Анализ свойств локальных моделей в задачах кластеризации квазипериодических временных рядов

Грабовой Андрей Валериевич

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель д.ф.-м.н. В. В. Стрижов

Москва,
2019г

Цель работы

Исследуется

Исследуется задача поиска характерных переодических структур внетри временного ряда.

Требуется

Требуется предложить алгоритм поиска характерных сегментов, который основывается на методе главных компонент для локального снижения размерности.

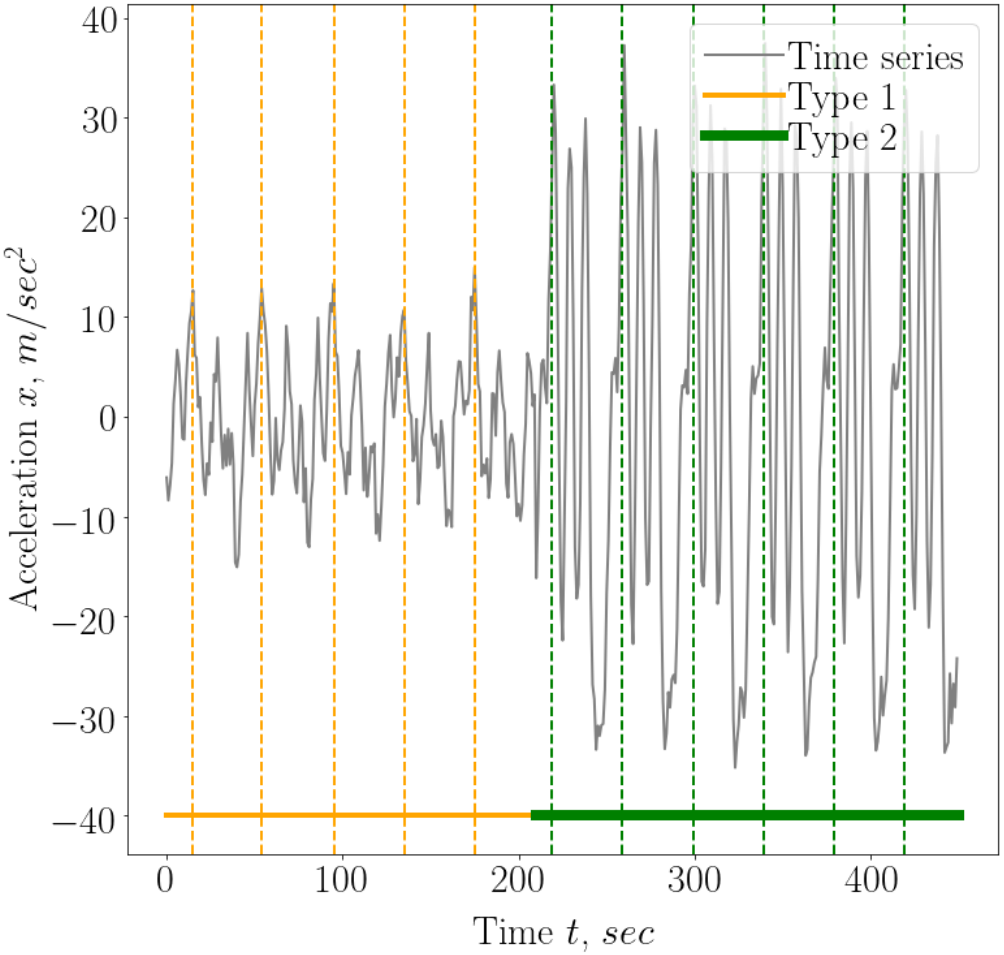
Проблемы

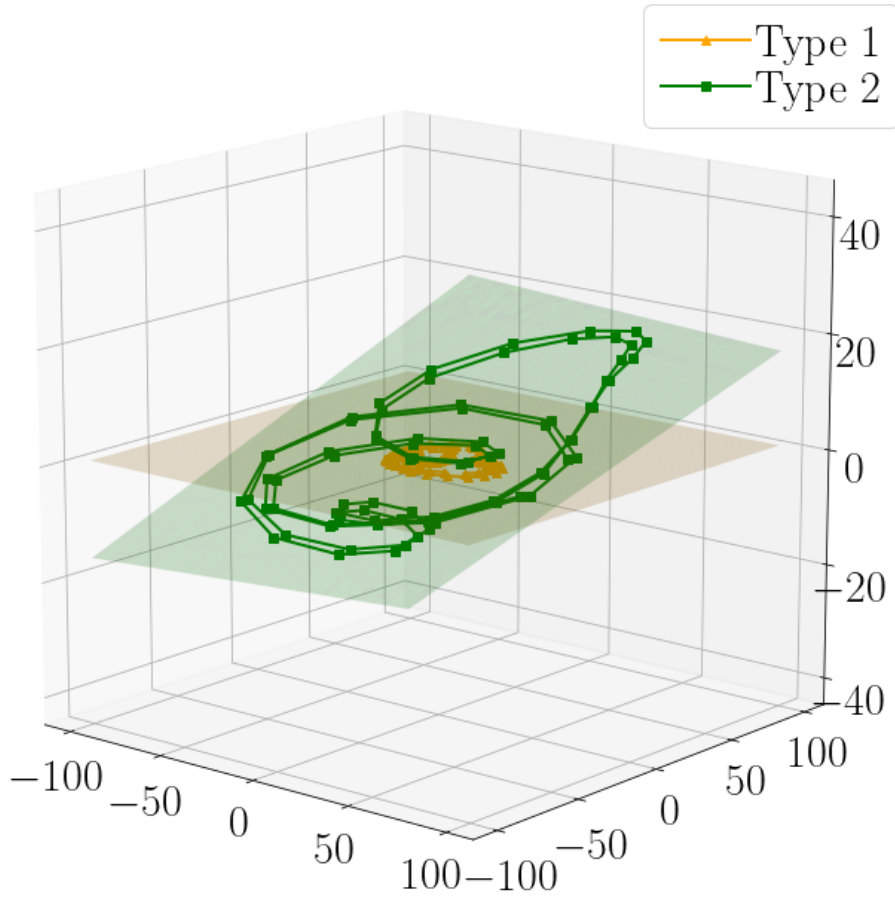
Построение признакового описания точек временного ряда низкой размерности.

Список литературы

- *И. П. Ивкин, М. П. Кузнецов* Алгоритм классификации временных рядов акселерометра по комбинированному признаковому описанию. // Машинное обучение и анализ данных, 2015.
- *V. V. Strijov, A. M. Katrutsa* Stresstes procedures for features selection algorithms. // Schemometrics and Intelligent Laboratory System, 2015.
- *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.
- *I. Borg, P. J. F. Groenen* Modern Multidimensional Scaling. — New York: Springer, 2005. 540 p.
- *Д. Л. Данилова, А. А. Жигловский* Главные компоненты временных рядов: метод "Гусеница". — СПбУ, 1997.

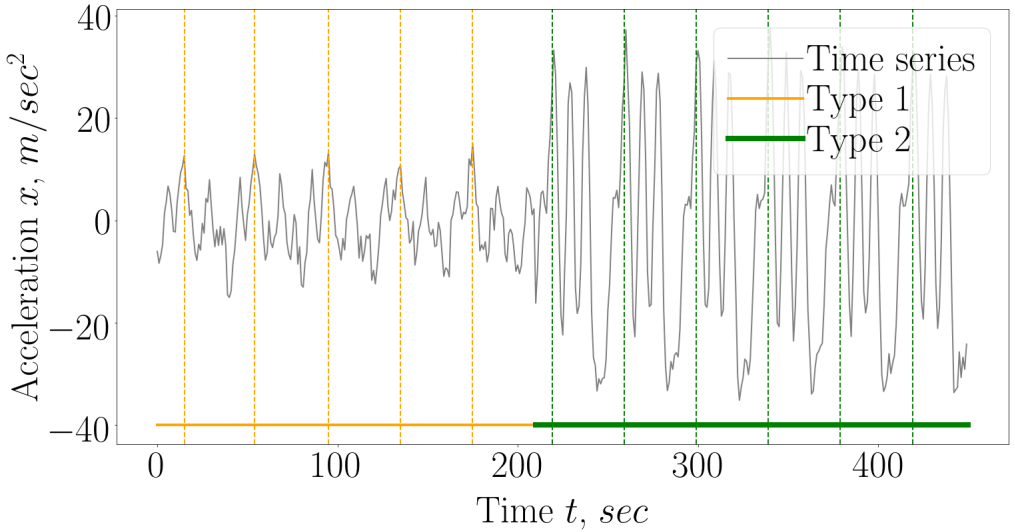
Илюстрация задачи





Временной ряд, с разметкой на кластеры: временной ряд с ассесорской разметкой на кластеры и выделением начала квазипериодического сегмента; проекция фазовых траекторий на первые две главные компоненты

Постановка задачи



Задан временной ряд:
$\mathbf{x} \in \mathbb{R}^{\mathrm{N}}, \quad \mathbf{x} = [\mathbf{v}_1, \cdots, \mathbf{v}_M], \quad \mathbf{v}_i \in \mathcal{V},$
где $\mathcal{V}$ множество возможных сегментов в ряде $\mathbf{x}$.

Предположения:

- $|\mathcal{V}| = K$,
- $\mathbf{v} \in \mathcal{V} \ |\mathbf{v}| \leq T$,
- для всех i либо $[\mathbf{v}_{i-1}, \mathbf{v}_i]$ либо $[\mathbf{v}_i, \mathbf{v}_{i+1}]$ является цепочкой действий,

где $|\mathcal{V}|$ мощность множества сигналов, а $|\mathbf{v}|$ длина сигнала.

Постановка задачи

Рассматривается отображение

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

где $t \in \{1, \cdots, N\}$ некоторый момент времени, на котором задан временной ряд. Требуется, чтобы отображение a удовлетворяло следующим свойствам:

$$\begin{cases} a(t_1) = a(t_2), & \text{если в моменты } t_1, t_2 \text{ совершается один тип действий} \\ a(t_1) \neq a(t_2), & \text{если в моменты } t_1, t_2 \text{ совершаются разные типы действий} \end{cases}$$

Пусть задана некоторая асессорская разметка временного ряда:

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

Тогда ошибка алгоритма a на временном ряде $\mathbf{x}$ представляется в следующем виде:

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

где t — момент времени, y_t асессорская разметка t-го момента времени для заданого временного ряда.

Кластеризация точек

Фазовая траектория ряда $\mathbf{x}$:

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

где $\mathbf{h}_t$ — точка фазовой траектории.

Фазовые подпространства:

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

где $\mathbf{s}_t$ — это сегмент фазовой траектории.

Множество базисов:

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

где $[\mathbf{w}_t^1, \mathbf{w}_t^2]$ и $[\lambda_t^1, \lambda_t^2]$ это базисные векторы и соответствующие им собственные числа для сегмента фазовой траектории $\mathbf{s}_t$.

Кластеризация точек

Расстояние между элементами $\mathbf{W}_{t_1}, \mathbf{W}_{t_2}$:

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

где $\mathbf{e}_i$ это базисный вектор пространства $\mathbf{W}_i$, а $d_i\left(\mathbf{e}\right)$ является расстоянием от вектора $\mathbf{e}$ до пространства $\mathbf{W}_i$.

Theorem

Пусть задано множество подпространств $\mathbb{W}$ пространства $\mathbb{R}^n$. Каждое подпространство которого задается базисом $\mathbf{W}_i \in \mathbf{W}$, тогда функция расстояния $\rho\left(\boldsymbol{W}_1, \boldsymbol{W}_2\right)$ является метрикой заданой на множестве базисов $\mathbf{W}$:

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

где $\boldsymbol{e}_i$ это базисный вектор из $\boldsymbol{W}_i$, а $d_i\left(\boldsymbol{e}\right)$ является расстоянием от вектора $\boldsymbol{e}$ до пространства заданого базисом $\boldsymbol{W}_i$.

Кластеризация точек

Расстояние между элементами $\mathbf{W}_{t_1}, \mathbf{W}_{t_2}$:

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

где $\mathbf{W}_1 \cup \mathbf{W}_2$ это объединение базисных векторов первого и второго пространства, $V\left(\mathbf{a}, \mathbf{b}, \mathbf{c}\right)$ — объем параллелепипеда построенного на векторах $\mathbf{a}, \mathbf{b}, \mathbf{c}$, которые являются столбцами матрицы $\mathbf{W}_1 \cup \mathbf{W}_2$.

Расстояние между элементами $\mathcal{L}$:

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

Расстояние между точками временного ряда:

$$\rho\left(t_1, t_2\right) = \rho\left(\mathbf{W}_1, \mathbf{W}_2\right) + \rho\left(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2\right).$$

Матрица попарных растояний:

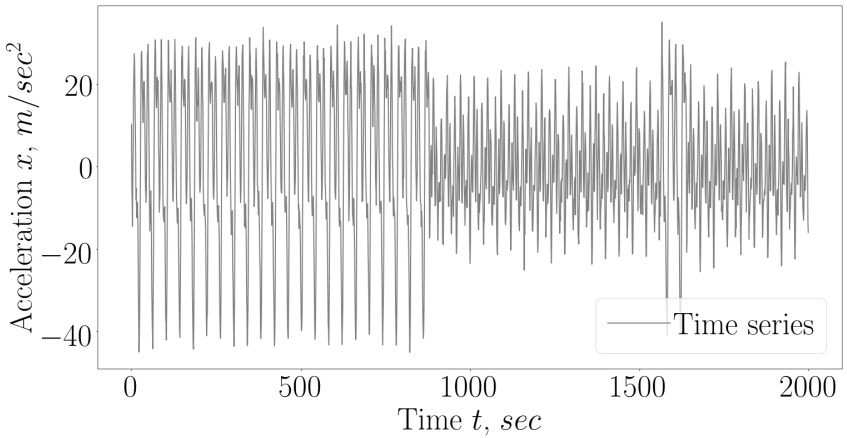
$$\mathbf{M} = \mathbb{R}_{+}^{N \times N}.$$

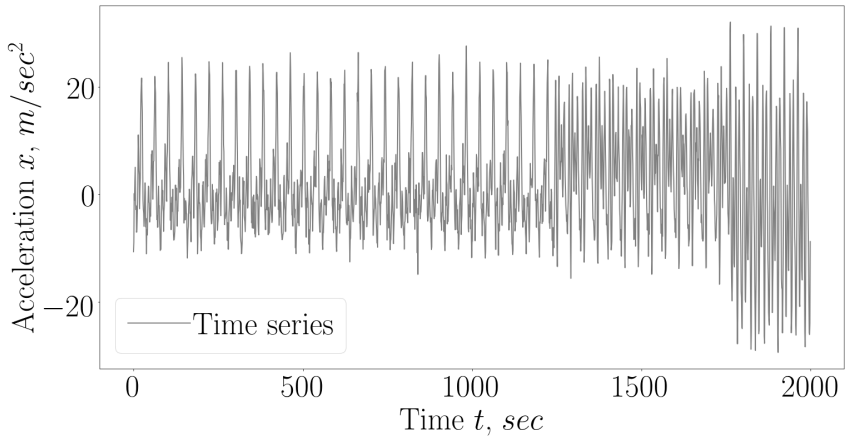
Описание временных рядов в эксперименте

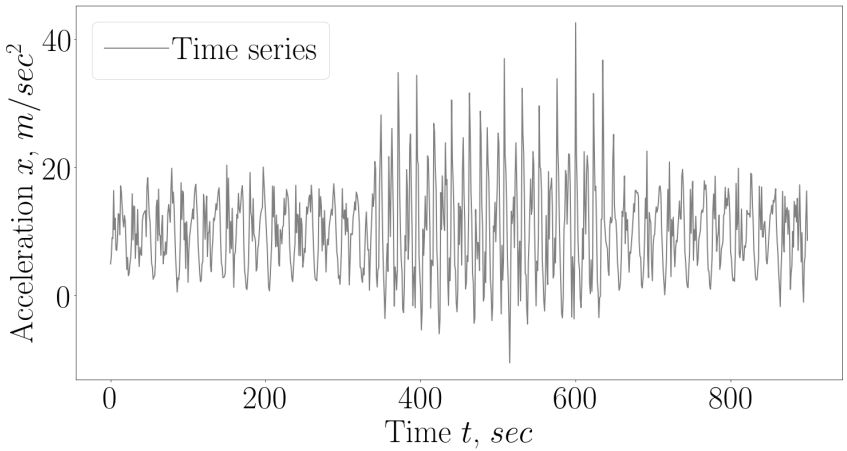
Ряд, $\mathbf{x}$	Точек, N	Сегментов, K	Период, T
Physical Motion 1	900	2	40
Physical Motion 2	900	2	40
Synthetic 1	2000	2	20
Synthetic 2	2000	3	20
Simple 1	1000	2	100

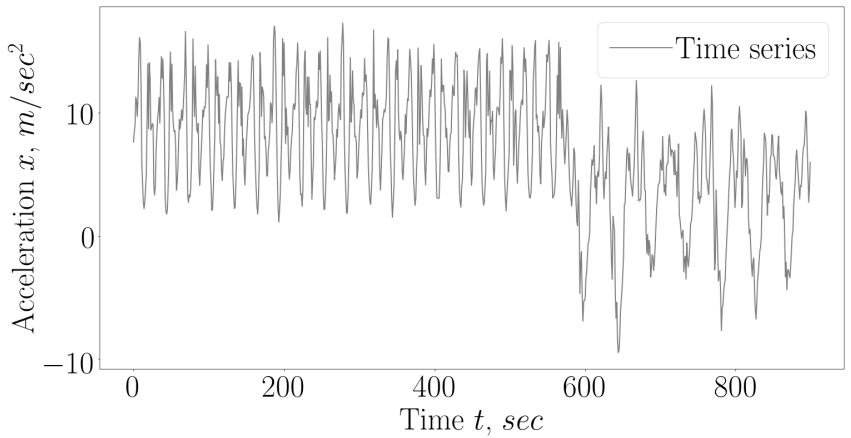
- N — число точек во временном ряде,
- K — число различных действий во временном ряде,
- T — максимальная длина сегмента.

Пример временных рядов



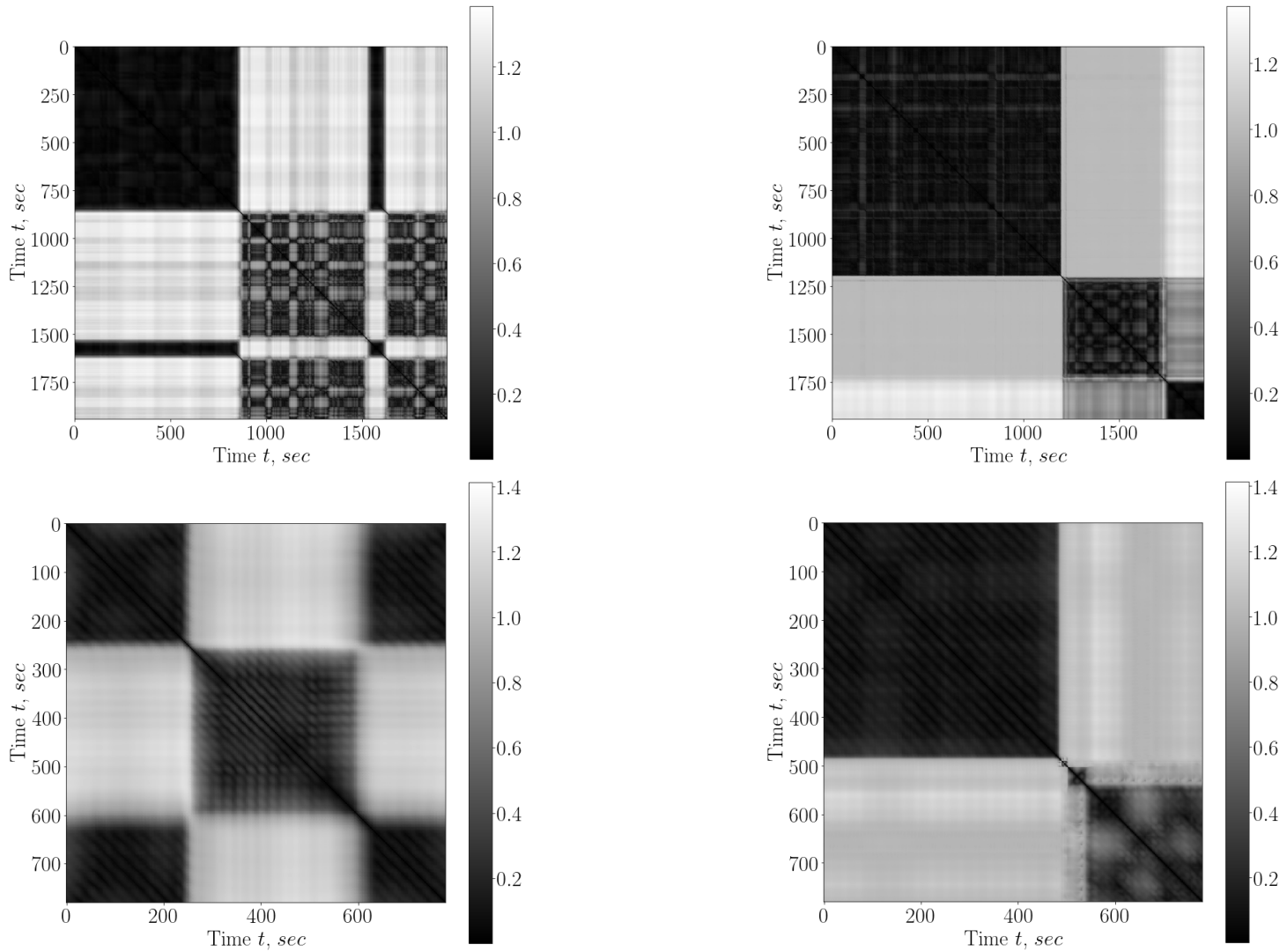






Временные ряды построенные синтетически, а также при помощи мобильного акселерометра.

Матрица попарных расстояний M



Матрицы попарных расстояний для временных рядов, построенных синтетически, а также при помощи мобильного акселерометра.

Проекция точек фазовой траектории на плоскость

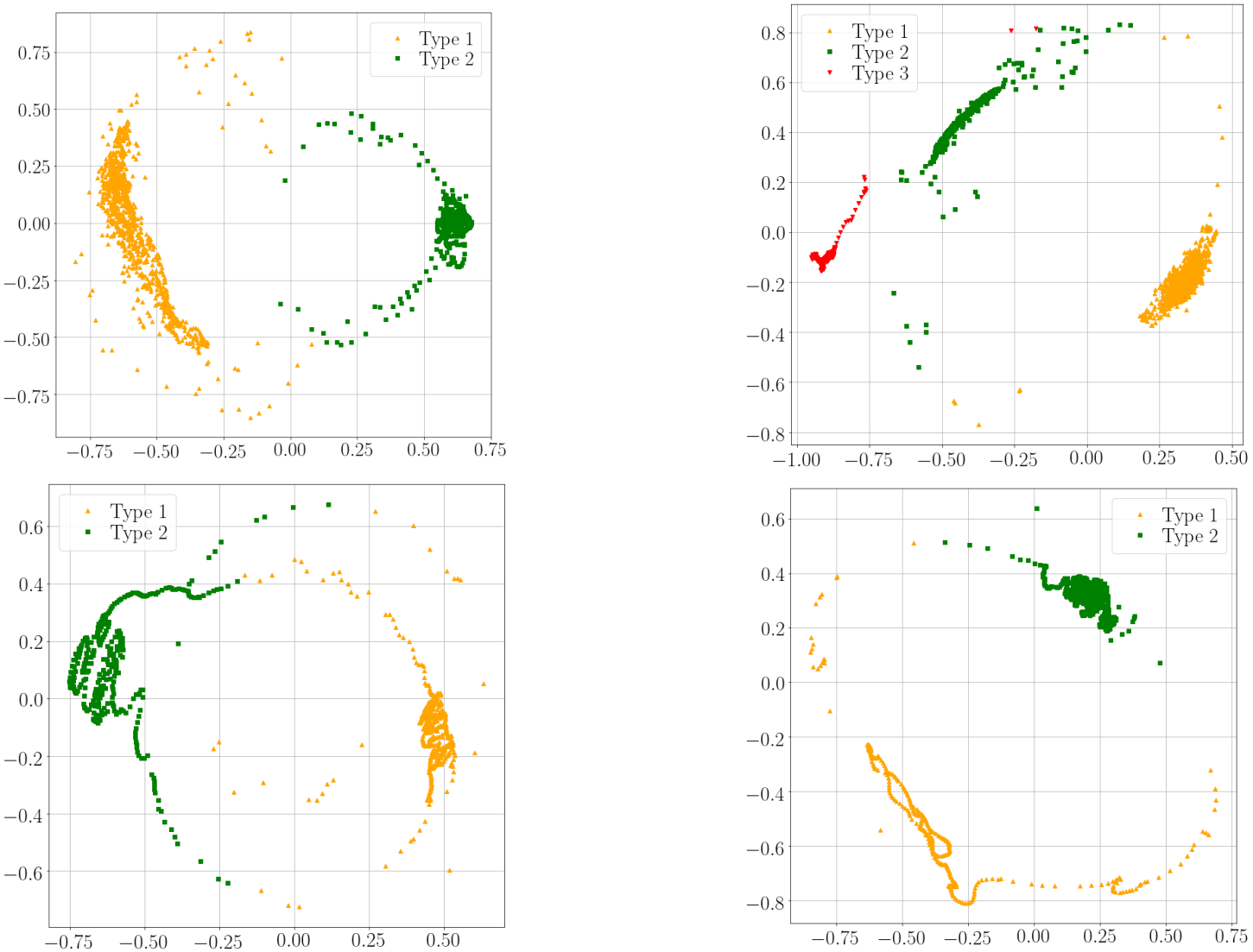
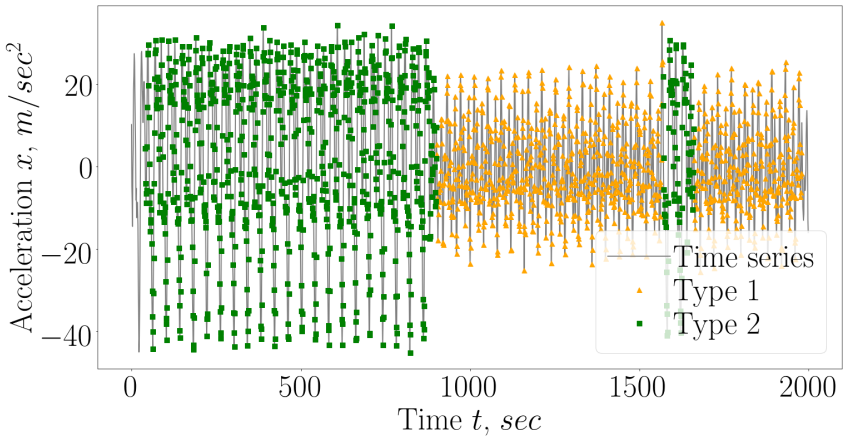
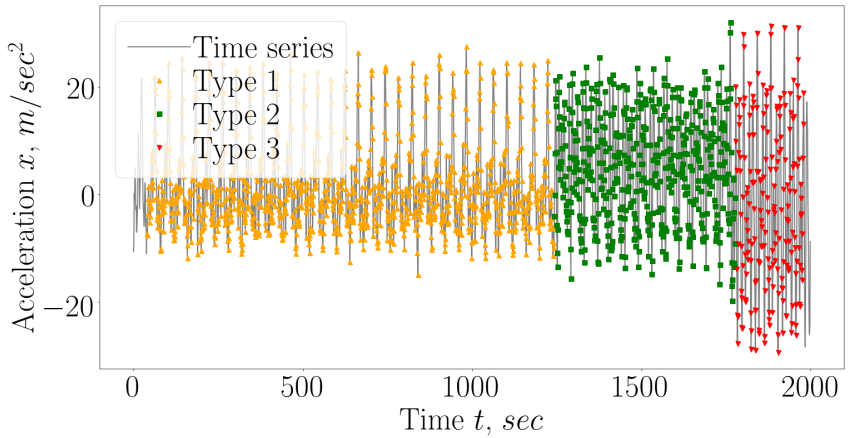
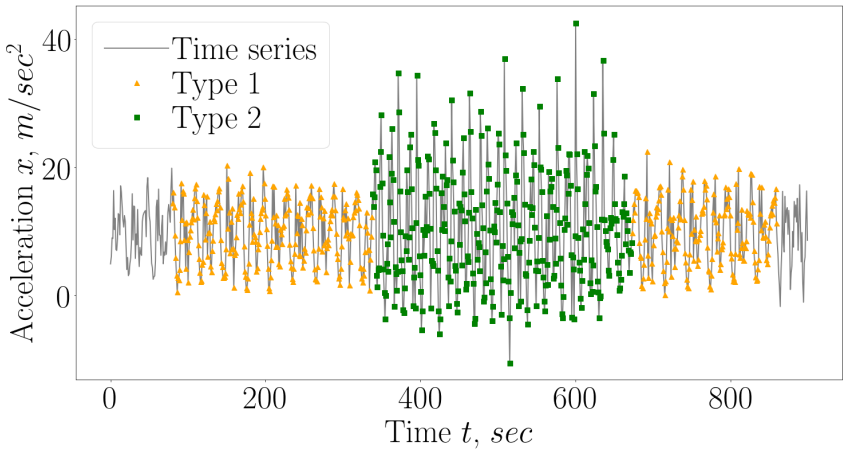


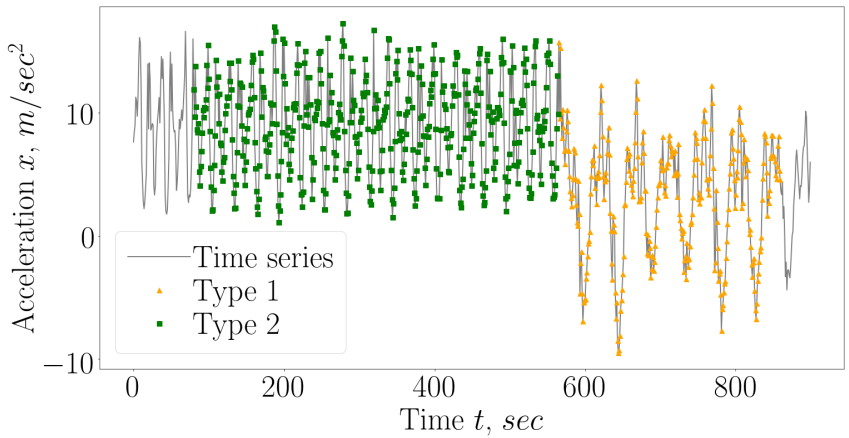
Иллюстрация проекции признакового описания точек временного ряда на плоскости для временных рядов, построенных синтетически, а также при помощи мобильного акселерометра.

Кластеризация точек временного ряда









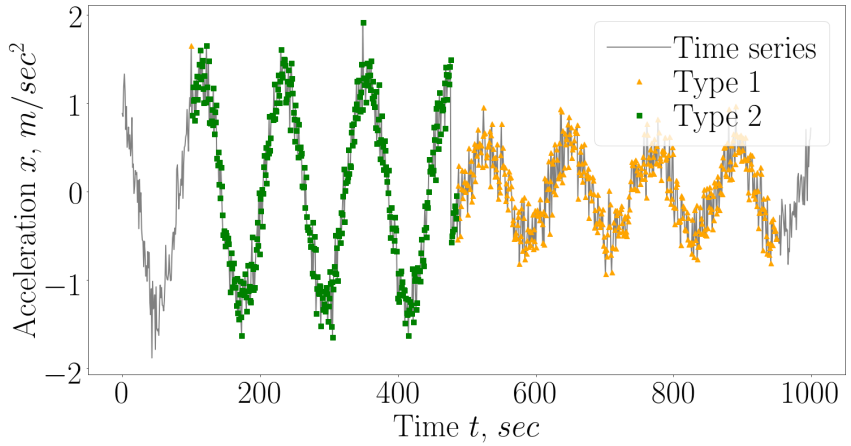
Результат кластеризации точек временных рядов, построенных синтетически, а также при помощи мобильного акселерометра.

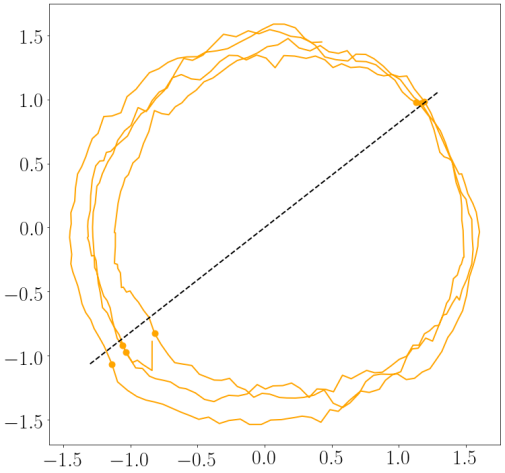
Результаты работы алгоритма кластеризации

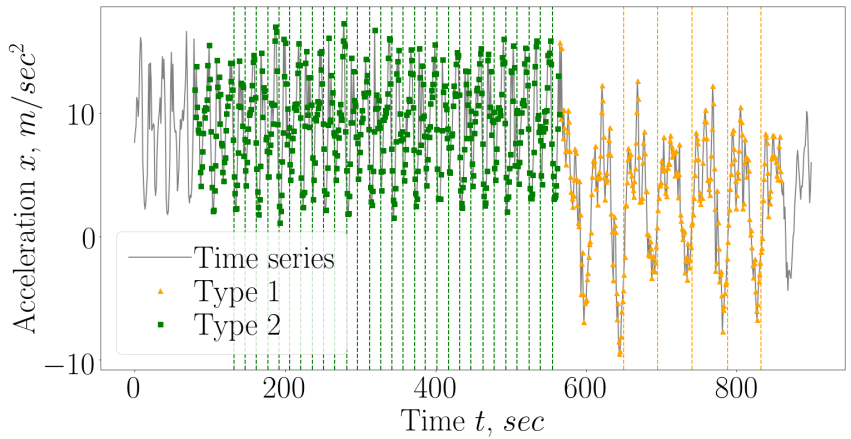
Ряд, $\mathbf{x}$	Длина, N	Сегментов, K	Длина, T	Ошибка, S
Phys. Motion 1	900	2	40	0.06
Phys. Motion 2	900	2	40	0.03
Synthetic 1	2000	2	20	0.04
Synthetic 2	2000	3	20	0.03

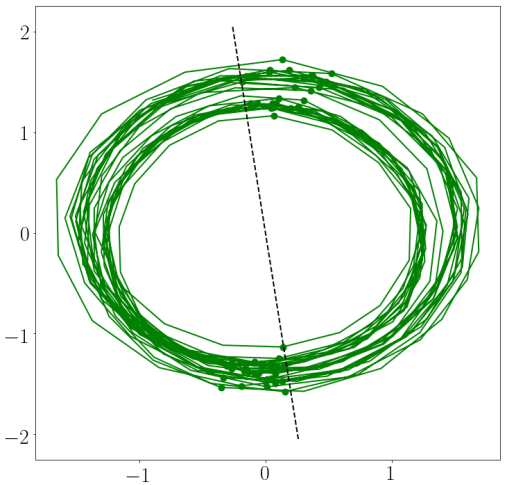
- N — число точек во временном ряде,
- K — число различных действий во временном ряде,
- T — максимальная длина сегмента,
- S — точность кластеризации.

Сегментация временных рядов









Результат сегментации временных рядов, в случае двух синусоидальных сигналов в произвольной частотой и амплитудой, а также в случае реальных данных, полученных при помощи акселерометра.

Заключение

- Предложен алгоритм поиска характерных сегментов, который основывается на методе главных компонент для локального снижения размерности
- Введена функция расстояния между локальными базисами в каждый момент времени, которые интерпретировались как признаковое описание точки временного ряда. Данная функция является метрикой.
- В ходе эксперимента, на реальных показаниях акселерометра, а также на синтетических данных, было показано, что предложенный метод измерение расстояния между базисами хорошо разделяет точки которые принадлежат различным действиям, что приводит к хорошей кластеризации объектов.
- Также в эксперименте была проведена полная сегментация временных рядов для каждого кластера по отдельности.

- Планируется решить задачу нахождения минимального размера фазового пространства, для которого фазовая траектория не имеет самопересечений.

Публикации

- *Грабовой А. В., Стрижов В. В.* Анализ свойств локальных моделей в задачах кластеризации квазипериодических временных рядов // (в процессе)
- *Грабовой А. В., Бахтеев О. Ю., Стрижов В. В.* Определение релевантности параметров нейросети // Информатика и ее применения, 2019, 13(2).
- *Гадаев Т. Т., Грабовой А. В., Мотренко А. П., Стрижов В. В.* Численные методы оценки объема выборки в задачах регрессии и классификации //(в процессе)
- *Бучнев Т. Т., Грабовой А. В., Гадаев Т. Т., Стрижов В. В.* Ранее прогнозирование достаточного объема выборки для обобщенно линейной модели // (в процессе)