

УДК: 51-76, 519.25

Поиск точек разладки в биометрических данных: ретроспективные непараметрические методы сегментации на основе динамического программирования и скользящих окон

А. И. Шестопёров^{1,а}, А. В. Ивченко^{1,2}, Е. В. Фомина²

¹Институт прикладной математики им. М. В. Келдыша РАН,
Россия, 125047, г. Москва, Миусская пл., д. 4

²ГНЦ РФ – Институт медико-биологических проблем РАН,
Россия, 123007, г. Москва, Хорошёвское ш., д. 76а

E-mail: ^а alex.shestopero@yandex.ru

Получено 06.07.2024, после доработки – 18.08.2024.

Принято к публикации 29.08.2024.

Работа посвящена анализу медико-биологических данных, получаемых с помощью локомоторных тренировок и тестирований космонавтов, проводимых как на Земле, так и во время полета. Данные эксперименты можно описать как движение космонавта по беговой дорожке согласно прописанному регламенту в различных скоростных режимах, во время которых не только записывается скорость, но и собирается ряд показателей, включающих частоту сердечных сокращений, величину давления на опору и пр. С целью анализа динамики состояния космонавта на протяжении длительного времени, для независимой оценки целевых показателей необходимо проводить качественную сегментацию режимов его движения. Особую актуальность данная задача приобретает при разработке автономной системы жизнеобеспечения космонавтов, которая будет действовать без сопровождения персонала с Земли. При сегментации целевых данных сложность заключается в наличии различных аномалий, включая отход испытуемого от заранее прописанного регламента, переходы между режимами движения произвольного вида и длительности, аппаратные сбои и пр. Статья включает в себя подробный обзор ряда современных ретроспективных (оффлайн) непараметрических методов поиска многократных разладок во временном ряде, где под разладкой понимается резкое изменение свойств наблюдаемого ряда, происходящее в неизвестный заранее момент времени. Особое внимание уделено алгоритмам и статистическим показателям, которые определяют степень однородности данных, а также способам поиска точек разладки. В данной работе рассматриваются подходы, основанные на методах динамического программирования и скользящего окна. Вторая часть статьи посвящена численному моделированию представленных методов на характерных примерах экспериментальных данных, включающих как простые, так и сложные скоростные профили движения. Проведенный анализ позволил выделить методы, которые в дальнейшем будут проанализированы на полном корпусе данных. Предпочтение отдается методам, обеспечивающим близость разметки к заданному эталону, потенциально позволяющим детектировать обе границы переходных процессов, а также обладающим робастностью относительно внутренних параметров.

Ключевые слова: космическая медицина, локомоторное тестирование, временные ряды, точка разладки, сегментация, непараметрический, ретроспективный, динамическое программирование, скользящее окно

Работа выполнена при поддержке гранта РФФИ, проект № 23-71-10055.

UDC: 51-76, 519.25

Changepoint detection in biometric data: retrospective nonparametric segmentation methods based on dynamic programming and sliding windows

A. I. Shestoperov^{1,a}, A. V. Ivchenko^{1,2}, E. V. Fomina²

¹Keldysh Institute of Applied Mathematics of RAS,
4 Miusskaya sq., Moscow, 125047, Russia

²State Research Center – Institute of Biomedical Problems of RAS,
76a Khoroshevskoe sh., Moscow, 123007, Russia

E-mail: ^a alex.shestoperov@yandex.ru

*Received 06.07.2024, after completion – 18.08.2024.
Accepted for publication 29.08.2024.*

This paper is dedicated to the analysis of medical and biological data obtained through locomotor training and testing of astronauts conducted both on Earth and during spaceflight. These experiments can be described as the astronaut's movement on a treadmill according to a predefined regimen in various speed modes. During these modes, not only the speed is recorded but also a range of parameters, including heart rate, ground reaction force, and others, are collected. In order to analyze the dynamics of the astronaut's condition over an extended period, it is necessary to perform a qualitative segmentation of their movement modes to independently assess the target metrics. This task becomes particularly relevant in the development of an autonomous life support system for astronauts that operates without direct supervision from Earth. The segmentation of target data is complicated by the presence of various anomalies, such as deviations from the predefined regimen, arbitrary and varying duration of mode transitions, hardware failures, and other factors. The paper includes a detailed review of several contemporary retrospective (offline) nonparametric methods for detecting multiple changepoints, which refer to sudden changes in the properties of the observed time series occurring at unknown moments. Special attention is given to algorithms and statistical measures that determine the homogeneity of the data and methods for detecting change points. The paper considers approaches based on dynamic programming and sliding window methods. The second part of the paper focuses on the numerical modeling of these methods using characteristic examples of experimental data, including both "simple" and "complex" speed profiles of movement. The analysis conducted allowed us to identify the preferred methods, which will be further evaluated on the complete dataset. Preference is given to methods that ensure the closeness of the markup to a reference one, potentially allow the detection of both boundaries of transient processes, as well as are robust relative to internal parameters.

Keywords: aerospace medicine, locomotion test, time series, changepoint, segmentation, nonparametric, retrospective, dynamic programming, sliding window

Citation: *Computer Research and Modeling*, 2024, vol. 16, no. 5, pp. 1295–1321 (Russian).

The work was supported by grant of the Russian Science Foundation, project No. 23-71-10055.

1. Введение

На сегодняшний день все более актуальной становится проблема освоения дальнего космоса. При длительном космическом полете человек адаптируется к условиям невесомости [Shelhamer et al., 2020; Norsk, 2020; Hagio et al., 2022]. В его организме происходят гипогравитационные нарушения, изменяется поведение кардиореспираторной, костной, двигательной и мышечной систем, влиянию подвергается пространственная ориентация [Фомина, Кукоба, 2019].

Высокая автономность будущих экспедиций ставит задачу создания системы медицинского обеспечения без сопровождения специалистов с Земли. Для ее решения большим подспорьем служат данные, полученные с международной космической станции, поскольку уже там проявляет себя большинство негативных эффектов межпланетного полета. В основу разработки его медико-биологического обеспечения и ложатся результаты анализа полученных данных [Фомина и др., 2020].

Подготовка к освоению дальнего космоса поставила новые задачи в части совершенствования системы медико-биологического обеспечения межпланетных полетов. Качество выполнения операторских задач на поверхности планеты будет являться залогом успешности всей миссии. Состояние здоровья космонавта после длительного полета определяется эффективностью профилактических мероприятий, направленных на купирование всего комплекса рисков таких миссий. Уроки орбитальных полетов должны лечь в основу системы медицинского обеспечения межпланетных миссий, так как основной комплекс негативных факторов межпланетного полета уже моделируется в условиях МКС. Переход к персонифицированным мероприятиям, направленным на профилактику негативных влияний факторов космического полета, должен осуществляться с привлечением методов машинного обучения, что позволит быстро анализировать информацию об эффективности выполненных мероприятий, ответе физиологических систем на предлагаемое воздействие и реализовать автоматизированную систему поддержки принятия решений в автономном полете.

Для купирования рисков здоровью, возникающих в течение полета, на Земле космонавт выполняет персональные тренировки, которые корректируются согласно его физиологическим показателям. Последние анализируются с помощью различных методик, одной из которых является локомоторное тестирование [Фомина и др., 2016; Фомина и др., 2021]. При локомоторном тестировании космонавт, в соответствии с априорно прописанным временным регламентом, меняет скоростные режимы своего движения по беговой дорожке, происходит сбор его биометрических данных. Корректная идентификация скоростных режимов космонавта является важным фактором при обработке других целевых показателей, например ЧСС. Также указанная процедура предвосхищает построение профилей шага, характерных для каждого режима перемещения.

Представленная статья видит своей задачей сегментацию скоростных профилей движения космонавта, сформированных в рамках локомоторного тестирования. В каждом отдельном случае сегментация может быть осуществлена за счет эмпирического подбора граничных значений скоростей. Однако указанный подход оказывается неэффективным (а возможно, и приводящим к неправильным заключениям) при наличии сложных профилей. К сложным в работе причисляются профили, содержащие длительные процессы выхода на рабочую скорость, сочетающие участки произвольной длительности с ускоренным движением и движением с постоянной средней скоростью, в связи с чем стационарные участки данных являются трудно локализуемыми. К таким типам профилей могут приводить отсутствие строгого следования регламенту эксперимента, постепенность и изменчивость характера перехода между прописанными режимами движения, вынуждающие выделять переходные процессы в отдельную категорию, а также потенциальное наличие технических неполадок, приводящих к искажению или отсутствию сигнала с одного или нескольких датчиков. В таком случае на помощь приходят математические

методы поиска точек разладки временных рядов. К настоящему моменту разработано огромное количество алгоритмов сегментации временных рядов, которые условно делятся по двум основным признакам: онлайн-алгоритмы [Tartakovsky, Nikiforov, Basseville, 2014; Xie et al., 2021; Desobry, Davy, Doncarli, 2005] и офлайн-алгоритмы (ретроспективные) [Truong, Oudre, Vayatis, 2020], параметрические [Picard et al., 2011; Frick, Munk, Sieling, 2014; Wang, Samworth, 2018] и непараметрические [Brodsky, Darkhovsky, 1993] алгоритмы.

В статье анализируется возможность применения некоторых ретроспективных непараметрических методов поиска неизвестного числа точек разладки к сегментации профилей скоростей движения. Указанный класс методов выбран в соответствии с требованиями прикладной задачи. Ретроспективные методы применяются в ситуациях, когда в распоряжении имеется весь массив исследуемых данных. В отличие от параметрических непараметрические методы не требуют априорных знаний о распределении данных, что может быть полезно при сложной нестационарной природе последних.

В основании любой процедуры поиска многократных разладок лежит оптимизационная задача, решать которую можно различными способами. В текущей статье будут описаны и апробированы на реальных данных методы двух типов: первые — оптимальные методы, основанные на методе динамического программирования [Беллман, 1960; Auger, Lawrence, 1989; Jackson et al., 2005]; вторые — приближенные методы, использующие концепцию скользящих окон [Truong, Oudre, Vayatis, 2020]. Большинство изложенных методов работает с многомерными данными, что в перспективе позволит обрабатывать несколько биометрических показателей одновременно (например, частоты сердечных сокращений (ЧСС), величины давления на опору беговой дорожки и т. п.).

Статья имеет следующую структуру. § 1 — введение, объясняющее актуальность решаемой задачи. Постановка задачи помещена в § 2. В нем также описывается характер профилей движения, с которыми приходится сталкиваться на практике. § 3 содержит описание ряда методов поиска точек разладки, которые впоследствии (§ 4) используются для сегментации скоростных профилей движения космонавта по беговой дорожке. Основным выводам посвящено заключение представленной статьи.

2. Описание исследуемых данных и постановка задачи

Прежде чем перейти к постановке задачи, с помощью ряда примеров создадим представление о характерных профилях скорости движения космонавта по беговой дорожке, подлежащих сегментации. Стандартный профиль скорости, в данной работе называемый простым, изображен на рис. 1. Согласно регламенту локомоторного тестирования допускаются режимы движения как с постоянной скоростью, так и с ускорением. В данном примере представлены режимы движения с постоянной скоростью 3 км/ч и 6 км/ч, а также бег с ускорением до достижения скорости в 17 км/ч. Переход из режима в режим относительно быстрый, т. е. размер интервалов, на которых скорость меняется от 3 км/ч до 6 км/ч или обратно, меньше, нежели размер интервалов с постоянной средней скоростью. Кроме того, форма таких переходов может быть аппроксимирована прямой. Как правило, их форма является следствием режима проведения эксперимента, при котором полотно дорожки приводится в движение мотором. Во втором режиме эксперимента мотор выключен, и космонавт самостоятельно выходит на рабочую скорость движения, отслеживая ее на мониторе, что приводит к более сложным профилям движения. Режимы проведения эксперимента могут перемежаться в одном и том же тестировании или во время одной и той же тренировки космонавта.

Необходимо отметить, что для локализации режимов движения в приведенном простом примере достаточно ввести эмпирическое правило с верхними и нижними границами скоростей, а для выделения переходных процессов — оценить распределение их длин и отмечать, например,

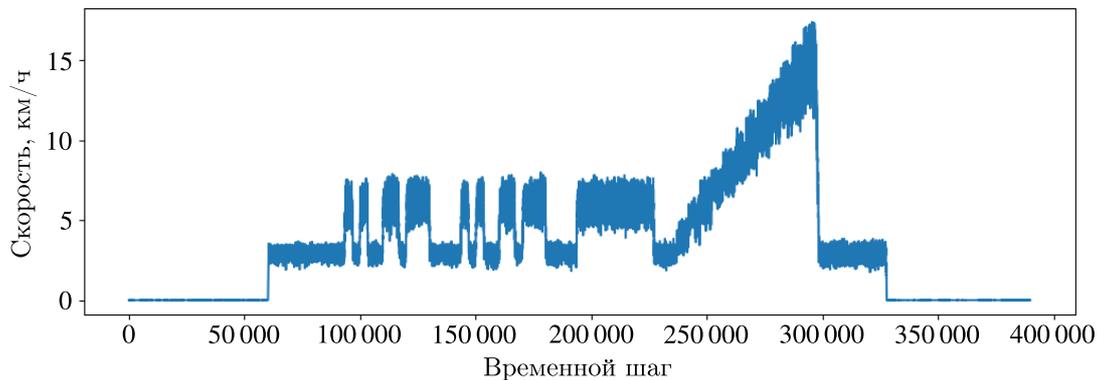


Рис. 1. Профиль скорости движения космонавта во время локомоторного тестирования, аномалии при переходных процессах отсутствуют. Шаг временного отчета — 0,003 с

интервалы длиной в два медианных значения между режимами. Однако уже при анализе режима ускоренного движения, начинающегося в окрестности точки 230 000 на рис. 1, выделение ступенек и интервалов-переходов между ними становится менее очевидной задачей.

Рассмотрим примеры экспериментальных данных, содержащих сложные переходные процессы и аномалии различной природы. На рис. 2, *а* представлены высокоамплитудные осцилляции относительно сохраняющейся средней скорости 7 км/ч, затем в окрестности точки 27 500 наблюдается ее резкий спад до 4 км/ч (возможно, испытуемый споткнулся), далее находится участок движения с ускорением вплоть до точки 37 000, в ходе которого испытуемый заново выходит на рабочую скорость движения.

Здесь мы наблюдаем незапланированное нарушение скоростного режима, приведшее к возникновению участка нестационарности.

На рис. 2, *б* представлен мгновенный переход с бега на 13 км/ч к нулевой скорости, после чего идет выход на рабочую скорость в 8 км/ч через два плато и три интервала ускоренного движения. В окрестности точки 216 000 и далее представлены два интервала с ненулевым ускорением (точки 216 000–224 000, 224 000–234 000) и переход в шаг на 5 км/ч, что можно интерпретировать как завершение эксперимента, в ходе которого испытуемый уже вне скоростного регламента постепенно замедляется.

На рис. 2, *в* представлена часть режима с бегом на 5 км/ч до точки 166 000, далее располагается участок бега с ускорением для набора рабочей скорости в 10 км/ч, бег в данном режиме (при несоблюдении сохранения постоянной средней скорости), далее в окрестности точки 171 500 расположен нестационарный участок с замедлением и ускорением движения, и в окрестности точки 173 000 происходит выход на режим бега на 8 км/ч (в ходе которого также не соблюдено сохранение средней скорости).

На рис. 2, *г* представлены множественные потери сигнала (скорость равна нулю), кроме того, в окрестностях точек 35 000, 62 000 и 102 000 зафиксирована скорость ниже нуля. Данное поведение вызвано техническими неполадками.

На рис. 2, *д* представлен режим бега на скорости 5 км/ч, который на интервале с 174 000 до 188 000 отсчета прерывается нестационарным движением с двумя спадами скорости до нуля, участками ускоренного и замедленного бега. В отличие от профиля на рис. 2, *з*, где переход в нулевую скорость происходит мгновенно, здесь переход имеет наклон и может быть вызван остановкой испытуемого и дальнейшим продолжением бега.

На рис. 2, *е* представлен режим бега на скорости 10 км/ч, располагающийся в интервале 11 800–150 000. Перед ним расположен участок из комбинаций плато и ускоренного движения,

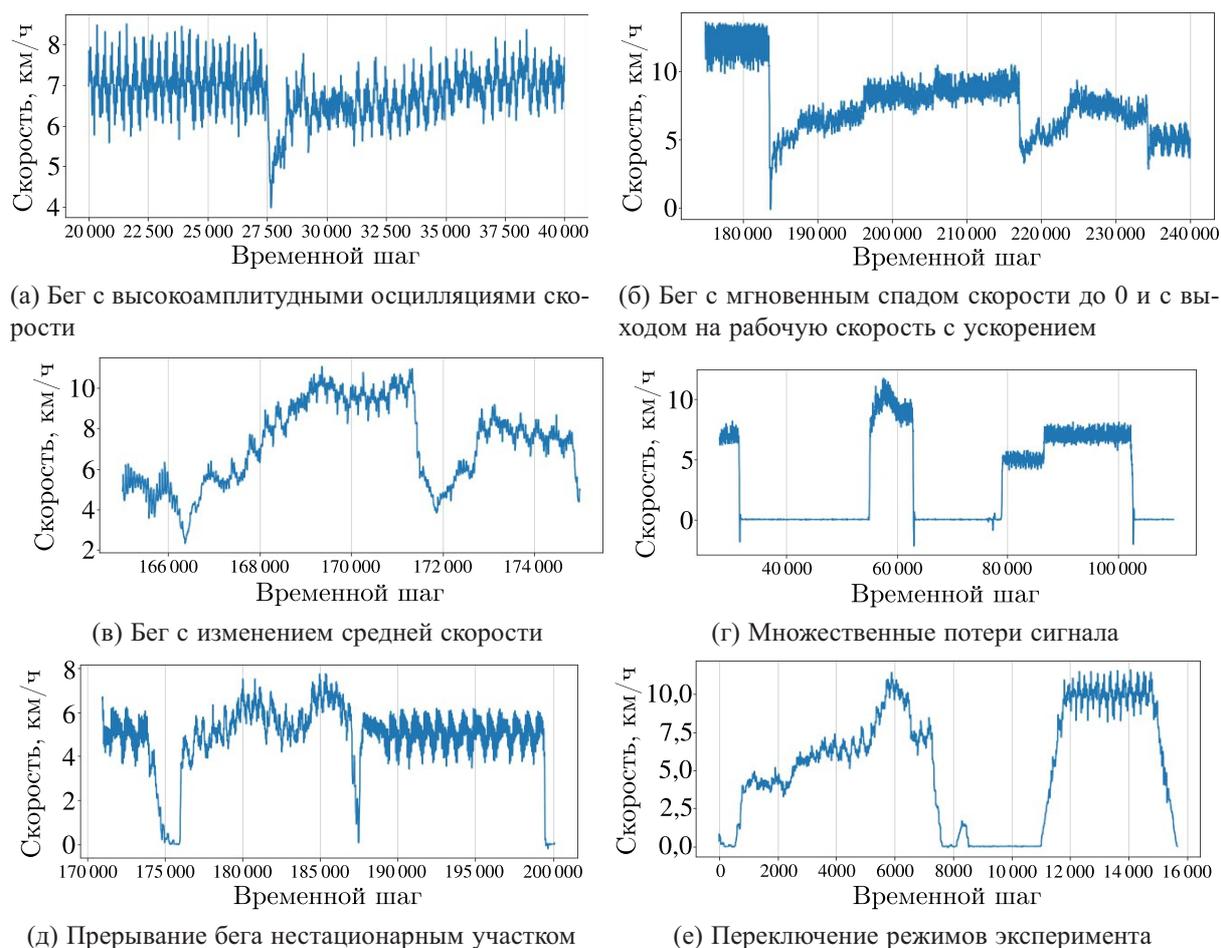


Рис. 2. Примеры переходных процессов и аномалий различной природы в скоростных профилях движения космонавтов. Шаг временного отчета — 0,003 с

далее следует интервал с нулевой скоростью с относительно коротким повышением в окрестности 8200. Учитывая наличие наклона при переходе с нулевой скорости и возвращении к ней, данный профиль можно характеризовать как бег вне регламента эксперимента, затем остановку движения и включение мотора, приводящего дорожку в движение.

Таким образом, при наличии разного рода трудно интерпретируемых переходных процессов движения, локализация протокольных границ режимов движения требует решения математической задачи о поиске точек разладки временного ряда, которая в условиях текущей работы формулируется следующим образом.

Задан упорядоченный набор наблюдений

$$\{x_i = (x_{i1}, \dots, x_{id})^T \in \mathbb{R}^d, i = \overline{1, n}\} \quad (1)$$

независимых случайных величин $\{X_i \in \mathbb{R}^d\}_{i=1}^n$ с неизвестными распределениями $\{\mathbb{P}_i\}_{i=1}^n$. Для чисел $a \in \mathbb{N}_0$ и $b \in \mathbb{N}$, $a < b$, обозначим $\llbracket a, b \rrbracket \triangleq [a, b] \cap \mathbb{N}_0$, где $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. Упорядоченный набор из $(M - 1)$ индексов $\tau = \{\tau_m\}_{m=1}^{M-1} \subset \llbracket a, b \rrbracket$, $a = \tau_0 < \tau_1 < \dots < \tau_{M-1} < \tau_M = b$, назовем *множеством точек разладки*, или *сегментацией* на $M_\tau \triangleq M$ частей, если для него выполняется условие $\mathbb{P}_i \neq \mathbb{P}_{i+1}$, $i = \overline{1, n-1}$, тогда и только тогда, когда $i \in \{\tau_m\}_{m=1}^{M-1}$. Множество

$$\mathcal{T}^M \triangleq \{\tau = (\tau_0, \dots, \tau_M) \in \mathbb{N}^{M+1} : a = \tau_0 < \tau_1 < \dots < \tau_{M-1} < \tau_M = b\} \quad (2)$$

назовем *множеством сегментаций* на M частей, где $\tau_0 = a$ и $\tau_M = b$. С помощью $\tau \in \mathcal{T}^M$ данные разделяются на M сегментов

$$\mathbf{x}_{\tau_{i-1}+1:\tau_i} \triangleq \left\{ \mathbf{x}_{\tau_{i-1}+1}, \dots, \mathbf{x}_{\tau_i} \right\}, \quad i = \overline{1, M}. \quad (3)$$

§ 3 посвящен изложению нескольких современных *ретроспективных непараметрических* методов локализации *многократных* точек разладки. Методы, описанные в подпараграфах 3.1–3.3, опираются на методологию динамического программирования (ДП), в то время как в подпараграфах 3.4 и 3.5 для поиска точек разладки используется метод скользящих окон. В случае заранее заданного числа разладок рекурсивная процедура ДП приводит к единственной оптимальной сегментации временного ряда. Введение штрафных функций позволяет распространить указанную методику на случай неизвестного числа точек разладки. Метод скользящих окон – приближенный метод поиска многократных разладок, суть которого заключается в оценке меры расхождения между выборками данных на двух смежных окнах, сканирующих набор наблюдений (1). Разладка детектируется, если величина используемой аппроксимирующей статистики превосходит заданное пороговое значение. В [Truong, Oudre, Vayatis, 2020] содержатся псевдокоды, описывающие обе указанные процедуры (алгоритмы 1 и 3 соответственно).

3. Обзор ретроспективных непараметрических методов сегментации

3.1. Методы NMCD и ED-PELT

Для набора *одномерных* наблюдений в [Zou et al., 2014] был разработан непараметрический метод NMCD (Nonparametric Multiple Change-point Detection), в основу которого положен интегральный критерий качества, сформулированный в терминах эмпирических функций распределения сегментов данных. Поиск точек разладки осуществлялся с помощью байесовского информационного критерия (BIC) [Schwarz, 1978]. Метод ДП, призванный минимизировать соответствующий функционал качества, реализовывался на наборе наблюдений, предварительно отобранных сканирующей процедурой.

Пусть в (1) $d = 1$ и набор данных разделен на $m + 1$ сегмент (3) точками разладки τ_i , $i = \overline{1, m}$, причем $0 = \tau_0 < \tau_1 < \tau_2 < \dots < \tau_m < \tau_{m+1} = n$. В [Zou et al., 2014] был построен функционал качества

$$c_{NMCD}(\mathbf{x}_{\tau_{i-1}+1:\tau_i}) = - \int_{-\infty}^{+\infty} L(\mathbf{x}_{\tau_{i-1}+1:\tau_i} | t) dw(t) \quad (4)$$

сегмента $\mathbf{x}_{\tau_{i-1}+1:\tau_i}$. Дифференциал весовой функции $w(t)$ определялся через функцию распределения всего временного ряда $F(t)$ соотношением $dw(t) = (\widehat{F}(t)(1 - \widehat{F}(t)))^{-1} d\widehat{F}(t)$, а логарифмическая функция правдоподобия

$$L(\mathbf{x}_{\tau_{i-1}+1:\tau_i} | t) = (\tau_i - \tau_{i-1}) \times (\widehat{F}_i(t) \log \widehat{F}_i(t) + (1 - \widehat{F}_i(t)) \log(1 - \widehat{F}_i(t)))$$

строилась на основе эмпирической функции распределения $\widehat{F}_i(t)$ сегмента $\mathbf{x}_{\tau_{i-1}+1:\tau_i}$:

$$\widehat{F}_i(t) = \frac{1}{\tau_i - \tau_{i-1}} \left(\sum_{j=\tau_{i-1}+1}^{\tau_i} \mathbf{1}\{x_j < t\} + \frac{\mathbf{1}\{x_j = t\}}{2} \right).$$

Итоговый функционал качества на наборе (1) определялся суммой слагаемых (4) по $m + 1$ сегменту данных.

Для нахождения оптимальной сегментации при фиксированном числе разладок m минимизировалась целевая функция [Zou et al., 2014]

$$Q_{NMCD}(\tau_{1:m} | x_{1:n}) = -n \sum_{i=1}^{m+1} \sum_{t=1}^n (\tau_i - \tau_{i-1}) \times \frac{\widehat{F}_i(t) \log \widehat{F}_i(t) + (1 - \widehat{F}_i(t)) \log(1 - \widehat{F}_i(t))}{(t - \frac{1}{2})(n - (t - \frac{1}{2}))}, \quad (5)$$

полученная в результате аппроксимации интегралов (4) по всем точкам временного ряда. При неизвестном m задача сегментации формулировалась в терминах ВИС:

$$\text{BIC} = \min_{\tau_1, \dots, \tau_m} \{Q_{NMCD}(\tau_{1:m} | x_{1:n}) + m\xi_n\}, \quad (6)$$

где $Q_{NMCD}(\tau_{1:m} | x_{1:n})$ определена в (5) и ξ_n — последовательность, стремящаяся к бесконечности. В [Zou et al., 2014] рекомендовалось брать $\xi_n = \frac{(\log n)^{2+c}}{2}$ с $c = 0,1$. Для минимизации целевой функции из (6) авторы [Zou et al., 2014] использовали алгоритм SNS (Segment Neighbourhood Search) [Auger, Lawrence, 1989], основанный на методе ДП, для всех значений m в диапазоне от 1 до некоторого заранее заданного числа M . Последнее характеризует максимально допустимое число точек разладки во временном ряде. При указанной схеме расчета алгоритм сегментации обладал сложностью $O(n^3 + Mn^2)$, что становится проблемой при продолжительном периоде наблюдений.

Для ускорения вычислений в [Zou et al., 2014] была предложена процедура, которая, предваряя процесс оптимизации функционала качества (6), сканировала временной ряд на предмет отсеивания наблюдений, малозначимых в контексте вероятности возникновения разладки. Она включала в себя следующие шаги.

1. Задавались скользящее окно размером $2N_I$ и пустое множество индексов потенциальных точек разладки. Рекомендовалось брать $N_I = \lceil 0,5 \cdot (\log n)^{1,5} \rceil$, где $\lceil x \rceil$ — наименьшее целое число, превосходящее x .
2. В процессе скольжения для каждого из $(n - 2N_I + 1)$ окон рассчитывалась двухвыборочная статистика Крамера – фон Мизеса. Полученные значения привязывались к центрам соответствующих окон. Первым $(N_I - 1)$ точкам, а также N_I последним присваивались нулевые значения.
3. Точка добавлялась в множество потенциальных точек разладки, если присвоенное ей значение является максимальным в пределах соответствующего ей окна.

В результате метод оптимизации применялся на полученном множестве потенциальных точек разладки и, следовательно, на пространстве меньшей размерности. Хотя итоговая сегментация и переставала гарантировать глобальную оптимизацию функции затрат. Следует отметить, что изложенный подход подразумевает, что существует по крайней мере одна точка разладки. На практике данный факт требует проверки, которая может быть проведена, в частности, с помощью тестов, предложенных в [Einmahl, McKeague, 2003; Zou et al., 2007].

Иной способ борьбы с высокой вычислительной сложностью алгоритма решения оптимизационной задачи (6) из [Zou et al., 2014] был разработан в [Haynes, Fearnhead, Eckley, 2017]. Во-первых, был пересмотрен способ аппроксимации функционала (4). А именно, интеграл аппроксимировался не по всем точкам ряда, а по $K \ll n$ неравномерно распределенным точкам, положение которых придавало больший вес значениям в хвосте распределения данных. В результате функционал качества сегмента принял вид

$$c_K(x_{\tau_{i-1}+1:\tau_i}) = -\frac{2c}{K} \sum_{k=1}^K L(x_{\tau_{i-1}+1:\tau_i} | t_k). \quad (7)$$

В [Haynes, Fearnhead, Eckley, 2017] полагали $K = \lceil 4 \log n \rceil$. Во-вторых, оптимизационная задача в

$$Q_{\text{PELT}}(x_{1:n} | \xi_n) = \min_{\tau_1, \dots, \tau_m} \left\{ \sum_{i=1}^{m+1} \left(c_K(x_{\tau_{i-1}+1:\tau_i}) + \xi_n \right) \right\}$$

была решена благодаря использованию PELT-алгоритма (Pruned Exact Linear Time) [Killick, Fearnhead, Eckley, 2012]. Возможность его применения доказана авторами для функционалов качества (7), записанных в терминах эмпирических функций распределения. По этой причине алгоритм был назван ED-PELT (Empirical Distribution PELT).

Одним из недостатков ED-PELT по сравнению с NMCD является то, что ED-PELT производит единственную сегментацию, которая оптимальна для заранее выбранного значения штрафа ξ_n . Однако априорное задание функции штрафа может ухудшить качество сегментации реальных данных. В [Haynes, Eckley, Fearnhead, 2017] был предложен метод CROPS (Changepoints over a Range Of Penalties), который определяет оптимальные сегментации временного ряда для нескольких значений штрафа, лежащих в заданном диапазоне. Лучшая из них выбирается с помощью эмпирического метода, предложенного в [Lavielle, 2005], суть которого состоит в следующем. График зависимости минимальных значений статистики T от числа точек разладки M разбивался на два участка — быстрого и медленного убывания — при различных числах M . На каждом из них данные аппроксимировались прямой с помощью метода наименьших квадратов. При искомом числе M сумма остаточных сумм квадратов, вычисленных для обеих частей графика, должна быть минимальной.

Предложенная в [Zou et al., 2014] процедура была применена для выявления изменений в структуре изохор [Oliver et al., 2004]. В [Haynes, Fearnhead, Eckley, 2017] решалась задача обнаружения изменений частоты сердечных сокращений во время бега.

3.2. Method MultiRank

В [Lung-Yut-Fong, Lévy-Leduc, Carré, 2015] локализация точек разладки производилась с помощью ранговой статистики. Точнее, строился ранг j -го члена x_{jk} , $j = 1, n$, в наборе k -х компонент $\{x_{1k}, \dots, x_{nk}\}$ наблюдений из (1), т. е. $R_j^{(k)} = \sum_{i=1}^n \mathbf{1}\{x_{ik} \leq x_{jk}\}$. При условии, что данные имеют $M \geq 0$ точек разладки, их положение выявлялось с помощью тестовой статистики

$$T(\tau_1, \dots, \tau_{M-1}) = \frac{4}{n^2} \sum_{m=0}^{M-1} (n_{m+1} - n_m) \bar{\mathbf{R}}_m^T \Sigma_n^{-1} \bar{\mathbf{R}}_m, \quad (8)$$

которая позволила распространить на многомерный случай классический критерий Краскела – Уоллиса, используемый для одномерных наблюдений и являющийся обобщением критерия Уилкоксона – Манна – Уитни на случай более двух выборок. В (8) вектор-столбец $\bar{\mathbf{R}}_m = \left[\bar{R}_m^{(k)} - \frac{n+1}{2} \right]_{k=1}^d$ с $\bar{R}_m^{(k)} = (n_{m+1} - n_m)^{-1} \sum_{j=n_m+1}^{n_{m+1}} R_j^{(k)}$ — среднее значение ранга в m -м сегменте данных, а через $\Sigma_n = \left[\Sigma_n^{k,k'} \right]_{k,k'=1}^d$ обозначена d -мерная эмпирическая ковариационная матрица с компонентами

$$\Sigma_n^{k,k'} = \frac{4}{n} \sum_{i=1}^n \left(\widehat{F}_{nk}(x_{ik}) - \frac{1}{2} \right) \left(\widehat{F}_{nk'}(x_{ik'}) - \frac{1}{2} \right), \quad 1 \leq k, k' \leq d,$$

где $\widehat{F}_{nk}(t) = n^{-1} \sum_{j=1}^n \mathbf{1}\{x_{jk} \leq t\}$ — эмпирическая функция распределения k -й координаты наблюдений.

В [Lung-Yut-Fong, Lévy-Leduc, Carré, 2015] рассматривались три ситуации.

1. M известно, и $M > 0$.

Поскольку матрица Σ_n в (8) является общей для всех сегментов данных, функционал качества имеет аддитивную структуру. Благодаря этому неизвестные местоположения точек изменения оценивались путем решения оптимизационной задачи

$$(\widehat{\tau}_1, \dots, \widehat{\tau}_{M-1}) = \arg \max_{1 \leq \tau_1 < \dots < \tau_{M-1} \leq n} T(\tau_1, \dots, \tau_{M-1})$$

с помощью метода ДП [Kau, 1993; Bai, Perron, 2003].

2. M неизвестно, и $M > 0$.

Для оценки числа точек разладки была использована та же эвристическая процедура [Lavielle, 2005], что была использована в рамках алгоритма ED-PELT (подпараграф 3.1).

3. M неизвестно, и $M \geq 0$.

В этом случае проводилась дополнительная проверка на наличие хотя бы одной потенциальной точки разладки, которая заключается в определении степени значимости наибольшего значения модифицированной статистики (8) путем вычисления соответствующего асимптотического p -значения.

Предложенный алгоритм был апробирован на реальном наборе данных микрочипов, а именно aCGH-сигнала [Vert, Bleakley, 2010], который состоит из записей вариаций числа копий ДНК. В рамках задачи сегментации были обнаружены участки аномальных изменений данного показателя, характерные при наличии рака у пациента.

3.3. Методы КСР и Kernseg

В работе [Narchaoui, Carré, 2007] был предложен ядерный метод сегментации данных, ключевая идея которого состояла в том, чтобы перейти от поиска точек разладки в терминах распределений выборок на наборе наблюдений (1) к обнаружению изменений средних значений случайных векторов в некотором гильбертовом пространстве с воспроизводящим ядром (RKHS). В [ArLOT, Celisse, Narchaoui, 2019] он был развит благодаря введению новой теоретически обоснованной функции штрафа, позволившей определять число потенциальных точек разладки. Итоговая ретроспективная непараметрическая процедура получила название КСР (Kernel-based Change-Point). Наряду с классическими пространствами \mathbb{R}^d , $d \geq 1$, в качестве исходных данных КСР допускает наборы графов, строк текстов, последовательностей ДНК и графических изображений. Следует отметить, что алгоритм КСР можно рассматривать как ядерный аналог метода, предложенного в [Lebarbier, 2005] и призванного обнаруживать точки изменения среднего значения сигнала, искаженного гауссовским шумом.

Помимо последовательности *независимых* случайных величин X_1, \dots, X_n , имеющих распределения \mathbb{P}_i , $i = \overline{1, n}$, для работы алгоритма КСР задаются следующие объекты.

- А. Положительно полуопределенное ядро $k: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$, т.е. такая измеримая функция, что $\forall \mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^d$, из (1), матрица Грамма $\mathbf{K} = [\mathbf{K}_{ij}]_{i,j=1}^n \triangleq [k(\mathbf{x}_i, \mathbf{x}_j)]_{i,j=1}^n$ является положительно полуопределенной.
- Б. Константа M_{\max} , определяющая максимальное число точек разладки временного ряда, допускаемое алгоритмом. Со ссылкой на эксперимент авторами [ArLOT, Celisse, Narchaoui, 2019] рекомендовано брать $M_{\max} \approx \frac{n}{\log n}$.

В. Константы $c_1, c_2 > 0$, характеризующие модель функции штрафа. Для их поиска предлагается [Arlot, Celisse, Narchaoui, 2019] использовать одну из реализаций эвристики наклона (slope heuristics). Обзор методов, базирующихся на данном подходе, можно найти в работе [Arlot, 2019].

Алгоритм КСР включает в себя два этапа.

Этап 1. Среди сегментаций \mathcal{T}^M из множества (2), где $a = 0$ и $b = n$, с фиксированным числом точек разладки $M = \overline{1, M_{\max}}$ ищется наилучшая

$$\widehat{\tau}(M) \in \arg \min_{\tau \in \mathcal{T}^M} \{\widehat{\mathcal{R}}_n(\tau)\}, \quad (9)$$

в смысле статистики [Narchaoui, Carré, 2007]:

$$\widehat{\mathcal{R}}_n(\tau) = \frac{1}{n} \sum_{l=1}^M C_{\tau_{l-1}, \tau_l},$$

где показатель качества сегмента имеет вид

$$C_{\tau_{l-1}, \tau_l} = \sum_{i=\tau_{l-1}+1}^{\tau_l} k(X_i, X_i) - \frac{1}{\tau_l - \tau_{l-1}} \sum_{i=\tau_{l-1}+1}^{\tau_l} \sum_{j=\tau_{l-1}+1}^{\tau_l} k(X_i, X_j). \quad (10)$$

Далее рассчитывается оптимальное значение статистики $\widehat{\mathcal{R}}_n(\widehat{\tau}(M))$, соответствующее полученному разбиению.

Этап 2. Используя штрафную функцию [Arlot, Celisse, Narchaoui, 2019]

$$\text{pen}(\tau) = \frac{1}{n} \left[c_1 \log \left(\frac{n-1}{M-1} \right) + c_2 M \right],$$

оптимальное число точек разладки определяем как

$$\widehat{M} \in \arg \min_{1 \leq M \leq M_{\max}} \{\widehat{\mathcal{R}}_n(\widehat{\tau}(M)) + \text{pen}(\tau)\}.$$

На выходе имеем искомый набор точек разладки $\widehat{\tau} = \widehat{\tau}(\widehat{M})$.

Решение оптимизационной задачи на этапе 1 обладает высокой вычислительной сложностью. Как и в [Zou et al., 2014] (см. подпараграф 3.1), авторы [Narchaoui, Carré, 2007] осуществили минимизацию (9) с помощью SNS-метода [Auger, Lawrence, 1989]. В алгоритме Kernseg [Celisse et al., 2018] как пространственная, так и временная сложности указанного шага алгоритма КСР были значительно уменьшены благодаря установлению рациональной взаимосвязи между рекурсией ДП и вычислением матрицы качества $[C_{i,j}]_{i,j=0}^n$ с элементами (10). А именно:

- 1) была преобразована циклическая структура метода ДП, использовавшегося в [Narchaoui, Carré, 2007; Arlot, Celisse, Narchaoui, 2019], что позволило хранить в памяти не всю матрицу качества $[C_{i,j}]_{i,j=0}^n$ целиком, а лишь ее столбцы $C_{\bullet,j}$ последовательным образом;
- 2) была предложена эффективная итерационная процедура вычисления столбца $C_{\bullet,j+1}$ матрицы качества через столбец $C_{\bullet,j}$.

3.4. Method DensRatio

В [Liu et al., 2013] оценивалась степень расхождения данных на смежных сегментах скользящего окна. Метод детектирования разладок, который здесь назван *DensRatio* (Density Ratio), заключался в сравнении полученных оценок с некоторым заданным пороговым значением μ . Ядром метода являлись разработанные в [Sugiyama et al., 2008; Kanamori, Hido, Sugiyama, 2009; Yamada et al., 2013] алгоритмы оценки отношения плотностей распределения выборок.

Введем ряд обозначений. Выборку длиной k из набора (1), начинающуюся с t -го наблюдения, представим в виде вектора $\mathbf{X}_t \triangleq [\mathbf{x}_t^T, \mathbf{x}_{t+1}^T, \dots, \mathbf{x}_{t+k-1}^T]^T \in \mathbb{R}^{dk}$. Сегментом выборок назовем множество из m последовательных выборок $\mathcal{X}_t^m \triangleq \{\mathbf{X}_t, \mathbf{X}_{t+1}, \dots, \mathbf{X}_{t+m+1}\}$ с началом в точке t .

Пусть $\mathcal{P}^{rf} \triangleq \mathcal{P}_t$ и $\mathcal{P}^{te} \triangleq \mathcal{P}_{t+m}$ — распределения вероятностей выборок \mathbf{X} на смежных эталонном (reference) \mathcal{X}_t^m и тестовом (test) \mathcal{X}_{t+m}^m сегментах скользящего окна шириной $2m$, а $p^{rf}(\mathbf{X})$ и $p^{te}(\mathbf{X})$ — соответствующие плотности распределения. Согласно [Liu et al., 2013] момент \hat{t} признавался точкой разладки, если $\mathcal{D}(\hat{t}) \geq \mu$, где расстояние $\mathcal{D}(t)$ определяет степень расхождения распределений \mathcal{P}^{rf} и \mathcal{P}^{te} в момент t . Классическими примерами могут служить расстояния Кульбака – Лейблера:

$$KL(\mathcal{P}^{rf} \parallel \mathcal{P}^{te}) = \int p^{rf}(\mathbf{X}) \log \left(\frac{p^{rf}(\mathbf{X})}{p^{te}(\mathbf{X})} \right) d\mathbf{X} \quad (11)$$

и Пирсона:

$$PE(\mathcal{P}^{rf} \parallel \mathcal{P}^{te}) = \frac{1}{2} \int p^{te}(\mathbf{X}) \left(\frac{p^{rf}(\mathbf{X})}{p^{te}(\mathbf{X})} - 1 \right)^2 d\mathbf{X}. \quad (12)$$

Как видно из (11) и (12), они зависят от неизвестных плотностей $p^{rf}(\mathbf{X})$ и $p^{te}(\mathbf{X})$, подлежащих оценке по выборкам данных.

В серии работ [Sugiyama et al., 2008; Kanamori, Hido, Sugiyama, 2009; Yamada et al., 2013; Liu et al., 2013] было предложено оценивать непосредственно отношение плотностей вероятностей $w(\mathbf{X}) = \frac{p^{rf}(\mathbf{X})}{p^{te}(\mathbf{X})}$, а не функции $p^{rf}(\mathbf{X})$ и $p^{te}(\mathbf{X})$ отдельно друг от друга. В этом случае алгоритм DensRatio включает в себя два основных этапа [Liu et al., 2013]:

- 1) построение оценки отношения плотностей $\hat{w}(\mathbf{X})$ по имеющимся выборкам,
- 2) построение оценки выбранного расстояния $\mathcal{D}(t)$ на основе $\hat{w}(\mathbf{X})$.

В [Liu et al., 2013] приведены три метода оценки отношения плотностей $\hat{w}(\mathbf{X})$:

- 1) KLIEP (the Kullback – Leibler Importance Estimation Procedure) [Sugiyama et al., 2008],
- 2) uLSIF (the unconstrained Least-Squares Importance Fitting) [Kanamori, Hido, Sugiyama, 2009],
- 3) RuLSIF (Relative uLSIF) [Yamada et al., 2013; Liu et al., 2013].

В каждом из них отношение плотностей представлялось в виде линейной комбинации

$$w(\mathbf{X}, \boldsymbol{\theta}) = \sum_{l=1}^m \theta_l k_G(\mathbf{X}, \mathbf{X}_l) \quad (13)$$

гауссовых ядер

$$k_G(\mathbf{X}, \mathbf{X}_l) = \exp \left(-\frac{\|\mathbf{X} - \mathbf{X}_l\|^2}{2h^2} \right) \quad (14)$$

со стандартным отклонением $h > 0$ и центрами в точках тестового сегмента \mathcal{X}_{t+m}^m . Основное отличие трех приведенных подходов состоит в способах настройки вектора параметров $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)^T$ по набору данных. Поскольку uLSIF устраняет ряд недостатков KLIEP [Kanamori,

Hido, Sugiyama, 2009], а RuLSIF является модификацией uLSIF [Yamada et al., 2013], изложим ниже лишь метод RuLSIF.

Для начала вместо тестовой плотности $p^{te}(\mathbf{X})$ вводятся α -смешанная плотность $p_\alpha^{te}(\mathbf{X}) = \alpha p^{rf}(\mathbf{X}) + (1 - \alpha)p^{te}(\mathbf{X})$ и, как следствие, α -относительное отношение плотностей $r_\alpha(\mathbf{X}) = \frac{p^{rf}(\mathbf{X})}{p_\alpha^{te}(\mathbf{X})}$. Такой подход решает проблему потенциальной неограниченности $w(\mathbf{X})$, оказывающей влияние на скорость сходимости алгоритма [Yamada et al., 2013]. Вектор $\boldsymbol{\theta}$ определялся путем минимизации квадратичной функции потерь [Yamada et al., 2013; Liu et al., 2013]

$$J(\boldsymbol{\theta}) \triangleq \mathbb{E}_{p_\alpha^{te}}(r_\alpha(\mathbf{X}) - w(\mathbf{X}, \boldsymbol{\theta}))^2 = \int p_\alpha^{te}(\mathbf{X})(r_\alpha(\mathbf{X}) - w(\mathbf{X}, \boldsymbol{\theta}))^2 d\mathbf{X}.$$

Благодаря аппроксимации математического ожидания эмпирическим средним задача минимизации функционала $J(\boldsymbol{\theta})$ сводится к оптимизационной задаче:

$$\widehat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta} \in \mathbb{R}^m} \left[\frac{1}{2} \boldsymbol{\theta}^T \widehat{\mathbf{H}} \boldsymbol{\theta} - \widehat{\mathbf{h}}^T \boldsymbol{\theta} + \frac{\lambda}{2} \boldsymbol{\theta}^T \boldsymbol{\theta} \right],$$

в которой λ играет роль множителя регуляризации, матрица $\widehat{\mathbf{H}} = \{H_{lk}\}_{l,k=1}^m$ состоит из элементов

$$H_{lk} = \frac{\alpha}{m} \sum_{i=1}^m k_G(\mathbf{X}_i, \mathbf{X}_l) k_G(\mathbf{X}_i, \mathbf{X}_k) + \frac{1 - \alpha}{m} \sum_{j=1}^m k_G(\mathbf{X}'_j, \mathbf{X}_l) k_G(\mathbf{X}'_j, \mathbf{X}_k),$$

а вектор $\widehat{\mathbf{h}} = \{h_l\}_{l=1}^m$ — из $h_l = m^{-1} \sum_{i=1}^m k_G(\mathbf{X}_i, \mathbf{X}_l)$. Решение данной задачи было получено в аналитическом виде:

$$\widehat{\boldsymbol{\theta}} = (\widehat{\mathbf{H}} + \lambda I_m)^{-1} \widehat{\mathbf{h}}, \quad (15)$$

что является достоинством метода RuLSIF. В (15) I_m — единичная матрица размером m . Отметим, что как параметр h из (14), так и множитель λ , влияющие на качество оценки, определяются в ходе процедуры перекрестной проверки, описанной в [Kanamori, Hido, Sugiyama, 2009].

При детектировании разладок [Liu et al., 2013] методом DensRatio полученная в (13) и (15) оценка отношения плотностей $\widehat{w}(\mathbf{X}) = w(\mathbf{X}, \widehat{\boldsymbol{\theta}})$ применялась для построения эмпирической меры расхождения между распределениями \mathcal{P}^{rf} и \mathcal{P}^{te} . В совокупности с методом RuLSIF авторы использовали симметричное расстояние $\mathcal{D}_{PE_\alpha}(t) = PE_\alpha(\mathcal{P}^{rf} \parallel \mathcal{P}^{te}) + PE_\alpha(\mathcal{P}^{te} \parallel \mathcal{P}^{rf})$, где

$$PE_\alpha(\mathcal{P}^{rf} \parallel \mathcal{P}^{te}) \triangleq PE(\mathcal{P}^{rf} \parallel \alpha \mathcal{P}^{rf} + (1 - \alpha) \mathcal{P}^{te}) = \mathbb{E}_{p_\alpha^{te}}(r_\alpha(\mathbf{X}) - 1)^2 = \frac{1}{2} \int p_\alpha^{te}(\mathbf{X})(r_\alpha(\mathbf{X}) - 1)^2 d\mathbf{X}. \quad (16)$$

С учетом (13) и (15) эмпирический аналог (16) принимает вид [Liu et al., 2013]

$$\widehat{PE}_\alpha = -\frac{\alpha}{2m} \sum_{i=1}^m \widehat{w}(\mathbf{X}_i)^2 - \frac{1 - \alpha}{2m} \sum_{i=1}^m \widehat{w}(\mathbf{X}'_i)^2 + \frac{1}{m} \sum_{i=1}^m \widehat{w}(\mathbf{X}_i) - \frac{1}{2}.$$

Последнее выражение позволяет построить окончательную оценку $\widehat{\mathcal{D}}_{PE_\alpha}$ расстояния между сегментами \mathcal{X}_t^m и \mathcal{X}_{t+m}^m , с помощью которой и производилось определение степени значимости потенциальных точек разладки в исходном наборе данных (1).

Сравнительный анализ реализаций метода DensRatio в совокупности с процедурами KLIER, uLSIF и RuLSIF проводился на трех наборах данных. Первый [Hasc Challenge 2011], полученный в рамках программы Human Activity Sensing Consortium (HASC) challenge 2011, представляет информацию о деятельности человека, собранную портативными трехосными акселерометрами. Ставилась задача сегментации временного ряда с целью определения шести режимов

движения. Вторая задача состояла в извлечении речевых фрагментов человеческого голоса, записанного в шумной среде. Для этого использовался набор данных [Speech Resources Consortium] IPSJ SIG-SLP Corpora and Environments for Noisy Speech Recognition (CENSREC). Последняя задача сегментации временного ряда имела дело с архивом сообщений Twitter, собранных с февраля по октябрь 2010 года (набор Twitter CMU7) [Tweets от Noah's ARK]. С учетом частоты использования заданных ключевых слов оценивалась степень популярности той или иной темы.

3.5. Метод NP-MOJO

Еще один метод, использующий технику скользящих окон в задаче сегментации данных, был предложен в статье [McGonigle, Cho, 2023]. В ней изучается следующая модель многомерного временного ряда $\{X_t\}_{t=1}^n \subset \mathbb{R}^d$:

$$X_t = \sum_{j=0}^q X_t^{(j)} \cdot \mathbf{1}\{\tau_j + 1 \leq t \leq \tau_{j+1}\},$$

состоящего из $q + 1$ стационарных сегментов. Здесь $\tau_0 = 0$ и $\tau_{q+1} = n$. Предполагается, что для каждой последовательности $\{X_t^{(j)} : t \geq 1\}$, $j = \overline{0, q}$, в \mathbb{R}^d существует функция $g^{(j)}(\cdot)$ такая, что $X_t^{(j)} = g^{(j)}(F_t)$, где $F_t = \sigma(\varepsilon_s : s \leq t)$, а ε_t — независимые одинаково распределенные величины. В точках разладки $\{\tau_j\}_{j=1}^q$ выполняется $g^{(j-1)} \neq g^{(j)}$. Приведенная постановка задачи не исключает зависимость между сегментами.

Разработанный в [McGonigle, Cho, 2023] метод NP-MOJO (a NonParametric MOving sum procedure for detecting changes in the JOint characteristic function) представляет собой непараметрический аналог процедуры скользящей суммы [Chu, Hornik, Kaun, 1995; Huskova, Slaby, 2001] и позволяет определить многократные разладки в многомерных временных рядах. Для этого метод скользящей суммы используется в совокупности со статистикой, детектирующей изменения совместной характеристической функции последовательности $\{X_t^{(j)}\}$ при временном лаге в $\ell \geq 0$ отсчетов, которая имеет вид

$$\phi_t^{(j)}(u, v) = E \left[\exp \left(\iota \langle u, X_t^{(j)} \rangle + \iota \langle v, X_{t+\ell}^{(j)} \rangle \right) \right], \quad (17)$$

где $0 \leq j \leq q$ и $\iota^2 = -1$. Как утверждается, это позволяет обнаруживать изменения в совместном распределении $X_t, X_{t+\ell}$.

Мера расхождения между распределениями на соседних окнах шириной $G \in \mathbb{N}$, не превышающей половину длины ряда, имеет вид

$$\mathcal{D}_\ell(G, k) = \sum_{j=0}^q \left(\frac{G - \ell - |k - \tau_j|}{G - \ell} \right)^2 d_\ell^{(j)} \cdot \mathbf{1}\{|k - \tau_j| \leq G - \ell\}, \quad (18)$$

где для заданного лага $\ell \geq 0$ формула

$$d_\ell^{(j)} = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \left| \phi_t^{(j)}(u, v) - \phi_t^{(j-1)}(u, v) \right|^2 \omega(u, v) du dv, \quad 1 \leq j \leq q, \quad (19)$$

определяет взвешенную L_2 -норму между локальными совместными характеристическими функциями (17) сегментов $\{X_t^{(j-1)}\}$ и $\{X_t^{(j)}\}$. В (19) $\omega(u, v)$ — весовая функция, принимающая положительные значения.

Процедура детектирования точек разладки требует статистической аппроксимации выражения (18), а точнее интеграла из (19). В [McGonigle, Cho, 2023] для этого использовалась детектирующая двухвыборочная V -статистика, которая для момента временного ряда $k = \overline{G, n - G}$ задается формулой

$$T_l(G, k) = \frac{1}{(G - \ell)^2} \left(\sum_{s,t=k-G+1}^{k-\ell} h(Y_s, Y_t) + \sum_{s,t=k+1}^{k+G-\ell} h(Y_s, Y_t) - 2 \sum_{s=k-G+1}^{k-\ell} \sum_{t=k+1}^{k+G-\ell} h(Y_s, Y_t) \right). \quad (20)$$

В (20) функция h является ядром преобразования, связанного с весовой функцией $\omega(u, v)$.

Как и в подпараграфе 3.4, момент $\widehat{\tau}$ признается точкой разладки, если локально выполняется условие $T_l(G, \widehat{\tau}) > \mu(n, G)$, где $\mu(n, G)$ — пороговое значение, а

$$\widehat{\tau} = \arg \max_{k: |k-\widehat{\tau}| \leq \eta G} T_l(G, k).$$

Здесь $\eta \in (0, 1)$ — некоторая фиксированная константа. Для выбора $\mu(n, G)$ авторы адаптировали процедуру дикого бутстрепа (wild bootstrap), предложенную в [Leucht, Neumann, 2013]; ширина окна G при моделировании полагалась равной $\lfloor \frac{n}{6} \rfloor$.

Суть метода NP-MOJO заключается в грамотном подборе величины лага ℓ , с помощью которой имеется возможность увеличения значения статистики $T_l(G, k)$, по максимуму которой ищутся точки разладки.

4. Сегментация реальных профилей скорости

Этот параграф посвящен детектированию точек разладки на примерах экспериментальных данных с помощью программных реализаций методов, описанных в § 3, а также анализу полученных результатов. Целью анализа является выбор методов, которые в дальнейшем будут применены к полному корпусу данных.

В предлагаемом эксперименте каждый алгоритм применяется к трем скоростным профилям движения космонавта (рис. 3):

- простому профилю: стандартный профиль перехода из режима «шаг, 3 км/ч» в режим «бег, 6 км/ч» и обратно (рис. 3, а);
- сложному профилю № 1: содержит резкий переход из режима бега со скоростью 8 км/ч к режиму быстрого шага со скоростью 5 км/ч, после чего происходит выход на рабочую скорость 9 км/ч примерно к 530-му отсчету через два плато в интервалах с 118-го отсчета по 175-й отсчет и с 456-го отсчета по 517-й отсчет, между которыми также можно выделить три скоростных интервала (отсчеты 180–246, 246–357, 357–448), кроме того, ряд меняет поведение также в точке 549 (рис. 3, б);
- сложному профилю № 2: переход с шага со скоростью 4 км/ч к быстрому шагу на 6 км/ч, который в окрестности 350-го отсчета качественно меняется (начинается бег на той же скорости), затем наблюдается ускорение в интервале с 514-го отсчета по 575-й отсчет, плавное замедление с 575-го отсчета по 649-й отсчет, далее выход на короткое плато в интервале с 662-го отсчета по 727-й отсчет и остановка движения (рис. 3, в).

Отметим, что простой профиль характерен для варианта эксперимента, в котором дорожка приводится в движение мотором. Профили на рис. 3, б и 3, в характерны для режима эксперимента, в котором мотор не используется, и космонавт самостоятельно выходит на рабочую скорость и поддерживает ее.

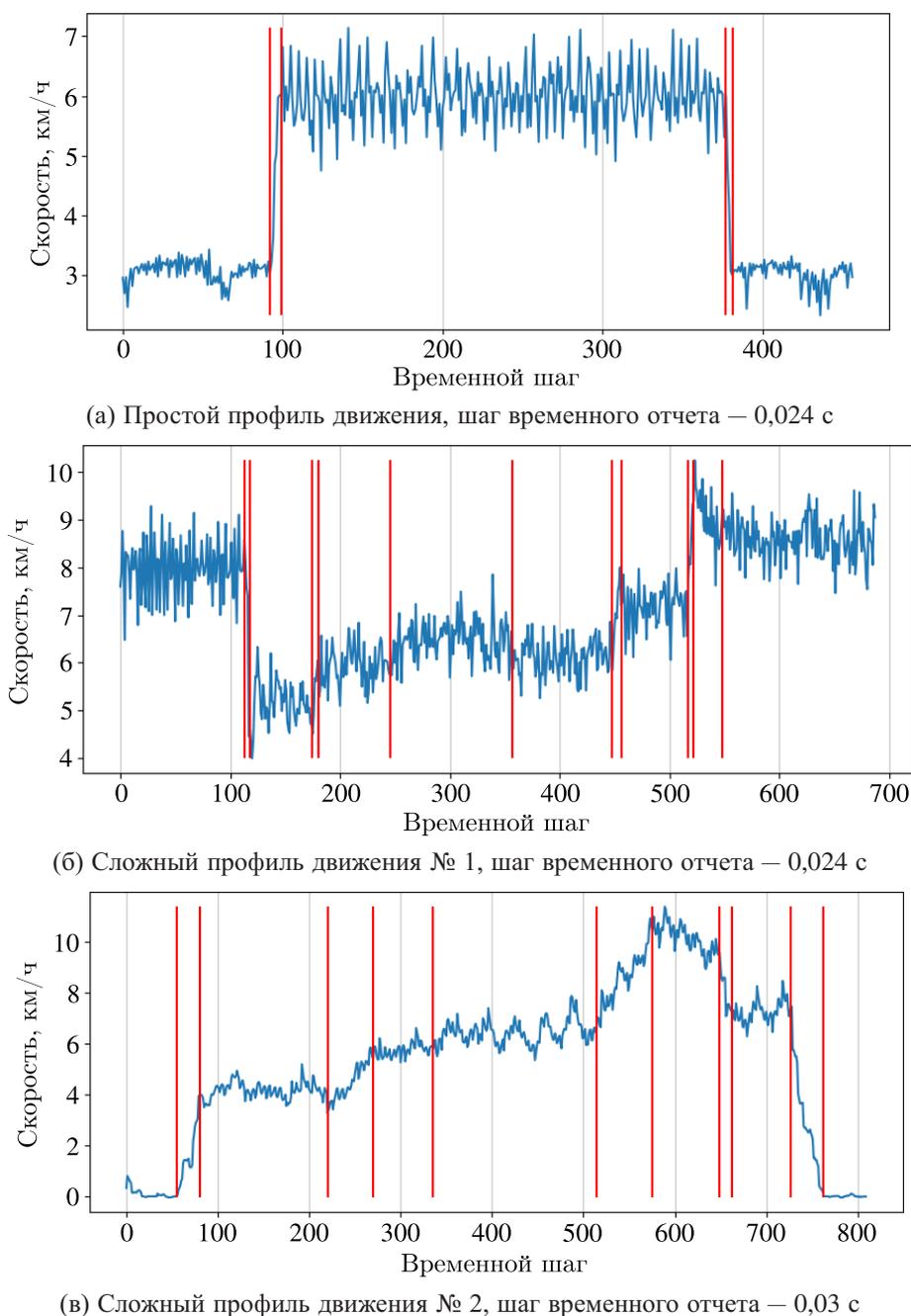


Рис. 3. Характерные скоростные профили движения космонавтов, полученные в рамках локомоторного тестирования. Эталонная разметка (вертикальные красные линии) соответствует протокольным моментам переключения скоростных режимов эксперимента, а также оценкам начала и конца интервалов перехода между скоростными режимами

Моделирование проводилось на ПК с 8Гб ОЗУ и процессором Intel i5 2 ГГц с операционной системой Linux. Использованные языки программирования — Python 3, R. В ходе моделирования использовались программные реализации методов с открытым исходным кодом.

Во время моделирования использовались разряженные временные ряды. На рис. 3, а и 3, б временной интервал составляет 0,24 с, на рис. 3, в — 0,03 с. Оценка качества детектирования разладок производится в контексте соответствия эталонной разметке (рис. 3) и характеризует

возможность методов выделять в типичных экспериментальных данных стационарные и нестационарные участки. Читатель может отметить, что итоговой, целевой задачей, согласно §§ 1 и 2, является выделение режимов движения и переходных процессов между ними.

По степени качества описанные методы условно разделяются на две группы:

- подходящие: близость к эталонной разметке, робастность по параметрам метода, потенциальное детектирование обеих границ интервалов;
- неподходящие: в противном случае.

При моделировании использовались параметры по умолчанию, параметры из примеров запусков, а также подбирались комбинации, реализующие близкое к эталонному разбиение. Проверялась стабильность разбиения всех трех рядов при различных комбинациях параметров метода. Конкретнее: если при изменениях параметров качество разбиения существенно не менялось (смещение точек разладки на 3–5 отсчетов, изменение числа точек разладки на одну-две единицы), вводится гипотеза о существовании множества оптимальных комбинаций параметров метода.

В данном параграфе рассматриваются найденные в открытом доступе реализации рассмотренных выше алгоритмов. В случае если открытая реализация не была найдена, проверка не проводилась.

4.1. Метод ED-PELT

Реализация метода ED-PELT [Haynes, Fearnhead, Eckley, 2017] представлена в пакете [changepoint.np] (2022) на языке R. В процессе моделирования использовалась функция *cpt.np*. На каждом из трех профилей метод демонстрирует близкую к эталонной сегментацию рядов (рис. 4), однако стоит отметить чувствительность метода к незначительным для задачи изменениям временного ряда, о чем свидетельствуют лишние точки разладки. Так, на рис. 4, *a* в режиме медленного шага найдены лишние точки разладки в отсчетах 6, 51, 62, 427, 444, на рис. 4, *b* — в отсчетах 9, 21, 523, 606, 731, 786, 788.

Отдельные интервалы переходных процессов на сложных профилях были правильно локализованы: интервалы 113–118, 517–549 на рис. 4, *b*, интервалы 55–80, 506–575, 727–763 на рис. 4, *b*. Однако остальные переходные процессы определены единственной точкой разладки, расположенной ближе к центру нестационарного интервала: на рис. 4, *a* — оба перехода с шага на быстрый шаг и обратно; на рис. 4, *b* — посередине интервала 448–456, на рис. 4, *b* — посередине интервала 220–270. При этом стоит отметить верное детектирование горба, расположенного в интервале 246–357 на рис. 4, *b*.

Метод детектирует дополнительные сегменты, которые необходимо объединить на этапе постобработки. У половины нестационарных интервалов определены начало и конец, что допускает возможность подбора параметров для более точной сегментации.

При варьировании параметров метод демонстрирует робастное поведение, сегментация воспроизводится с точностью до изменения расположения разладок на 2–5 временных шагов, а также появления или исчезновения нескольких точек разладки относительно эталонного разбиения на рис. 3.

Для решения задачи сегментации биомедицинских данных метод ED-PELT признается подходящим. Однако он может применяться лишь для одномерных данных.

4.2. Метод MultiRank

Программной реализации алгоритма MultiRank [Lung-Yut-Fong, Lévy-Leduc, Cappé, 2015] не найдено. В пакете [ruptures] имеется [Truong, Oudre, Vayatis, 2020] подходящий функционал качества.

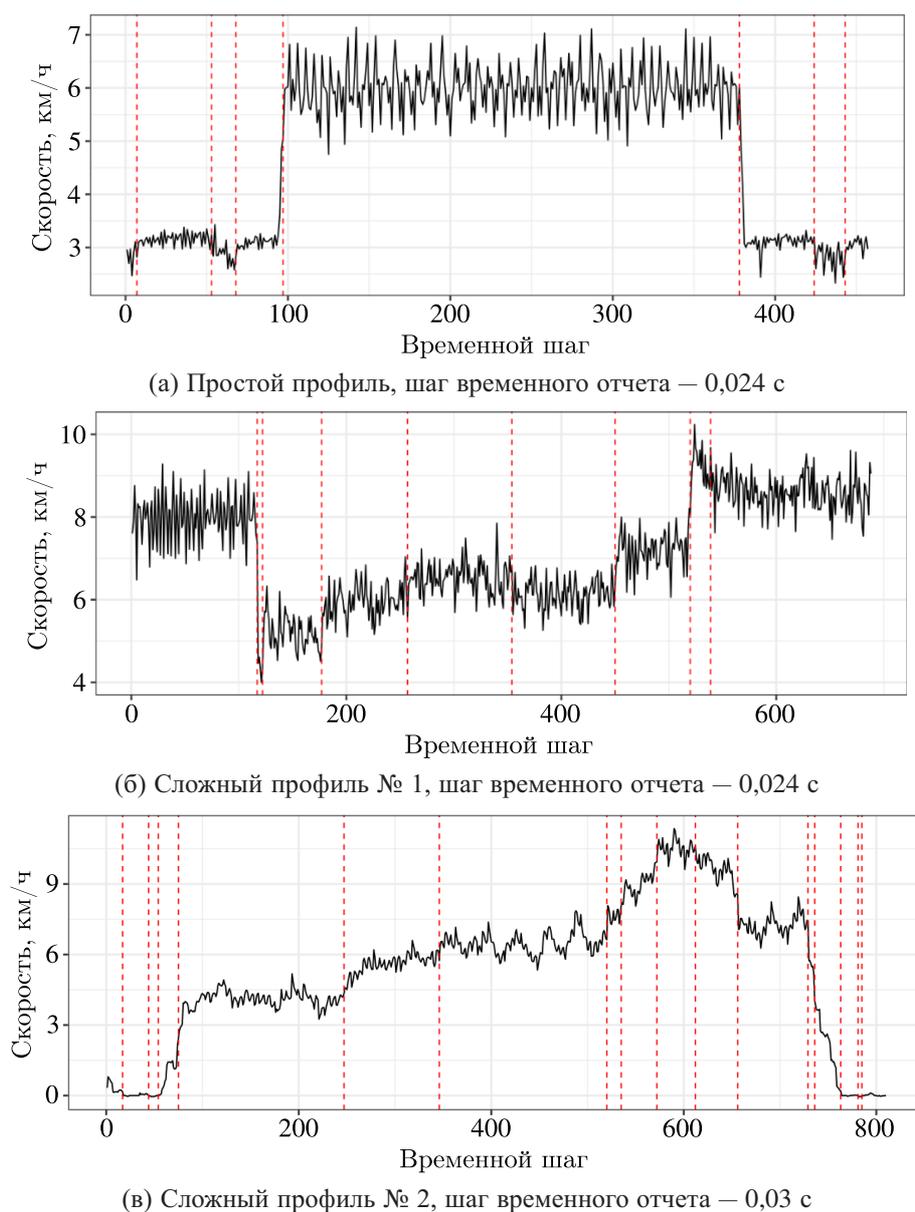


Рис. 4. Сегментация скоростных профилей движения методом ED-PELT. Вертикальными красными линиями отмечены полученные точки разладки. Присутствуют лишние точки разладки внутри скоростных режимов, полученные дополнительные сегменты возможно объединить на этапе постобработки

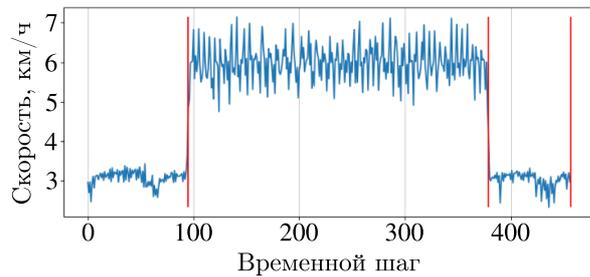
4.3. Методы KCP и Kernseg

Реализация алгоритма KCP на языке Python доступна в пакетах ruptures [ruptures; Truong, Oudre, Vayatis, 2020] и Chapydette [Jones, Harchaoui, 2019]. Реализация алгоритма из [Celisse et al., 2018], названного авторами [KernSeg], содержится в R-пакете Kernseg [Marot, Celisse, Rigaiil, 2018]. В пакете ruptures для случая заранее неизвестного количества разладок реализована оптимизационная процедура из [Killick, Fearnhead, Eckley, 2012], в Chapydette — из [Arlot, Celisse, Harchaoui, 2019]. В настоящей работе осуществлено моделирование методами из пакетов Kernseg и ruptures.

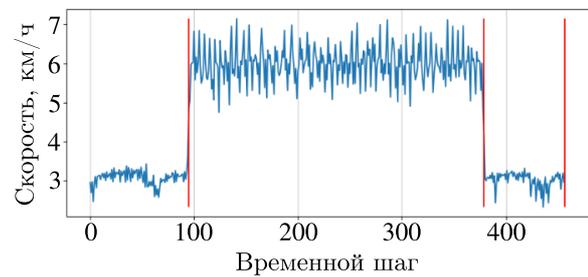
Согласно [Arlot, Celisse, Harchaoui, 2019] для обнаружения изменений в распределении данных в целом следует использовать характеристические ядра [Sriperumbudur, Fukumizu,

Lanckriet, 2010], в роли которых часто берутся гауссовы ядра (14) [Arlot, Celisse, Harchaoui, 2019; Cabrieto et al., 2017] с дисперсией, рассчитанной с помощью медианной эвристики [Gretton et al., 2012]. В свою очередь, нехарактеристические ядра также могут быть полезны в определенных задачах сегментации данных. Например, в работе [Lebarbier, 2005] линейное ядро $k_{lin}(x, y) = xy$ успешно помогло обнаружить изменения средних значений случайных величин $X_1, \dots, X_n \in \mathbb{R}$. В соответствии с указанными рекомендациями в настоящей работе моделирование было проведено на гауссовом и линейном ядрах.

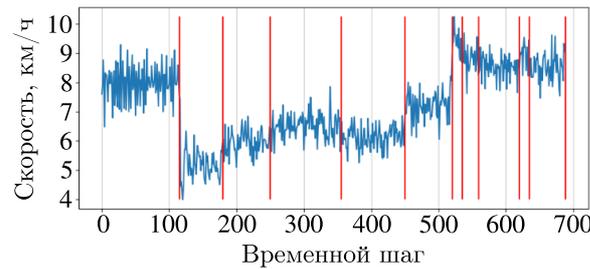
На обоих ядрах метод КСР сегментировал простой профиль достаточно близко к эталону (рис. 3). Отличие заключается в лишней точке разладки в крайней правой точке ряда (рис. 6, а, рис. 5, а); эта проблема наблюдается и у обоих сложных профилей. Переходные процессы детектированы одной точкой разладки, находящейся по центру переходов.



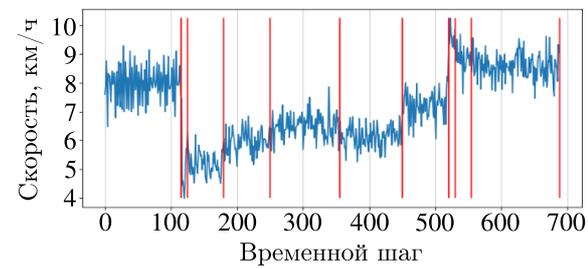
(а) Простой профиль, шаг временного отчета — 0,024 с



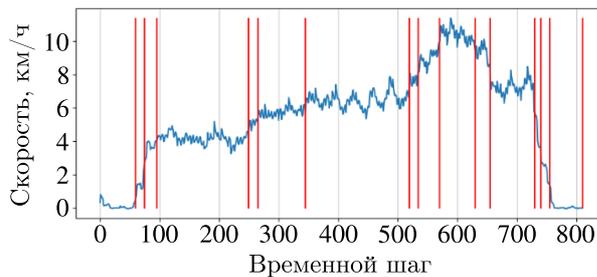
(а) Простой профиль, шаг временного отчета — 0,024 с



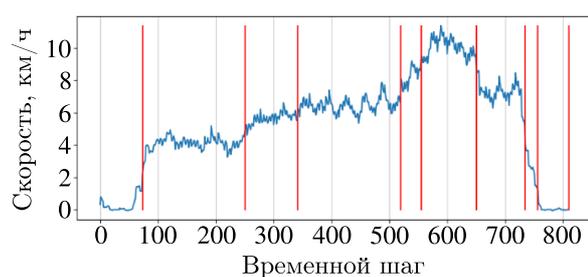
(б) Сложный профиль № 1, шаг временного отчета — 0,024 с



(б) Сложный профиль № 1, шаг временного отчета — 0,024 с



(в) Сложный профиль № 2, шаг временного отчета — 0,03 с

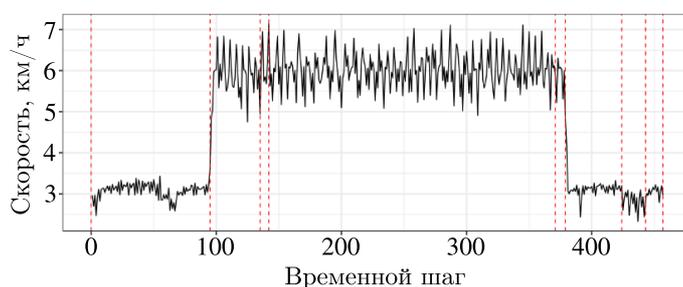


(в) Сложный профиль № 2, шаг временного отчета — 0,03 с

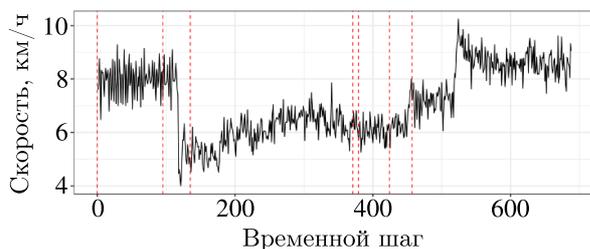
Рис. 5. Сегментация скоростных профилей движения методом КСР на линейном ядре. Вертикальными красными линиями отмечены полученные точки разладки. Большая часть переходных процессов определена единственной точкой разладки, находящейся посередине интервала перехода, детектированы дополнительные сегменты

Рис. 6. Сегментация скоростных профилей движения методом КСР на гауссовом ядре. Вертикальными красными линиями отмечены полученные точки разладки. По сравнению с линейным ядром — более точное детектирование переходных процессов и меньшее количество дополнительных сегментов

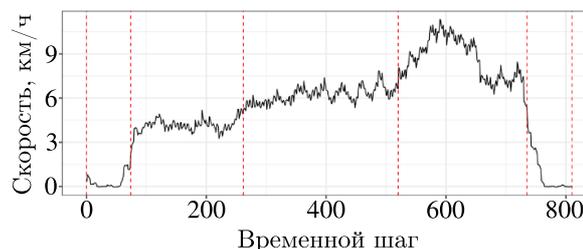
В случае сложного профиля № 1 и гауссового ядра полученная сегментация практически идентична эталонной. Точки разладки при переходе 113–118 смещены вправо на 4 отсчета; то же смещение наблюдается при переходе 517–522 (рис. 6, б). Переход 448–456 отмечен центром этого интервала, также найдена лишняя точка разладки в точке 552. В случае же линейного ядра детектируются лишние точки разладки также в точках 534, 620, 631; все переходные процессы найдены, но отмечены в центре интервалов переходов (рис. 5, б). Сложный профиль № 2 в случае гауссового ядра в целом также имеет близкую к эталонной разметку (рис. 6, в), однако интервалы 55–80, 220–270, 649–662 детектированы центрами переходов, точки разладки в переходах 502–575, 727–763 смещены к центру данных интервалов. Линейное ядро добавляет лишние точки разладки: в крайнем левом переходе определены левая и правая границы, но также и точка посередине интервала; в крайнем правом переходе также найдена точка 739 примерно посередине интервала, правая граница интервала (отсчет 751) смещена левее истинной границы.



(а) Простой профиль, шаг временного отчета — 0,024 с



(б) Сложный профиль № 1, шаг временного отчета — 0,024 с



(в) Сложный профиль № 2, шаг временного отчета — 0,03 с

Рис. 7. Сегментация скоростных профилей движения методом KernSeg на гауссовом ядре. Вертикальными красными линиями отмечены полученные точки разладки. Сегментация недостаточно подробна для «сложных» профилей

Метод Kernseg, реализованный в функции *KernSeg_MultiD*, моделируется с гауссовым ядром. На рис. 7, а представлены результаты сегментации простого профиля. В сравнении с эталонным расположением разладок найдены лишние точки разладки в крайних точках ряда, в точках 140, 144. Как и в методе ED-PELT, отмечены разладки в отсчетах 427, 444. Переход с 6 км/ч к 3 км/ч отмечен двумя разладками, но эти точки смещены влево относительно эталонной разметки.

На рис. 7, б сегментация существенно отличается от эталонной. Имеется лишняя точка разладки в начале ряда, переход с 8 км/ч к 3 км/ч отмечен, но левая и правая границы должны быть ближе к самому интервалу перехода (детектированные точки 98 и 141), а также 366 и 370. Интервал перехода на 5 км/ч в окрестности точки 350 отмечен, верно отмечена его правая граница, однако левая граница детектирована существенно левее эталонной.

На рис. 7, в полученная сегментация недостаточно подробна. Интервалы переходов 55–80, 220–270 определены одной центральной точкой, в интервале 727–763 определена только левая

граница. Пропущены точки 575, 649, 662. Детектированы лишние точки разладки в крайних точках ряда.

Метод КСР демонстрирует сегментацию профилей скоростей близкую к эталонной и робастность при переборе параметров. Определение границ нестационарных режимов определяется в большинстве случаев верно, в перспективе точность может быть улучшена более точным подбором параметров для полного корпуса данных. Метод определяется как подходящий для решения задачи сегментации профилей скоростей.

Метод Kernseg демонстрирует сегментацию простого профиля скоростей, близкую к эталонной, и робастность при переборе параметров. Однако в случае сложных профилей сегментация недостаточно подробна, пропущены точки разладки. Метод определяется как неподходящий.

4.4. Метод DensRatio

Программной реализации методов, основанных на оценке отношения плотностей распределения, не найдено. В пакетах [ruptures] (Python) [Truong, Oudre, Vayatis, 2020], [densratio], [RuLSIF] (Matlab) имеется подходящий функционал для оценки отношения плотностей распределения на основе ограниченного числа наблюдений. С помощью функции под названием *densratio*, т. е. Density Ratio Estimation, возможна реализация метода DensRatio (подпараграф 3.4).

4.5. Метод NP-MOJO

Реализация метода NP-MOJO [McGonigle, Cho, 2023] на языке R доступна в пакете [CptNonPar]. Согласно рекомендациям в [McGonigle, Cho, 2023] на первом этапе для сегментации была использована реализация указанного метода со стандартным гауссовым ядром, а также с ядром h_2 , представленном в [McGonigle, Cho, 2023]. Для любых $\delta > 0$, $x = (x_1, \dots, x_{2p})^T$ и $y = (y_1, \dots, y_{2p})^T$, ядро $h_2: \mathbb{R}^{2p} \times \mathbb{R}^{2p} \rightarrow [-2e^{-2/3}, 1]$, $p \in \mathbb{N}$, задается формулой

$$h_2(x, y) = \prod_{r=1}^{2p} \frac{(2\delta - (x_r - y_r)^2) \exp\left(-\frac{(x_r - y_r)^2}{4\delta}\right)}{2\delta}.$$

Моделирование показало не робастное поведение метода NP-MOJO при переборе параметров, которое наблюдалось на обоих типах ядер на каждом из трех экземпляров данных (рис. 8–11). Лишь редкие комбинации последних приводили к детектированию более двух точек разладки (рис. 9, б).

На качество сегментации данных с помощью метода NP-MOJO влияют размер скользящего окна и временной лаг последовательности данных, т. е. величина ее смещения относительно самой себя. При поиске точек разладки на рис. 10, б величина лага равнялась нулю, на

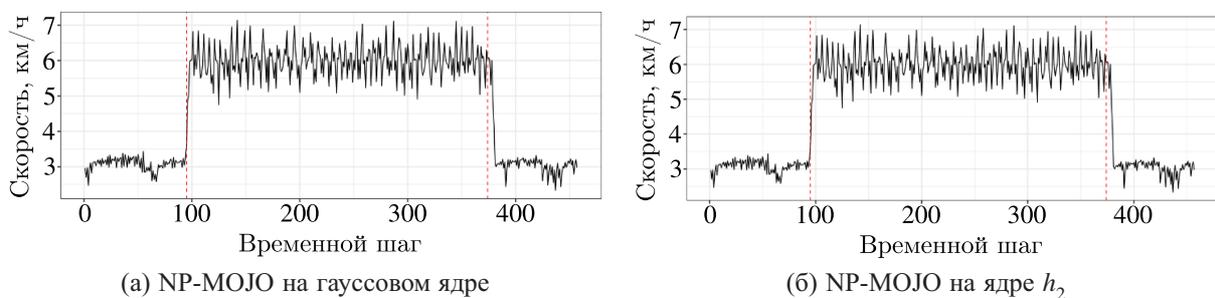


Рис. 8. Сегментация простого скоростного профиля движения методом NP-MOJO на различных ядрах. Вертикальными красными линиями отмечены полученные точки разладки, шаг временного отчета — 0,024 с. Переходные процессы детектированы только одной точкой разладки

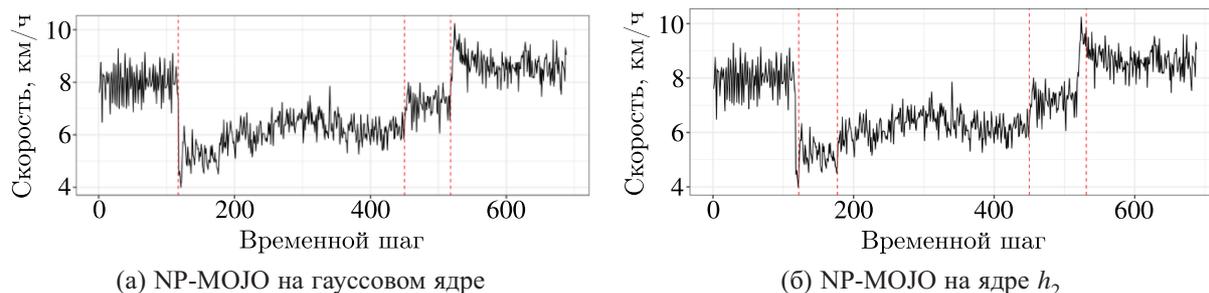


Рис. 9. Сегментация сложного скоростного профиля движения № 1 методом NP-MOJO на различных ядрах. Вертикальными красными линиями отмечены полученные точки разладки, шаг временного отчета — 0,024 с. Переходные процессы детектированы только одной точкой разладки, сегментация недостаточно подробна

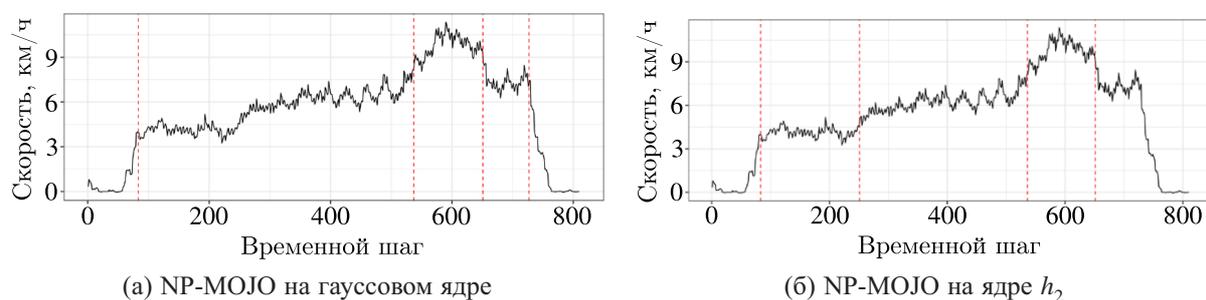


Рис. 10. Сегментация сложного скоростного профиля движения № 2 методом NP-MOJO на различных ядрах. Вертикальными красными линиями отмечены полученные точки разладки, шаг временного отчета — 0,03 с. Переходные процессы детектированы только одной точкой разладки, сегментация недостаточно подробна

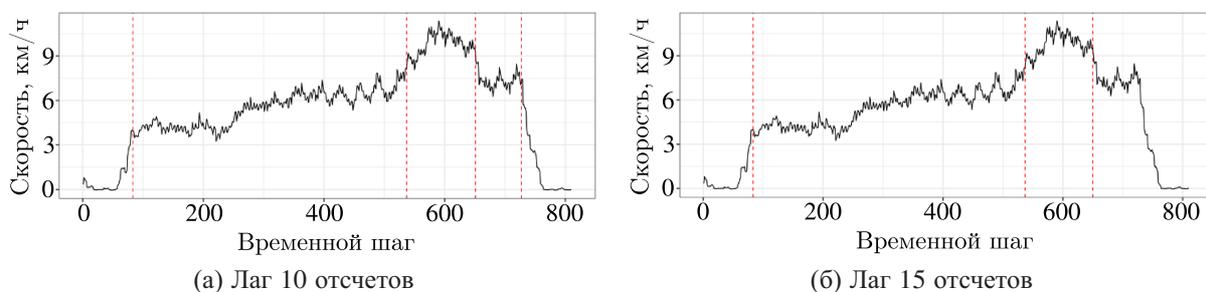


Рис. 11. Сегментация сложного скоростного профиля движения № 2 методом NP-MOJO на ядре h_2 при различной величине временного лага. Вертикальными красными линиями отмечены полученные точки разладки, шаг временного отчета — 0,03 с. Результат не соответствует эталону и неробастен при изменении параметров

рис. 11, *a* — 10, на рис. 11, *б* — 15. При величине лага, равной 20, не детектируется ни одной точки. Изменение иных параметров либо не изменяло положение точек разладки и их количество, либо смещало на 1–3 отсчета. Иные параметры не менялись. Как видно из приведенных графиков, и число, и положение найденных точек разладки существенно отличаются. К аналогичным результатам приводит моделирование с гауссовым ядром по трем скоростным профилям. Таким образом, улучшение качества процесса сегментации за счет изменения величины временного лага не наблюдалось.

Помимо рекомендованных, поиск точек разладки был произведен на следующих ядрах: *eucclidean* — взвешенное евклидово ядро, h_3 в [McGonigle, Cho, 2023]; *laplace* — ядро на основе

взвешенного расстояния Лапласа, *kernel 2* в [Fan et al., 2017]; *sine* — ядро на основе синусоидальной весовой функции, *kernel 4* в [Fan et al., 2017]. Результаты моделирования на указанных ядрах согласуются с рекомендациями авторов: количество найденных точек разладки было меньше (вплоть до их отсутствия) по сравнению с сегментацией на гауссовом и h_2 -ядрах.

В итоге метод NP-MOJO был определен как неподходящий для сегментации профиля скорости в связи с несоответствием результата работы эталонной разметке, а также отсутствием робастности результатов его применения в зависимости от используемых при моделировании наборов параметров.

5. Заключение

Была поставлена задача анализа медико-биологических данных космонавтов с целью разработки системы медицинского обеспечения без сопровождения специалистов с Земли. Данные представляют собой результаты локомоторных тестов, включающих движение человека по беговой дорожке в различных режимах (шаг, бег, бег с ускорением). Как было продемонстрировано в § 2, профиль скорости может обладать нетривиальной формой, особенно в вопросе переходов между режимами движения. Примером служит ситуация, когда космонавт самостоятельно набирает рабочую скорость и поддерживает ее. Неверная сегментация профиля скорости приведет к внесению ошибки в дальнейший анализ целевых показателей.

Для решения этой задачи были выбраны ретроспективные непараметрические методы, которые были классифицированы по типу оптимизационной процедуры на три группы. В настоящей статье исследованы методы, основанные на динамическом программировании и методе скользящего окна.

Найденные программные реализации рассмотренных методов были протестированы на трех размеченных характерных эталонных примерах, охватывающих как простые профили, так и сложные. Помимо соответствия разметке, проверялась также робастность метода.

Открытых реализаций метода MultiRank и метода, основанного на оценке отношения плотностей распределения (DensRatio), не было найдено. Метод NP-MOJO признан неподходящим, поскольку продемонстрировал неробастную сегментацию, существенно отличающуюся от эталонной.

В настоящей работе метод ED-PELT признан подходящим. Алгоритм обладает робастностью относительно внутренних параметров. С его помощью получена разметка, близкая к эталонной, зачастую детектировались обе границы переходных участков, а не только их центры. Однако в ряде случаев ED-PELT требует постобработки данных для устранения побочных точек, принадлежащих регламентированному участкам движения.

Также среди рассмотренных методов подходящим признан метод КСР, как демонстрирующий робастную сегментацию, близкую к эталонной. Верно определяя участки стационарного движения, метод КСР верно детектирует как центры переходных процессов, так и большую часть границ переходов. Несмотря на то что метод KernSeg является модификацией метода КСР, его результаты оказались менее точны относительно КСР. Вероятно, он требует более точного подбора параметров под конкретные данные.

Так как имеются различия при сегментации участков нестационарности с помощью различных методов, признанных подходящими, в дальнейшем улучшить полученные результаты можно будет с помощью создания их ансамбля. Если результаты работы методов будут не коррелированы, то их можно объединить, тем самым также повышая итоговое качество сегментации.

Необходимо отметить, что требования реализаций методов по памяти и вычислительным ресурсам существенно зависят от их программных реализаций. Так, в работе [Yudin et al., 2021] показано ускорение нескольких широко распространенных методов машинного обучения на два

порядка при одновременной оптимизации количества используемой памяти. Наилучший метод (или их комбинацию) планируется ускорить путем реализации на языке С.

В дальнейшем признанные подходящими методы планируется проверить на существенно большем количестве эталонов и ввести численные метрики качества сегментации, а также провести полный перебор параметров для оптимальной настройки методов под имеющиеся данные.

Список литературы (References)

- Беллман Р. Динамическое программирование. — М.: Иностранная литература, 1960.
Bellman R. Dynamic programming. — Princeton, New Jersey: Princeton University Press, 1957. (Russ. ed.: *Bellman R. Dinamicheskoe programmirovaniye.* — Moscow: Inostrannaya literatura, 1960.)
- Набор данных от Speech Resources Consortium (NII-SRC). — [Электронный ресурс]. — <http://research.nii.ac.jp/src/en/> (дата обращения: 05.12.2023).
 Dataset from Speech Resources Consortium (NII-SRC). — [Electronic resource]. — <http://research.nii.ac.jp/src/en/> (accessed: 05.12.2023).
- Набор данных соревнования Hasc Challenge 2011. — [Электронный ресурс]. — <http://hasc.jp/hc2011/> (дата обращения: 05.12.2023).
 Dataset of Hasc Challenge 2011. — [Electronic resource]. — <http://hasc.jp/hc2011/> (accessed: 05.12.2023).
- Набор данных Tweets от Noah's ARK. — [Электронный ресурс]. — <http://www.ark.cs.cmu.edu/> (дата обращения: 05.12.2023).
 Dataset Tweets from Noah's ARK. — [Electronic resource]. — <http://www.ark.cs.cmu.edu/> (accessed: 05.12.2023).
- Программный пакет changepoint.np. — [Электронный ресурс]. — <http://cran.r-project.org/web/packages/changepoint.np/index.html> (дата обращения: 17.06.2024).
 Package changepoint.np. — [Electronic resource]. — <http://cran.r-project.org/web/packages/changepoint.np/index.html> (accessed: 17.06.2024).
- Программный пакет CptNonPar. — [Электронный ресурс]. — <http://github.com/EuanMcGonigle/CptNonPar> (дата обращения: 17.06.2024).
 Package CptNonPar. — [Electronic resource]. — <http://github.com/EuanMcGonigle/CptNonPar> (accessed: 17.06.2024).
- Программный пакет densratio. — [Электронный ресурс]. — http://github.com/hoho-m/densratio_py (дата обращения: 17.06.2024).
 Package densratio. — [Electronic resource]. — http://github.com/hoho-m/densratio_py (accessed: 17.06.2024).
- Программный пакет KernSeg. — [Электронный ресурс]. — <http://rdrr.io/rforge/KernSeg> (дата обращения: 17.06.2024).
 Package KernSeg [Electronic resource]: <http://rdrr.io/rforge/KernSeg> (accessed: 17.06.2024).
- Программный пакет RuLSIF. — [Электронный ресурс]. — <http://riken-yamada.github.io/RuLSIF.html> (дата обращения: 17.06.2024).
 Package RuLSIF. — [Electronic resource]. — <http://riken-yamada.github.io/RuLSIF.html> (accessed: 17.06.2024).
- Программный пакет ruptures. — [Электронный ресурс]. — <http://centre-borelli.github.io/ruptures-docs> (дата обращения: 17.06.2024).
 Package ruptures. — [Electronic resource]. — <http://centre-borelli.github.io/ruptures-docs> (accessed: 17.06.2024).
- Фомина Е. В., Кукоба Т. Б. Основы применения результатов космической деятельности в интересах медицины: учебное пособие. — М.: МПГУ, 2019.
Fomina E. V., Kukoba T. B. Osnovy primeneniya rezul'tatov kosmicheskoi deyatel'nosti v interesakh meditsiny: uchebnoye posobie [Fundamentals of the application of the results of space activities in the interests of medicine: a textbook]. — Moscow: MPGU, 2019 (in Russian).
- Фомина Е. В., Лысова Н. Ю., Савинкина А. О., Жедяев Р. Ю., Сенаторова Н. А., Кукоба Т. Б. Роль стимуляции рецепторов опоры в локомоторных тренировках для профилактики гипогравитационных нарушений // Физиология человека. — 2021. — Т. 47, № 3. — С. 88–97.
Fomina E. V., Lysova N. Yu., Savinkina A. O., Zhedyaev R. Yu., Senatorova N. A., Kukoba T. B. Role of support receptor stimulation in locomotor training for the prevention of hypogravitational disorders // Human Physiology. — 2021. — Vol. 47, No. 3. — P. 320–327. (Original Russian paper: *Fomina E. V., Lysova N. Yu., Savinkina A. O., Zhedyaev R. Yu., Senatorova N. A., Kukoba T. B. Rol' stimulyatsii retseptorov opory v lokomotornykh trenirovkakh dlya profilaktiki gipogravitatsionnykh narushenii // Fiziologiya cheloveka.* — 2021. — Vol. 47, No. 3. — P. 88–97.)
- Фомина Е. В., Лысова Н. Ю., Чернова М. В., Хустнудинова Д. Р., Козловская И. Б. Сравнительный анализ профилактической эффективности различных режимов локомоторных тренировок в условиях космического полета // Физиология человека. — 2016. — Т. 42, № 5. — С. 84–91.

- Fomina E. V., Lysova N. Yu., Chernova M. V., Hustnudinova D. R., Kozlovskaya I. B.* Comparative analysis of preventive efficacy of different modes of locomotor training in space flight // *Human Physiology*. — 2016. — Vol. 42, No. 5. — P. 539–545. (Original Russian paper: *Fomina E. V., Lysova N. Yu., Chernova M. V., Hustnudinova D. R., Kozlovskaya I. B.* Sravnitel'nyi analiz profilakticheskoi effektivnosti razlichnykh rezhimov lokomotornykh trenirovok v usloviyakh kosmicheskogo poleta // *Fiziologiya cheloveka*. — 2016. — Vol. 42, No. 5. — P. 84–91.)
- Фомина Е. В., Сенаторова Н. А., Кириченко В. В., Вагнер И. В.* МКС — платформа для разработки системы профилактики гипогравитационных нарушений в межпланетных миссиях // *Воздушно-космическая сфера*. — 2020. — Т. 105, № 4. — С. 8–17.
- Fomina E. V., Senatorova N. A., Kirichenko V. V., Vagner I. V.* MKS — platforma dlya razrabotki sistemy profilaktiki gipogravitatsionnykh narushenii v mezhplanetnykh missiyakh [ISS — A platform for developing a system for hypogravitational disturbances countermeasures in interplanetary missions] // *Vozdushno-kosmicheskaya sfera*. — 2020. — Vol. 105, No. 4. — P. 8–17 (in Russian).
- Arlot S.* Minimal penalties and the slope heuristics: a survey // *Journal de la Société Française de Statistique*. — 2019. — Vol. 160, No. 3. — P. 1–106.
- Arlot S., Celisse A., Harchaoui Z.* A kernel multiple change-point algorithm via model selection // *Journal of Machine Learning Research*. — 2019. — Vol. 20, No. 162. — P. 1–56.
- Auger I. E., Lawrence C. E.* Algorithms for the optimal identification of segment neighborhoods // *Bulletin of Mathematical Biology*. — 1989. — Vol. 51, No. 1. — P. 39–54.
- Bai J., Perron P.* Computation and analysis of multiple structural change models // *Journal of Applied Econometrics*. — 2003. — Vol. 18, No. 1. — P. 1–22.
- Brodsky B. E., Darkhovsky B. S.* Nonparametric methods in change point problems. — Netherlands: Springer, 1993.
- Cabrieto J., Tuerlinckx F., Kuppens P., Grassmann M., Ceulemans E.* Detecting correlation changes in multivariate time series: A comparison of four non-parametric change point detection methods // *Behavior Research Methods*. — 2017. — Vol. 49, No. 3. — P. 988–1005.
- Celisse A., Marot G., Pierre-Jean M., Rigail G.* New efficient algorithms for multiple change-point detection with reproducing kernels // *Computational Statistics & Data Analysis*. — 2018. — Vol. 128. — P. 200–220.
- Chu C.-S. J., Hornik K., Kaun C.-M.* MOSUM tests for parameter constancy // *Biometrika*. — 1995. — Vol. 82, No. 3. — P. 603–617.
- Desobry F., Davy M., Doncarli C.* An online kernel change detection algorithm // *IEEE Transactions on Signal Processing*. — 2005. — Vol. 53, No. 8. — P. 2961–2974.
- Einmahl J. H. J., McKeague I. W.* Empirical likelihood based hypothesis testing // *Bernoulli*. — 2003. — Vol. 9. — P. 267–290.
- Fan Y., de Micheaux P. L., Penev S., Salopek D.* Multivariate nonparametric test of independence // *Journal of Multivariate Analysis*. — 2017. — Vol. 153. — P. 189–210.
- Frick K., Munk A., Sieling H.* Multiscale change point inference // *Journal of the Royal Statistical Society. Series B (Methodological)*. — 2014. — Vol. 76, No. 3. — P. 495–580.
- Gretton A., Sejdinovic D., Strathmann H., Balakrishnan S., Pontil M., Fukumizu K., Sriperumbudur B. K.* Optimal kernel choice for large-scale two-sample tests // *Proc. of 26th Annual Conference on Neural Information Processing Systems*. — Lake Tahoe, US, 2012. — P. 1205–1213.
- Hagio S., Ishihara A., Terada M., Tanabe H., Kibushi B., Higashibata A., Yamada S., Furukawa S., Mukai C., Ishioka N., Kouzaki M.* Muscle synergies of multidirectional postural control in astronauts on Earth after a long-term stay in space // *Journal of Neurophysiology*. — 2022. — Vol. 127, No. 5. — P. 1230–1239.
- Harchaoui Z., Cappé O.* Retrospective multiple change-point estimation with kernels // *Proc. of IEEE/SP 14th workshop on statistical signal processing SSP'07*. — Piscataway, NJ: IEEE, 2007. — P. 768–772.
- Haynes K., Eckley I. A., Fearnhead P.* Computationally efficient changepoint detection for a range of penalties // *Journal of Computational and Graphical Statistics*. — 2017. — Vol. 26, No. 1. — P. 134–143.

- Haynes K., Fearnhead P., Eckley I.A. A computationally efficient nonparametric approach for changepoint detection // *Statistics and Computing*. — 2017. — Vol. 27, No. 5. — P. 1293–1305.
- Huskova M., Slaby A. Permutation tests for multiple changes // *Kybernetika*. — 2001. — Vol. 37, No. 5. — P. 605–622.
- Jackson B., Sargle J.D., Barnes D., Arabhi S., Alt A., Gioumousis P., Gwin E., Sangtrakulcharoen P., Tan L., Tsai T.T. An algorithm for optimal partitioning of data on an interval // *IEEE Signal Processing Letters*. — 2005. — Vol. 12. — P. 105–108.
- Jones C., Harchaoui Z. Project chapydetete. — [Electronic resource]. — 2019. — <https://github.com/cjones6/chapydetete> (accessed: 05.12.2023).
- Kanamori T., Hido S., Sugiyama M. A least-squares approach to direct importance estimation // *Journal of Machine Learning Research*. — 2009. — Vol. 10. — P. 1391–1445.
- Kay S. *Fundamentals of statistical signal processing: detection theory*. — Prentice-Hall Inc., 1993.
- Killick R., Fearnhead P., Eckley I.A. Optimal detection of changepoints with a linear computational cost // *Journal of the American Statistical Association*. — 2012. — Vol. 107. — P. 1590–1598.
- Lavielle M. Using penalized contrasts for the change-points problems // *Signal Processing*. — 2005. — Vol. 85, No. 8. — P. 1501–1510.
- Lebarbier E. Detecting multiple change-points in the mean of Gaussian process by model selection // *Signal Processing*. — 2005. — Vol. 85, No. 4. — P. 717–736.
- Leucht A., Neumann M.H. Dependent wild bootstrap for degenerate U- and V-statistics // *Journal of Multivariate Analysis*. — 2013. — Vol. 117. — P. 257–280.
- Liu S., Yamada M., Collier N., Sugiyama M. Change-point detection in time-series data by relative density-ratio estimation // *Neural Networks*. — 2013. — Vol. 43. — P. 72–83.
- Lung-Yut-Fong A., Lévy-Leduc C., Cappé O. Homogeneity and change-point detection tests for multivariate data using rank statistics // *Journal de la Société Française de Statistique*. — 2015. — Vol. 156, No. 4. — P. 133–162.
- Marot G., Celisse A., Rigai G. R-package KernSeg 2018. — [Electronic resource]. — https://r-forge.r-project.org/R/?group_id=2300 (accessed: 05.12.2023).
- McGonigle E.T., Cho H. Nonparametric data segmentation in multivariate time series via joint characteristic functions. — 2023. — <https://doi.org/10.48550/arXiv.2305.07581>.
- Norsk P. Adaptation of the cardiovascular system to weightlessness: Surprises, paradoxes and implications for deep space missions // *Acta Physiologica*. — 2020. — Vol. 228, No. e13434. — P. 1–16.
- Oliver J.L., Carpena P., Hackenberg M., Bernaola-Galvan P. IsoFinder: Computational prediction of isochores in genome sequences // *Nucleic Acids Research*. — 2004. — Vol. 32. — P. 287–292.
- Picard F., Lebarbier E., Budinska E., Robin S. Joint segmentation of multivariate Gaussian processes using mixed linear models // *Computational Statistics and Data Analysis*. — 2011. — Vol. 55, No. 2. — P. 1160–1170.
- Schwarz G. Estimating the dimension of a model // *The Annals of Statistics*. — 1978. — Vol. 6, No. 2. — P. 461–464.
- Shelhamer M., Bloomberg J., LeBlanc A.G., Prisk K., Sibonga J., Smith S.M., Zwart S.R., Norsk P. Selected discoveries from human research in space that are relevant to human health on Earth // *npj Microgravity*. — 2020. — Vol. 6, No. 5. — P. 1–5.
- Sriperumbudur B.K., Fukumizu K., Lanckriet G.R. On the relation between universality, characteristic kernels and RKHS embedding of measures // *Proc. of 13th International Conference on Artificial Intelligence and Statistics*. — Sardinia, Italy, 2010. — P. 773–780.
- Sugiyama M., Suzuki T., Nakajima S., Kashima H., von Buena P., Kawanabe M. Direct importance estimation for covariate shift adaptation // *Annals of the Institute of Statistical Mathematics*. — 2008. — Vol. 60, No. 4. — P. 699–746.

- Tartakovsky A. G., Nikiforov I. V., Basseville M.* Sequential analysis hypothesis testing and changepoint detection. — Boca Raton, London, New York: Chapman and Hall/CRC Press, 2014.
- Truong C., Oudre L., Vayatis N.* Selective review of offline change point detection methods // *Signal Processing*. — 2020. — Vol. 167. — P. 107299.
- Vert J., Bleakley K.* Fast detection of multiple change-points shared by many signals using group LARS // *Advances in Neural Information Processing Systems*. — 2010. — P. 1–9.
- Wang T., Samworth R. J.* High dimensional change point estimation via sparse projection // *Journal of the Royal Statistical Society: Series B*. — 2018. — Vol. 80, No. 1. — P. 57–83.
- Xie L., Zou S., Xie Y., Veeravalli V. V.* Sequential (quickest) change detection: classical results and new directions // *IEEE Journal on Selected Areas in Information Theory*. — 2021. — Vol. 2, No. 2. — P. 494–514.
- Yamada M., Suzuki T., Kanamori T., Hachiya H., Sugiyama M.* Relative density-ratio estimation for robust distribution comparison // *Neural Computation*. — 2013. — Vol. 25, No. 5. — P. 1324–1370.
- Yudin N., Kamzolov D., Sinolits V., Golovkin P., Erchenko A.* Embedded online machine learning // *Proc. of 2021 International Conference Engineering and Telecommunication (En&T)*. — Dolgoprudny, Russia, 2021. — P. 1–5.
- Zou C., Liu Y., Qin P., Wang Z.* Empirical likelihood ratio test for the change-point problem // *Statistics & Probability Letters*. — 2007. — Vol. 77. — P. 374–382.
- Zou C., Yin G., Feng L., Wang Z.* Nonparametric maximum likelihood approach to multiple change-point problems // *The Annals of Statistics*. — 2014. — Vol. 42, No. 3. — P. 970–1002.