ИНДИКАТОР РАЗЛАДКИ ДЛЯ НЕСТАЦИОНАРНЫХ СЛУЧАЙНЫХ ПРОЦЕССОВ

А. А. Кислицын^{1,*}, А. Б. Козлова², М. Б. Корсакова², Ю. Н. Орлов¹

Представлено академиком РАН Б.Н. Четверушкиным 16.08.2018 г.

Поступило 30.08.2018 г.

В работе исследуются свойства статистики, называемой согласованным уровнем стационарности нестационарных временных рядов. Показывается, что её изменение можно трактовать как разладку. Рассматриваются такие задачи, как оценка уровня значимости принятия решений, выделение характерных периодов в нестационарном случайном процессе, определение оптимальной длины выборки для построения индикаторов в задачах стохастического управления. В качестве практической реализации исследуется индикатор разладки на данных электроэнцефалограмм у больных эпилепсией.

Ключевые слова: точки стационарности, уровень значимости, индекс нестационарности, индикатор разладки, эпилепсия, согласованный уровень стационарности, ЭЭГ, эквивалентность по Чернову.

DOI: <https://doi.org/10.31857/S0869-56524844393-396>

В настоящей работе строится статистический индикатор разладки для нестационарного случайного процесса. В отличие от большинства статистических критериев, оперирующих либо непосредственно с элементами временных рядов, либо предполагающих стационарность соответствующих функций распределения (см., например, [1, 2]), введённый нами функционал представляет собой определённый квантиль выборочной функции распределения расстояний в норме C между выборочными функциями распределения (ВФР), построенными по непересекающимся выборкам временного ряда. Предлагаемый подход эффективно применяется к выборкам больших длин (от нескольких миллионов данных).

Цель построения индикатора разладки применительно к большим данным состоит в том, чтобы на его основе иметь возможность предсказывать изменение состояния анализируемой системы, предполагая, что первые признаки такого изменения могут быть выявлены посредством анализа эволюции ВФР [3]. Характерным практическим примером служат ряды показаний электроэнцефалограмм (ЭЭГ) у больных

эпилепсией [4, 5]. Эти данные образуют многомерный (по числу отведений), обычно 64-мерный массив векторов длиной от 10 до 100 млн каждый. Традиционно применяемые в задаче предсказания приступа методы, основанные на анализе спектра мощности сигнала ЭЭГ [4, 6], дают весьма большую ошибку, связанную с тем, что данные методы корректно применимы к стационарным случайным процессам, тогда, как будет показано ниже, ряды ЭЭГ сильно нестационарны. В этом случае разладкой следует считать изменение уровня нестационарности временного ряда, если предположить, что для каждого состояния пациента (сон, бодрствование, болезнь, определённое занятие) характерен свой уровень нестационарности ЭЭГ.

Как известно [7, 8], решение задачи о принадлежности двух ВФР одной генеральной совокупности использует непараметрический критерий Колмогорова–Смирнова. Строится статистика

$$S_N = \sup_x |F_{1,N}(x) - F_{2,N}(x)|, \quad (1)$$

для которой справедлива асимптотика

$$\lim_{N \rightarrow \infty} P \left\{ 0 < \sqrt{\frac{N}{2}} S_N < z \right\} = K(z), \quad (2)$$

где $K(z)$ есть табулированная функция Колмогорова, а N — длина выборки. Неубывающая непрерывная функция $F_N(x)$ есть ВФР случайной величины ξ , значения которой x наблюдаются в виде последовательности событий во временном

¹ Федеральный исследовательский центр
Институт прикладной математики им. М.В. Келдыша
Российской Академии наук, Москва

² Научный медицинский исследовательский центр
нейрохирургии им. Н.Н. Бурденко Минздрава России,
Москва

* E-mail: alexey.kislitsyn@gmail.com

окне длины N . В формуле (2) уровень значимости Q , на котором принимается решение, считается равным $1-K(z)$. Пусть ε — расстояние между выборками в норме C , определяемое формулой (1). Рассмотрим стационарную точку уровня значимости, определяемую соотношением

$$Q(\varepsilon) \equiv 1 - K\left(\sqrt{\frac{N}{2}}\varepsilon\right) = \varepsilon. \quad (3)$$

Для стационарных распределений положение стационарной точки $\varepsilon = \varepsilon_0(N)$ как решения уравнения (3) не зависит от вида распределения F . Это решение единственно в силу монотонности функции $K(z)$. Табулированные значения $\varepsilon_0(N)$ приведены в [9].

Для нестационарных ВФР распределение расстояний между выборками заданной длины отличается от статистики (2). Обозначим через $G_N(\rho)$ эмпирическую функцию распределения расстояний $\rho(N)$ между двумя непересекающимися выборками длины N :

$$\rho(N) = \|F_{1,N}(x) - F_{2,N}(x)\|_C. \quad (4)$$

Численное решение относительно ρ уравнения

$$G_N(\rho) = 1 - \rho \quad (5)$$

даёт значение согласованного уровня стационарности (СУС) $\rho^*(N)$ как вероятности того, что расстояние между выборками длины N больше ρ^* . Если оказалось, что $\rho^*(N) > \varepsilon_0(N)$, то ВФР считается нестационарной. Величина $\rho^*(N)$ есть уровень значимости, на котором можно принимать решение об однородности выборок.

Если считать, что величина СУС представляет собой характеристику данного случайного процесса, то изменение его можно было бы трактовать как разладку. Практический вопрос, возникающий в таком случае, состоит в длине окна, где считается индекс.

Пусть по длине L_{tot} фактически известно-го фрагмента временного ряда построен СУС как решение уравнения (5) для встык-выборок длины N . Следует учесть, что $L_{\text{tot}} \gg 2N$, так что на длине L_{tot} укладывается достаточно много длин N , чтобы можно было набрать статистику для построения распределения $G_N(\rho)$. На этой длине L_{tot} в доле случаев $\rho^*(N)$ расстояние между встык-выборками больше, чем $\rho^*(N)$. Если теперь в определённом скользящем окне длины L доля расстояний между встык-выборками длины N оказалась больше, чем вычисленная ранее величина $\rho^*(N)$, то в этом окне фиксируется разладка. Для такого вывода требуется, чтобы СУС

нестационарного ряда был случайной величиной, имеющей стационарное распределение.

Исследуя локальное значение СУС — т.е. значение, полученное по промежутку времени L , существенно меньшему, чем первоначальный объём данных L_{tot} , естественно возникает вопрос о флуктуациях этого локального СУС относительно базового значения, отвечающего множеству в целом. Важно понимать, что СУС всего множества $\rho_{\text{tot}}^*(N)$ не является средним значением последовательности СУС $\rho_n^*(N, L)$, построенной для встык-выборок длины N в окнах длины L . Рассмотрим функцию

$$\Psi_N(\rho) = 1 - G_N(\rho). \quad (6)$$

Предположим, что ВФР $G_N(\rho)$ аппроксимирует дифференцируемую функцию $G(\rho)$ соответствующей генеральной совокупности. Тогда функция $\Psi_N(\rho)$ обладает свойствами $\Psi_N(0) = 1$, $\Psi'_N(0) = -a_N \leq 0$. В таком случае существует предел

$$\lim_{n \rightarrow \infty} \left(\Psi_N\left(\frac{\rho}{n}\right) \right)^n = e^{-\rho a_N} \equiv \Phi_N(\rho). \quad (7)$$

Следуя определениям, данным в [10–12] для операторных функций, будем называть предельную функцию $\Phi_N(\rho)$ эквивалентной по Чернову функции $\Psi_N(\rho)$ (т.е. уровню значимости распределения $G_N(\rho)$). Будем обозначать эту эквивалентность как $\Phi_N(\rho) \propto_{\text{Ch}} \Psi_N(\rho)$. Очевидно, $|\Phi_N(\rho) - \Psi_N(\rho)| = o(\rho)$.

Согласно [11], если имеется (конечный или бесконечный) набор функций $\Psi_N^k(\rho)$ вида (6), каждая из которых эквивалентна в смысле (7) функции $\Phi_N^k(\rho)$ с коэффициентом $-a_N^k$ в показателе экспоненты, и если задан соответствующий набор неотрицательных коэффициентов p_k таких, что $\sum_k p_k = 1$, то средняя функция $\bar{\Psi}_N(\rho) = \sum_k p_k \Psi_N^k(\rho)$ эквивалентна в смысле (7) функции $\Phi_N(\rho) = e^{-\rho \bar{a}_N}$, где

$$\bar{a}_N = \sum_k p_k a_N^k. \quad (8)$$

Рассмотрим последовательность непересекающихся промежутков длины L . Для каждого k -го промежутка построим эмпирическое распределение $G_N^k(\rho, L)$ расстояний между встык-выборками длины N . Пусть таких промежутков n , так что $nL = L_{\text{tot}}$ — полная длина изучаемого фрагмента ряда. Тогда распределение расстояний,

построенное по всему объёму данных, есть среднее распределение, полученное усреднением по отдельным выборкам:

$$G_N(\rho, L_{\text{tot}}) = \frac{1}{n} \sum_{k=1}^n G_N^k(\rho, L), \quad (9)$$

$$\bar{\Psi}_N(\rho) = \sum_k p_k \Psi_N^k(\rho) = 1 - G_N(\rho, L_{\text{tot}}), \quad p_k = \frac{1}{n}.$$

Используя результаты [11], можно доказать следующую теорему о стационарных точках распределений с непрерывными плотностями.

Т е о р е м а . Пусть распределения случайных величин имеют непрерывные плотности и пусть на множестве этих случайных величин задана некоторая неотрицательная мера. Тогда стационарная точка функции, эквивалентной по Чернову среднему уровню значимости данных распределений, с точностью до бесконечно малой второго порядка совпадает с обратной величиной к среднему значению обратных величин стационарных точек функций, эквивалентных по Чернову уровням значимости данных распределений:

$$\frac{1}{\tilde{\rho}(N)} = \sum_{k=1}^n \frac{p_k}{\tilde{\rho}_k(N)} \equiv \frac{1}{\bar{\rho}(N)}. \quad (10)$$

Формула (10) принципиальна, поскольку она позволяет значительно уменьшить количество вычислительных процедур, если требуется изучить поведение СУС на множествах, представляющих собой объединение некоторых других множеств, в задачах анализа больших данных.

Применительно к задаче анализа ЭЭГ для унификации работы с данными при проведении вычислений используется модуль библиотеки —EDF-browser, который служит для конвертации исходных файлов в формат ASCII. Алгоритм, вычисляющий СУС и строящий эквивалентные по Чернову стационарные точки выборочных распределений, реализован на языке C++ в среде разработки MSVS 2017. Поскольку обрабатывается большое количество векторов данных, вычисления распараллеливаются с использованием интерфейса OpenMP.

Опишем кратко результаты численного анализа данных для реальных пациентов. Характерно, что СУС данных ЭЭГ для них в 4–7 раз превышал стационарный уровень значимости $\varepsilon_0(N)$. Будем следовать работе [5], где минимальная длина окна, в котором вычислялся индикатор разладки, была принята равной 2 мин (при частоте 250 Гц в этом промежутке укладывалось $L = 30$ тыс. данных ЭЭГ по каждому отведению). Оптимальная длина встык-выборок, между которыми вычислялось расстояние

ρ , была принята равной $N = 5000$. Расстояния между встык-выборками вычислялись в скользящем окне длиной 10 000 с шагом 500. Таким образом, для данной длины L получилось 40 вычисленных расстояний между ВФР встык-выборок. Общий объём данных, отвечающих визуально не меняющемуся состоянию бодрствования пациента, составил 2,5 ч, т.е. примерно 2,3 млн данных, или $n = 78$ независимых двухминутных отрезков. Для каждого из этих двухминутных промежутков и для каждого отведения по имеющимся 40 данным вычислялся СУС $\rho_k^*(5000, 30\,000)$, $k = 1, 2, \dots, n = 78$.

Для стационарных выборок длиной 40 данных теоретический согласованный уровень значимости составляет $\varepsilon_0(40) = 0,25$. В приведённом примере (как и для всех 64 отведений рассматриваемых данных ЭЭГ) СУС меняется от 0,245 до 0,255. Это означает, что ряд СУС стационарный, хотя сам исходный ряд показаний ЭЭГ, для которых этот СУС посчитан, не является стационарным.

В работе [5] было выяснено, что лишь некоторые отведения (порядка 10) могут служить индикаторами-предвестниками приступа эпилепсии. Это те отведения, для которых относительное превышение индекса нестационарности в последние две минуты перед приступом по сравнению со спокойным состоянием составляет величину не менее, чем СУС. В работе продемонстрировано, что по каждому типу электродов — индикативному и не индикативному — распределение СУС является стационарным. Иными словами, тип нестационарного поведения временного ряда не меняется в течение того промежутка времени, пока состояние пациента визуально считается стабильным. При этом детерминация аппроксимаций рассмотренных зависимостей для “обычных” электродов заметно выше, чем для индикативных, т.е. эти “обычные” электроды порождают более стационарный ряд данных.

Были рассмотрены уровни нестационарности для состояний, которые предположительно можно считать различными: это, во-первых, состояния бодрствования и сна без приступов эпилепсии и, во-вторых, состояния, предшествующие двум приступам эпилепсии. Выяснилось, что распределения СУС этих состояний различаются значительно сильнее, чем это можно ожидать для выборок из одной генеральной совокупности. При этом существенно, что для каждого состояния показания ЭЭГ имеют стационарно распределённый ряд значений СУС.

СПИСОК ЛИТЕРАТУРЫ

1. *Королук В.С., Портенко Н.И., Скороход А.В., Турбин А.Ф.* Справочник по теории вероятностей и математической статистике. М.: Наука, 1985. 640 с.
2. *Кобзарь А.И.* Прикладная математическая статистика. М.: Физматлит, 2006. 816 с.
3. *Орлов Ю.Н.* Кинетические методы исследования нестационарных временных рядов. М.: МФТИ, 2014. 276 с.
4. Нейрофизиологические исследования в клинике / Под ред. Г. А. Щекутева. М.: Антидор, 2001. 236 с.
5. *Kislitsyn A.A., Kozlova A.B., Masherov E.L., Orlov Yu.N.* Numerical Algorithm for Self-Consistent Stationary Level for Multidimensional Non-Stationary Time-Series. Keldysh Inst. Prepr. M., 2017. № 124. 14 p.
6. *The Fourier Transform in Biomedical Engineering. T.M. Peters, J. Williams (Eds.).* Basel: Birkhäuser, 1998. 199 p.
7. *Kolmogoroff A.N.* Sulla determinazione empirica di una legge di distribuzione // *Giornale Ist. Italiano degli Attuari.* 1933. V. 4. № 1. P. 83–91.
8. *Гнеденко Б.В.* Курс теории вероятностей. М.: Физматлит, 1961. 406 с.
9. *Кислицын А.А., Козлова А.Б., Корсакова М.Б., Машеров Е.Л., Орлов Ю.Н.* Стационарная точка уровня значимости для нестационарных функций распределения. Препр. ИПМ им. М.В. Келдыша. М., 2018. № 113. 20 с.
10. *Smolyanov O.G., Weizsacker H., Wittin O.* Chernoff's Theorem and Discrete time Approximation of Brownian Motion on Manifolds // *Potential Anal.* 2007. V. 26. P. 1–29.
11. *Орлов Ю.Н., Сакбаев В.Ж., Смолянов О.Г.* Формулы Фейнмана как метод усреднения случайных гамильтонианов // *Тр. Мат. ин-та РАН.* 2014. Т. 285. С. 232–243.
12. *Бутко Я.А., Смолянов О.Г.* Формулы Фейнмана в стохастической и квантовой динамике. В сб.: *Современные проблемы математики и механики.* М., 2011. Т. 6. № 1. С. 61–75.

DISORDER INDICATOR FOR NONSTATIONARY STOCHASTIC PROCESSES

A. A. Kislitsyn, A. B. Kozlova, M. B. Korsakova, Yu. N. Orlov

Presented by Academician of the RAS B.N. Chetverushkin August 16, 2018

Received August 30, 2018

The properties of a statistic called a self-consistent stationarity level of nonstationary time series are examined in this work. It is shown that a change in this statistic can be treated as a disorder in the nonstationarity properties of the series. The significance level of decision making is estimated, characteristic periods in a non-stationary stochastic process are detected, and an optimal sample size for constructing indicators in stochastic control problems is determined. A disorder indicator for electroencephalogram data from epilepsy patients is studied as a practical application.

Keywords: stationary points, significance level, non-stationary index, disorder indicator, epileptic seizure, self-consistent stationary level, electroencephalogram, Chernoff equivalence.