На правах рукописи

Артёмов Алексей Валерьевич

МАТЕМАТИЧЕСКИЕ МОДЕЛИ ВРЕМЕННЫХ РЯДОВ С ТРЕНДОМ В ЗАДАЧАХ ОБНАРУЖЕНИЯ РАЗЛАДКИ

05.13.18 — математическое моделирование,
численные методы и комплексы программ

АВТОРЕФЕРАТ
диссертации на соискание учёной степени
кандидата физико-математических наук

МОСКВА — 2016

Работа выполнена в лаборатории №10 «Интеллектуальный анализ данных и предсказательное моделирование» ФГБУ науки Института проблем передачи информации им. А. А. Харкевича Российской академии наук.

Научный руководитель: **Бурнаев Евгений Владимирович,**
кандидат физико-математических наук, доцент,
заведующий лабораторией №10 Института проблем передачи информации им. А. А. Харкевича РАН.

Официальные оппоненты: **Стрижов Вадим Викторович,**
доктор физико-математических наук,
научный сотрудник Федерального государственного учреждения «Федеральный исследовательский центр «Информатика и управление» РАН;

Житлухин Михаил Валентинович,
кандидат физико-математических наук,
научный сотрудник отдела теории вероятностей и математической статистики Математического института им. В. А. Стеклова РАН.

Ведущая организация: ФГБУ науки Институт проблем управления им. В. А. Трапезникова РАН.

Защита состоится «____» _____ 2017 г. в 11:00 на заседании диссертационного совета Д 002.073.04 при Федеральном государственном учреждении «Федеральный исследовательский центр «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН) по адресу: 117312, Москва, пр. 60-летия Октября, 9 (конференц-зал, 1 этаж).

С диссертацией можно ознакомиться в библиотеке ФИЦ ИУ РАН, Москва, ул. Вавилова, д. 40.

Электронные версии диссертации и автореферата расмещены на официальном сайте ФИЦ ИУ РАН `http://www.frccsc.ru`.

Электронная версия автореферата размещена на официальном сайте ВАК Министерства образования и науки РФ по адресу `http://vak.ed.gov.ru`.

Отзывы и замечания по автореферату в двух экземплярах, заверенные печатью, просьба высылать по адресу 117312, Москва, пр. 60-летия Октября, 9, ФИЦ ИУ РАН, диссертационный совет Д 002.073.04.

Автореферат разослан ____.____._____.

Телефон для справок: +7 (499) 135-51-64.

Учёный секретарь диссертационного совета Д 002.073.04,
доктор технических наук, профессор — Крутько В. Н.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность проблемы. В последние десять лет при активном развитии информационно-коммуникационных технологий возник новый тип высокотехнологичных систем: системы с интенсивным программным обеспечением (software-intensive systems[1]). К этому типу относятся цифровые системы широкополосной связи, интернет-системы (включая устройства, сети передачи данных и интернет-сервисы), информатизированные центры обслуживания и колл-центры, интеллектуальные транспортные системы, платформы электронной торговли, автоматизированные системы в здравоохранении и др. Для всех этих систем характерны: большой объем накапливаемых и обрабатываемых данных, сложная взаимозависимость программных компонент и огромное количество аппаратных элементов, а также чрезвычайно большое число людей, использующих систему для различных целей. Как показывает изучение существующих аппаратно-программных комплексов, в настоящее время одной из их центральных проблем является низкая надежность их эксплуатации: неизбежные при крупном масштабе программные, аппаратные и антропогенные отказы являются на практике нормой, а не исключением[2]. Согласно исследованиям[3], доминирующей причиной системных отказов является именно возникновение сбоев программного обеспечения. Таким образом, обеспечение бесперебойной и эффективной эксплуатации высокотехнологичных систем с интенсивным ПО представляет собой крупную проблему, для устранения которой необходимо прежде всего предотвращение отказов их программной составляющей, в частности, их быстрое и точное обнаружение.

В ряде работ предложены подходы к обнаружению проблемного поведения ПО (дефектов, вредоносного вмешательства и т. д.) на основании данных, собираемых при эксплуатации системы — измерений количества обработанных запросов и средней длительности ожидания в единицу времени, измерения объема переданного сетевого трафика и т. д. Среди таких работ отметим работу Casas, 2010[4], где измерения объема передаваемого сетевого трафика используются для обнаружения перегрузок сетевого ядра, работу Tartakovsky, 2013[5],

[1]Система, функциональность которой определяется главным образом ее программными средствами, согласно стандарту ISO/IEC/IEEE, “Systems and software engineering – Architecture description,” in ISO/IEC/IEEE 42010:2011(E) (Revision of ISO/IEC 42010:2007 and IEEE Std 1471-2000), 2011.

[2] Yigitbasi N. et al. Analysis and modeling of time-correlated failures in large-scale distributed systems //Grid Computing (GRID), 2010 11th IEEE/ACM International Conference on. — IEEE, 2010. — Pp. 65–72.

[3] Northrop L. et al. Ultra-large-scale systems: The software challenge of the future. — Carnegie Mellon Software Engineering Institute, Ultra-Large-Scale Systems Study Report, 2006.

[4]Casas P. et al. Optimal volume anomaly detection and isolation in large-scale IP networks using coarse-grained measurements //Computer Networks. — 2010. — Vol. 54. — no. 11. — Pp. 1750–1766.

[5]Tartakovsky A. G., Polunchenko A. S., Sokolov G. Efficient computer network anomaly detection by changepoint detection methods //Selected Topics in Signal Processing, IEEE Journal of. — 2013. — Vol. 7. — no. 1. — Pp. 4–11.

в которой измерения профиля сетевого трафика применяются для детектирования внедрений в компьютерные сети. Рассмотренные в этих работах задачи сводятся к выявлению момента резкого изменения некоторых характеристик рассматриваемой системы на основе наблюдаемых статистических данных о других характеристиках этой системы. Задачи такого типа *(задачи о разладке)* были рассмотрены А. Н. Колмогоровым, А. Н. Ширяевым[6,7,8] и др. Однако на практике соответствующие методы детектирования разладок обладают рядом ограничений ввиду следующих особенностей сигналов реальных систем.

Будучи системами массового обслуживания, системы с интенсивным ПО испытывают антропогенные циклы нагрузки на ряде масштабов времени (день, неделя, год). В силу изменчивости на большом масштабе времени основной цикл будет стохастическим. Поэтому для успешного решения задачи обнаружения разладок сложных систем необходим эффективный аппарат математического моделирования и оценивания квазипериодических сигналов.

Значимой характеристикой потоков данных в информационных системах является также длинная память (long-range dependence). Длинная память является основной причиной возникновения всплесков нагрузки и присутствует на чрезвычайно большом диапазоне масштабов времени; известно ее значительное влияние на эффективность систем массового обслуживания[9]. Таким образом, для идентификации и оценивания реальных сигналов, порожденных системами с интенсивным ПО, необходимо использование специальных стохастических моделей, позволяющих моделировать длинную память.

Для решения задач обнаружения отказов реальных информационных систем естественно использовать специальные статистические процедуры обнаружения разладки, такие как метод кумулятивных сумм[10], метод контрольных карт[11],

[6]Ширяев А. Н. Задача скорейшего обнаружения нарушения стационарного режима // Доклады Академии наук. — 1961. — Т. 138, № 5. — С. 1039–1042.

[7]Ширяев А. Н. Обнаружение спонтанно возникающих эффектов // Доклады Академии наук. — 1961. — Т. 138. — С. 799–801.

[8]Колмогоров А. Н., Прохоров Ю. В., Ширяев А. Н. Вероятностно-статистические методы обнаружения спонтанно возникающих эффектов // Теория вероятностей, теория функций, механика, Сборник обзорных статей 5. К 50-летию Института. — Труды Математического Института им. В.А.Стеклова, Т. 182. — М.: Наука, 1988. — С. 4–23.

[9]Erramilli A., Narayan O., Willinger W. Experimental queueing analysis with long-range dependent packet traffic //IEEE/ACM Transactions on Networking (TON). — 1996. — Vol. 4. — no. 2. — Pp. 209–223.

[10]Page E. S. Continuous inspection schemes //Biometrika. — 1954. — Pp. 100–115.

[11]Shewhart W. A. Economic control of quality of manufactured product. — ASQ Quality Press, 1931.

процедуру Ширяева-Робертса[12,13], Байесовские подходы[14,15] и т. п., поскольку для этих процедур существуют теоретические результаты об эффективности обнаружения разладки. В свою очередь, для применения таких процедур требуется определить математическую модель возникающего отказа в терминах распределений наблюдаемых характеристик. На практике сделать это часто невозможно, так как типы возникающих отказов и сопутствующие им изменения статистических характеристик априори произвольны; как следствие, в этих задачах могут быть неэффективны даже теоретически оптимальные методы обнаружения разладки[16].

В области машинного обучения широко известен подход на основе алгоритмической композиции или ансамбля, который заключается в совместном использовании множества «слабых» алгоритмов для получения лучшей предсказательной силы[17]. Согласно композиционному подходу, процедуры обнаружения разладки, для которых сигналы тревоги слабо (однако больше, чем просто случайно) коррелируют с истинными разладками, естественно рассматривать как «слабые» детекторы разладки. В этих условиях для эффективного обнаружения разладки достаточно использовать ее стандартную математическую модель[18] и для каждого класса наблюдений, представленного обучающей выборкой, выбрать наиболее эффективную композицию.

В последние десять лет возникли существенно новые практические условия, в которых беспрецедентные объемы данных обостряют проблему высокоэффективного автоматизированного обнаружения разладок современных больших систем[2]. В этих условиях возникают и новые усиленные требования к методологии и алгоритмике решения описываемых задач. До сих пор не было предложено единой архитектуры, пригодной для обнаружения разладок сложных естественных и инженерных систем крупного размера.

Таким образом, для обнаружения отказов крупных систем с интенсивным ПО актуально исследование методов моделирования сигналов с квазиперио-

---

[12]Ширяев А. Н. Об оптимальных методах в задачах скорейшего обнаружения //Теория вероятностей и ее применения. — 1963. — Т. 8. — № 1. — С. 26–51.

[13]Roberts S. W. A comparison of some control chart procedures //Technometrics. — 1966. — Т. 8. — № 3. — С. 411–430.

[14]Girshick M. A., Rubin H. A Bayes approach to a quality control model //The Annals of mathematical statistics. — 1952. — Pp. 114–125.

[15]Ширяев А. Н. Задача скорейшего обнаружения нарушения стационарного режима //Докл. АН СССР. — 1961. — Т. 138. — №. 5. — С. 1039–1042.

[16]Lai T. L., Xing H. Sequential change-point detection when the pre- and post-change parameters are unknown //Sequential Analysis. — 2010. — Vol. 29. — no. 2. — Pp. 162–175.

[17]Schapire R. E., Freund Y. Boosting: Foundations and algorithms. — MIT press, 2012.

[18]В литературе, как правило, стандартная модель разладки заключается в изменении среднего значения стационарной гауссовской случайной последовательности. В этом случае наблюдаемый процесс $\xi = (\xi_t)_{t\geqslant 0}$ имеет вид $\xi_t = \mu \mathbb{1}_{\{t\geqslant\theta\}}(t) + \nu_t$, где $\mu \in \mathbb{R}$ — магнитуда разладки, $\theta \geqslant 0$ — момент появления разладки, и $\nu = (\nu_t)_{t\geqslant 0}$ — последовательность независимых стандартно нормально распределенных случайных величин.

дическим трендом и с шумовой компонентой, обладающей длинной памятью, исследование методов обнаружения разладки в случае нарушения стандартных предположений о ее модели, а также разработка единой масштабируемой программной архитектуры для обнаружения разладок и аномалий в условиях больших объемов данных.

Целью работы являются разработка и исследование математических методов, алгоритмов и комплексов программ обнаружения разладок и аномалий больших динамических систем при наличии квазипериодических трендов, шумовой компоненты с длинной памятью, в случае нарушения стандартных предположений о модели разладки. Для достижения поставленной цели в работе рассматривались следующие **задачи исследования:**

- разработка и исследование математических методов оценки параметров сигнала по данным измерений, выполненных во фрактальном шуме;
- разработка и исследование алгоритма обнаружения разладки на основе ансамбля «слабых» детекторов для повышения эффективности обнаружения разладки в случае нарушения стандартных предположений о ее модели;
- разработка математических моделей и алгоритмов оценивания сигналов с трендом (в частности, квазипериодического сигнала) и обнаружения разладок и аномалий на фоне тренда;
- создание комплекса программ, реализующих разработанные методы для решения модельных и реальных задач обнаружения разладки.

Общая методика исследования. В диссертационной работе используются подходы стохастического анализа, теории непараметрического оценивания сигналов, методы численной оптимизации выпуклых функций. Комплекс программ, реализующий алгоритмы фильтрации и методы обнаружения разладки, выполнен на языке `python` в виде модульной системы с использованием подходов объектно-ориентированного программирования.

Научная новизна результатов, полученных в диссертационной работе, состоит в том, что в ней

1. Впервые поставлены и решены задачи фильтрации сигнала, представляемого в виде разложения по заданной системе функций, по данным его регистрации во фрактальном шуме и при различных типах дополнительной информации о сигнале.
2. Впервые разработан и исследован алгоритм обнаружения разладки временного ряда, основанный на совместном использовании множества процедур обнаружения разладки.

3. Предложены и исследованы математические модели временных рядов с трендом (в частности, квазипериодического временного ряда) и обнаружения разладок и аномалий на фоне тренда.
4. Создано и внедрено в производство в компании «Яндекс» новое программное обеспечение, реализующее методы оценки параметров и процедуры обнаружения разладок реальных сигналов.

Теоретическая значимость работы. Результаты диссертационной работы, имеющие теоретический характер, относятся к теории оптимальной фильтрации фрактальных динамических систем. Они позволяют теоретически исследовать фильтры, основанные на конкретных системах функций и могут применяться при построении и оценке эффективности компонент автоматизированных информационных систем, используемых для решения задач прогнозирования сигналов. Результаты диссертационной работы, относящиеся к методам оценивания квазипериодических трендов и обнаружения краткосрочных разладок и аномалий, имеют **практическую значимость** и были успешно применены для решения следующих прикладных задач:

1. Задача оценки параметров наблюдаемых сигналов больших информационных систем компании «Яндекс» в режиме реального времени.
2. Задача обнаружения отказов программного обеспечения больших информационных систем компании «Яндекс» в режиме реального времени.
3. Задача оценки нагрузки сети передачи данных Абилин на основе измерений объема передаваемого между узлами сети трафика.

На защиту выносятся следующие **научные результаты,** носящие теоретический и прикладной характер:

1. Разработаны новые математические методы оценки параметров сигнала по данным измерений, выполненным во фрактальном шуме, в том числе:
   - получена оценка максимального правдоподобия параметра сигнала;
   - получены оптимальные Байесовские оценки для случаев нормального и равномерного априорных распределений параметра сигнала;
   - охарактеризован оптимальный момент остановки измерений сигнала для случая нормального априорного распределения параметра сигнала.
2. Разработан и исследован алгоритм обнаружения разладки временного ряда на основе ансамбля процедур обнаружения разладки, предложен метод настройки параметров ансамбля.
3. Предложена и исследована методология моделирования квазипериодических сигналов и обнаружения их разладок, в том числе:
   - предложена математическая модель квазипериодического временного ряда на основе разложения по заданной системе функций и вычис-

лительный алгоритм оценки ее параметров на основе оптимального фильтра п. 1;
- предложена многокомпонентная математическая модель квазипериодического временного ряда и вычислительный алгоритм оценки ее параметров на основе непараметрической регрессии;
- предложена математическая модель краткосрочной разладки квазипериодического временного ряда и процедура обнаружения этой разладки на основе ансамблей «слабых» детекторов.

4. Создан комплекс программ, реализующий предложенные в диссертационной работе вычислительные алгоритмы фильтрации тренда фрактального случайного сигнала, оценивания квазипериодического сигнала, настройки параметров ансамбля и обнаружения разладки временного ряда на основе ансамбля.

Научная обоснованность и достоверность полученных результатов гарантируется использованием строгих доказательств, основанных на хорошо изученных методах стохастического анализа; совпадением полученных оценок с известными результатами в частных случаях линейных задач; описаниями проведенных экспериментов, допускающими их воспроизводимость; успешным применением результатов исследования в реальных задачах обнаружения программных отказов систем с интенсивным ПО.

Апробация работы. Результаты работы докладывались и обсуждались на следующих научных и технических конференциях и семинарах:

1. Научный семинар кафедры математического моделирования и информатики физического факультета МГУ им. М. В. Ломоносова под руководством профессора Ю. П. Пытьева (05.03.2015).
2. Научный семинар «Математические методы в естественных науках» физического факультета МГУ им. М. В. Ломоносова под руководством профессора А. Н. Боголюбова (26.03.2015).
3. XXII международная научная конференция студентов, аспирантов и молодых учёных «Ломоносов-2015», 13–17 апреля 2015 г., Москва, Россия.
4. Научный семинар «Practical Machine Learning» компании «Яндекс» под руководством к. ф.-м. н. М. А. Ройзнера (04.06.2015).
5. Научный семинар «Математические модели информационных технологий» департамента анализа данных и искусственного интеллекта Высшей школы экономики под руководством профессора С. О. Кузнецова (18.06.2015).
6. Научный семинар отдела Интеллектуальных систем ВЦ РАН под руководством члена-корреспондента РАН К. В. Рудакова (24.06.2015).

7. Научный семинар «Случайные процессы и стохастический анализ» кафедры теории вероятностей механико-математического факультета МГУ им. М. В. Ломоносова под руководством академика РАН А. Н. Ширяева (23.09.2015)
8. Научный семинар Yandex Data Factory под руководством к. ф.-м. н. Е. А. Рябенко (09.10.2015).
9. Научный семинар лаборатории математического моделирования сложных естественных и инженерных систем МГУ им. М. В. Ломоносова под руководством доцента Е. А. Грачева (06.11.2015).
10. The 8th International Conference on Machine Vision, 19–21 November 2015, Barcelona, Spain.
11. 58-я научная конференция МФТИ, 23–28 ноября 2015 г., г. Долгопрудный, Россия.
12. Общемосковский постоянный научный семинар «Теория автоматического управления и оптимизации» ИПУ РАН им. В. А. Трапезникова под руководством профессора Б. Т. Поляка (11.12.2015).
13. Deep Machine Intelligence Workshop, Skolkovo Institute of Science and Technology, 4–5 June 2016, Moscow, Russia.
14. Международная конференция по стохастическим методам, 27 мая–03 июня 2016 г., пос. Абрау-Дюрсо, г. Новороссийск, Россия.
15. Международная конференция по алгебре, анализу и геометрии, 26 июня–2 июля 2016 г., г. Казань, Россия.
16. 9th European Summer School in Financial Mathematics, 29 August–2 September 2016, Pushkin, St. Petersburg, Russia.
17. Регулярный семинар «Структурные модели и глубинное обучение» ИППИ РАН им. А. А. Харкевича под руководством доцента Е. В. Бурнаева и профессора В. Г. Спокойного (18.10.2016).

Личный вклад автора в работах, выполненных с соавторами, состоит в следующем:

1. В работах [1,5–7] предложены модели квазипериодических сигналов и алгоритмы оценивания их параметров, проведены вычислительные эксперименты для оценки качества предложенной методологии обнаружения разладок.
2. В работе [2] проведен теоретический подсчет структуры оптимальных фильтров во всех случаях, а также численное исследование функции штрафа для случая нормального априорного распределения.
3. В работах [3, 4] предложен критерий качества процедур обнаружения разладки и алгоритм оптимизации этого критерия для ансамблей «слабых»

детекторов, проведены вычислительные эксперименты для оценки качества ансамблей.

Публикации. По теме диссертационной работы опубликовано 7 печатных работ, в том числе 1 работа в журнале из списка ВАК и 3 работы в журналах из списка Scopus. Список публикаций приведен в конце автореферата.

Структура и объем диссертации. Диссертация состоит из введения, пяти глав, заключения и списка литературы, включающего N наименования. Работа изложена на M страницах и содержит K рисунков.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** характеризуется проблематика исследования; приводится анализ известных результатов, связанных с темой диссертационной работы, и излагаются доводы в пользу актуальности последней; формулируются цели и задачи диссертационной работы; приводятся ее краткое содержание и основные результаты.

Первая глава посвящена исследованию задачи оценки параметров сигнала по данным его измерений, выполненным во фрактальном шуме (шуме с длинной памятью).

В **первом разделе** содержится постановка задачи оценивания тренда случайного процесса, управляемого фрактальным броуновским движением, и приводится обзор известных из литературы результатов по фильтрации тренда.

Стандартное фрактальное броуновское движения $B^H = \left(B_t^H\right)_{0\leqslant t\leqslant T}$ на $[0,T]$ с параметром $H \in (0,1)$ — это гауссовский процесс с непрерывными траекториями, такой, что

$$B_0^H = 0, \quad \mathrm{E}\, B_t^H = 0, \quad \mathrm{E}\, B_s^H B_t^H = \frac{1}{2}\left(t^{2H} + s^{2H} - |t-s|^{2H}\right).$$

Пусть на фильтрованном вероятностном пространстве $\left(\Omega,\mathcal{F},(\mathcal{F}_t)_{t\geqslant 0},\mathrm{P}\right)$ задан случайный процесс $\xi = (\xi_t)_{0\leqslant t\leqslant T}$, имеющий представление

$$\xi_t = f(t) + \sigma(t) B_t^H, \tag{1}$$

где $B^H = \left(B_t^H\right)_{0\leqslant t\leqslant T}$ — стандартное фрактальное броуновское движение с параметром $H \in (0,1)$, а коэффициенты сноса $f(t)$ и диффузии $\sigma(t)$ удовлетворяют условиям $\int_0^T |f(t)|dt < \infty$ и $\int_0^T |\sigma(t)|^2 dt < \infty$ соответственно, причем функция $\sigma(t)$ предполагается известной. Принимается, что коэффициент сноса $f(t)$

можно представить в виде разложения

$$f(t) = \sum_{i=0}^{n_\theta} \theta_i g_i(t) \tag{2}$$

по заданной системе функций $g_i(t)$, таких, что $\int_0^T |g_i(t)|dt < \infty, i = 0, \ldots, n_\theta$, а параметры $\theta_i, i = 0, \ldots, n_\theta$ — неизвестны. Принимаются векторные обозначения $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_{n_\theta})^\intercal$, $\boldsymbol{g}(t) = (g_0(t), \ldots, g_{n_\theta}(t))^\intercal$, в терминах которых

$$f(t) = \boldsymbol{\theta}^\intercal \boldsymbol{g}(t). \tag{3}$$

Рассматривается задача нахождения последовательной оценки значения $\boldsymbol{\theta}$ по наблюдениям $\{\xi_s, 0 \leqslant s \leqslant t\}$, доступным до момента времени t. Предлагается рассматривать оценку максимального правдоподобия и последовательную Байесовскую оценку. В случае оценки максимального правдоподобия $\boldsymbol{\theta}$ считается неизвестным детерминированным вектором параметров, и требуется отыскать оценку $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}} = \widehat{\boldsymbol{\theta}}_{\mathrm{ML}}(t)$, максимизирующую правдоподобие наблюдений. В случае Байесовской оценки предполагается, что $\boldsymbol{\theta}$ — случайный элемент $\mathbb{R}^{n_\theta+1}$, имеющий известное априорное распределение $p^{\boldsymbol{\theta}}(\mathbf{x}), \mathbf{x} \in \mathbb{R}^{n_\theta+1}$, и рассматривается задача нахождения такого последовательного правила оценивания $\widehat{\delta}_{\mathrm{BAYES}} = (\widehat{\varrho}_{\mathrm{BAYES}}, \widehat{\boldsymbol{\theta}}_{\mathrm{BAYES}})$, что

$$\mathrm{E}\left[c\widehat{\varrho}_{\mathrm{BAYES}} + \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\mathrm{BAYES}}\|^2\right] = \inf_{\delta \in \mathbb{D}} \mathrm{E}\left[c\varrho + \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|^2\right], \tag{4}$$

где $\mathbb{D} = \left\{\delta : \delta = \left(\varrho, \widehat{\boldsymbol{\theta}}\right)\right\}$ — класс правил оценивания с конечными моментами остановки $\varrho \leqslant T < \infty$ относительно фильтрации $\mathcal{F}_t^\xi = \sigma\left(\{\xi_s, 0 \leqslant s \leqslant t\}\right)$, а $c > 0$ — заданная постоянная, интерпретируемая как плата за длительность наблюдений. Байесовская стратегия последовательного оценивания $\boldsymbol{\theta}$ заключается в том, что наблюдения останавливаются в момент $\widehat{\varrho}_{\mathrm{BAYES}}$, и $\widehat{\boldsymbol{\theta}}_{\mathrm{BAYES}}$ принимается оптимальной оценкой значения $\boldsymbol{\theta}$.

Во **втором разделе** описывается структура оценки максимума правдоподобия параметра тренда $\boldsymbol{\theta} \in \mathbb{R}^{n_\theta+1}$ для процесса (1). Для этого определяются вспомогательный случайный процесс $M^H = \left(M_t^H\right)_{0 \leqslant t \leqslant T}$, удовлетворяющий равенству

$$M_t^H \equiv \int_0^t k_H(t,s) d\xi_s,$$

где $k_H(t,s) = \kappa_H^{-1} s^{1/2-H} (t-s)^{1/2-H}$, $\kappa_H = 2H\Gamma\left(\frac{3}{2} - H\right)\Gamma\left(\frac{1}{2} + H\right)$, и вспомогательная функция $w_H(t)$ согласно равенству

$$w_H(t) = \lambda_H^{-1} t^{2-2H},$$

где $\lambda_H = 2H\Gamma(3-2H)\,\Gamma\left(\frac{1}{2}+H\right)\left(\Gamma\left(\frac{3}{2}-H\right)\right)^{-1}$. Дифференциал dw_t^H понимается следующим образом: $dw_t^H = \lambda_H^{-1}(2-2H)t^{1-2H}dt$.

Структура оценки максимального правдоподобия параметра тренда описывается теоремой 1.

Теорема 1. *Пусть коэффициент сноса $f(t)$ фрактального броуновского движения имеет вид (2)–(3). Тогда оценка $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}$ максимального правдоподобия параметра $\boldsymbol{\theta}$ имеет вид*

$$\widehat{\boldsymbol{\theta}}_{\mathrm{ML}} = \boldsymbol{R}_H^{-1}(t)\boldsymbol{\psi}_t^H, \tag{5}$$

где компоненты $(n+1)$-мерного случайного процесса $\boldsymbol{\psi}^H = \left(\boldsymbol{\psi}_t^H\right)_{0\leqslant t\leqslant T}$ и элементы неслучайной матрицы $\boldsymbol{R}_H(t)$ равны

$$(\boldsymbol{\psi}_t^H)_i = \int\limits_0^t \psi_i(s)dM_s^H \quad и \quad (\boldsymbol{R}_H(t))_{ij} = \int\limits_0^t \psi_i(s)\psi_j(s)dw_s^H, \tag{6}$$

$i,j = 0,\ldots,n_\theta$, соответственно, а функции $\psi_i(t), i = 0,\ldots,n$ задаются соотношениями

$$\psi_i(t) = \frac{d}{dw_t^H}\int\limits_0^t k_H(t,s)\frac{dg_i(s)}{ds}ds, \qquad i = 0,\ldots,n_\theta. \tag{7}$$

Третий раздел посвящен отысканию оптимальной Байесовской оценки параметра тренда рассматриваемого случайного процесса в предположении, что параметр тренда является векторнозначной случайной величиной, имеющей многомерное нормальное распределение. Доказывается следующая

Теорема 2. *Пусть $\boldsymbol{\theta}$ — нормальный случайный вектор с математическим ожиданием $\boldsymbol{m}$ и ковариационной матрицей $\boldsymbol{\Sigma}$. Тогда оптимальной в среднем квадратичном Байесовской оценкой значения $\boldsymbol{\theta}$ является апостериорное среднее*

$$\widehat{\boldsymbol{\theta}}_{\mathrm{BAYES}} = \mathrm{E}\left[\boldsymbol{\theta}\middle|\mathcal{F}_t^\xi\right] = \left(\boldsymbol{R}_H(t) + \boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\boldsymbol{\psi}_t^H + \boldsymbol{\Sigma}^{-1}\boldsymbol{m}\right). \tag{8}$$

Величина условной среднеквадратичной ошибки оценивания $\mathrm{E}\left[\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\mathrm{BAYES}}\|^2\middle|\mathcal{F}_t^\xi\right]$ определяется следом условной ковариационной матрицы

$$\mathrm{cov}\left[\boldsymbol{\theta}\middle|\mathcal{F}_t^\xi\right] = \left(\boldsymbol{R}_H(t) + \boldsymbol{\Sigma}^{-1}\right)^{-1} \tag{9}$$

С применением теоремы 2 демонстрируется, что в случае полиномиального сноса функция штрафа

$$F_H(t) = ct + \mathrm{E}\left[\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\mathrm{BAYES}}\|^2\middle|\mathcal{F}_t^\xi\right]$$

имеет единственный минимум при $t \in [0,T]$. Этот результат иллюстрируется рисунком 2, содержащим график функции $F_H(t)$ для случая кубического сноса и значений параметров $H = 0.8, c = 0.02$.

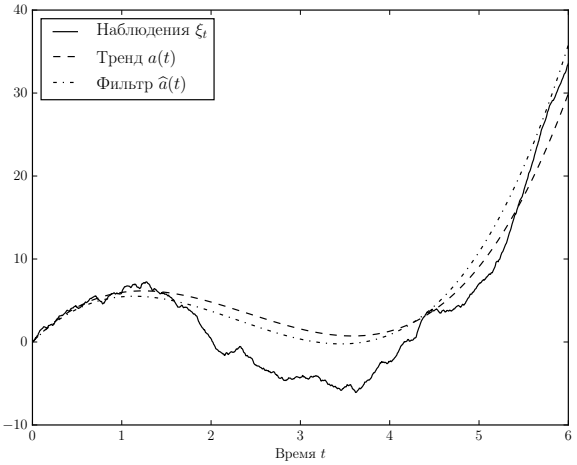

Рисунок 1: Траектории результата наблюдения ξ_t, тренда $f(t) = \sum_{k=0}^{3} \theta_i t^i$ и фильтра $\widehat{f}(t) = \sum_{k=0}^{3} (\widehat{\boldsymbol{\theta}}_{\mathrm{BAYES}})_i t^i$, $0 \leqslant t \leqslant T$, в модельной задаче выделения кубического тренда при значении параметра $H = 0.8$.

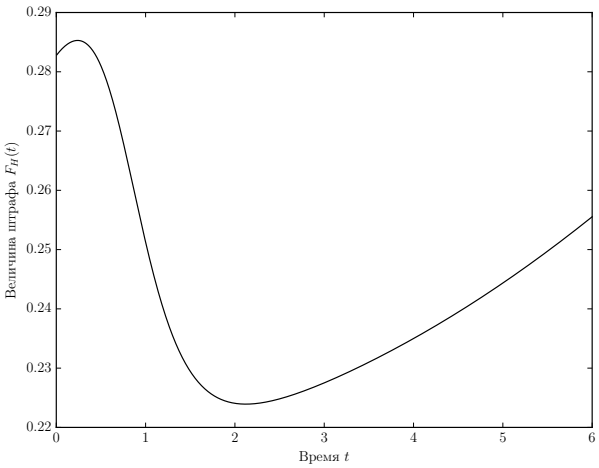

Рисунок 2: Значения функции штрафа $F_H(t)$ в модельной задаче выделения полиномиального (кубического) тренда при значениях параметров $H = 0.8, c = 0.02$.

В **четвертом разделе** подсчитывается и исследуется оптимальная Байесовская оценка параметра линейного тренда процесса (1) в предположении равномерного априорного распределения параметра тренда.

Поскольку аналитический расчет для общего случая $\boldsymbol{\theta} \in \mathbb{R}^{n_\theta+1}$ труден, оценка подсчитывается для важного частного случая линейного сноса, в котором наблюдаемый процесс ξ определяется стохастическим дифференциальным уравнением

$$d\xi_t = \theta_1 dt + \sigma dB_t^H, \tag{10}$$

где $\theta_1 \sim U(a, b)$. Результат подсчета в этой задаче составляет

Теорема 3. *Пусть в* (10) θ_1 *— равномерно распределенная на* $[a,b]$ *случайная величина, не зависящая от* B_t^H. *Тогда оптимальная в среднеквадратичном Байесовская оценка параметра* θ_1 *имеет вид*

$$(\widehat{\theta}_1)_{\mathrm{BAYES}} = m_t^H + \big[Z_t^H w_H(t)\big]^{-1}\big[\Lambda_t^H(a) - \Lambda_t^H(b)\big], \tag{11}$$

а условная среднеквадратичная погрешность оценивания равна

$$\begin{aligned} \gamma_t^H &= \mathrm{E}\big[\|\theta_1 - (\widehat{\theta}_1)_{\mathrm{BAYES}}\|^2 \big| \mathcal{F}_t^\xi\big] = \big[w_H(t)\big]^{-1} + \\ &+ \big[Z_H(t) w_H(t)\big]^{-1}\big[\Lambda_t^H(a)(a - m_t^H) - \Lambda_t^H(b)(b - m_t^H)\big] - \\ &- \big[Z_H(t) w_H(t)\big]^{-2}\big[\Lambda_t^H(a) - \Lambda_t^H(b)\big]^2, \end{aligned} \tag{12}$$

где

$$Z_t^H = \sqrt{\frac{2\pi}{w_H(t)}}\exp\left\{\frac{1}{2}\left(m_t^H\right)^2 w_H(t)\right\}C_t^H, \tag{13}$$
$$C_t^H = \Phi\left(\left(b-m_t^H\right)\sqrt{w_H(t)}\right) - \Phi\left(\left(a-m_t^H\right)\sqrt{w_H(t)}\right).$$

Во **второй главе** разрабатываются и исследуются алгоритмы обнаружения разладки временного ряда на основе ансамблей «слабых» детекторов в условиях нарушения стандартных предположений о модели разладки.

В **первом разделе** приводится обзор известных из литературы постановок задачи о разладке случайной последовательности и соответствующих процедур обнаружения разладки, а также рассматривается модель разладки с конечной длительностью. Пусть наблюдаемый случайный процесс $\xi = \left(\xi_t\right)_{t\geqslant 0}$ имеет структуру

$$\xi_t = \begin{cases} \xi_t^{\infty}, & \text{если } t \in \mathcal{T}_{\infty}, \\ \xi_t^{0}, & \text{если } t \in \mathcal{T}_{0}, \end{cases}$$

где случайные процессы $\xi^{\infty} = \left(\xi_t^{\infty}\right)_{t\geqslant 0}$ и $\xi^{0} = \left(\xi_t^{0}\right)_{t\geqslant 0}$ имеют (одномерные) плотности распределения $p_{\infty}(\cdot)$ и $p_0(\cdot)$ соответственно, а множества $\mathcal{T}_{\infty} \subseteq [0,\infty)$ и $\mathcal{T}_0 \subseteq [0,\infty)$ соответствуют промежуткам времени, в течение которых процесс X находится в состояниях без разладки *(нормальном)* и с разладкой *(аномальном)*, соответственно. Когда $\mathcal{T}_{\infty} = [0,\theta)$ и $\mathcal{T}_0 = [\theta,\infty)$, появление разладки соответствует установлению в момент θ нового режима наблюдений, причем последний имеет бесконечную продолжительность. В диссертационной работе рассматривается ситуация кратковременного изменения, в которой $\mathcal{T}_{\infty} = [0,\theta)\cup[\theta+\Delta,\infty)$ и $\mathcal{T}_0 = [\theta,\theta+\Delta)$, предполагающая длительность Δ аномального состояния конечной: $\Delta < \infty$. Пока наблюдения за процессом ξ согласуются с нормальным состоянием, требуется продолжать наблюдения. Если состояние изменяется, требуется обнаружить изменение как можно скорее, избегая ложных тревог.

Во **втором разделе** вводится понятие *ансамбля* процедур обнаружения разладки и приводятся примеры конкретных реализаций ансамблей. Пусть $\Pi_1,\ldots,\Pi_{n_\Pi}$ — множество процедур обнаружения разладки, причем процедура Π_k предписывает подавать сигнал тревоги в момент τ_k первого выхода некоторой статистики $s^k = \left(s_t^k\right)_{t\geqslant 0}$ на заданный уровень $h_k > 0$, $k = 1,\ldots,n_\Pi$: $\tau_k = \inf\{t \geqslant 0 : s_t^k \geqslant h_k\}$. Пусть $\mathbf{S}_t^k = \left\{s_u^k, 0 \leqslant u \leqslant t\right\}$ — история сигнала s_t^k до момента времени t.

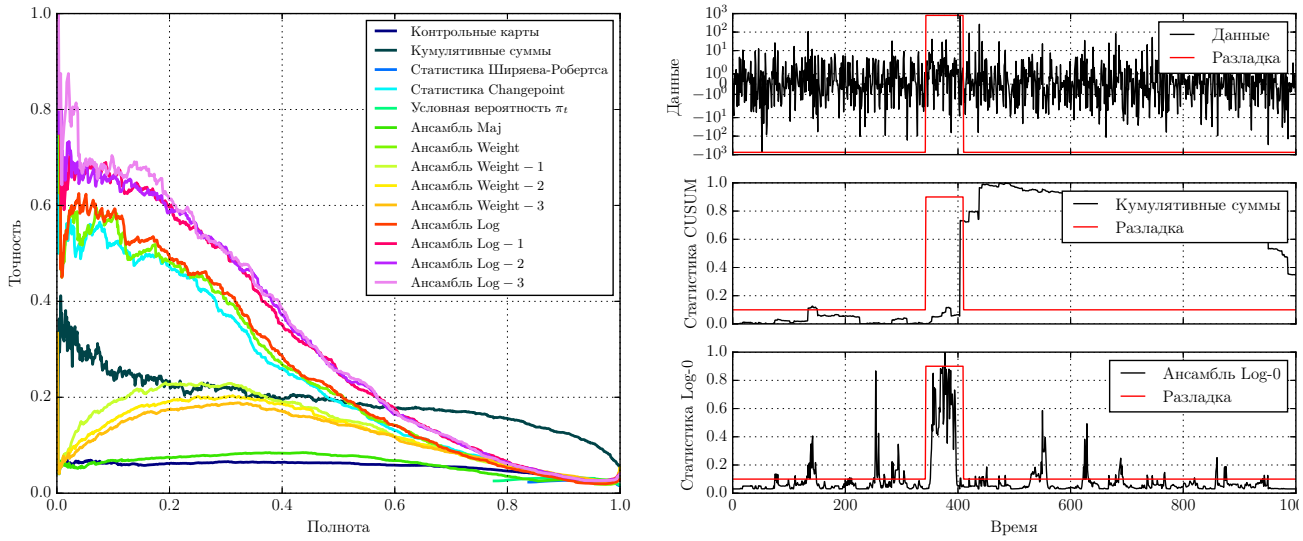

Рисунок 3: Слева: Кривые «точность–полнота» для пяти «слабых» детекторов и девяти рассматриваемых ансамблей для эксперимента по обнаружению разладки процесса Коши. Справа: траектория наблюдений (верхний рисунок), статистики кумулятивных сумм (средний рисунок) и ансамбля Log-0^{19}для эксперимента по обнаружению разладки процесса Коши.

Определение 1. *Ансамблем процедур обнаружения разладки назовем процедуру* A *обнаружения разладки, предписывающую подавать сигнал тревоги в момент* τ_{A} *первого выхода некоторого процесса* $a = \big(a_t\big)_{t\geqslant 0}$ *на заданный уровень* h_{A}*:* $\tau_{\mathrm{A}} = \inf\{t \geqslant 0 : a_t \geqslant h_{\mathrm{A}}\}$*. Процесс* a_t *строится как функция траекторий* $\mathbf{S}_t^1, \ldots, \mathbf{S}_t^{n_{\Pi}}$ *сигналов* $s^1, \ldots, s^{n_{\Pi}}$*:*

$$a_t = \psi(\mathbf{S}_t^1, \ldots, \mathbf{S}_t^{n_{\Pi}}, \boldsymbol{\theta}), \tag{14}$$

где $\boldsymbol{\theta} \in \mathbb{R}^d (d \geqslant n_{\Pi})$ *— параметры ансамбля, а функция* $\psi(\cdot)$ *— некоторая заданная агрегирующая функция.*

С помощью конкретного выбора агрегирующей функции вводятся ансамбль на основе голосования большинством, ансамбль на основе взвешенного голосования, а также ансамбль на основе логистической регрессии.

Третий раздел содержит определение нового критерия эффективности процедур обнаружения разладки и описание вычислительного алгоритма обучения ансамбля по множеству размеченных траекторий $\mathbf{X}^{\ell} = \{(X^i, Y^i)\}_{i=1}^{\ell}$, в котором каждая точка (X^i, Y^i) состоит из пары «наблюдение–разметка», причем процесс $X^i = \big(X_t^i\big)_{t\geqslant 0}$ соответствует наблюдениям, а процесс $Y^i = \big(Y_t^i\big)_{t\geqslant 0}$ — индикатору аномального состояния: $Y_t^i = \mathbb{1}_{\mathcal{T}_0^i}(t), t \geqslant 0$.

Пусть процедура обнаружения разладки Π предписывает подавать сигнал тревоги при выполнении условия выхода некоторой статистики $s = \big(s_t\big)_{t\geqslant 0}$ на заданный уровень $h > 0$, т. е. при тех $t \geqslant 0$, для которых $s_t \geqslant h$.

Определение 2. *Математическое ожидание потерь, свойственных процедуре обнаружения разладки* Π*,*

$$\mathbf{F}(\Pi) = c_\infty \mathrm{E}_\infty \left[\frac{\sum \mathbb{1}_{\{s_t \geqslant h\}}(t) \mathbb{1}_{\mathcal{T}_\infty}(t)}{\sum \mathbb{1}_{\mathcal{T}_\infty}(t)} \right] + c_0 \mathrm{E}_0 \left[\frac{\sum \mathbb{1}_{\{s_t < h\}}(t) \mathbb{1}_{\mathcal{T}_0}(t)}{\sum \mathbb{1}_{\mathcal{T}_0}(t)} \right], \qquad (15)$$

где c_0 *и* c_∞ *суть потери за единицу времени, сопутствующие ошибочным решениям о наличии и отсутствии разладки, соответственно.* Согласно этому определению, процедура обнаружения разладки Π тем лучше, чем меньше сопутствующие ей ожидаемые потери.

Оптимизация штрафа, заданного в (15), позволяет осуществить выбор параметров $\boldsymbol{\theta} \in \mathbb{R}^d$ ансамбля A в (14) и получить ансамбль A*, для которого $\mathbf{F}(\mathrm{A}^*) = \inf\limits_{\boldsymbol{\theta} \in \mathbb{R}^d} \mathbf{F}(\mathrm{A})$. Поскольку прямое вычисление математических ожиданий в (15) в общем случае невозможно, в диссертационной работе рассматривается аппроксимация $\mathbf{F}_{\mathrm{EMP}}(\Pi)$ функции потерь $\mathbf{F}(\Pi)$, называемая эмпирическим риском:

$$\begin{aligned} \mathbf{F}_{\mathrm{EMP}}(\Pi) &= c_\infty \frac{1}{\ell} \sum_{i=1}^{\ell} \left[\frac{\sum \mathbb{1}_{\{s_t^i \geqslant h\}}(t) \mathbb{1}_{\mathcal{T}_\infty^i}(t)}{\sum \mathbb{1}_{\mathcal{T}_\infty^i}(t)} \right] + c_0 \frac{1}{\ell} \sum_{i=1}^{\ell} \left[\frac{\sum \mathbb{1}_{\{s_t^i < h\}}(t) \mathbb{1}_{\mathcal{T}_0^i}(t)}{\sum \mathbb{1}_{\mathcal{T}_0^i}(t)} \right] \\ &= \frac{1}{\ell} \sum_{i=1}^{\ell} \left\{ \frac{c_\infty}{T_\infty^i} \sum_{t \in \mathcal{T}_\infty^i} \mathbb{1}_{\{s_t^i \geqslant h\}}(t) + \frac{c_0}{T_0^i} \sum_{t \in \mathcal{T}_0^i} \mathbb{1}_{\{s_t^i < h\}}(t) \right\}, \qquad (16) \end{aligned}$$

где $s^i = \left(s_t^i\right)_{t \geqslant 0}$ — траектория процесса s, подсчитанная по наблюдениям X^i, $\mathcal{T}_\infty^i$ и $\mathcal{T}_0^i$ суть промежутки времени нормального и аномального состояний в точке (X^i, Y^i), T_∞^i и T_0^i — длительности этих промежутков, соответственно. Согласно классическому подходу статистической теории обучения[20], минимизация эмпирического риска $\mathbf{F}_{\mathrm{EMP}}(\Pi)$ дает процедуру обнаружения разладки Π^*_{EMP}, для которой ожидаемые потери $\mathbf{F}(\Pi^*)$ близки к своему минимуму. Ввиду разрывности градиента прямая оптимизация эмпирического риска $\mathbf{F}_{\mathrm{EMP}}(\Pi)$ трудна (если вообще возможна); по этой причине в диссертационной работе предлагается рассматривать сглаженную версию эмирического риска $\mathbf{F}_{\mathrm{EMP}}(\Pi)$, задаваемую соотношением

$$\mathbf{F}_{\mathrm{DIFF}}(\Pi) = \frac{1}{\ell} \sum_{i=1}^{\ell} \left\{ \frac{c_\infty}{T_\infty^i} \sum_{t \in \mathcal{T}_\infty^i} \sigma(s_t^i - h) + \frac{c_0}{T_0^i} \sum_{t \in \mathcal{T}_0^i} \sigma(h - s_t^i) \right\}, \qquad (17)$$

---

[19] Ансамбль Log-0 задается агрегирующей функцией $\psi_{\mathrm{Log}-0}(\boldsymbol{\theta}; \mathbf{S}_t^1, \ldots, \mathbf{S}_t^{n_\Pi}) = \sum\limits_{k=1}^{n_\Pi} \theta_k s_t^k - \theta_0$ и соответствует классификатору на основе логистической регрессии.

[20] Vapnik V. Principles of risk minimization for learning theory // In Moody, J. E., Hanson, S. J. & Lippmann, R. P. (eds.), Advances In Neural Information Processing Systems 4, pp. 831–838. Morgan Kaufman, San Mateo, CA.

где $\sigma(x) = 1/(1+e^{-x})$ — логистическая функция. Так определенная функция риска является дифференцируемой по параметрам ансамбля $\boldsymbol{\theta} \in \mathbb{R}^d$, и может быть оптимизирована градиентными методами.

Четвертый раздел содержит результаты эмпирического анализа эффективности ансамблей «слабых» детекторов для различных агрегирующих функций, полученные в вычислительном эксперименте при моделировании данных различной природы. Рис. 3 представляют результат вычислительного эксперимента по обнаружению разладки сигнала с тяжелыми хвостами.

Третья глава посвящена разработке и исследованию математических моделей и алгоритмов обнаружения разладок квазипериодических временных рядов.

Первый раздел содержит постановку задачи оценивания параметров квазипериодического тренда по данным зашумленных измерений. Предполагается, что наблюдения $\xi = \big(\xi_t\big)_{t\geqslant 0}$ выполнены согласно общей модели

$$\xi_t = f(t) + \nu_t, \qquad t \geqslant 0, \tag{18}$$

где $f(t)$ — гладкая функция (тренд), наблюдаемая в шуме ν_t, $\mathrm{E}\,\nu_t = 0$. По данным зашумленных измерений $\mathbf{X}^\ell = \{(X_k, t_k)\}_{k=1}^{\ell}$, $X_k = \xi_{t_k}$, выполненным согласно (18), требуется оценить значение $f(t) = \mathrm{E}\,\xi_t$ для каждого $t \geqslant 0$.

Во **втором разделе** рассматривается алгоритм оценивания параметров тренда на основе фильтра, разработанного в первой главе.

1. Выбирается окно наблюдений $W(a,b) = \{(X_k, t_k) : a \leqslant t_k \leqslant b\}$ в окрестности некоторого $t_0 \in [a,b]$.
2. В окрестности $t \in [a,b]$ значения t_0 рассматривается аппроксимация гладкого тренда кубическим полиномом, а в качестве модели шума рассматривается процесс фрактального гауссовского шума $Z^H = \big(Z_t^H\big)_{t\geqslant 0}$, так что модель данных принимает вид

$$X_k = \sum_{i=0}^{3} \theta_i (t_k - t_0)^i + \sigma Z_k^H, \qquad (X_k, t_k) \in W(a,b), \tag{19}$$

   где значение дисперсии $\sigma > 0$ принимается известным, а $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_3)$ является неизвестным параметром. Для оценивания значения $\boldsymbol{\theta}$ используется метод максимального правдоподобия и оценка $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}$ из первой главы.
3. Оценка $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}$ используется для вычисления оценки траектории с помощью соотношения $\widehat{f}_{[a,b]}(t) = \sum_{i=0}^{3} (\widehat{\boldsymbol{\theta}}_{\mathrm{ML}})_i (t-t_0)^i$ для всех $t \in [a,b]$. Итоговая оценка $\widehat{f}(t)$ для всех $t \geqslant 0$ вычисляется усреднением согласно соотношению $\widehat{f}(t) = (n_{[a,b]}(t))^{-1} \sum\limits_{(a,b):a\leqslant t\leqslant b} \widehat{f}_{[a,b]}(t)$ по $n_{[a,b]}(t) = \sum\limits_{(a,b)} \mathbb{1}_{[a,b]}(t)$ локальных оценок.

В **третьем разделе** рассматривается многокомпонентная математическая модель квазипериодического сигнала и алгоритм оценивания его параметров на основе метода непараметрической регрессии. Принимается, что равенство

$$\xi_t = f(t) + \nu_t, \qquad t \geqslant 0, \tag{20}$$

задает модель сигнала ξ, в которой моделью тренда служит соотношение $f(t) = Q_t S_t$, где $Q = \big(Q_t\big)_{t\geqslant 0}$ — случайный процесс, а $S_t = S(\varphi(t))$ — неслучайная функция; модель случайной помехи ν_t задается равенством $\nu_t = \sigma_t \varepsilon_t$, где $\sigma_t = \sigma_{\varphi(t)}$ — неслучайная функция, а $\varepsilon = (\varepsilon_t)_{t\geqslant 0}$ — процесс стандартного гауссовского белого шума. Величина $\varphi(t) = 2\pi\{t/T\}$ имеет смысл фазы (известного) периода T, соответствующего моменту времени t (где $\{x\} = x - \lfloor x \rfloor$ — дробная часть x).

Факторизация представлений тренда и помехи в (20) позволяет выразить важные свойства реальных сигналов систем с интенсивным ПО, такие как медленный рост числа обрабатываемых запросов и его флуктуации в течение суток. В модели тренда $Q = \big(Q_t\big)_{t\geqslant 0}$ и $S = \big(S_t\big)_{t\geqslant 0}$ интерпретируются как ненаблюдаемые амплитуда и сезонная составляющая, соответственно.

Рассматривается алгоритм оценивания значений $\widehat{S}_{\psi_j}$ и $\widehat{\sigma}^2_{\psi_j}$ для каждой фазы $\psi_j = \varphi(t_j)$, где $t_j = j\Delta, \Delta = T/p$, $j = 1, \ldots, p$, по наблюдениям $\mathbf{X}^\ell = \{(X_k, t_k)\}_{k=1}^{\ell}$.

Инициализация. Величина $\widehat{Q}_k = \widehat{Q}(t_k)$ полагается равной 1 для каждого $k = 1, \ldots, \ell$, а величина $\widehat{\sigma}^2_{\psi_j}$ полагается равной дисперсии наблюдений $X_1, \ldots, X_\ell$ для каждого $j = 1, \ldots, p$.

Итерации. Повторяются следующие шаги:

1. С использованием оценки Надарая-Ватсона переоценивается $\widehat{S}_{\psi_j}$:

$$\widehat{S}_{\psi_j} = \frac{\sum\limits_{k=1}^{\ell} w_k X_k / \widehat{Q}_k K_h(\varphi_k, \psi_j)}{\sum\limits_{k=1}^{\ell} w_k K_h(\varphi_k, \psi_j)} \tag{21}$$

где $\varphi_k = \varphi(t_k)$ суть фазы в моменты времени $t_k, k = 1, \ldots, \ell$, $K_h(\varphi,\psi)$ — ядро ширины $h > 0$, и w_k — вес k-го измерения.

2. С использованием оценки Надарая-Ватсона переоценивается $\widehat{\sigma}^2_{\psi_j}$:

$$\widehat{\sigma}^2_{\psi_j} = \frac{\sum\limits_{k=1}^{\ell} \big(X_k - \widehat{X}_k\big)^2 K_h(\varphi_k,\psi_j)}{\sum\limits_{k=1}^{\ell} K_h(\varphi_k,\psi_j)}. \tag{22}$$

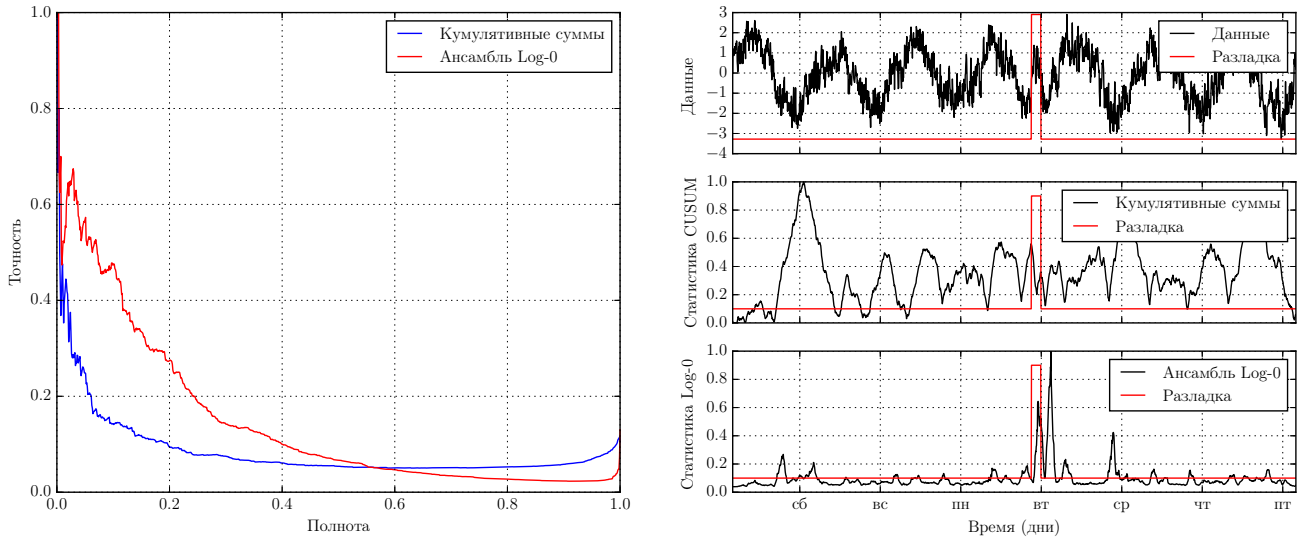

Рисунок 4: Слева: кривая «точность–полнота» для искусственных данных. Справа сверху: пример недельного профиля нагрузки в искусственных данных и индикатора аномального состояния. Справа в центре: траектория статистики кумулятивных сумм и индикатор аномального состояния. Справа внизу: траектория статистики ансамбля Log−0 и индикатор аномального состояния.

3. Переоценивается $\widehat{X}_k = \widehat{X}(t_k)$ и $\widehat{Q}_k = \widehat{Q}(t_k)$, $k = 1, \ldots, \ell$. Для вычисления прогноза $\widehat{X}_k$ значения X_k выбирается некоторое $H > 0$ и рассматриваются моменты времени $t_{k-p}, \ldots, t_k$, где $t_k - H \leqslant t_{k-p} < \ldots < t_k$. Согласно (20) в предположении локально постоянной амплитуды

$$X_i = Q_k \widehat{S}_{\varphi(t_i)} + \nu_i, \qquad i = k-p, \ldots, k, \tag{23}$$

где значение $\widehat{S}_{\varphi(t_i)}$ получено кубической интерполяцией значений $\widehat{S}_{\psi_j}$ по четырем ближайшим к $\varphi(t_i)$ точкам сетки $\psi_1, \ldots, \psi_n$. В предположении $\nu_i \sim \mathcal{N}(0, \widehat{\sigma}^2_{\psi_i})$ амплитуда Q_k в (23) оценивается методом взвешенной линейной регрессии с весами $\lambda_i = 1/\widehat{\sigma}^2_{\psi_i}, i = k-p, \ldots, k$. Прогноз $\widehat{X}$ значения X_k вычисляется согласно $\widehat{X}_k = \widehat{Q}_k \widehat{S}_{\varphi(t_k)}$.

В **четвертом разделе** содержится постановка задачи обнаружения разладки квазипериодического сигнала и метод ее решения, основанный на применении ансамблей «слабых» детекторов, предложенных во второй главе. Рассматривается модель краткосрочной разладки в (18):

$$\nu_t = \mu \mathbb{1}_{[\theta, \theta+\Delta t]}(t) + Z_t, \qquad t \geqslant 0, \tag{24}$$

где θ — неизвестный момент разладки, μ — неизвестная величина разладки, Δt — неизвестная длительность разладки, и $Z = \big(Z_t\big)_{t \geqslant 0}$ — гауссовский процесс белого шума. Задается процесс

$$R_t = \sigma^{-1}(X_t - \widehat{X}_t), \qquad t \geqslant 0, \tag{25}$$

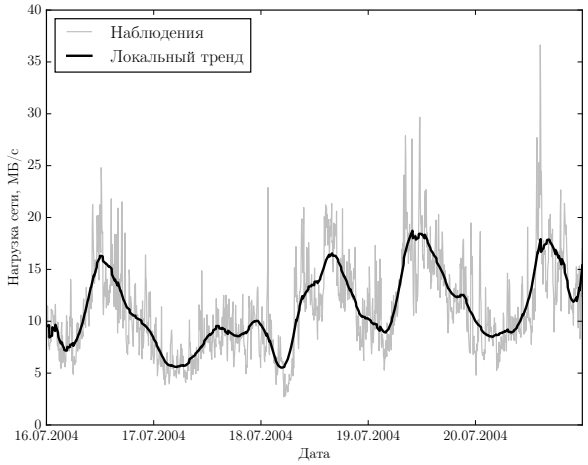

Рисунок 5: Результат аппроксимации локального тренда нагрузки соединения Хьюстон–Чикаго (по данным измерений нагрузки сети Абилин).

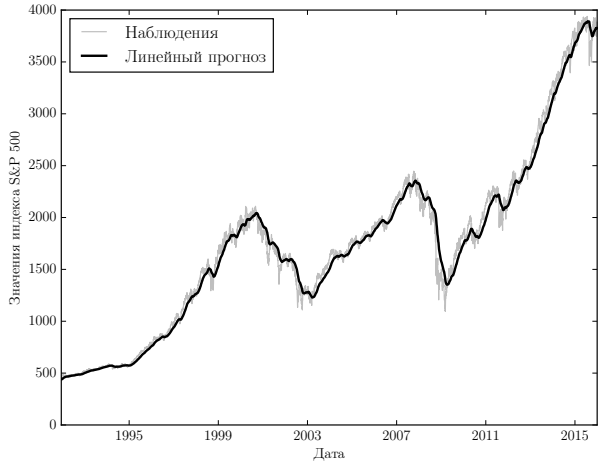

Рисунок 6: Результат решения задачи прогнозирования значения индекса S&P 500 на один день в предположении линейности тренда.

где X — наблюдаемый в (18) процесс с (известной или оцениваемой) дисперсией σ, а значение $\widehat{X}_t$ процесса $\widehat{X}$ является оценкой значения X_t процесса X, полученной одним из алгоритмов, описанных выше. Показывается, что при нормальном режиме наблюдений $\mathrm{E}\, R_t \approx 0$, а при аномальном $\mathrm{E}\, R_t \approx \mu$, и для обнаружения разладки процесса R используется ансамбль «слабых» детекторов, параметры которого подбираются по множеству размеченных траекторий процесса X.

В **пятом разделе** представлены описание и результаты вычислительных экспериментов с использованием искусственных данных для оценки качества обнаружения разладки сигнала с квазипериодическим трендом. Рис. 4 представляет результат вычислительного эксперимента по обнаружению разладки квазипериодического сигнала, наблюдаемого в шуме с длинной памятью.

Четвертая глава содержит описание структуры и функционала разработанного комплекса программ. В качестве платформы для реализации разработанных математических методов и алгоритмов используется язык программирования `python`, библиотека математических функций `numpy` и библиотека научных расчетов `scipy`. Комплекс программ включает 5 основных пакетов:

1. Пакет, реализующий алгоритмы оптимального оценивания параметров тренда сигнала, наблюдаемого во шуме с длинной памятью.
2. Пакет, реализующий вычислительный алгоритм оценивания компонент квазипериодической модели (20).
3. Пакет для работы с ансамблями «слабых» детекторов, в том числе:
   - модуль численной оптимизации сглаженного эмпирического риска (17) по заданной обучающей выборке;

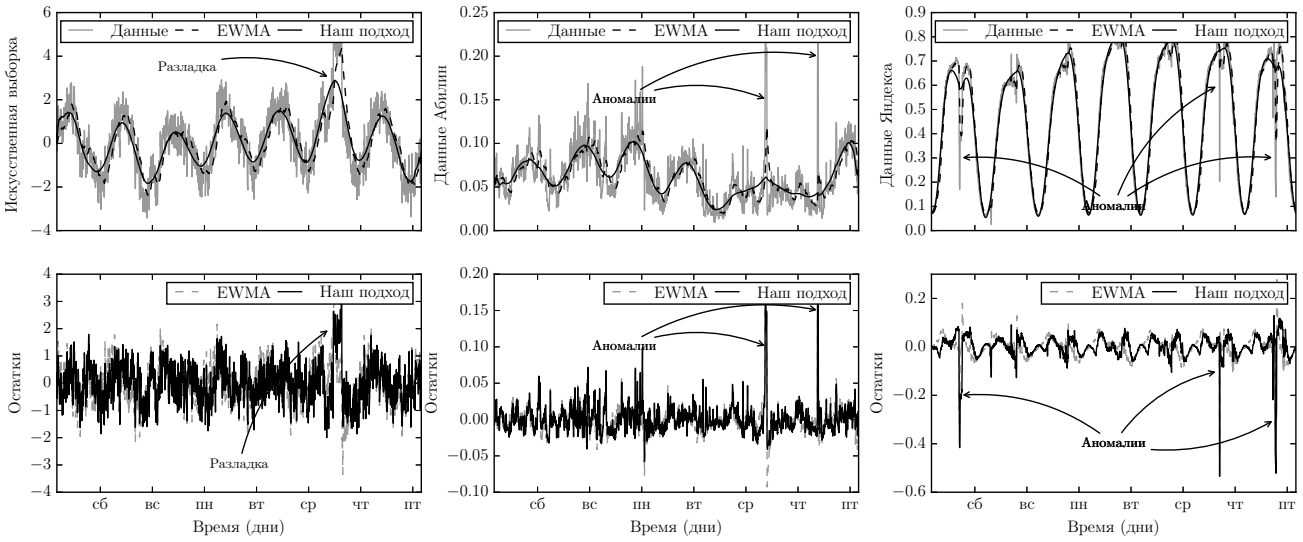

Рисунок 7: Верхний ряд: пример искусственных данных (слева), данных сети Абилин (в центре) и данных «Яндекса» (справа), а также результат выделения тренда с использованием экспоненциального сглаживания и подхода, предложенного в работе. Нижний ряд: остатки, полученные вычитанием тренда из наблюдаемого сигнала. Стрелки указывают на обнаруженные разладки.

– модуль обнаружения кратковременной разладки случайного сигнала на основе ансамбля (14) (в том числе в режиме реального времени).

4. Пакет моделирования реализаций случайных сигналов с заданными статистическими параметрами, такими как кратковременные разладки, квазипериодические тренды и длинная память.
5. Пакет оценивания эффективности исследуемых алгоритмов и визуализации данных.

В **пятой главе** излагаются результаты применения разработанных математических методов в задачах анализа реальных сигналов. В **первом разделе** описываются результаты решения задачи оценки профиля нагрузки реальной компьютерной сети Абилин[21] по данным измерений объема трафика, переданного между узлами сети. Для решения этой задачи был использован алгоритм 1 из второго раздела третьей главы. Во **втором разделе** рассматривается задача прогнозирования значений экономических и финансовых показателей на момент закрытия («цена закрытия») по историческим данным на один день вперед. В заключительном **третьем разделе** описываются результаты решения задачи обнаружения разладок реальной информационной системы в режиме реального времени. В последней задаче для обнаружения разладки неоходимо было строить оценку тренда квазипериодического сигнала по измерениям в шуме с длинной

[21] Данные измерений нагрузки сети Абилин за 2004 г. находятся в публичном доступе по ссылке `http://www.cs.utexas.edu/~yzhang/research/AbileneTM` (проверена 12.10.2016).

или короткой памятью[22]. В разделе описываются два подхода к решению этой задачи, основанные на предложенной в работе методологии.

В **заключении** сформулированы основные результаты работы:

1. Разработаны новые математические методы оценки параметров сигнала по данным измерений, выполненным во фрактальном шуме, в том числе:
   - получена оценка максимального правдоподобия параметра сигнала;
   - получены оптимальные Байесовские оценки для случаев нормального и равномерного априорных распределений параметра сигнала;
   - охарактеризован оптимальный момент остановки измерений сигнала для случая нормального априорного распределения параметра сигнала.
2. Разработан и исследован алгоритм обнаружения разладки временного ряда на основе ансамбля процедур обнаружения разладки, предложен метод настройки параметров ансамбля.
3. Предложена и исследована методология моделирования квазипериодических сигналов и обнаружения их разладок, в том числе:
   - предложена математическая модель квазипериодического временного ряда на основе разложения по заданной системе функций и вычислительный алгоритм оценки ее параметров на основе оптимального фильтра п. 1;
   - предложена многокомпонентная математическая модель квазипериодического временного ряда и вычислительный алгоритм оценки ее параметров на основе непараметрической регрессии;
   - предложена математическая модель краткосрочной разладки квазипериодического временного ряда и процедура обнаружения этой разладки на основе ансамблей «слабых» детекторов.
4. Создан комплекс программ, реализующий предложенные в диссертационной работе вычислительные алгоритмы фильтрации тренда фрактального случайного сигнала, оценивания квазипериодического сигнала, настройки параметров ансамбля и обнаружения разладки временного ряда на основе ансамбля.
5. С помощью разработанного программного комплекса решен ряд прикладных задач обнаружения отказов промышленных программных систем.

[22] Во всех прикладных задачах для оценивания значения показателя Херста использовался подход, предложенный в работе Dubovikov M. M., Starchenko N. V., Dubovikov M. S. Dimension of the minimal cover and fractal analysis of time series //Physica A: Statistical Mechanics and its Applications. — 2004. — Vol. 339. — no. 3. — Pp. 591–608.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Artemov A. V. Effective signal extraction via local polynomial approximation under long-range dependency conditions // Accepted for publication in Lobachevskii Journal of Mathematics. — 2016. — Vol. 37. — Issue 1.
2. Артёмов А. В., Бурнаев Е. В. Оптимальное оценивание сигнала, наблюдаемого во фрактальном гауссовском шуме // Теория вероятностей и ее применения. — 2015. — Т. 60. — №. 1. — С. 163-171.
3. Artemov A. V., Burnaev E. V. Ensembles of Detectors for Online Detection of Transient Changes // Proceedings of The 8th International Conference on Machine Vision (ICMV 2015). — 2015. — P. 98751Z–98751Z-5.
4. Артёмов А. В., Бурнаев Е. В. Исследование процедуры обнаружения разладки временного ряда на основе взвешенного голосования // Материалы XXII международной научной конференции студентов, аспирантов и молодых учёных «Ломоносов». Москва, 2015. С. 34–35.
5. Artemov A. V., Burnaev E. V. Nonparametric Decomposition of Quasi-periodic Time Series for Change-point Detection // Proceedings of The 8th International Conference on Machine Vision (ICMV 2015). — 2015. — P. 987520–987520-5.
6. Артёмов А. В. Масштабируемая архитектура оценивания сигналов для мониторинга крупных квазипериодических систем // Труды 58-й научной конференции МФТИ «Современные проблемы фундаментальных и прикладных наук». — Москва–Долгопрудный: ФУПМ, 2015. — С. 11–14.
7. Бурнаев Е. В., Артемов А. В. О выделении тренда из шума с длинной памятью и обнаружении разладок на его фоне. Труды Международной конференции по стохастическим методам, 2016.