Moscow, Russia; darbor2004@mail.ru

* Correspondence: alpiryat@sfsu.edu

† These authors contributed equally to this work.

Abstract: We consider a retrospective change-point detection problem for multidimensional time series of arbitrary nature (in particular, panel data). Change-points are the moments at which the changes in generating mechanism occur. Our method is based on the new theory of ϵ -complexity of individual continuous vector functions and is model-free. We present simulation results confirming the effectiveness of the method.

Keywords: ϵ -complexity; change-point detection; model-free segmentation



Citation: Piryatinska, A.; Darkhovsky, B. Retrospective Change-Points Detection for Multidimensional Time Series of Arbitrary Nature: Model-Free Technology Based on the ϵ -Complexity Theory. *Entropy* **2021**, *23*, 1626. <https://doi.org/10.3390/e23121626>

Academic Editor: Jiri Petrzelá

Received: 22 October 2021

Accepted: 24 November 2021

Published: 2 December 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In retrospective studies, all observations are collected a priori. A retrospective analysis of multivariate time series begins by checking their homogeneity. We call data homogeneous if the same mechanism generates them. When the homogeneity assumption is violated (i.e., the data generation mechanism changes during their collection), we must perform segmentation of the data into homogeneous increments. In cases of *stochastic* data generating mechanism, the segmentation problem is well-known as the “change-point detection” problem in the retrospective setting. A vast amount of literature is devoted to the change-point detection for stochastic processes in both “off-line” and “on-line” formulation, see e.g., [1–3]. In this case, the change-points are the moments of changes in their probabilistic characteristics.

Segmentation problems arise in econometrics and financial mathematics. In these areas, the change-points are called structural breaks. In the last 20 years, detecting structural breaks in the so-called panel data has attracted the attention of many researchers. Panel data is data that contains observations about different cross-sections across time. Groups that may make up panel data series include countries, firms, individuals, or demographic groups. The primary difference between panel data models and time series models is that panel data models allow for heterogeneity across groups and introduce individual-specific effects. Panel data are usually high-dimensional (have hundreds of components). In the literature, there are many interesting publications on such problems (see e.g., [4–7]). In such studies, only the stochastic models were used to model panel data.

However, in many applications, the data are more complex and cannot always be modeled as stochastic processes. There is a large class of complex systems that, being *deterministic, exhibit stochastic behavior*. Such systems are called chaotic. The existing mathematical theory of chaotic systems (see, for example, [8]) suggests that they should be described by an *unchanging equation of evolution*. Meanwhile, in real chaotic systems, changes in parameters can occur, resulting in conditions in which the system can pass from one regime to another. This is how the *multifractality* phenomenon can arise, which is

currently receiving much attention in the literature (see, for example, [9]). It follows that the problem of checking homogeneity and segmentation is no less critical for the analysis of chaotic systems.

These considerations lead to a broader interpretation of the term “change-point” (or “moments of disorder”). Namely, we mean by this term *the moment of change in the generating mechanism of a multidimensional time series, regardless of what its nature is*.

All currently known methods for solving problems on the change-point detection of *stochastic processes* in one way or another rely on their models (i.e., on the knowledge about the model that generates data). However, information about data generation mechanisms is not always available. A typical example here is the EEG signal, which, according to most experts, is one of the most demanding physical processes to study, and there is no generally accepted model of such a process. The situation is similar in this sense in applications to financial series, some biological problems, etc., where there are no firmly established models of the observed processes. The situation with chaotic systems seems to be even more complicated in this respect. Thus, in many applications, a *vicious circle* situation arises: for adequate data segmentation, it is required to know the model of this data, and the model can be built only after data segmentation into “homogeneous” fragments.

In this paper, we propose a *model-free method* for retrospective detection of multiple change-points in multidimensional time series. This method is based on the theory of the ϵ -complexity of continuous vector-functions. The theory of the ϵ -complexity of continuous functions was developed in our recent works [10,11]. It enables us to develop a model-free method for detection of change-points (i.e., in our terminology, *moments of change in the generating mechanism*) for multivariate time series of arbitrary nature (stochastic, deterministic, or mixed). We demonstrate the effectiveness of the method with the help of simulations. In our simulations, we consider examples of multivariate time series, which are generated by a multidimensional stochastic process with dependent components (vector autoregressive model); multidimensional chaotic deterministic processes, with some dependent components; and the mixed process, which has some stochastic components and some chaotic deterministic components.

The proposed method is a modification of the method published in [12]. In that paper, we presented the method for retrospective detection of change-points in a time series of small dimensions (6 independent components of a vector series). The main differences between this article and [12] are as follows:

- (a) Here, we rely on more exact definitions and formulations of the theory of ϵ -complexity (which did not lead to a change in the basic relations), given in [11]. (The general idea of ϵ -complexity was created about 8 years ago. This idea was quickly implemented in computer algorithms and successfully used to solve problems of disorder detection and classification. However, as is often the case with new ideas, rigorous mathematical formulations took more time. The definitions and results of the theory are given in [11] and, in this paper, we rely on the improved definitions and theory.)
- (b) We made a change in the algorithm for calculating the complexity coefficients (see Section 2 below), which made it possible to detect changes in mean value and variance.
- (c) For simulations of stochastic components, we used a general multidimensional linear model of high dimension (50 components of the vector series, interconnected by linear relations) and investigated the possibility of detecting changes in the matrices of this model and the mean value of the vector series.
- (d) We investigated the effectiveness of our method in the case when abrupt changes did not occur in all components of the time series.

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 [11]. Section 3 describes the method of retrospective detection of changes in the mean value, variance, and parameters of chaotic processes in multidimensional time series of an arbitrary nature. Section 4 shows the results of the simulations. Section 5 provides conclusions.

2. Brief Description of the Results of the Theory of ϵ -Complexity

In this section, we present the results of the theory of the ϵ -complexity of continuous vector functions at the meaningful level. Mathematically rigorous definitions and results are given in [11].

Let a continuous vector function $x(\cdot) = (x_1(\cdot), \dots, x_m(\cdot))$ be defined on a finite time interval. Denote $R_i = \max_t |x_i(t)|, i \in I = (1, \dots, m)$ and we will assume that $R = \min_{i \in I} R_i > 0$.

Without loss of generality, we assume that vector function $x(\cdot)$ is defined on $[0, 1]$. Consider a uniform grid on $[0, 1]$ with some step $1 > h > 0$. We call an arbitrary Borel function that transfers a finite set of discrete vector function values (i.e., m -dimensional vectors, the number of which is determined by the value h) into some bounded vector function on $[0, 1]$ the *method of recovery (approximation) of a continuous vector function* (a uniform metric is introduced in the space of bounded vector functions).

Let us fix an arbitrary countable set of Borel vector functions with values in the space of bounded vector functions depending, respectively, on $1, 2, 3, \dots$ arguments. We call a *list* the union of these countable sets. The list contains a countable set of recovery methods for all $h > 0$.

Let us fix some list \mathbb{F} of recovery methods. Throughout what follows, the symbol \mathcal{F} denotes an arbitrary nonempty subset of \mathbb{F} containing some collection of Borel vector functions from $1, 2, 3, \dots$ arguments.

The sets \mathcal{F} (and, accordingly, the lists \mathbb{F} for $\mathcal{F} = \mathbb{F}$) are *admissible* if they contain methods of approximation by piecewise constant (stepwise) vector functions and power polynomials.

The recovery methods are “physically realizable” if they can be represented as computer programs. Such recovery methods contain a finite set of bounded piecewise continuous vector functions of a finite number of variables with values in the space of bounded vector functions. Note that any finite set of “physically realizable” recovery methods is included in some admissible list.

We set

$$\delta_i^{\mathcal{F}}(h) = \inf_{\hat{x}_{i,h}(\cdot) \in \mathcal{F}} \sup_{t \in [0,1]} |\hat{x}_{i,h}(t) - x_i(t)|, i = 1, 2, \dots, m.$$

Here, the symbol $\hat{x}_{i,h}(\cdot) \in \mathcal{F}$ denotes the estimates of the i -th component of the vector function $x(\cdot)$ by its finite set of values with step h obtained by methods of the family \mathcal{F} . In the case when $\mathcal{F} = \mathbb{F}$, all functions included in \mathbb{F} are used for evaluation.

Lemma 1 (Density lemma). *Let \mathbb{F} be an arbitrary fixed admissible list. The set of continuous vector functions that cannot be precisely reconstructed from a finite number of functions’ values by the methods from the list \mathbb{F} is everywhere dense in the space of all continuous vector functions.*

Vector functions that *cannot be exactly reconstructed* by methods of an arbitrary nonempty admissible subset $\mathcal{F} \subseteq \mathbb{F}$ we call *\mathcal{F} -nontrivial*.

Let \mathbb{F} be a fixed admissible list and $\mathcal{F} \subseteq \mathbb{F}$ be an arbitrary nonempty admissible subset. Let $x(t)$ be \mathcal{F} -nontrivial vector function. For sufficiently small $\epsilon > 0$, put

$$h_i^*(\epsilon, \mathcal{F}) = \begin{cases} \inf\{h \leq 1 : \frac{\delta_i^{\mathcal{F}}(h)}{R_i} > \epsilon\}, & \text{if } x_i(\cdot) \text{ is } \mathcal{F}\text{-nontrivial} \\ 1, & \text{in opposite case} \end{cases}$$

Set

$$h_x^*(\epsilon, \mathcal{F}) = \prod_{i=1}^m h_i^*(\epsilon, \mathcal{F})$$

Definition 1. *The (ϵ, \mathcal{F}) -complexity of a continuous vector function $x(\cdot)$ is the value $S_x(\epsilon, \mathcal{F}) = -\log h_x^*(\epsilon, \mathcal{F})$.*

If a vector function is not \mathcal{F} -nontrivial (i.e., it can be reconstructed exactly from a finite number of its values), then we assume that its (ϵ, \mathcal{F}) -complexity is zero (see definition above). Thus, the Density Lemma implies that “almost all” continuous vector functions have nonzero (ϵ, \mathcal{F}) -complexity for any $\mathcal{F} \subseteq \mathbb{F}$ for an arbitrary fixed admissible list \mathbb{F} .

Note that $h_i^*(\epsilon, \mathcal{F}) > 0$ for $\epsilon > 0$ and $\lim_{\epsilon \rightarrow 0} h_i^*(\epsilon, \mathcal{F}) = 0$ if x_i is \mathcal{F} -nontrivial. On the other hand, $\lim_{h \rightarrow 0} \max_i \delta_i^{\mathcal{F}}(h) = 0$. Therefore, for any (sufficiently small) $\epsilon > 0$, there exists $\eta(\epsilon) > 0$, $\eta(\epsilon) \rightarrow 0$ for $\epsilon \rightarrow 0$ such that $\max_i \delta_i^{\mathcal{F}}(h_x^*(\cdot)) \leq \eta(\epsilon)$.

Considering that $1/h_x^*(\epsilon, \mathcal{F})$ is an estimate of the number of values of a vector function, we obtain that the (ϵ, \mathcal{F}) -complexity is (logarithm) of the number of its values that are required for its reconstruction by methods of the family \mathcal{F} with a relative error of at most $R^{-1}\eta(\epsilon)$. In other words, we can say that this is the *shortest description of the vector function* by these methods with a given precision. In this sense, our definition is consistent with the main idea of A.N. Kolmogorov that the complexity of an object should be measured by the length of its shortest description.

In most modern 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 vector function on some uniform grid*, we can extend the theory of ϵ -complexity to this case.

Let the number $0 < S < 1$ be chosen. Let us discard some part of the initial n values of the vector function so that after discarding $[Sn]$, values will retain (discarding the sample points should be done in such a way that the remaining sample points are approximately evenly spaced). Thus, S is the fraction (of the total n) of sample points that remain after discarding.

Denote by $\epsilon_i(n, \mathcal{F}, S) \stackrel{\text{def}}{=} \epsilon_i(\cdot)$ *minimal* (by all methods of the collection \mathcal{F}) recovery error for the i -th components of the vector function $x(\cdot)$ (now it is a multidimensional vector time series) by the remaining $[Sn]$ time points. The recovery error can be measured in any finite dimensional standard norm.

We set

$$\log \rho = \sum_i \left(\log \frac{\epsilon_i}{R_i} + \log \epsilon_i \right) \tag{1}$$

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

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

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

The richer set of approximation methods \mathcal{F} , and the greater the number of function values n on a fixed time interval, the more accurate the recovery is. (In our paper [13], relation (2) was given for the case when the sum in relationship (1) contained only the first term. However, the general theory of ϵ -complexity implies that the addition of the second term in (1) does not fundamentally change relation (2). The need to introduce the second term in (1) is caused by the desire to capture changes in *the mean and the variance*. Without this term, such changes may not be detected using the complexity coefficients).

The coefficients $A(n), B(n)$ in (2) will be called the *ϵ -complexity coefficients*. The complexity coefficients have nothing to do with the time series generation mechanism (i.e., the model that generates them). Therefore, any method that utilizes these coefficients will be automatically model-free. *The method for detecting change-points in a multidimensional time series of an arbitrary nature, described below, is based on the ϵ -complexity coefficients.*

3. Method for Detection of Changes in Generating Mechanism in Multidimensional Time Series of Arbitrary Nature

The main idea of our methodology for retrospective detection of change-points in multidimensional time series of arbitrary nature is as follows.

Let $X = \{x(t)\}_{t=1}^N$ be a time series with unknown moments of change in the generation mechanism (MCGM) $t_i, i = 2, \dots, k$ (such moments may not be present). We emphasize that the mechanisms for generating the series *are unknown and can be stochastic, deterministic, or mixed*.

Segments of the series $[t_i, t_{i+1}], t_1 = 1, t_{k+1} = N$, which are generated by the same mechanism, we call *homogeneous* and assume that the homogeneity segments are sufficiently long.

As shown in Section 2, the ϵ -complexity of a segment is determined by the parameters $\mathbb{R} = (A, B)$. Notice that in relationship (2), A and B depend on n ; further, the windows size n will be fixed, and therefore, $A(n) = A$ and $B(n) = B$.

Let us choose a window of size n (it is assumed that $n \ll \min_i(t_{i+1} - t_i)$) and for each segment of the series $x(t), t \in [jn + 1, (j + 1)n], j = 0, 1, \dots, [N/n]$, we will calculate the complexity coefficients $\mathbb{R}(j + 1)$. As a result, we obtain a new *diagnostic* vector sequence $\{\mathbb{R}(j)\}_{j=1}^{j=[N/n]}$.

The key idea of the proposed method is the following *hypothesis*: on the i -th homogeneity segment $[t_i, t_{i+1}]$ of the time series X for $t_i \leq t, (t + n) < t_{i+1}$ (and for corresponding intervals of the diagnostic sequence), the complexity coefficients satisfy the relation

$$\mathbb{R}(j) = \mathbb{R}_i + \zeta^i(j),$$

where $\zeta^i(j)$ is a random vector sequence with zero mathematical expectation.

Note that when the moving window crosses any moment of the MCGM (if our hypothesis is true), the mathematical expectation of the sequence \mathbb{R} changes according to some transient process from one constant to another. However, since, by assumption, the window size is significantly less than the length of any homogeneity segment, such a transient will not significantly affect the estimates of the MCGM.

Thus, if the given hypothesis is valid, the problem of time series segmentation is reduced to the change-point detection problem with the change in the mean values in the diagnostic vector sequence $\mathbb{R}(j)$.

To detect change-points in diagnostic sequences, we use a 3-step procedure (introduced in [2]) based on the family of statistics

$$Y(s, \delta) = \left((\mathbb{N} - s)s/\mathbb{N}^2 \right)^\delta \left(s^{-1} \sum_{k=1}^s z(k) - (\mathbb{N} - s)^{-1} \sum_{k=n+1}^{\mathbb{N}} z(k) \right),$$

where $0 \leq \delta \leq 1, 1 \leq s \leq \mathbb{N} - 1, \mathbb{N} = [N/n], Z = \{z(k)\}_{k=1}^{\mathbb{N}}$ —implementation of the components of the diagnostic sequence $\mathbb{R}(j)$.

The first version of this family of statistics was proposed in [14]; a short description of the 3-step detection procedure can be found in [13].

It can be shown (see [2] for details) that under broad assumptions about random sequences $\{\zeta^i(j)\}$, the statistic leads to asymptotically (for $N \rightarrow \infty$) minimax estimates for the moment of change in the generation mechanism.

So, our method for detecting MCGM in a multidimensional time series is as follows:

1. Choose the size of the disjoint intervals or sliding window for the considered time series.
2. Calculate complexity coefficients for each window. For this purpose, the parameter S in (2) is assigned different values S_1, \dots, S_k ; for each value of $S_j, j = 1, \dots, k$, the value $\log \rho_j$ is determined (in this case, averaging over all possible locations of the row counts remaining after discarding) and then using the set of pairs $(\log \rho_j, \log S_j)$, the complexity coefficients A, B for the window under consideration are calculated using the standard least squares method. The scheme of these calculations is described in detail in [12]. It is necessary to take into account the replacement of the error appearing there by the value ρ from (1).
3. The above 3-step change-point detection procedure is applied to each component of the sequence of complexity coefficients.

4. We combine detected change-points from both components of the complexity coefficients sequence. As a result, we obtain the estimates of MCGM.

4. Simulations

In this section, we present our simulations, which demonstrate the performance of our method.

4.1. Stochastic and Deterministic Processes Used in the Simulations

Lets us first describe the processes that we employed in our simulation study.

The *stochastic process* we utilize here is vector autoregressive process of order p (denoted by $VAR(p)$). It is given as follows.

$$x_t = \mu + \Theta_1 x_{t-1} + \dots + \Theta_p x_{t-p} + u_t, t = 0, \pm 1, \pm 2, \dots, \tag{3}$$

where x_t, u_t , and μ are $(K \times 1)$ vectors and Θ_i are $(K \times K)$ matrices for each $i = 1, \dots, p$. In addition, the error term u_t is a white noise random vector such that $E(u_t) = 0, E(u_t u_t') = \Sigma_u$, and $E(u_t u_s') = 0$ for $s \neq t$, where Σ_u is a $(K \times K)$ positive definite matrix. Such model is often used to simulate panel data and investigate structural breaks in panel data, see e.g., [15]. This model can be rewritten in a compact form (see e.g., [16]),

$$X_t = \mu + \Theta X_{t-1} + U_t, \quad t = 0, 1, 2, \dots \tag{4}$$

where $X_t = [x_t', x_{t-1}', \dots, x_{t-p+1}']', \mu = [\mu', 0, \dots, 0]', U_t = [u_t', 0', \dots, 0']'$ are $(Kp \times 1)$ vectors and

$$\Theta = \begin{bmatrix} \Theta_1 & \Theta_2 & \dots & \Theta_{p-1} & \Theta_p \\ I_k & 0 & \dots & 0 & 0 \\ 0 & I_k & \dots & 0 & 0 \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & I_k & 0 \end{bmatrix} \tag{5}$$

is a $(Kp \times Kp)$ matrix. The model (5) is stable if and only if $|\lambda_{\max}(\Theta)| < 1$, where $|\lambda_{\max}(\Theta)|$ denotes the largest absolute value of the eigenvalues of the matrix Θ .

Using this model, we simulate multivariate time series with dependent components. Our segments of multivariate time series have either different variance covariance matrices Σ_{u_1} and Σ_{u_2} or different lag matrices Θ_1 and Θ_2 . The mean values of the components of our processes change too. In our simulations, we report spectral norms of the matrices $\Sigma_{u_i}, \Theta_i, i = 1, 2$.

Let us remind that the spectral norm of a matrix D is the largest singular value of the matrix D , i.e., the square root of the largest eigenvalue of the matrix D^*D , where D^* denotes the conjugate transpose of D :

$$\|D\|_2 = \sqrt{\lambda_{\max}(D^*D)}$$

(see e.g., [17]).

We also consider chaotic deterministic processes in discrete time. The change in generating mechanisms in some of these processes will correspond to the change in the parameters. In another case, we concatenate different chaotic processes, where changes are the points of the concatenation.

All processes that we consider are as follows: $x_t = f(x_{t-1}), t = 1, 2, \dots$. The functions $f(x)$ are described below.

We consider the following maps.

1. The logistic map

$$f(x) = ax(1 - x) \quad x_0 \in (0, 1) \tag{6}$$

The parameter for this process is α . In our simulations, we use $3.85 \leq \alpha \leq 4$. It is well known (see [18]) that under these parameter values, the corresponding processes exhibit chaotic behavior.

2. The quadratic map, see e.g., [19]

$$f(x) = c - x^2, \quad 0 < c \leq 2, \quad x_0 \in (0, 1). \quad (7)$$

3. Process 3

$$f(x) = 1 - |1 - 2x|, \quad x_0 \in (0, 1) \quad (8)$$

4. The Interval map, see e.g., [20]

$$f(x) = 2x(\text{mod } 1), \quad x_0 \in (0, 1). \quad (9)$$

Let us notice that the process 3 given by function (8) and the Interval map given by function (9) does not have parameters that can be changed.

We also consider two-dimensional maps of the following form $z_t = f(z_{t-1})$, where

$$z_t = \begin{pmatrix} x_t \\ y_t \end{pmatrix}.$$

5. The Hénon Map, see [21]

$$f(z) = \begin{pmatrix} 1 - ax^2 + y \\ bx \end{pmatrix} \quad (10)$$

6. The Ikeda map, see [22]

$$f(z) = \begin{pmatrix} 1 + \mu(x \cos \phi(x, y) - y \sin \phi(x, y)) \\ \mu(x \sin \phi(x, y) + y \cos \phi(x, y)) \end{pmatrix} \quad (11)$$

$$\text{Here, } \phi(x, y) = 0.4 - \frac{6}{1+x^2+y^2}$$

4.2. Results of Simulations

In each example, we simulate multidimensional time series. We take into account the fact that in chosen processes, stationary probability distributions are established sufficiently fast. Here, we discard the beginning of the simulated process before such stabilization. We will concatenate three or four homogeneous multidimensional time series. The length of each homogeneous component will be 5000. In some examples, we will change the coefficients in the models. In other examples, we will link different deterministic processes. After concatenation, we will separate each multidimensional time series into non-overlapping segments of length 100. For each segment, the ϵ -complexity coefficients will be calculated. As a result, we generate two-dimensional diagnostic sequences. For each component of a diagnostic sequence, we will apply the 3-step nonparametric change-point detection procedure of Brodsky and Darkhovsky. 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.

Example 1. Stochastic process, VAR(1).

In this example, we consider the VAR(1) process. We choose $K = 50$; as a result, we have 50 dimensional multivariate time series with dependent components. We simulated 5 different segments of length 5000, concatenated them, and obtained the time series of length 25,000 with four change-points (or MCGM). We performed 1000 replications of the experiment.

The description of the segments is provided in Table 1. The first column lists the type of matrices that define the model. The 2nd, 3rd, 4th, and 5th columns correspond to the specification for each segment. In the first row, one can see which model matrices

are the same and which are different for corresponding segments. In the second row, we provided the corresponding norms of the model matrices. In the third row, one can see which variance–covariance matrices are the same and which are different for corresponding segments. In the fourth row, we provided corresponding norms for variance–covariance matrices. The first MCGM corresponds to the change in the mean in half of the components. The second disorder corresponds to the change in model matrix Θ . The third MCGM corresponds to the change in variance–covariance matrix U . The fourth disorder is the change in the mean for all components.

Table 1. Example 1. Description of the segments in simulations.

	Segm 1	Segm 2	Segm 3	Segm 4	Segm 5
Mean	$\mu_{1:50} = 0$	$\mu_{1:25} = 2, \mu_{26:50} = 0,$	$\mu_{1:50} = 0$	$\mu_{1:50} = 0$	$\mu_{1:50} = 1.5$
Model Matrix	Θ_1	Θ_1	Θ_2	Θ_2	Θ_2
Norms of Θ	0.13	0.13	0.09	0.09	0.09
Variance–covariance	U_1	U_1	U_1	U_2	U_2
Norms of U	44.2	44.2	44.2	52.6	52.6

Figure 1 shows an example of the simulated process from Example 1. The Left plot shows all 50 components of the process. In this realization, each homogeneous segment (the one with the same generating mechanism) has a length of 500 points. The right plot shows only ten components. It allows us to see better the behavior of the process.

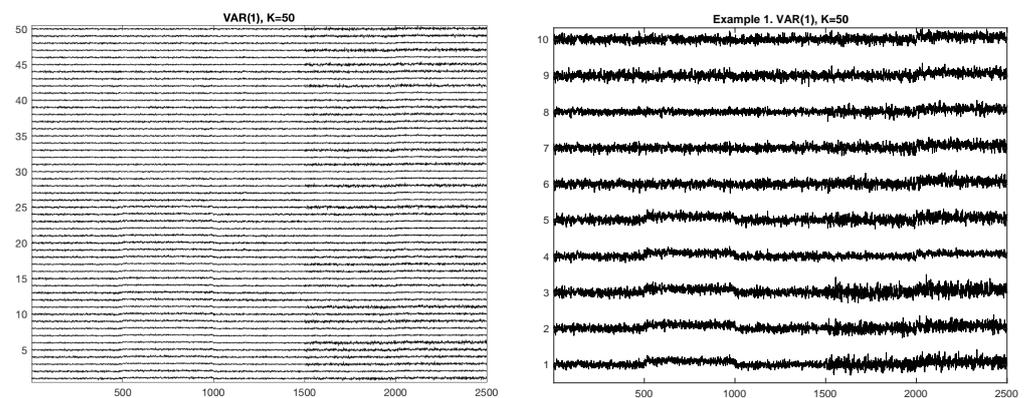


Figure 1. Example 1. Realization of the VAR(1) process. **Left:** All 50 components. **Right:** selected 10 components.

Figure 2 shows the examples of the diagnostic sequences A (Left plot) and B (Right plot) and detected change-points. Black solid lines correspond to the diagnostic sequences, horizontal blue lines correspond to the mean values between the detected change-points. The jump points correspond to the detected change-points. The vertical red lines correspond to the true change-points.

The numerical results are presented in Tables 2 and 3. The percentage of the number of detected points for diagnostic sequences of coefficients A and B are presented in Table 2. Let us notice that the coefficient B was not useful for detecting change-points in this example. The percentages of correctly found numbers of each of the four change-points (true positive rate) and corresponding bootstrap confidence intervals are presented in Table 3. To compare the new method with the old method, we present in Table 4 the percentages of correctly found numbers of each of the four change-points (true positive rate) and corresponding bootstrap confidence intervals for our old method.

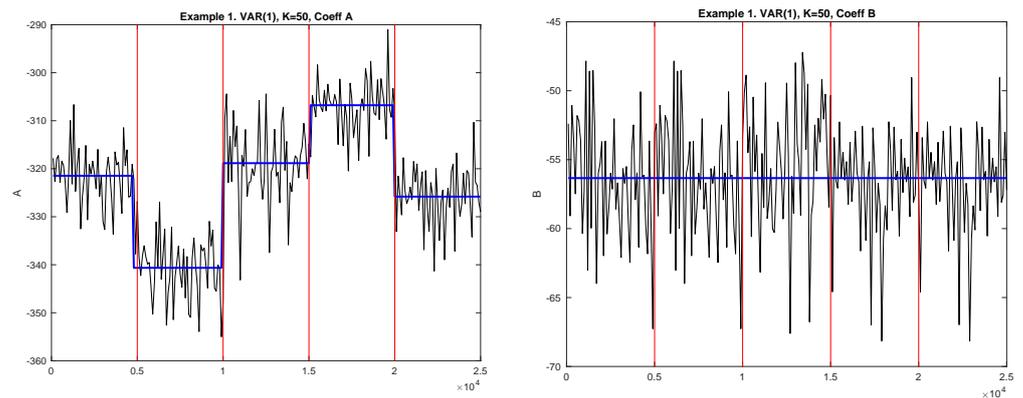


Figure 2. Example 1. Diagnostic sequences and detected MCGM. **Left:** Coefficient A; **Right:** Coefficient B. Black solid lines correspond to the diagnostic sequences; horizontal blue lines correspond to the mean values between the detected MCGM. The jump points correspond to the detected change-points. The vertical red lines correspond to the true change-points.

Table 2. Example 1. The percentage of the number of detected change-points for diagnostic sequences of coefficients A and B.

# of Detected Points	Coeff A(t)	Coeff B(t)
1	0%	32.3%
2	0%	65.9%
3	1%	1.6%
4	74.5%	0.2%
5	20%	0%
7	0.8%	0%

Table 3. Example 1. The percentages of correctly found numbers of each of the four change-points (true positive rate) and corresponding bootstrap confidence intervals using the proposed method.

Change-Point	True Positive Rate, Coeff A	CI, Coeff A	True Positive Rate, Coeff B	CI, Coeff B
1	96.6%	(4800, 5200)	0.1%	N/A
2	98.9%	(9800, 10,176)	26.1%	(9100, 10,900)
3	95.5%	(14,500, 15,300)	12.7	(14,100, 15,733)
4	78.3 %	(19,900, 20,100)	0	N/A

Table 4. Example 1. The percentages of correctly found numbers of each of the four change-points (true positive rate) and corresponding bootstrap confidence intervals using the old method.

Change-Point	True Positive Rate, Coeff A	CI, Coeff A	True Positive Rate, Coeff B	CI, Coeff B
1	1.6%	(4100, 5800)	2.3%	(4100,5700)
2	99.8%	(9900, 10,200)	89.9%	(9600, 10,750)
3	5.8%	(14,200, 15,900)	3.0	(14,125, 15,875)
4	56.4	(19,300, 20,700)	37.5	(19,300, 20,700)

In the first change-points, the change occurred in 50% of the components, and the size of the shift was approximately $0.77\sigma = 2$. Here, σ is the maximal standard deviation of the components. If we reduce the number of components or decrease the size of the shift, our accuracy will decrease. For the last change-point, we decrease the size of the change in the mean but have a change in the mean of all components. In this case, the shift was approximately 0.75σ .

To measure differences between the matrices and variance–covariance matrices for which we were able to detect changes, we report spectral norms $\|\Theta_1 - \Theta_2\|_2$ and $\|\Sigma_{u_1} - \Sigma_{u_2}\|_2$ and ratios $\|\Theta_1 - \Theta_2\|_2 / (0.5\|\Theta_1\|_2 + 0.5\|\Theta_2\|_2)$, $\|\Sigma_{u_1} - \Sigma_{u_2}\|_2 / (0.5\|\Sigma_{u_1}\|_2 + 0.5\|\Sigma_{u_2}\|_2)$. Thus, for the second change-point, when the change occurred in the model matrix, $\|\Theta_1 - \Theta_2\|_2 = 0.053$ and $\|\Theta_1 - \Theta_2\|_2 / (0.5\|\Theta_1\|_2 + 0.5\|\Theta_2\|_2) = 0.49$. For the third change-point, where change occurred in variance–covariance matrix, $\|\Sigma_{u_1} - \Sigma_{u_2}\|_2 = 12.64$ and $\|\Sigma_{u_1} - \Sigma_{u_2}\|_2 / (0.5\|\Sigma_{u_1}\|_2 + 0.5\|\Sigma_{u_2}\|_2) = 0.25$.

As one can see from Tables 3 and 4, we detected the first and third change-points with the proposed method, and the old approach did not detect them. In the case of the fourth change-point, the old method detected 56% of simulations while the proposed method detected 78.3% of simulations.

Example 2. Chaotic deterministic processes.

In this example, we created a seven-dimensional series with chaotic components. The processes and parameters for each process are presented in Table 5. In the first column, we gave the index of the component. In the second column, we presented the name of the process. In parentheses, we provided the reference to the equations that generate the process. For components 1–6 in columns 3, 4, and 5 (with titles Segment 1, Segment 2, Segment 3), we provided parameters of the processes used to generate corresponding segments. For component 7, the processes do not have parameters, and we provide the reference to the generating equation and its name. We generated segment 4 the same way as segment 3, but we added a shift of size 0.5 of the standard deviation of the components of segment 3.

Table 5. Example 2. Processes and changes in the parameters.

Component	Process	Segment 1	Segment 2	Segment 3
1	Logistic map, (6)	$\alpha = 3.94$	$\alpha = 4$	$\alpha_3 = 3.89$
2	Hénon Map, x (10)	$a_1 = 1.5$	$a_2 = 1.3$	$a_3 = 1.4$
		$b_1 = 0.2$	$b_2 = 0.2$	$b_3 = 0.2$
3	Hénon Map, y (10)			
4	Ikeda map, x (11)	$\mu_1 = 0.9$	$\mu_2 = 0.87$	$\mu_3 = 0.9$
		$c_1 = 1.97$	$c_2 = 1.99$	$c_3 = 1.97$
5	Ikeda map, y (11)			
6	Quadratic map (7)	$c_1 = 2$	$c_2 = 1.87$	$c_3 = 1.95$
7		Process 3 (8)	Interval map (9)	Interval map (9)

One can see that the first change in generating mechanism occurred in all components. The second change occurred in the parameters of the first six components. The third change in generating mechanism is due to the shift.

Figure 3 shows an example of the simulated process from Example 2. In this realization, each homogeneous segment has length 500 points.

Figure 4 shows the examples of the diagnostic sequences A (Left plot) and B (Right plot) and detected change-points for Example 2. Black solid lines correspond to the diagnostic sequences; horizontal blue lines correspond to the mean values between the detected change-points. The jump points correspond to the detected change-points. The vertical red lines correspond to the actual change-points.

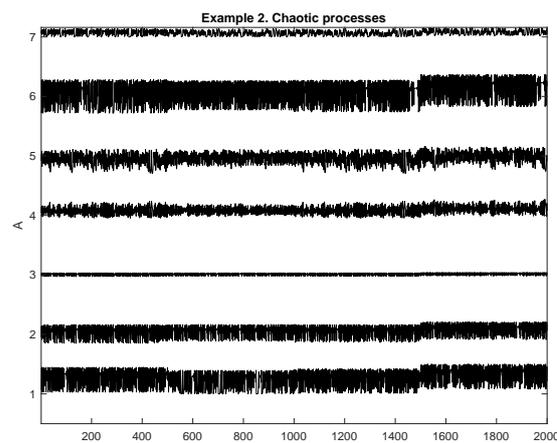


Figure 3. Example 2. Realization of the chaotic multidimensional process.

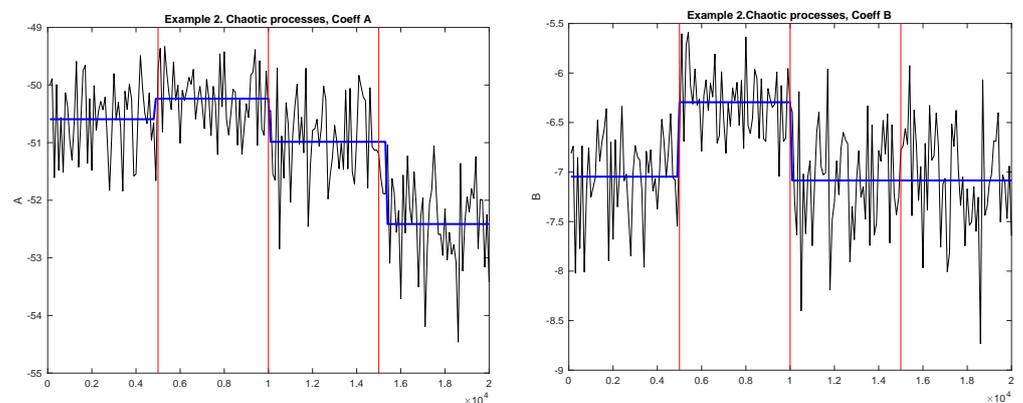


Figure 4. Example 2. Diagnostic sequences and detected MCGM. **Left:** Coefficient A; **Right:** Coefficient B. Black solid lines correspond to the diagnostic sequences; horizontal blue lines correspond to the mean values between the detected change-points. The jump points correspond to the detected change-points. The vertical red lines correspond to the true change-points.

The numerical results are presented in Tables 6 and 7. The percentage of the number of detected points for diagnostic sequences of coefficients A and B are presented in Table 6. The percentages of correctly found numbers of each of the three change-points (true positive rate) and corresponding bootstrap confidence intervals are presented in Table 7. The results for the old method are presented in Table 8. Let us notice that the coefficient B was more efficient for detecting changes in generating mechanism when one chaotic process changes by another one. In this case, we do have change in the Hölder constant and, therefore, coefficient B detection works best. Let us notice that the second MCGM was detected only in 75% of cases by coefficient B. In this case, it was no change for one of the components.

We observe that the first two points can be detected using coefficient B. It agrees with our hypothesis that for such processes, the Hölder constant changes. The shift cannot be detected using coefficient B. In terms of coefficient B, our proposed method and our old method detect a similar proportion of first and second change-points. However, we could not detect a change in the mean of the multivariate process using the old method (see, Table 8). The new method detected this change in 95.3% of the simulations.

Example 3. Mixed process.

In this example, we combine processes from the first examples and parametric processes from the second example. As a result, we obtained a multidimensional time series that has stochastic and deterministic components. In this example, we simulated 20 components of the multivariate stochastic process and eight components of deterministic

processes. We simulated five homogeneous segments of length 5000. The total length is 25,000. There are four MCGMs.

Table 6. Example 2. The percentage of the number of detected change-points for diagnostic sequences of coefficients A and B .

# of Detected Points	Coeff $A(t)$	Coeff $B(t)$
1	16%	0%
2	13.7%	99.5%
3	52%	0.5%
4	15.2%	0%
5	1.4%	0%
6	0.3%	0%
7	0.1%	0%

Table 7. Example 2. The percentages of correctly found numbers of each of the three change-points (true positive rate) and corresponding bootstrap confidence intervals, proposed method.

Change-Point	True Positive A	CI, Coeff A	True Positive B	CI, Coeff B
1	66.9%	(4900, 5000)	98.1%	(4900, 5000)
2	69.4%	(9700, 10,700)	75.1%	(10,000, 10,100)
3	95.3 %	(14,800, 15,800)	0.2%	N/A

Table 8. Example 2. The percentages of correctly found numbers of each of the three change-points (true positive rate) and corresponding bootstrap confidence intervals using the old method.

Change-Point	True Positive A	CI, Coeff A	True Positive B	CI, Coeff B
1	1.1%	(51,000, 5900)	98.8%	(4900, 5000)
2	5.8%	(9400, 10,100)	74.9%	(10,000, 10,100)
3	0 %	N/A	0.6%	N/A

The processes and parameters for each process are presented in Table 9.

In the first column, we present the index of the component. In the second column, we provide the name of the process. In parentheses, we provide the reference to the equations that generate the process. The first 20 components are trajectories of the VAR(1) model. In Table 9, one can see which matrices are the same and which are different for different segments. For components 21–28 in columns 3, 4, 5, and 6 (with titles Segment 1, Segment 2, Segment 3, Segment 4), we provide parameters of the processes we used to generate corresponding segments. We generated segment four in the same way as segment three, but for each component. We added shifts of size 0.3 of the standard deviation of the components for segment 4.

The first change in generating mechanism occurred only in the deterministic components. The second change occurred in model matrix Θ of the VAR(1) process. The third MCGM corresponds to the change in variance–covariance matrix of the VAR(1) process and one component of deterministic process. The fourth change is the change in the mean value for all components. Here, we keep parameters of each component as it is in segment 4 but added the shift 0.3 of standard deviation for segment 4 of the corresponding components.

Figure 5 shows an example of the simulated process from Example 3. The Left plot shows all 27 components of the process. In this realization, each homogeneous segment has a length of 500 points. The right plot shows only ten components (five are stochastic and five are deterministic) of the given process, which allows us to see the process' behavior better.

Table 9. Example 3. Processes and changes in the parameters.

Component	Process	Segm 1	Segm 2	Segm 3	Segm 4
1–20	VAR(1) norms Θ norms U	Θ_1, U_1 0.16 8.6	Θ_1, U_1 0.02 8.6	Θ_2, U_1 0.02 8.6	Θ_2, U_2 0.16 12.7
21	Logistic map, (6)	$\alpha = 4$	$\alpha = 3.98$	$\alpha_3 = 3.97$	$\alpha_4 = 3.98$
22	Hénon Map, x (10)	$a_1 = 1.5$ $b_1 = 0.2$	$a_2 = 1.3$ $b_2 = 0.2$	$a_3 = 1.4$ $b_3 = 0.2$	$a_4 = 1.4$ $b_4 = 0.2$
23	Hénon Map, y (10)				
24	Hénon Map, x (10)	$a_1 = 1.5$ $b_1 = 0.18$	$a_2 = 1.2$ $b_2 = 0.2$	$a_3 = 1.2$ $b_3 = 0.2$	$a_4 = 1.4$ $b_4 = 0.2$
25	Hénon Map, y (10)				
26	Ikeda map, x (11)	$\mu_1 = 0.9$ $c_1 = 1.97$	$\mu_2 = 0.86$ $c_2 = 1.995$	$\mu_3 = 0.86$ $c_3 = 1.995$	$\mu_4 = 0.86$ $c_4 = 1.995$
27	Ikeda map, y (11)				
28	Quadratic map (7)	$c_1 = 2$	$c_2 = 1.9$	$c_3 = 1.9$	$c_4 = 1.97$

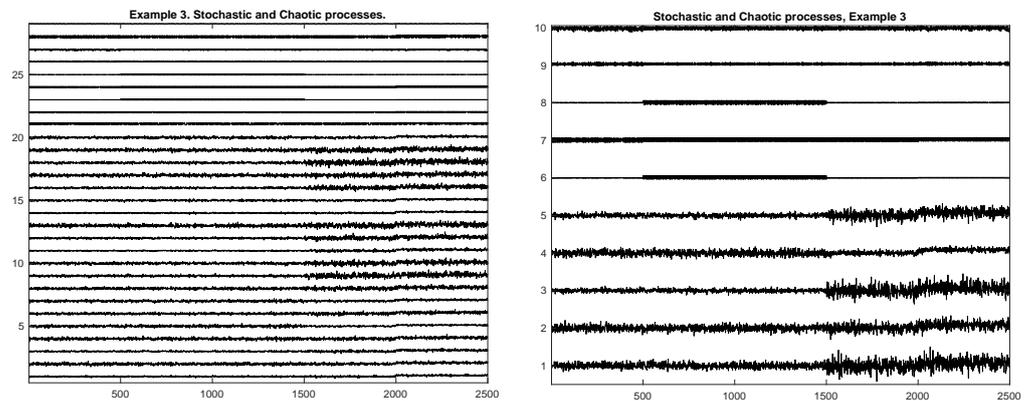


Figure 5. Example 3. Realization of the mixed process (some components are stochastic and some are deterministic). **Left:** All components. **Right:** selected 10 components.

Figure 6 shows the examples of the diagnostic sequences A (Left plot) and B (Right plot) and detected change-points for Example 3. Black solid lines correspond to the diagnostic sequences; horizontal blue lines correspond to the mean values between the detected change-points. The jump points correspond to the detected change-points. The vertical red lines correspond to the actual change-points.

The numerical results are presented in Tables 10 and 11. The percentage of the number of detected points for diagnostic sequences of coefficients A and B in 1000 replications are presented in Table 10. The percentages of correctly found numbers of each of the four change-points (true positive rate) and corresponding bootstrap confidence intervals are presented in Table 11. The results for the old method are presented in Table 12.

In this example, the diagnostic sequence of coefficient A works better for this example. Let us also observe that the second change in generating mechanism (change in the parameter α of the Logistic map and parameter a of the Hénon map) is better detected by the old method (see, Table 12). For other MCGMs, the new method works better.

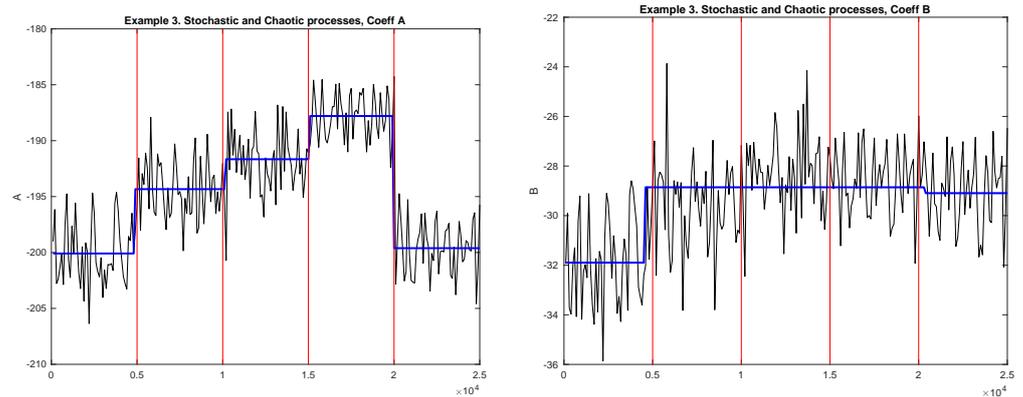


Figure 6. Example 3. Diagnostic sequences and detected MCGM. **Left:** Coefficient A; **Right:** Coefficient B. Black solid lines correspond to the diagnostic sequences; horizontal blue lines correspond to the mean values between the detected change-points. The jump points correspond to the detected change-points. The vertical red lines correspond to the true change-points.

Table 10. Example 3. The percentage of the number of detected change-points for diagnostic sequences of coefficients A and B.

# of Detected Points	Coeff A(t)	Coeff B(t)
1	7.3%	25.1%
2	13.6%	51%
3	60%	19.2%
4	11.6%	3.6%
5	4.6%	0.9%
6	2.2%	0.2%
7	0.7%	0%

Table 11. Example 3. The percentages of correctly found numbers of each of the four change-points (true positive rate) and corresponding bootstrap confidence intervals using the proposed method.

Change-Point	True Positive A	CI, Coeff A	True Positive B	CI, Coeff B
1	98.8.9%	(4700, 5400)	79.1.1%	(4300, 5770)
2	71.0%	(9600, 10,600)	51.5%	(9100, 10,700)
3	81.2%	(14,400, 15,300)	0.4%	(14,100, 15,900)
4	82.5%	(20,000, 20,100)	0.6%	N/A

Table 12. Example 3. The percentages of correctly found numbers of each of the four change-points (true positive rate) and corresponding bootstrap confidence intervals using the old method.

Change-Point	True Positive A	CI, Coeff A	True Positive B	CI, Coeff B
1	1.9%	(4200, 5900)	79.2.9%	(4200, 5800)
2	99.6.0%	(9900, 10200)	88.7.5%	(9600, 10,700)
3	5.4.2%	(14,100, 15,820)	3.5%	(14,300, 15,900)
4	58.5.5%	(19,200, 20,700)	45.5.6%	(19,200, 20,800)

5. Conclusions

In this paper, we proposed the model-free method for retrospective detection of moments of changes in generating mechanisms of multivariate time series. The detection

of moments of changes in the generating mechanism is important for the subsequent analysis of the collected data. It allows one to carry out segmentation of the data on homogeneous fragments.

In econometrics, the moments of changes in the generation mechanism of multidimensional data are called structural breaks. The problem of detection of changes in chaotic processes arises in the study of the phenomenon of multifractality.

However, often, the mechanism for generating a time series is either known inaccurately or entirely unknown. Typical examples here are multidimensional EGG signals, financial time series, some biological data, etc. Thus, it is essential to develop methods for detecting the moments of changes in the generation mechanism of time series that do not use models.

We proposed the method for the detection of changes regardless of the generating mechanisms of arbitrary nature. This method is an extension of our approach proposed in [12]. The given simulation results demonstrate the effectiveness of the new version of our method.

In our simulation study, we considered three examples. In the first example, we simulated the VAR(1) process with four change-points. The first one corresponds to the change of the mean values of 50% of the component. The change was approximately $0.77\sigma = 2$, where σ is the maximal standard deviation of the components. In case of change in the mean of all components, we detected a change of $0.77\sigma = 2$. The new method was able to catch them while our previous method did not detect this change. The new method detected the change in the variance–covariance matrix U but the old approach did not. Both methods were able to detect the change in the model matrix Θ . To measure differences between the matrices and variance–covariance matrices for which we were able to detect changes, we provided the following spectral norms $\|\Theta_1 - \Theta_2\|_2$ and $\|\Sigma_{u_1} - \Sigma_{u_2}\|_2$ and ratios $\|\Theta_1 - \Theta_2\|_2 / (0.5\|\Theta_1\|_2 + 0.5\|\Theta_2\|_2)$, $\|\Sigma_{u_1} - \Sigma_{u_2}\|_2 / (0.5\|\Sigma_{u_1}\|_2 + 0.5\|\Sigma_{u_2}\|_2)$.

In the second example, we detected changes in multivariate chaotic deterministic processes with some dependent components. In this case, we were able to detect a shift $0.5\sigma_i$, where σ_i is the standard deviation of the i -th component. We observed that old and new methods detected changes in the parameters of the chaotic deterministic processes with similar accuracy; however, only the new approach enabled us to detect changes in the mean of multivariate chaotic deterministic processes.

In the last example, we considered the process with stochastic and chaotic components. We observed that the new method was superior to detecting changes in the VAR(1) process and a change in the mean value of the process.

The limitation of our method is that it requires a relatively long sequence of multivariate time series. To calculate the complexity coefficient, we need at least 100 data 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 MCGM, the mathematical expectation of the sequence of complexity coefficient varies according to a certain 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 estimates of MCGM.

A fundamental feature of the proposed method is its independence from the model of the observed process. As far as we know, model-free methods for solving such problems have not been considered in the literature. 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.

Author Contributions: All authors contributed equally to the work. All authors have read and agreed to the published version of the manuscript.

Funding: Boris Darkhovsky's research was funded by No. RFFI 20-07-00221.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: We would like to thank the editors for the invitation to this special issue. We thank the anonymous reviewers for their careful reading of our manuscript and their insightful comments and suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

VAR Vector Autoregressive model
MCGM Moment of Changes in Generating Mechanism

References

1. Basseville, M.; Nikiforov, I.V. *Detection of Abrupt Changes: Theory and Application*; Prentice Hall Englewood Cliffs: Hoboken, NJ, USA, 1993; Volume 104.
2. Brodsky, E.; Darkhovsky, B.S. *Non-Parametric Statistical Diagnosis: Problems and Methods*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2000; Volume 509.
3. Csörgö, M.; Horváth, L. *Limit Theorems in Change-Point Analysis*; John Wiley & Sons Inc.: Hoboken, NJ, USA, 1997; Volume 18.
4. Horváth, L.; Hušková, M. Change-point detection in panel data. *J. Time Ser. Anal.* **2012**, *33*, 631–648. [[CrossRef](#)]
5. Aue, A.; Horváth, L. Structural breaks in time series. *J. Time Ser. Anal.* **2013**, *34*, 1–16. [[CrossRef](#)]
6. Baltagi, B.H.; Kao, C.; Liu, L. Estimation and identification of change points in panel models with nonstationary or stationary regressors and error term. *Econom. Rev.* **2017**, *36*, 85–102. [[CrossRef](#)]
7. Antoch, J.; Hanousek, J.; Horváth, L.; Hušková, M.; Wang, S. Structural breaks in panel data: Large number of panels and short length time series. *Econom. Rev.* **2019**, *38*, 828–855. [[CrossRef](#)]
8. Collet, P.; Eckmann, J.P. *Concepts and Results in Chaotic Dynamics: A Short Course*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2007.
9. Pavlov, A.N.; Anishchenko, V.S. Multifractal analysis of complex signals. *Phys. Uspekhi* **2007**, *50*, 819–834. [[CrossRef](#)]
10. Darkhovsky, B.; Piryatinska, A. New approach to the segmentation problem for time series of arbitrary nature. *Proc. Steklov Inst. Math.* **2014**, *287*, 54–67. [[CrossRef](#)]
11. Darkhovsky, B.S. On the complexity and dimension of continuous finite-dimensional maps. *Theory Probab. Its Appl.* **2020**, *65*, 375–387. [[CrossRef](#)]
12. Darkhovsky, B.; Piryatinska, A. Model-free offline change-point detection in multidimensional time series of arbitrary nature via ϵ -complexity: Simulations and applications. *Appl. Stoch. Model. Bus. Ind.* **2018**, *34*, 633–644. [[CrossRef](#)]
13. Darkhovsky, B.; Piryatinska, A. Model-free classification of panel data via the ϵ -complexity theory. In *Communications in Statistics-Simulation and Computation*; Taylor and Francis: London, UK, 2020; pp. 1–14.
14. Darkhovskii, B.; Brodskii, B. An Identification of the “Disorder” Time of the Random Sequence. *IFAC Proc. Vol.* **1979**, *12*, 373–379. [[CrossRef](#)]
15. Horváth, L.; Hušková, M.; Rice, G.; Wang, J. Asymptotic properties of the cusum estimator for the time of change in linear panel data models. *Econom. Theory* **2017**, *33*, 366–412. [[CrossRef](#)]
16. Lütkepohl, H. *New Introduction to Multiple Time Series Analysis*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2005.
17. Meyer, C.D. *Matrix Analysis and Applied Linear Algebra*; Siam: Philadelphia, PA, USA, 2000; Volume 71.
18. May, R.M. Simple mathematical models with very complicated dynamics. In *The Theory of Chaotic Attractors*; Springer: Berlin/Heidelberg, Germany, 2004; pp. 85–93.
19. Grebogi, C.; Ott, E.; Yorke, J.A. Crises, sudden changes in chaotic attractors, and transient chaos. *Phys. D Nonlinear Phenom.* **1983**, *7*, 181–200. [[CrossRef](#)]
20. Vulpiani, A.; Cecconi, F.; Cencini, M. *Chaos: From Simple Models To complex Systems*; World Scientific: Singapore, 2009; Volume 17.
21. Hénon, M. A two-dimensional mapping with a strange attractor. In *The Theory of Chaotic Attractors*; Springer: Berlin/Heidelberg, Germany, 1976; pp. 94–102.
22. Ikeda, K. Multiple-valued stationary state and its instability of the transmitted light by a ring cavity system. *Opt. Commun.* **1979**, *30*, 257–261. [[CrossRef](#)]