Analysis of the properties of local models in clustering problems of quasi-periodic time series*

A. V. Grabovoy[1], **V. V. Strijov**[2]

Annotation: The paper analysis periodic signals in the time series for recognize the physical actions of a person by using a mobile accelerometer. The method of clustering points of the time series is proposed for searching quasi-periodic segments of the time series. Time series are objects of a complex structure. There are not any feature vector for describe points of a time series. The paper proposes to consider principal components of the local neighborhood of the phase trajectory near some point for feature description of this point. The article proposes a distance function between points in the new space of objects. Clustering is carried out using a matrix of pairwise distances between the points in time series. The algorithm was tested on synthetic and real data. Real data was obtained by using a mobile accelerometer.

Key words: time series; clustering; segmentation; recognition of physical activity; principal component method.

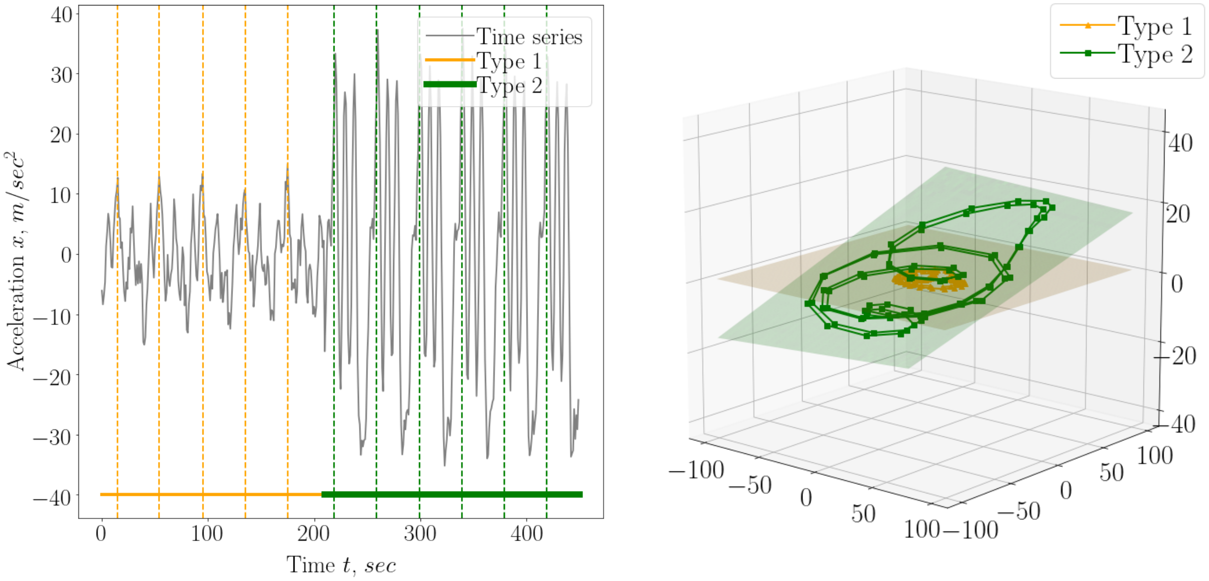
*This research was supported by RFBR, project ..., and by Government of the Russian Federation, agreement ...

[1] Moscow Institute of Physics and Technology, grabovoy.av@phystech.edu

[2] Moscow Institute of Physics and Technology, strijov@ccas.ru

1 Introduction

Analysis of a person's physical activity is carried out by using mobile phones, smart watches etc [1, 2]. These devices are using an accelerometer, gyroscope and magnetometer. The main purpose of this work is to markup and recognize human activity during the time [3, 4], and also search for the beginning of a periodic signal [6, 5]. Examples of an action segment is a step, a step of running, a single squat, a single jump, etc. Current work considers an sequences that consist of at least two consecutive segments that correspond to the same type of human activity.



(a)

(b)

Figure 1: A time series with clustering: a) time series with assessor markings on clusters and markings at beginning of each quasi-periodic segment; b) projection of phase trajectories on the first two principal components

Time series are objects of complex structure. The method of constructing feature vectors for points is very important for their clustering. In this article, the object of analysis, as well as clustering, is a point on the time axis. The paper investigates the problem of clustering points in a time series. *Clustering* is a process in which all points in a time series are labeled with a label from the finite set of labels. Each label corresponds to one characteristic physical action. A *segment* is a part of a time series that corresponds to one characteristic physical action, for example: a step with two legs when walking, or a step with two legs when running. A sequence of segments that correspond to one physical action form a *chain* of actions. It is assumed that the chain of actions forms a quasi-periodic sequence of values of the time series. A sequence of points $\{b_t\}_{t=1}^N$ is called *quasi-periodic* with period of T, if for all t there is a Δ, such that:

$$b_t \approx b_{t+T+\Delta}, \quad |\Delta| \ll T. \tag{1.1}$$

An example of clustering and splitting a series into segments is shown in fig. 1a. The time series is divided into two characteristic physical actions, which are marked Type 1 and Type 2. Also, this time series contains two quasi-periodic chains of actions.

The proposed solution of the clustering problem for points in a time series consists of two stages. First, an algorithm for local approximation of time series by using the principal component method [11] is proposed to obtain a feature description of points in the time series. *Local approximation* of the time series means that only a certain neighborhood of a point is used to describe the features of their points. The two main components of the phase trajectory segment are considered as a feature description of a time series point. The fig. 1b shows the first two principal components of the phase trajectories, and also shows the projection of the phase trajectories on these components. The trajectories relate to different physical actions, which are denoted Type 1 and Type 2, in the time series. The planes that are generated by these basic components are different. That means that Type 1 and Type 2 are two different actions. Secondly, the distance function between points in the new feature space is considered. This function is the distance between two bases of some subspaces within the phase space of a time series. It can be explained by using fig 1b. The function is considered between two planes, which are defined by two different bases for the Type1 and Type2 segments. Points can be clustered by using a pairwise distance matrix. The segmentation problem is solved by using the main components [6] of the phase trajectory in each cluster separately.

The method has several assumptions. It is assumed that the periods of different segments differ slightly. Minimum and maximum periods of segments are known. The number of different segments in the time series is known too. It is also assumed that the type of segments in time does not change often.

The quality analysis of the proposed clustering method is carried out on synthetic and real data. A synthetic data constructed by using the sum of the first few terms of the Fourier series with random coefficients. But the experiment on the segmentation of the time series was carried out on simple sinusoidal signals with random amplitude and frequency. Real data was received by using a mobile accelerometer, which took readings during exercise of a person.

2 Literature analysis

3 Statement

A time series

$$\mathbf{x} \in \mathbb{R}^N, \tag{3.1}$$

where N is a number of points in the time series. The time series consists of a sequence of segments:

$$\mathbf{x} = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_M], \tag{3.2}$$

where $\mathbf{v}_i$ is a segment from the set of segments $\mathbf{V}$, which are observed in the time series $\mathbf{x}$. For all i either the $[\mathbf{v}_{i-1}, \mathbf{v}_i]$ or the $[\mathbf{v}_i, \mathbf{v}_{i+1}]$ is a chain of action. Let the set of segments $\mathbf{V}$ satisfies the following properties:

$$|\mathbf{V}| = K, \quad \mathbf{v} \in \mathbf{V} \;\; |\mathbf{v}| \leq T, \tag{3.3}$$

where $|\mathbf{V}|$ is a number of different action in the set of segments $\mathbf{V}$, $|\mathbf{v}|$ is a length of segment, K is a number of different action in the time series $\mathbf{x}$ and T is a length of maximum segment.

Consider the mapping

$$a : t \to \mathbb{Y} = \{1, \cdots, K\}, \tag{3.4}$$

where $t \in \{1, \cdots, N\}$ is a point of time in which the time series is defined. Let the mapping a satisfy the following properties:

$$\begin{cases} a\left(t_1\right) = a\left(t_2\right), & \text{if time moments } t_1, t_2 \text{ relate to similar type of segments,} \\ a\left(t_1\right) \neq a\left(t_2\right), & \text{if time moments } t_1, t_2 \text{ relate to different type of segments.} \end{cases} \tag{3.5}$$

Consider following markup of the time series points:

$$\mathbf{y} \in \{1, \cdots, K\}^N. \tag{3.6}$$

The algorithm error for the time series can be calculated by the following formula:

$$S = \frac{1}{N}\sum_{t=1}^{N}[y_t = a\left(t\right)], \tag{3.7}$$

where t — is a point of time, y_t is a markup in the point of time t for time series $\mathbf{x}$.

4 Points clustering

Consider the phase trajectory of the time series $\mathbf{x}$:

$$\mathbf{H} = \{\mathbf{h}_t | \mathbf{h}_t = [x_{t-T}, x_{t-T+1}, \cdots, x_t],\ T \leq t \leq N\}, \tag{4.1}$$

where $\mathbf{h}_t$ is a phase trajectory point.

Information about the length of the maximum segment in the time series allows us to split the phase trajectory into segments.

$$\mathbf{S} = \{\mathbf{s}_t | \mathbf{s}_t = [\mathbf{h}_{t-T}, \mathbf{h}_{t-T+1}, \cdots, \mathbf{h}_{t+T-1}],\ T \leq t \leq N-T\}, \tag{4.2}$$

where $\mathbf{s}_t$ is a segment of phase trajectory. Segments have all local information about the time series, as it contains all the information on the period up to some time point t and information about the period after the time point t.

The principal components $\mathbf{W}_t$ for T-dimensional segments $\mathbf{s}_t$ are considered as a features description of a point t in a time series. The segment $\mathbf{s}_t$ is projected onto a subspace of dimension two by using the principal component method $\mathbf{z}_t = \mathbf{W}_t\mathbf{s}_t$:

$$\mathbf{W} = \{\mathbf{W}_t | \mathbf{W}_t = [\lambda_t^1\mathbf{w}_t^1, \lambda_t^2\mathbf{w}_t^2]\}, \quad \boldsymbol{\Lambda} = \{\boldsymbol{\lambda}_t | \boldsymbol{\lambda}_t = [\lambda_t^1, \lambda_t^2]\}, \tag{4.3}$$

where $[\mathbf{w}_t^1, \mathbf{w}_t^2]$ and $[\lambda_t^1, \lambda_t^2]$ are the basis vectors and eigenvalues obtained by using the principal component method for the phase trajectory segment $\mathbf{s}_t$.

Consider the distance function between points $\mathbf{W}_{t_1}, \mathbf{W}_{t_2}$ in the time series $\mathbf{x}$ for their clustering:

$$\rho\left(\mathbf{W}_1, \mathbf{W}_2\right) = \max\left(\max_{\mathbf{e}_2 \in \mathbf{W}_2} d_1\left(\mathbf{e}_2\right), \max_{\mathbf{e}_1 \in \mathbf{W}_1} d_2\left(\mathbf{e}_1\right)\right), \tag{4.4}$$

where $\mathbf{e}_i$ is the basic vector of space $\mathbf{W}_i$, and $d_i\left(\mathbf{e}\right)$ is the distance from vector $\mathbf{e}$ to the subspace $\mathbf{W}_i$.

If all subspaces $\mathbf{W}_t$ have dimension two, then the distance function $\rho\left(\mathbf{W}_1, \mathbf{W}_2\right)$ has the following interpretation:

$$\rho\left(\mathbf{W}_1, \mathbf{W}_2\right) = \max_{\{\mathbf{a},\mathbf{b},\mathbf{c}\} \subset \mathbf{W}_1 \cup \mathbf{W}_2} V\left(\mathbf{a}, \mathbf{b}, \mathbf{c}\right), \tag{4.5}$$

where $\mathbf{W}_1 \cup \mathbf{W}_2$ is a concatenation of bases, $V\left(\mathbf{a}, \mathbf{b}, \mathbf{c}\right)$ is the volume of parallelepiped built on vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}$, which are columns of matrix $\mathbf{W}_1 \cup \mathbf{W}_2$.

Consider the distance function between eigenvalues:

$$\rho\left(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2\right) = \sqrt{\left(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\right)^{\mathsf{T}}\left(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\right)}. \tag{4.6}$$

Consider the distance between two points in time t_1, t_2 by using equations (4.5-4.6), and consider the matrix of pairwise distances between pairs of points in the time series:

$$\rho\left(t_1, t_2\right) = \rho\left(\mathbf{W}_1, \mathbf{W}_2\right) + \rho\left(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2\right), \quad \mathbf{M} = \mathbb{R}^{N \times N}, \tag{4.7}$$

where $\mathbf{M}$ is a matrix of pairwise distances between all pairs of points t in the time series $\mathbf{x}$. The pairwise distance matrix $\mathbf{M}$ is used for clustering points t of the time series (3.4).

5 Experiment

5.1 Points clustering

In the experiment, the clustering of points in the time series was carried out using matrices of pairwise distances (4.7). The experiment was carried out on real and synthetic data, which are described in the table 1. ”Physical Motion” is a real time series obtained by using a mobile accelerometer. Synthetic time series were constructed by using the first few terms of the Fourier series with random coefficients from the standard normal distribution. At the first stage, short segments $\mathbf{v}$ were generated to build a set of all segments $\mathbf{V}$. The second stage is the following random process:

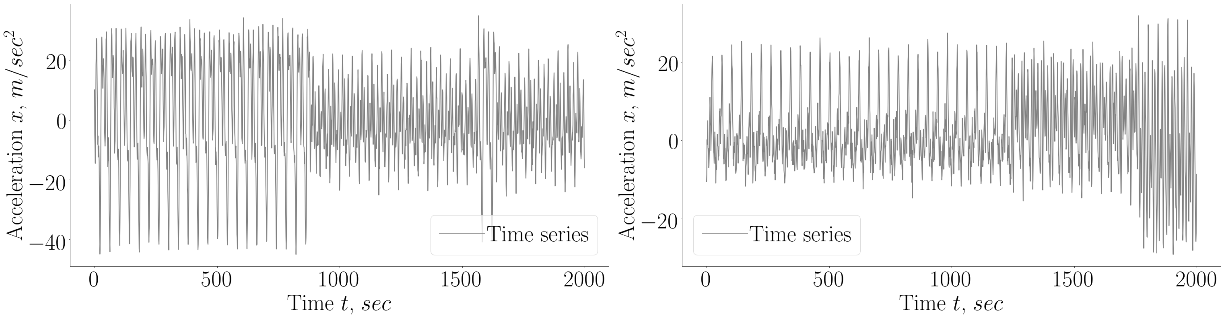
The generation of synthetic time series consisted of two stages.

$$\mathbf{x} = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_M] + \boldsymbol{\varepsilon}, \quad \begin{cases} \mathbf{v}_1 \sim \mathcal{U}(\mathbf{V}), & \\ \mathbf{v}_i = \mathbf{v}_{i-1}, & \text{with probability } \frac{3}{4}, \\ \mathbf{v}_i \sim \mathcal{U}(\mathbf{V}), & \text{with probability } \frac{1}{4} \end{cases} \tag{5.1}$$

where $\mathcal{U}(\mathbf{V})$ is a uniform distribution on objects from the set $\mathbf{V}$, and $\boldsymbol{\varepsilon}$ is gaussian noise.

Table 1: Description of time series in the experiment

Series, $\mathbf{x}$	Length, N	Number of segments, K	Period, T
Physical Motion 1	900	2	40
Physical Motion 2	900	2	40
Synthetic 1	2000	2	20
Synthetic 2	2000	3	20

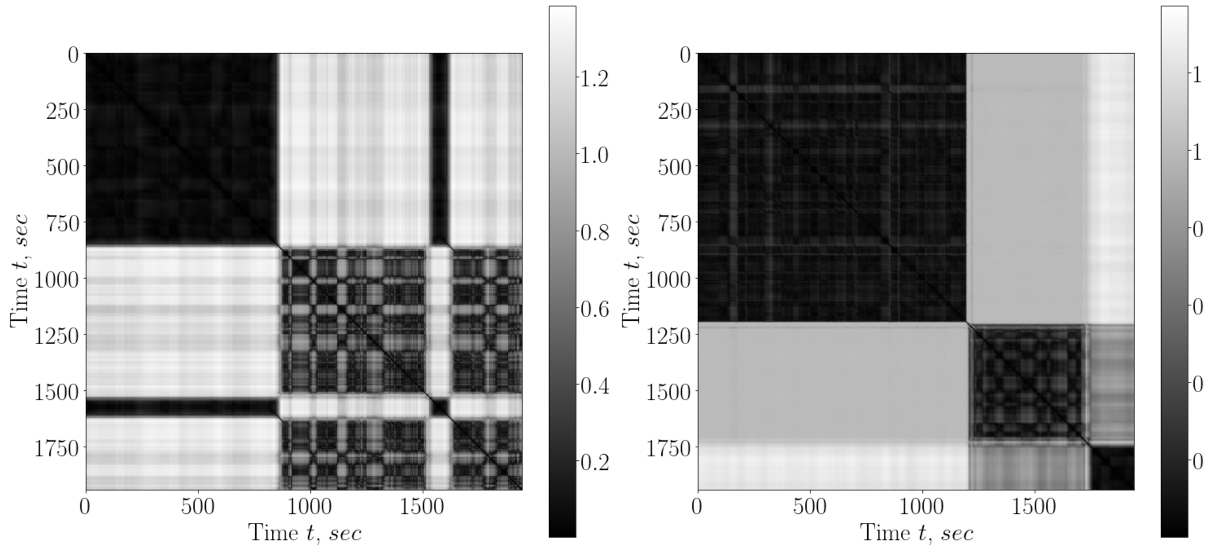


(a)

(b)

Figure 2: Example of synthetic time series: a) Synthetic 1; b) Synthetic 2

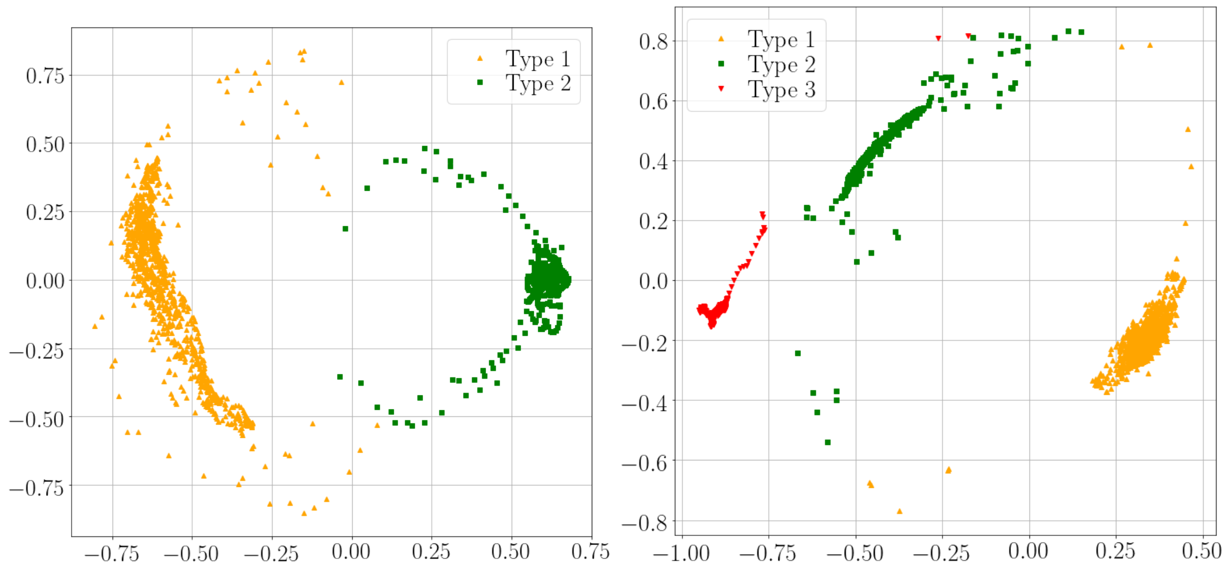
Synthetic data. The fig. 2 shows an example of synthetic time series. The fig. 2a shows an example of a time series in which the number of different segments is $K = 2$, and the



(a)

(b)

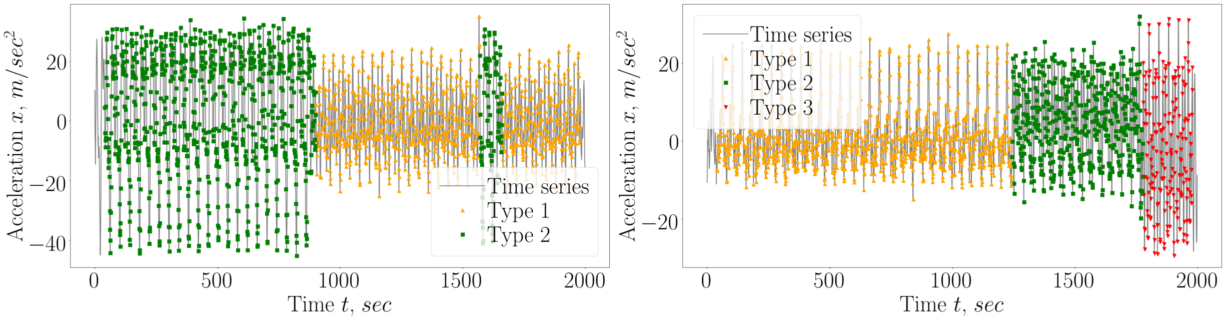
Figure 3: Pairwise distance matrix $\mathbf{M}$ between points of time series: a) Synthetic 1; b) Synthetic 2



(a)

(b)

Figure 4: The projection of points on a plane by using the pairwise distance matrix $\mathbf{M}$: a) Synthetic 1; b) Synthetic 2



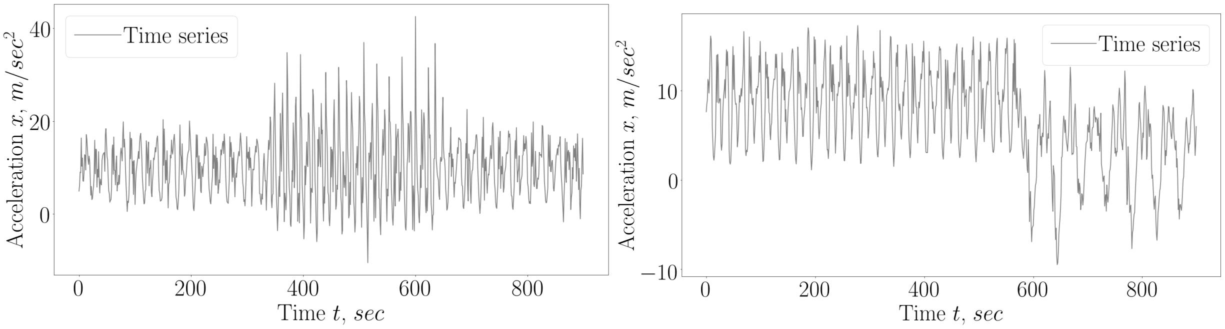
(a)

(b)

Figure 5: Time series points clustering: a) Synthetic 1; b) Synthetic 2

maximum length of segments is $T = 20$. The fig. 2b shows an example of a time series in which the number of different segments is $K = 3$, and the maximum length of segments is $T = 20$.

The fig. 3 illustrates the pairwise distance matrix $\mathbf{M}$ between all pairs of points t in the time series, which are constructed by using equention (4.7). Time series points can be easily visualized on a plane by using a pairwise distance matrix and a Multidimensional Scaling method [10]. The fig. 4 shows the visualization of points on the plane and performed their clustering by using the hierarchical clustering method.

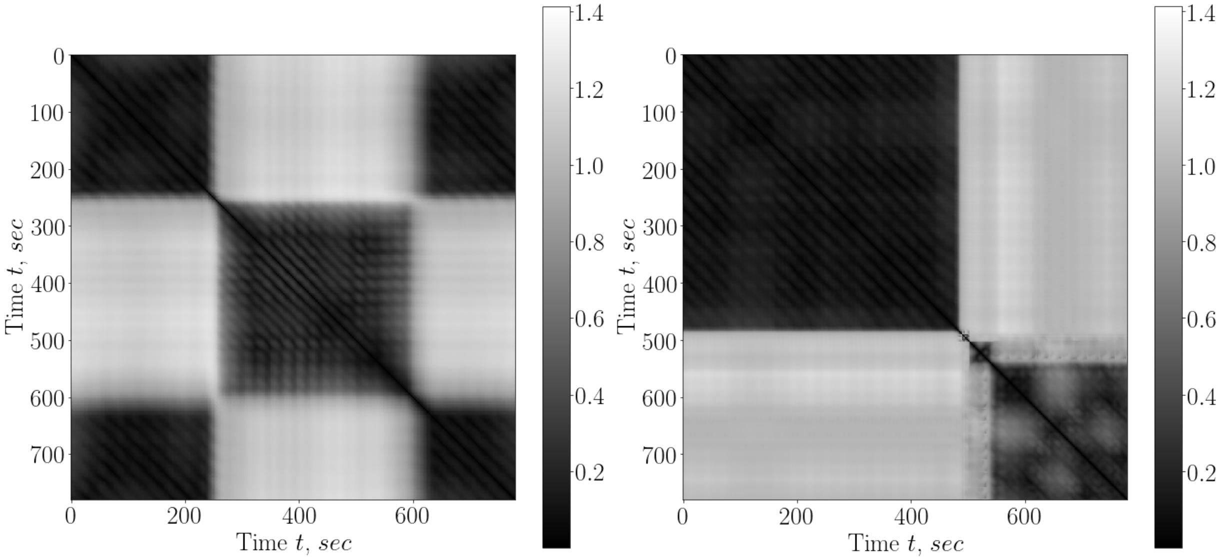


(a)

(b)

Figure 6: Example of real time series: a) Physical Motion 1; b) Physical Motion 2

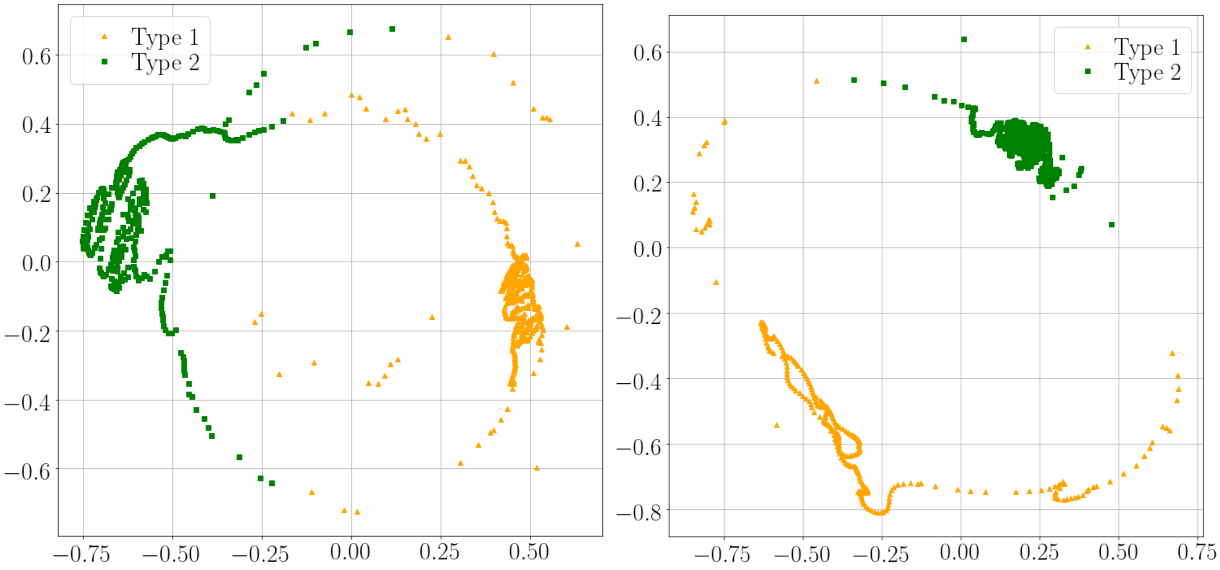
Real data. The fig. 6 shows an example of real time series obtained using a mobile accelerometer.



(a)

(b)

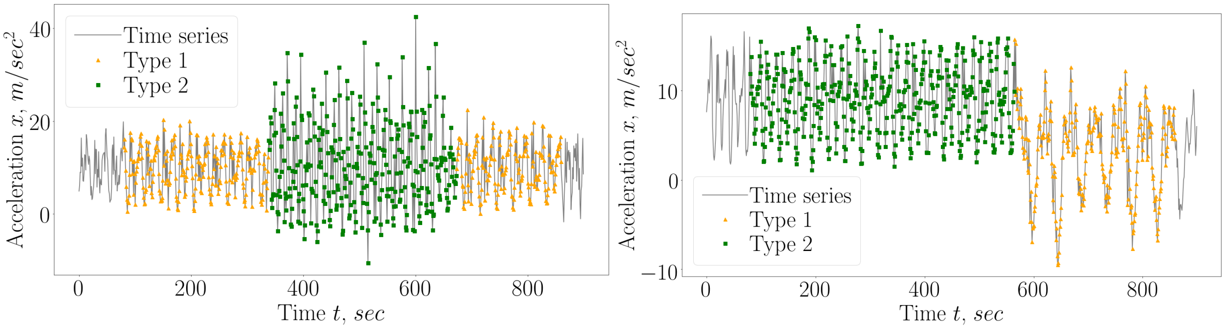
Figure 7: Pairwise distance matrix **M** between points of time series: a) Physical Motion 1; b) Physical Motion 2



(a)

(b)

Figure 8: The projection of points on a plane by using the pairwise distance matrix **M**: a) Physical Motion 1; b) Physical Motion 2



(a)

(b)

Figure 9: Time series points clustering: a) Physical Motion 1; b) Physical Motion 2

The fig. 7 illustrates the pairwise distance matrix $\mathbf{M}$ between all pairs of points t in the time series, which are constructed by using equention (4.7). Time series points can be easily visualized on a plane by using a pairwise distance matrix and a Multidimensional Scaling method [10]. The fig. 8 shows the visualization of points on the plane and performed their clustering by using the hierarchical clustering method.

5.2 Time series segmentation

Time series segmentation is carried out on synthetic and real data. A synthetic time series for this experiment were constructed by using the concatenation of two different sinusoidal signals with different amplitude and frequency. The experiment was carried out on real and synthetic data, which are described in the table 2.

Segmentation is carried out by using the method that is presented in the work [6]. The method is used for each action within the time series separately.

Table 2: Description of time series in the experiment of segmentation

Series, $\mathbf{x}$	Length, N	Number of segments, K	Period, T
Simple 1	1000	2	100
Physical Motion 2	900	2	40

Synthetic data. The fig. 10 shows the result of the segmentation for the Simple 1 time series. The algorithm is well marked the beginning of the segments.

The fig. 10 shows the projections of the phase spaces for both clusters onto their first two main components.

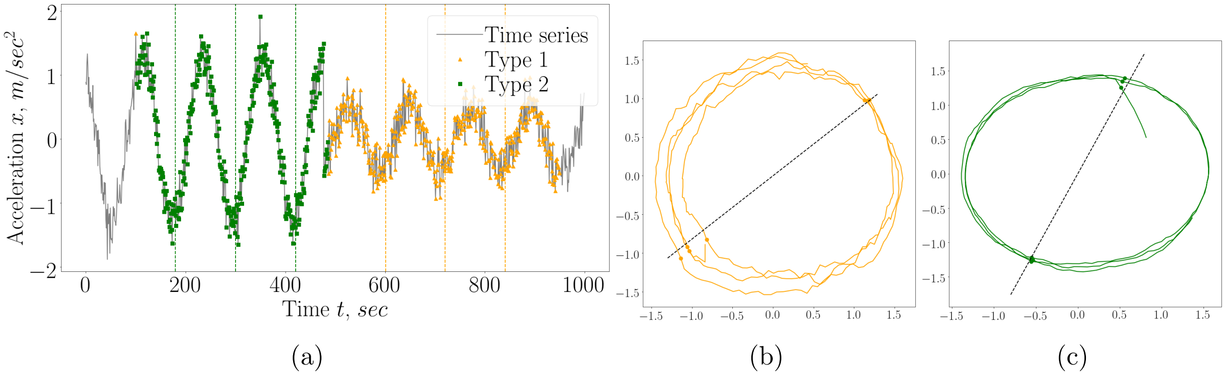
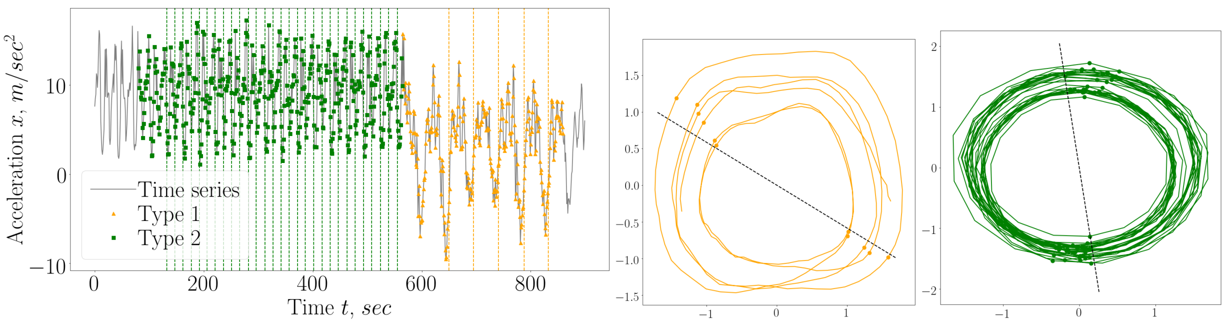


Figure 10: Segmentation for the Simple 1 time series: a) time series segmentation; b) the projection of the phase space on the first two principal components for the first cluster; c) the projection of the phase space on the first two principal components for the second cluster;

Real data. The fig. 10 shows the result of the segmentation for the Physical Motion 2 time series. The algorithm is well marked the beginning of the segments for Type 1 and bad for Type 2. The fig. 10 shows the projections of the phase spaces for both clusters onto their first two main components. Bad segmentation was obtained due to self-intersection of the phase trajectory.

6 Conclusion

The paper considered the problem of finding periodic structures within a time series. A method based on local reduction of the phase space dimension was considered. An algorithm for searching for segments was proposed. The algorithm is based on the principal component method for local dimension reduction. Also introduced is the function of the distance between local basis at each time instant. Local bases were interpreted as a features description of a point in the time series.



(a) (b) (c)

Figure 11: Segmentation for the Physical Motion 2 time series: a) time series segmentation; b) the projection of the phase space on the first two principal components for the first cluster; c) the projection of the phase space on the first two principal components for the second cluster;

Table 3: Algorithm analysis results

Series, $\mathbf{x}$	Length, N	Number of segments, K	Period, T	Error, S
Phys. Motion 1	900	2	40	0.06
Phys. Motion 2	900	2	40	0.03
Synthetic 1	2000	2	20	0.04
Synthetic 2	2000	3	20	0.03

During the experiment on the real and synthetic data showed that the proposed method for measuring the distance between the basis well separates points that are related to different type of action, which leads to good clustering of time series points. The results of the experiment are shown in the table 3. The experiment was carried out segmentation of time series by using the method [6] for each cluster separately.

The proposed method has few disadvantages associated with a large number of assumption on a time series. These restrictions will be relaxed in subsequent papers. It is also planned to solve the problem of finding the minimum dimension of the phase space for which the phase trajectory will not have self-intersections.

References

[1] *J. R. Kwapisz, G. M. Weiss, S. A. Moore* Activity Recognition using Cell Phone Accelerometers // Proceedings of the Fourth International Workshop on Knowledge Discovery from Sensor Data, 2010. Vol. 12. P. 74–82.

[2] *W. Wang, H. Liu, L. Yu, F. Sun* Activity Recognition using Cell Phone Accelerometers // Joint Conference on Neural Networks, 2014. P. 1185–1190.

[3] *A. D. Ignatov, V. V. Strijov* Human activity recognition using quasiperiodic time series collected from a single tri-axial accelerometer. // Multimedial Tools and Applications, 2015.

[4] *A. Olivares, J. Ramirez, J. M. Gorris, G. Olivares, M. Damas* Detection of (in)activity periods in human body motion using inertial sensors: A comparative study. // Sensors, 12(5):5791–5814, 2012.

[5] *Y. G. Cinar and H. Mirisaee* Period-aware content attention RNNs for time series forecasting with missing values // Neurocomputing, 2018. Vol. 312. P. 177–186.

[6] *A. P. Motrenko, V. V. Strijov* Extracting fundamental periods to segment biomedical signals // Journal of Biomedical and Health Informatics, 2015, 20(6). P. 1466 - 1476.

[7] *Y. P. Lukashin* Adaptive methods for short-term forecasting // Finansy and Statistik, 2003.

[8] *M. P. Kuznetsov, N. P. Ivkin* Time series classification algorithm using combined feature description // Machine Learning and Data Analysis, 2015, 11(1). P. 1471-1483.

[9] *V. V. Strijov, A. M. Katrutsa* Stresstes procedures for features selection algorithms. // Schemometrics and Intelligent Laboratory System, 2015.

[10] *I. Borg, P. J. F. Groenen* Modern Multidimensional Scaling. — New York: Springer, 2005. 540 p.

[11] *D. L. Danilov, A. A. Zhiglovsky* Main components of time series: method "Gesenitsa". — St. Petersburg University, 1997.