

Фильтрация сигналов с трендом в задачах обнаружения разладки

Артёмов Алексей Валерьевич

Диссертация на соискание ученой степени
кандидата физико-математических наук

05.13.18 — математическое моделирование,
численные методы и комплексы программ

Научный руководитель: к. ф.-м. н.,
доцент Сколковского института науки и технологий,
заведующий лабораторией ИППИ РАН Бурнаев Е. В.

ВВЕДЕНИЕ

Задачи обнаружения разладок
при обеспечении надежности систем
в программной инженерии

Моделирование сигналов в задачах обнаружения разладок

- Задачи обнаружения разладки: выявление моментов резкого изменения характеристик сигналов
- Применяются специализированные математические модели разладки и критерии качества ее обнаружения
- Современный контекст: обеспечение эффективной эксплуатации высокотехнологичных систем с интенсивным ПО (обнаружение разладок, сбоев, внедрений, атак и т. п.)
- Сигналы, регистрируемые в системах с интенсивным ПО, проявляют ряд характерных особенностей

Настоящая работа посвящена развитию эффективной методологии моделирования сигналов и обнаружения разладок систем с интенсивным ПО.

Типичная задача обнаружения разладки

- Структура наблюдаемого процесса X_t :

$$\underbrace{X_1, X_2, \dots, X_\theta}_{\sim f_\infty(\cdot)}, \underbrace{X_{\theta+1}, \dots}_{\sim f_0(\cdot)}$$

- θ : неизвестный момент появления разладки
- Пример:

$$X_t = \begin{cases} W_t, & 0 \leq t < \theta, \\ \mu + W_t, & \theta \leq t. \end{cases}$$

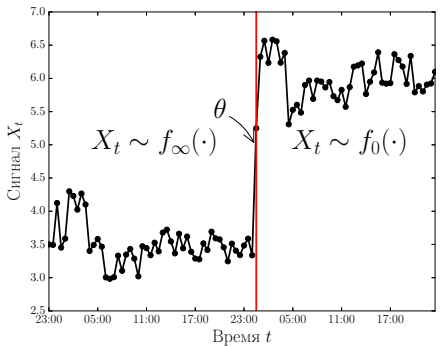


Рис.: Время получения первого байта при загрузке yandex.ru, мсек

- Тревога в момент $\tau = \inf\{t \geq 0 : s_t \geq h\}$
- Задача: построить такой процесс s_t , что задержка $\tau - \theta$ мала (при ограничениях на частоту ложных тревог)

Некоторые работы — теория обнаружения разладки

Основные процедуры обнаружения разладки, для которых были исследованы вопросы эффективности:

- Контрольные карты [Shewhart, 1931]
- Алгоритм кумулятивных сумм (CUSUM) [Page, 1954]
- Экспоненциально взвешенное скользящее среднее [Roberts, 1959]
- Фильтр Калмана [Kalman, 1960]
- Байесовские методы [Girshick & Rubin, 1952; Ширяев, 1961]
- Процедура Ширяева-Робертса [Ширяев, 1961; Roberts, 1966]
- Метод обобщенного отношения правдоподобия [Willsky, 1976]

Факторы, ограничивающие эффективность указанных процедур:

- Модель разладки известна заранее [Andersson et al., 2005; Du et al., 2015]
- Предположения о стационарности наблюдений [Basseville & Nikiforov, 1993]
- Предположения о бесконечной длительности разладки [Guépié et al., 2012]

АКТУАЛЬНОСТЬ

Характерные особенности сигналов реальных систем
с интенсивным ПО (software-intensive systems)

Большие системы с интенсивным ПО

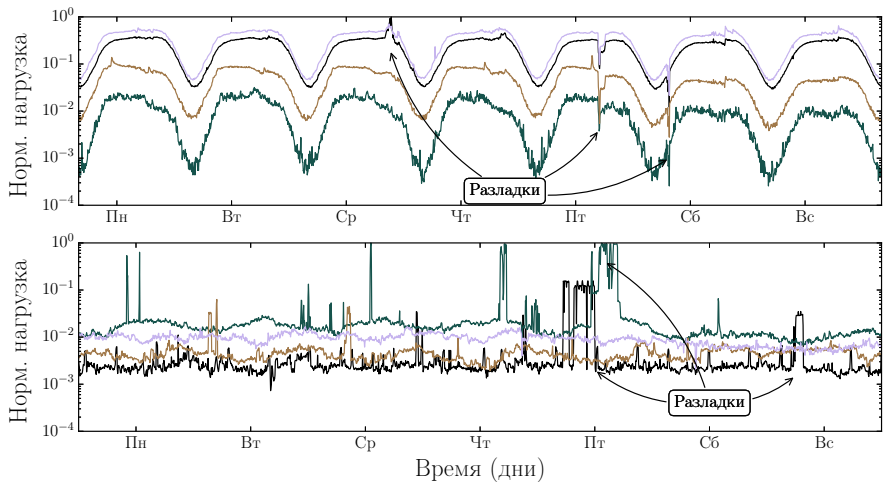


Рис.: Вверху: нагрузка сервисов Поиск, Карты, Почта, Портал компании Яндекс за неделю 10–17 февраля 2014 г. Внизу: нагрузка соединений Атланта–Лос Анджелес, Лос Анджелес–Чикаго, Сиэтл–Лос Анджелес, Чикаго–Лос Анджелес за неделю 14–21 июня 2004 г.

Некоторые работы — обнаружение отказов программных систем

Обнаружение разладок и аномалий в компьютерных сетях:

- Исследование метода случайных проекций на ортогональное дополнение к пространству наблюдений [Pham et al., 2014]
- Параметрическая модель сигнала и оптимальный алгоритм обнаружения аномалий объема трафика [Casas et al., 2010]
- Исследование PCA для разделения нормальных и аномальных сетевых условий [Lakhina et al., 2004]
- Исследование эффективности процедуры Ширяева-Робертса в постановке с перезапуском [Tartakovsky et al., 2005]

Ограничения указанных подходов:

- Нет явной модели циклов или длинной памяти
- Алгоритмическая вычислительная сложность $O(d)$, d — длина истории
- Нет теоретических результатов, касающихся эффективности обнаружения разладки

Большие системы с интенсивным ПО

1. Всплески нагрузки («длинная память»)
2. Стохастические циклы нагрузки (дневной, недельный, годовой)
3. Априори произвольные типы отказов
4. Огромные объемы данных (сотни тысяч или миллионы характеристик)

Игнорирование этих особенностей → **непрерывный поток ложных тревог!**

Цель работы: развитие эффективной методологии обнаружения разладок в указанных условиях

ЦЕЛЬ РАБОТЫ

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.

ЗАДАЧИ ИССЛЕДОВАНИЯ

1. Разработать метод оценки параметров сигнала по его измерениям во фрактальном шуме (шумовая компонента с **длинной памятью**).
2. Разработать методологию моделирования и оценивания сигналов с трендом (в т. ч. **квазипериодического сигнала**).
3. Разработать метод обнаружения разладки на основе ансамбля «слабых» детекторов для повышения эффективности обнаружения разладки в случае **нарушения стандартных предположений** о ее модели.
4. Реализовать разработанные математические методы и алгоритмы в составе **комплекса программ**.

ЗАДАЧА 1

Разработать метод оценки параметров сигнала по его измерениям во фрактальном шуме (шумовая компонента с **длинной памятью**)

Стохастическая модель длинной памяти

- Фрактальное броуновское движение $B^H = (B_t^H)_{0 \leq t \leq T}$ с параметром Хёрста $H \in (0, 1)$ — это гауссовский процесс с непрерывными траекториями такой, что

$$B_0^H = 0, \quad \mathbb{E} B_t^H = 0,$$

$$\mathbb{E} B_s^H B_t^H = \frac{1}{2}(t^{2H} + s^{2H} - |t - s|^{2H}).$$

- $H = 1/2 \implies$ обыкновенное броуновское движение.

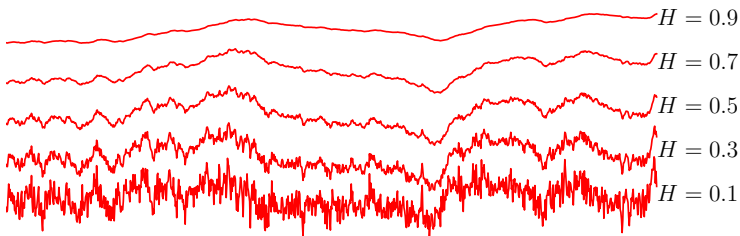


Рис.: Примеры реализаций fBM с различными значениями показателя H

Описание модели

- $(\Omega, \mathcal{F}, (\mathcal{F}_t^\xi)_{t \geq 0}, P)$: фильтрованное вероятностное пространство
- Наблюдаемый процесс $\xi = (\xi_t)_{0 \leq t \leq T}$ задан в виде

$$\xi_t = f(t) + \sigma(t)B_t^H,$$

где

- $B^H = (B_t^H)_{0 \leq t \leq T}$: fBM с $H \in (0, 1)$,
- коэффициент диффузии $\sigma(t)$ предполагается известным,
- коэффициент сноса $f(t)$ представляется в виде разложения

$$f(t) = \sum_{i=0}^n \theta_i g_i(t)$$

по заданному семейству функций $\{g_i(t)\}_{i=0}^{n_\theta}$,

- коэффициенты разложения $\theta_i, i = 0, \dots, n$ неизвестны

Обозначим $\boldsymbol{\theta} = (\theta_0, \dots, \theta_n)^\top$, $\boldsymbol{\varphi}(t) = (\varphi_0(t), \dots, \varphi_n(t))^\top$:

$$f(t) = \boldsymbol{\theta}^\top \mathbf{g}(t)$$

Постановка задачи

- Заданы наблюдения $\{\xi_s, 0 \leq s \leq t\}$ до момента t
- **Задача 1.** Предполагая $\theta \in \mathbb{R}^{n_\theta}$, найти ОМП
- **Задача 2.** Предполагая θ случайной величиной, распределение которой
 - нормальное $\mathcal{N}(\mathbf{m}, \Sigma)$, или
 - равномерное на $\mathbf{r} = [a_0, b_0] \times \cdots \times [a_{n_\theta}, b_{n_\theta}]$,
 найти такую стратегию оценивания $\delta^* = (\tau^*, \hat{\theta}^*)$, что

$$\inf_{\delta \in D} \mathbf{E} \left[c\tau + \|\theta - \hat{\theta}\|^2 \right] = \mathbf{E} \left[c\tau^* + \|\theta - \hat{\theta}^*\|^2 \right],$$

где

- $c \geq 0$: цена наблюдений,
- D : множество стратегий $\delta = (\tau, \hat{\theta})$ с конечными относительно \mathcal{F}_t^ξ моментами остановки

Обозначения

Обозначим

$$\begin{aligned}
 \kappa_H &= 2H\Gamma\left(\frac{3}{2} - H\right)\Gamma\left(\frac{1}{2} + H\right), \\
 k_H(t, s) &= \kappa_H^{-1} s^{1/2-H} (t-s)^{1/2-H}, \\
 \lambda_H &= \frac{2H\Gamma(3-2H)\Gamma\left(\frac{1}{2} + H\right)}{\Gamma\left(\frac{3}{2} - H\right)}, \\
 w_H(t) &= \lambda_H^{-1} t^{2-2H}, \\
 dw_t^H &= d(w_H(t)) = \lambda_H^{-1}(2-2H)t^{1-2H} dt
 \end{aligned}$$

Зададим процессы $M^H = (M_t^H)_{0 \leq t \leq T}$, $m^H = (m_t^H)_{0 \leq t \leq T}$ согласно

$$M_t^H \equiv \int_0^t k_H(t, s) d\xi_s, \quad m_t^H = M_t^H / w_H(t).$$

Теорема 1: оценка максимального правдоподобия

ОМП $\widehat{\theta}_{ML}$ для θ равна

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

где $\mathbf{R}_H(t)$ — неслучайная матрица с элементами

$$(\mathbf{R}_H(t))_{ij} = \int_0^t \psi_i(s) \psi_j(s) dw_s^H, \quad i, j = 0, \dots, n_\theta,$$

и $\boldsymbol{\psi}^H = (\boldsymbol{\psi}_t^H)_{0 \leq t \leq T}$ — \mathbb{R}^{n_θ} -мерный сл. процесс с координатами

$$(\boldsymbol{\psi}_t^H)_i = \int_0^t \psi_i(s) dM_s^H, \quad i = 0, \dots, n_\theta,$$

функции $\psi_i(t)$, $i = 0, \dots, n_\theta$, задаются равенствами

$$\psi_i(t) = \frac{d}{dw_t^H} \left(\int_0^t k_H(t, s) g_i'(s) ds \right), \quad i = 0, \dots, n_\theta.$$

Следствие: полиномиальный снос

Пусть $g_i(t) = t^i$, $i = 0, \dots, n_\theta$, $\sigma(t) \equiv \sigma$, тогда

- элементы ψ^H и $\mathbf{R}_H(t)$ задаются равенствами

$$(\psi_t^H)_i = \beta_H(i) \int_0^t s^{i-1} dM_s^H, \quad (\mathbf{R}_H(t))_{ij} = \alpha_H(i, j) t^{i+j-2H},$$

где

$$\alpha_H(i, j) = \lambda_H^{-1} \beta_H(i) \beta_H(j) \frac{2 - 2H}{i + j - 2H},$$

$$\beta_H(i) = i \frac{2 - 2H + i - 1}{2 - 2H} \frac{\Gamma(3 - 2H)}{\Gamma(3 - 2H + i - 1)} \frac{\Gamma(3/2 - H + i - 1)}{\Gamma(3/2 - H)},$$

- ОМП $\hat{\theta}_{ML}$: решение системы $\psi_t^H - \mathbf{R}_H(t)\theta = 0$.

Теорема 2: нормальное априорное распределение

Пусть θ — случайная величина $\mathcal{N}(\mathbf{m}, \Sigma)$. Тогда оптимальная Байесовская оценка $\hat{\theta}_{\text{BAYES}}$ равна

$$\hat{\theta}_{\text{BAYES}} = \mathbb{E}[\theta | \mathcal{F}_t^\xi] = (\mathbf{R}_H(t) + \Sigma^{-1})^{-1}(\psi_t^H + \Sigma^{-1}\mathbf{m}).$$

Условная погрешность оценивания $\mathbb{E}(\|\theta - \hat{\theta}_{\text{BAYES}}\|^2 | \mathcal{F}_t^\xi)$ определяется следом условной ковариационной матрицы

$$\text{cov}[\theta | \mathcal{F}_t^\xi] = (\mathbf{R}_H(t) + \Sigma^{-1})^{-1}.$$

Следствие: оптимальный момент остановки

- Оптимальный момент остановки

$$\begin{aligned}\hat{\tau}_{\text{BAYES}} &= \arg \inf_{\tau \in D} \mathbb{E} \left[c\tau + \mathbb{E} \left(\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\text{BAYES}}\|^2 \mid \mathcal{F}_{\tau}^{\xi} \right) \right] \\ &= \arg \inf_{t \in [0, T]} F_H(t),\end{aligned}$$

где функция

$$\begin{aligned}F_H(t) &= ct + \mathbb{E}(\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\text{BAYES}}\|^2 \mid \mathcal{F}_t^{\xi}) \\ &= ct + \text{tr} \left((\mathbf{R}_H(t) + \boldsymbol{\Sigma}^{-1})^{-1} \right), \quad t \in [0, T],\end{aligned}$$

- Пусть $g_i(t) = t^i$, $i = 0, \dots, n_{\theta}$, $\sigma(t) \equiv \sigma$, $\boldsymbol{\Sigma} = \text{diag}(\gamma_0^2, \dots, \gamma_{n_{\theta}}^2)$, тогда функция $F_H(t)$ имеет единственный минимум $t \in [0, T]$

Теорема 3: равномерное априорное распределение

Пусть $d\xi_t = \theta_1 dt + \sigma dB_t^H$, где $\theta_1 \sim U(a, b)$. Тогда оптимальная Байесовская оценка имеет вид

$$(\hat{\theta}_1)_{\text{BAYES}} = m_t^H + [Z_t^H w_H(t)]^{-1} [\Lambda_t^H(a) - \Lambda_t^H(b)],$$

условная с.к. погрешность оценивания равна

$$\begin{aligned} \gamma_t^H &= \text{E} \left((\theta_1 - (\hat{\theta}_1)_{\text{BAYES}})^2 \mid \mathcal{F}_t^\xi \right) = [w_H(t)]^{-1} + \\ &+ [Z_H(t)w_H(t)]^{-1} [\Lambda_t^H(a)(a - m_t^H) - \Lambda_t^H(b)(b - m_t^H)] \\ &- [Z_H(t)w_H(t)]^{-2} [\Lambda_t^H(a) - \Lambda_t^H(b)]^2, \end{aligned}$$

где

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

Оптимальный момент остановки

- Для подсчета оптимального момента остановки необходимо решить задачу на минимум

$$\begin{aligned}\hat{\tau}_{\text{BAYES}} &= \arg \inf_{\tau} E \left[c\tau + E \left((\theta_1 - (\hat{\theta}_1)_{\text{BAYES}})^2 \mid \mathcal{F}_{\tau}^{\xi} \right) \right] = \\ &= \arg \inf_{\tau} E [c\tau + \gamma_{\tau}^H].\end{aligned}$$

- Не представляется возможным найти аналитическое описание оптимального момента остановки даже в случае линейного тренда

РЕЗУЛЬТАТ

Разработаны **новые математические методы** оценки параметров сигнала по данным измерений, выполненных во фрактальном шуме, в том числе:

1. Получена оценка максимального правдоподобия параметров сигнала.
2. Получены оптимальные Байесовские оценки для случаев нормального и равномерного априорных распределений параметра сигнала.
3. Найден оптимальный момент остановки измерений сигнала для случая нормального априорного распределения параметра тренда.

ЗАДАЧА 2

Разработать методологию моделирования и оценивания сигналов с трендом (в т. ч. квазипериодического сигнала)

Оценивание локального тренда на основе ОМП

- Наблюдения порождены моделью

$$\xi_t = f(t) + \nu_t^H,$$

где:

- $f(t)$: гладкая n_θ раз дифференцируемая на \mathbb{R}^+ функция
- $\nu_t^H = \sigma Z_t^H$, $Z^H = (Z_t^H)_{t \geq 0}$ — процесс fGN, $\sigma > 0$
- Рассмотрим аппроксимацию $f(t)$ с помощью полинома $\sum_{i=0}^{n_\theta} \theta_i (t - t_0)^i$ в окрестности любого $t_0 > 0$
- **Цель:** оценка тренда $f(t) = \mathbb{E} \xi_t$ для каждого $t \in [t_1, t_\ell]$ по выборке $\{(X_k, t_k)\}_{k=1}^\ell$

Алгоритм оценивания тренда с поправкой на длинную память

1. Рассмотрим $W(a, b) = \{(X_k, t_k) : a \leq t_k \leq b\}$
2. Предполагая локально полиномиальную зависимость

$$X_k \approx \sum_{d=0}^3 \theta_d (t_k - t_0)^d + \sigma Z_k^H, \quad (X_k, t_k) \in W(a, b),$$

где $t_0 = (a + b)/2$, $H = 1/2$, вычислим оценку $\hat{\theta}_{ML}$.

3. Вычислим оценку тренда для каждого $t \in [a, b]$ согласно $\hat{f}(t) = \sum_{d=0}^3 (\hat{\theta}_{ML})_d (t - t_0)^d$.
4. Вычислим оценку дисперсии $\hat{\sigma}$ как выборочную дисперсию разностей $Y_k = X_k - \hat{f}(t_k)$, $(X_k, t_k) \in W(a, b)$.
5. Вычислим оценку \hat{H} по выборке $\{(R_k = Y_k / \hat{\sigma}, t_k) : (X_k, t_k) \in W(a, b)\}$.
6. Используя оценку \hat{H} , вычислим в п. 2 оценку тренда с поправкой на длинную память.

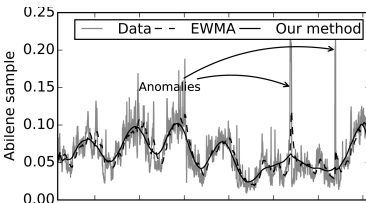
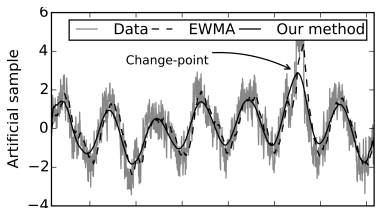


Рис.: Результаты оценивания тренда для выборок с длинной памятью.

Слева: искусственные данные $\{(X_k, t_k)\}_{k=1}^{2016}$, «измеренные» с интервалом 5 мин согласно модели $X_k = A \sin(2\pi t_k / T) + \nu_{t_k}^H$, где $A = 1.5$, $T = 288$, и ν_t^H — процесс шума с длинной памятью, $H = 0.8$.

Справа: нагрузка соединения Хьюстон-Чикаго за неделю 14–21 июня 2004 г.

Модель с явным учетом сезонности

- Наблюдения порождены моделью

$$\xi_t = f(t) + \nu_t,$$

где:

- $f(t) = Q_t S_{\varphi(t)}$: квазипериодический тренд
- $Q = (Q_t)_{t \geq 0}$: случайная амплитуда
- $\varphi(t) = 2\pi\{t/T\}$: фаза, T : период цикла
- $S_{\varphi(t)}$: неслучайный профиль цикла
- $\nu_t = \sigma_{\varphi(t)} \varepsilon_t$: случайный шум
- $\sigma_{\varphi(t)}$: неслучайная дисперсия шума
- $\varepsilon = (\varepsilon_t)_{t \geq 0}$: стандартный процесс белого шума
- **Цель:** оценить Q_t , S_{ψ} , σ_{ψ}^2 для каждого значения фазы $\psi \in [0, 2\pi]$ по выборке $\{(X_k, t_k)\}_{k=1}^{\ell}$

Алгоритм оценивания квазипериодического тренда

Шаг 0. Положим $\widehat{Q}_{t_k} = 1$, $\widehat{\sigma}_{\psi_j}^2 = \text{var}(X_1, \dots, X_\ell)$.

Шаг 1. Обновим $\widehat{S}_{\psi_j} = \frac{\sum_{k=1}^{\ell} w_k X_k / \widehat{Q}_k K_h(\varphi_k, \psi_j)}{\sum_{k=1}^{\ell} w_k K_h(\varphi_k, \psi_j)}$.

Шаг 2. Обновим $\widehat{\sigma}_{\psi_j}^2 = \frac{\sum_{k=1}^{\ell} (X_k - \widehat{X}_k)^2 K_h(\varphi_k, \psi_j)}{\sum_{k=1}^{\ell} K_h(\varphi_k, \psi_j)}$.

Шаг 3. Обновим $\widehat{X}(t_n) = \widehat{Q}_n \widehat{S}_{\varphi(t_n)}$, где \widehat{Q}_n — оценка методом наименьших квадратов из

$$X_i = Q_n \widehat{S}_{\varphi(t_i)} + \nu_i, \quad \nu_i \sim \mathcal{N}(0, \widehat{\sigma}_{\psi_i}^2)$$

РЕЗУЛЬТАТ

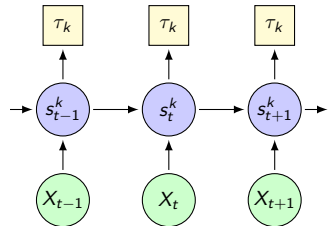
1. Разработана **новая методология моделирования** сигналов с трендом (в частности, квазипериодическим).
2. Предложены и исследованы **новые математические модели** квазипериодического сигнала и **вычислительные алгоритмы** оценивания тренда, в том числе:
 - 2.1 Локально полиномиальная модель сигнала, наблюдаемого в шуме с длинной памятью, и алгоритм оценивания тренда с поправкой на длинную память.
 - 2.2 Модель с явным учетом сезонности и алгоритм оценивания тренда на основе метода непараметрической регрессии.

ЗАДАЧА 3

Разработать метод обнаружения разладки
на основе ансамбля «слабых» детекторов
для повышения эффективности обнаружения разладки
в случае **нарушения стандартных предположений**
о ее модели

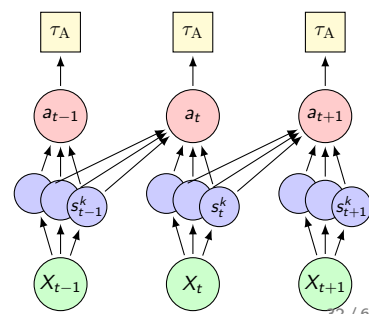
Ансамбль процедур обнаружения разладки

- Π_k : момент тревоги
 $\tau_k = \inf\{t \geq 0 : s_t^k \geq h_k\}$
- Π_k : «слабый» детектор
 (ограничительные предположения на модели сигнала и разладки)



- **Ансамбль**: момент тревоги
 $\tau_A = \inf\{t \geq 0 : a_t \geq h_A\}$
 $a_t = \psi(\lambda; \mathbf{S}_t^1, \dots, \mathbf{S}_t^n)$
- $\mathbf{S}_t^k = \{s_u^k, 0 \leq u \leq t\}$: траектория s^k
- $\lambda \in \mathbb{R}^m$: параметры
- Логистическая регрессия:

$$\psi(\lambda; \mathbf{S}) = \sigma\left(\sum_{j=0}^p \sum_{k=1}^n \lambda_{kj} s_{t-j}^k - \lambda_0\right)$$



Обучение ансамбля «слабых» детекторов

- $\mathbf{X}^\ell = \{(X_t^i, Y_t^i)\}_{i=1}^\ell$: обучающая выборка
- Модель разладки:

$$X_t^i = \begin{cases} (X_t^\infty)^i, & \text{если } t \in \mathcal{T}_\infty^i = \{t \geq 0 : Y_t^i = 0\}, \\ (X_t^0)^i, & \text{если } t \in \mathcal{T}_0^i = \{t \geq 0 : Y_t^i = 1\} \end{cases}$$

- Процессы X^∞ и X^0 : стандартная модель разладки

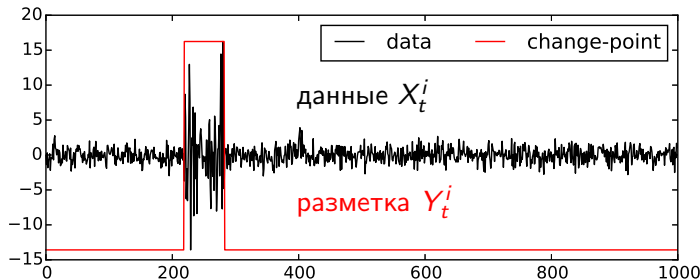


Рис.: Элемент (X_t^1, Y_t^1) , $t = 1, \dots, 1000$, искусственной выборки GARCH1

Обучение ансамбля «слабых» детекторов

- Средние потери, свойственные процедуре Π

$$\mathbf{F}(\Pi) = c_\infty E_\infty \left[\frac{\overbrace{\int_{\mathcal{T}_\infty} \mathbb{1}_{\{s_t \geq h\}}(t) dt}^{\text{длительность ложного сигнала тревоги}}}{\underbrace{\int_0^T \mathbb{1}_{\mathcal{T}_\infty}(t) dt}_{\text{длительность состояния без разрядки}}} \right] + c_0 E_0 \left[\frac{\overbrace{\int_{\mathcal{T}_0} \mathbb{1}_{\{s_t < h\}}(t) dt}^{\text{длительность ложного молчания при разрядке}}}{\underbrace{\int_0^T \mathbb{1}_{\mathcal{T}_0}(t) dt}_{\text{длительность состояния с разрядкой}}} \right]$$

- Оптимальная процедура: $\mathbf{F}(\Pi) \rightarrow \min_s$
- Обучение ансамбля: оптимизация потерь $\mathbf{F}(\Lambda) \rightarrow \min_{\Lambda \in \mathbb{R}^m}$
- Дифференцируемая аппроксимация эмпирических потерь на выборке \mathbf{X}^ℓ :

$$\widehat{\mathbf{F}}_D(\Pi) = \frac{1}{\ell} \sum_{i=1}^{\ell} \left\{ \frac{1}{T_\infty^i} \sum_{t \in \mathcal{T}_\infty^i} \sigma(s_t - h) + \frac{1}{T_0^i} \sum_{t \in \mathcal{T}_0^i} \sigma(h - s_t) \right\}$$

Эффективность ансамбля. Процесс fGn

Пример: изменение среднего значения процесса с длинной памятью

$$X_t = \begin{cases} W_t^H, & \text{если } t \notin [\theta, \theta + \Delta], \\ \mu + W_t^H, & \text{если } t \in [\theta, \theta + \Delta] \end{cases}$$

- $W^H = (W_t^H)_{0 \leq t \leq T}$: фрактальный гауссовский шум с показателем Хёрста $H = 0.95$
- величина разладки $\mu > 0$ неизвестна
- момент разладки $\theta \in [0, T]$ неизвестен

«Слабый» детектор предполагает модель

$$X_t = \mu \mathbb{1}_{\{t \geq \theta\}}(t) + W_t, \quad \mu = 1$$

Эффективность ансамбля. Процесс fGn

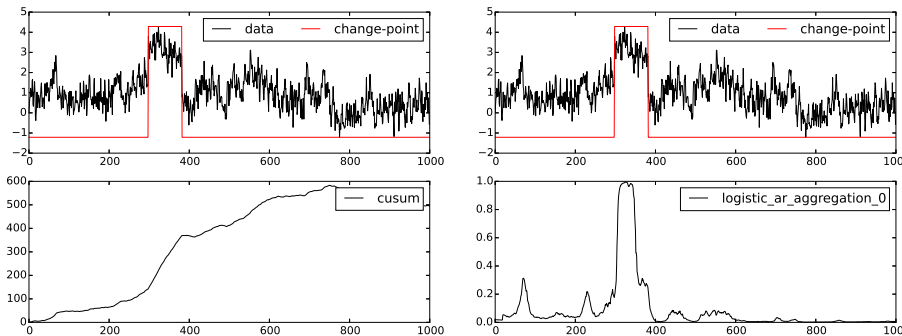


Рис.: Сверху: реализация наблюдений и разметки $(X_t^{15}, Y_t^{15}), t = 1, \dots, 1000$, искусственной выборки Fractal. Слева внизу: траектория $\{T_t, t = 1, \dots, 1000\}$ статистики кумулятивных сумм для реализации X^{15} . Справа внизу: траектория $\{a_t, t = 1, \dots, 1000\}$ статистики ансамбля LOG-1 для реализации X^{15} .

Эффективность ансамбля. Кривые качества

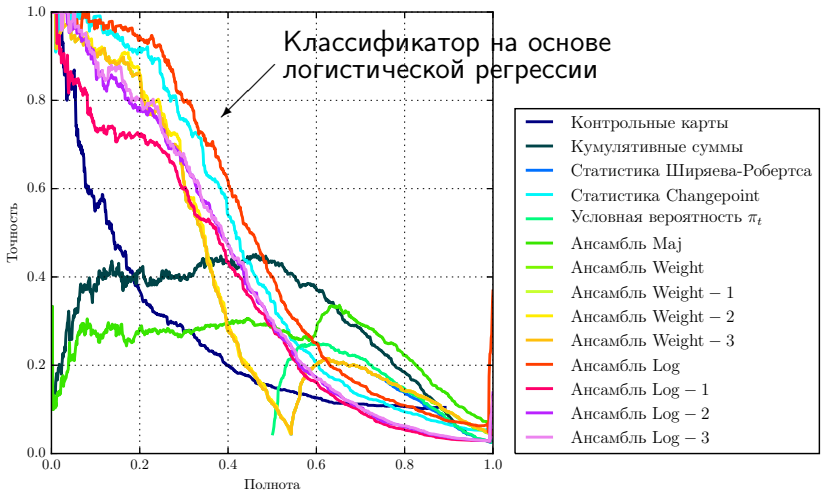


Рис.: Кривые «точность–полнота» для пяти «слабых» детекторов и девяти вариантов ансамблей, подсчитанные по отложенному подмножеству искусственной выборки Fractal.

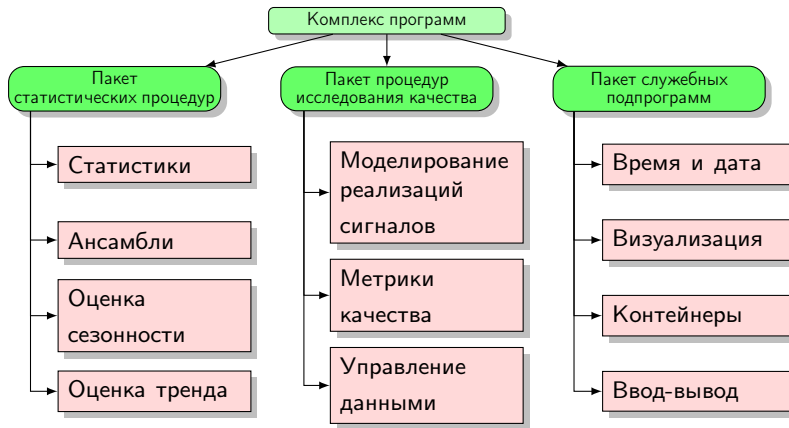
РЕЗУЛЬТАТ

1. Разработан новый **математический метод** обнаружения разладки временного ряда на основе ансамблей «слабых» детекторов.
2. Предложен новый функционал эффективности процедур обнаружения разладки.
3. Предложен **вычислительный алгоритм** настройки параметров ансамбля по выборке размеченных траекторий.

ЗАДАЧА 4

Реализовать разработанные математические методы
и алгоритмы в составе **комплекса программ**

Разработанный комплекс программ



- Разработанный программный комплекс реализован на языке python
- Реализованные алгоритмы включены в ядро комплекса мониторинга компании Яндекс

Автоматизированный модуль обнаружения разладки

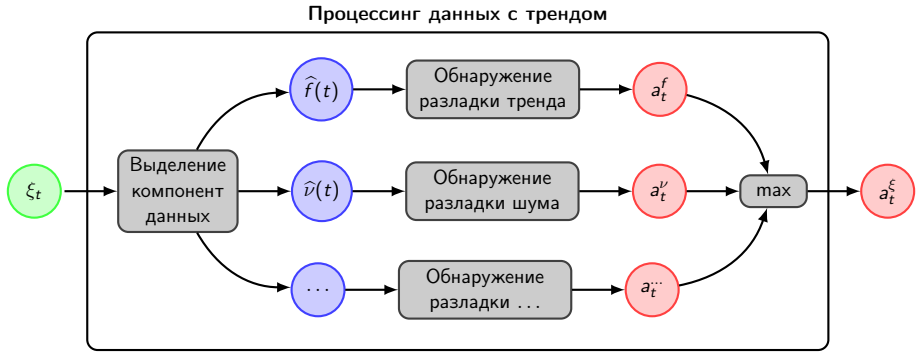


Рис.: Принципиальная схема программного модуля процессинга данных с трендом

Архитектура распределенного комплекса мониторинга

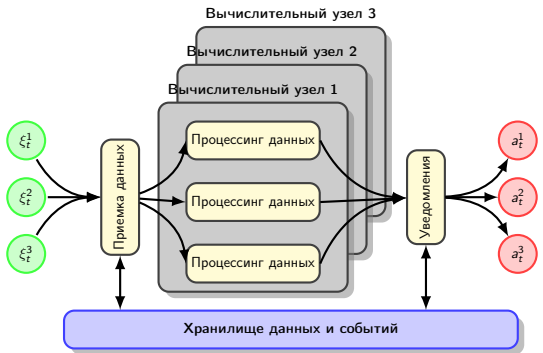


Рис.: Принципиальная схема распределенной системы мониторинга

- 3,5 млн точек в день
- 150 тыс. точек в час
- 400 ГБайт — объем базы данных
- 0.1-0.4 секунды обработка одного запроса
- Более 40 подключенных проектов

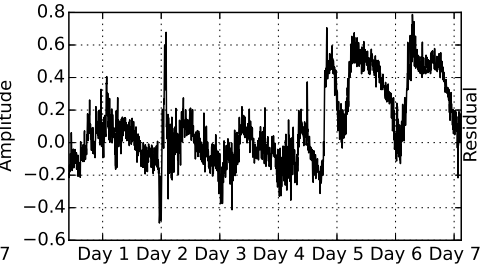
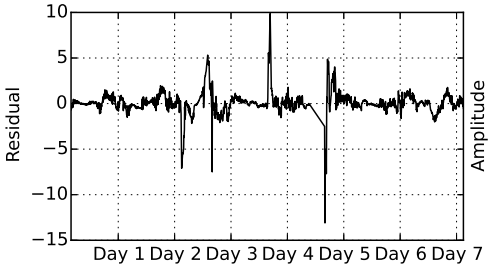
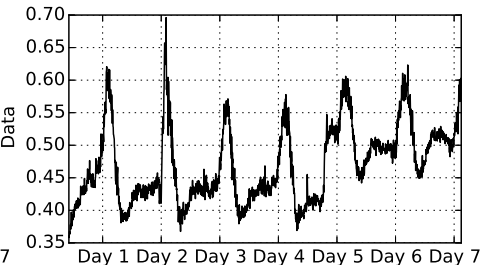
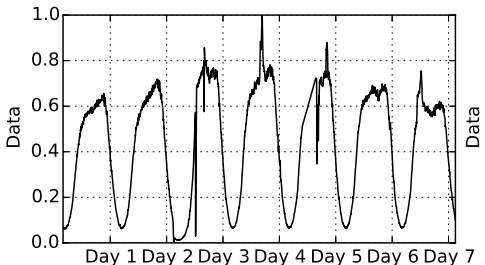
Некоторые сервисы Яндекса, использующие комплекс мониторинга

- Atom
- Yandex.Browser
- Портальная дистрибуция
- Эксперименты ABT
- Антифрод
- Авто-приемка
- Антиспам
- Авиабилеты
- AWAPS
- Яндекс.Бар
- Блендер
- CDN
- Контент-сервисы
- CORBA
- Крипта
- EntitySearch
- Админка экспериментов
- Внешний TDI
- Новостные метрики ранжирования
- Доля Яндекса по провайдерам
- OTRS
- Oxygen
- Портальные метрики
- Расписания
- Переранжирования
- Связные запросы
- Robot Coverage Metrics
- Показы объектного ответа на поисковой выдаче
- Производительность Серпа
- SERP perf testing
- Короткий клюв
- Снимпеты
- Страна по кнопке
- Саджест

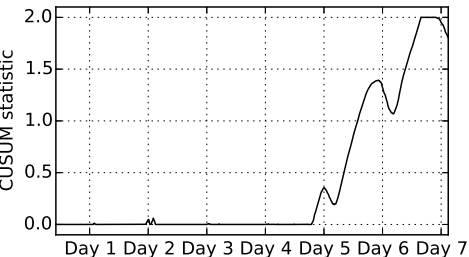
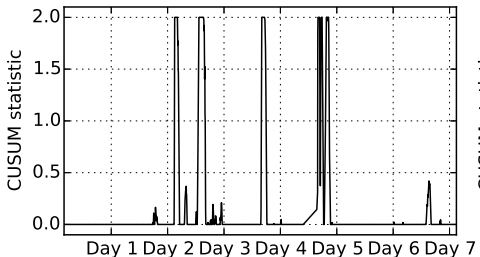
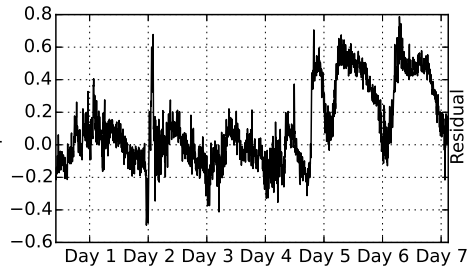
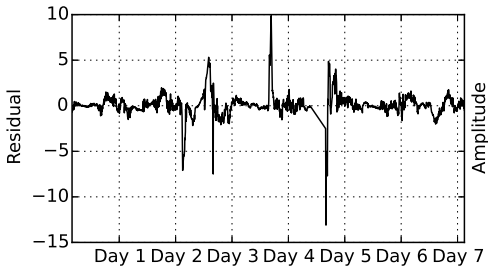
Кейс: детектирование отказов Поиска Яндекса

- 8147 одновременно наблюдаемых характеристик
- Временное разрешение — 30 мин или 24 часа
- Суточная цикличность получасовых данных (длина периода $T = 48$ точек/день)
- Экспертная разметка наличия отказов

Кейс: детектирование отказов Поиска Яндекса



Приложение: обнаружение разладок реальных сигналов



РЕЗУЛЬТАТ

1. Создан программный комплекс, реализующий разработанные математические методы и алгоритмы.
2. С использованием разработанного программного комплекса решен ряд задач в области обеспечения эксплуатации высокотехнологичных систем с интенсивным ПО.

ПОЛОЖЕНИЯ, ВЫНОСИМЫЕ НА ЗАЩИТУ





1. Разработаны новые **математические методы** оценки параметров сигнала по данным измерений, выполненным во фрактальном шуме:
 - 1.1 Найдена оценка максимального правдоподобия параметра сигнала.
 - 1.2 Найдены оптимальные Байесовские оценки для случаев нормального и равномерного априорных распределений параметра сигнала.
 - 1.3 Подсчитан оптимальный момент остановки измерений сигнала для случая нормального априорного распределения параметра тренда.

2. Предложены и исследованы **новые математические модели** квазипериодического сигнала и **вычислительные алгоритмы** оценивания тренда:
 - 2.1 Модель с явным учетом сезонности и алгоритм оценивания тренда на основе метода непараметрической регрессии.
 - 2.2 Локально полиномиальная модель сигнала, наблюдаемого в шуме с длинной памятью, и алгоритм оценивания тренда с поправкой на длинную память.

ПОЛОЖЕНИЯ, ВЫНОСИМЫЕ НА ЗАЩИТУ

- 3. Разработана новая **методология моделирования** сигналов с трендом (в частности, квазипериодическим) и обнаружения разладок и аномалий на фоне тренда.
- 4. Разработан и исследован новый **математический метод** обнаружения разладки на основе ансамбля «слабых» детекторов:
 - 4.1 Предложен новый функционал качества процедур обнаружения разладки.
 - 4.2 Предложен вычислительный алгоритм настройки параметров ансамбля по выборке размеченных траекторий.
- 5. Создан **комплекс программ**, реализующий разработанные алгоритмы.
- 6. С использованием разработанного программного комплекса решен ряд задач в области обеспечения эксплуатации высокотехнологичных систем с интенсивным ПО.

ПУБЛИКАЦИИ

-  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. (Индекс. Scopus)
-  Артемов 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.
-  Artemov A., Burnaev E. Ensembles of detectors for online detection of transient changes // In proc. of Eighth International Conference on Machine Vision. 2015. С. 98751Z–98751Z-5. (Индекс. Scopus)
-  Artemov A., Burnaev E. Nonparametric decomposition of quasi-periodic time series for change-point detection // In proc. of Eighth International Conference on Machine Vision. 2015. С. 987520–987520-5. (Индекс. Scopus)

КОНФЕРЕНЦИИ

Результаты работы докладывались и обсуждались на следующих российских и международных конференциях:

- XXII международная научная конференция студентов, аспирантов и молодых ученых «Ломоносов-2015» (2015, Москва).
- The 8th International Conference on Machine Vision (2015, Barcelona, Spain).
- 58-я научная конференция МФТИ (2015, Долгопрудный).
- Международная конференция по стохастическим методам (2016, пос. Абрау-Дюрсо, Новороссийск).
- Deep Machine Intelligence Workshop, Skolkovo Institute of Science and Technology (2016, Moscow, Russia).
- Международная конференция по алгебре, анализу и геометрии, секция теории вероятностей и математической статистики (2016, Казань).
- 9th European Summer School in Financial Mathematics (2016, St. Petersburg).

ПУБЛИКАЦИИ В СБОРНИКАХ ТРУДОВ КОНФЕРЕНЦИЙ



Артемов А. В. Масштабируемая архитектура оценивания сигналов для мониторинга крупных квазипериодических систем // Труды 58-й научной конференции МФТИ «Современные проблемы фундаментальных и прикладных наук». 2015. С. 11–14.



Артемов А. В., Бурнаев Е. В. Исследование процедуры обнаружения разладки временного ряда на основе взвешенного голосования // Материалы XXII международной научной конференции студентов, аспирантов и молодых ученых «Ломоносов». 2015. С. 34–35.



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.

ЗАДАЧА 2

Разработать методологию моделирования и оценивания сигналов с трендом (в т. ч. квазипериодического сигнала)

Оценка параметров квазипериодического сигнала

- Разность фаз: $K_h(\varphi, \psi) = K_h(\delta)$

$$\delta = \delta(\varphi, \psi) = \min(|\varphi - \psi|, |\varphi - \psi - 2\pi|, |\varphi - \psi + 2\pi|).$$

- Сглаживание и веса: вес X_k :

$$w_k = (1 - \alpha)^{-(t-t_k)/\Delta},$$

$\alpha \in (0, 1)$: вес новых данных, $\Delta = \text{const}$.

- Выбор ядра: гауссовское, лапласовское и косинусное

$$K_{\text{GAUSS}}(\delta) = \exp\left\{-\frac{\delta^2}{(2h)^2}\right\},$$

$$K_{\text{LAPLACE}}(\delta) = \exp\left\{-|\delta|/h\right\},$$

$$K_{\text{COS}}(\delta) = \frac{2 \cos(\delta/(2h)) - \cos(\delta/h)}{\pi \delta^2/h^2}$$

Оценивание квазипериодического тренда. Пример

Пример: искусственный сигнал с гладкой квазипериодической составляющей

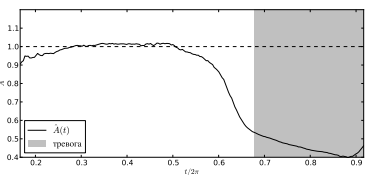
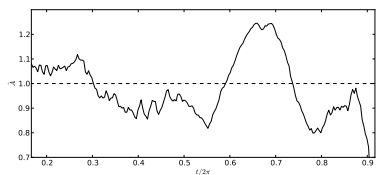
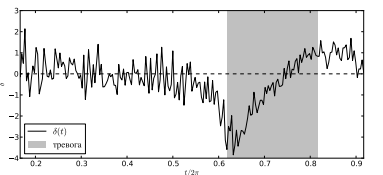
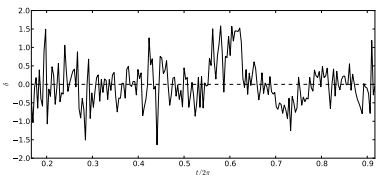
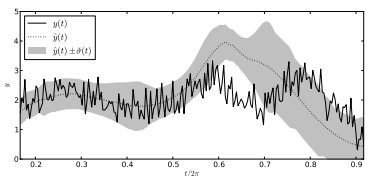
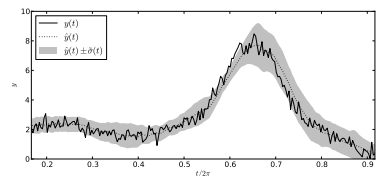
$$y_t = g(t) + f(t) + \varepsilon_t,$$

где

$$g(t) = 2 + \sin(t - 11\pi/16) + \sin(2t - 3\pi/8),$$

$$f(t) = \begin{cases} 2\mathbb{1}_{\{ \frac{9\pi}{8} < t < \frac{13\pi}{8} \}} (1 + \sin(4t - \pi)) & (\text{OK}), \\ -\mathbb{1}_{\{ \frac{9\pi}{8} < t < \frac{13\pi}{8} \}} (1 + \sin(4t - \pi)) & (\text{разладка}). \end{cases}$$

Разладка квазипериодических сигналов. Пример



Численное моделирование

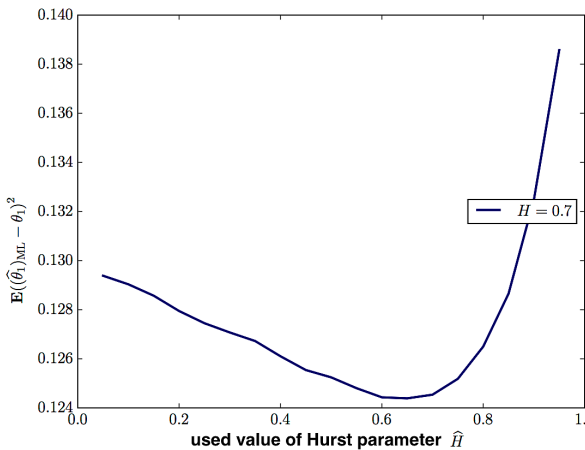


Рис.: Чувствительность алгоритма оценивания тренда к выбору показателя H

Численное моделирование

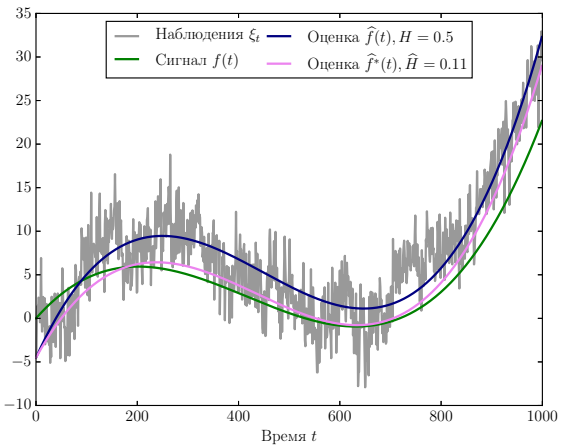


Рис.: Влияние поправки на точность оценивания тренда, $H = 0.1$

Сравнительный анализ точности выделения тренда

| МЕТОД | АППРОКСИМАЦИЯ | ПРОГНОЗ НА ОДНУ |
|----------------------------|---------------|-----------------|
| | ТРЕНДА | ТОЧКУ ВПЕРЕД |
| EWMA | 7.84 | 7.34 |
| РСА | 8.96 | 5.65 |
| РСА-РЕТРАИΝΕD | 5.58 | 3.80 |
| ПРЕДЛОЖЕННЫЙ ПОДХОД | 5.72 | 3.06 |

Таблица: Точность выделения тренда искусственного набора данных в терминах относительной с.к. погрешности (%) для процедур EWMA, РСА и предложенного подхода.

Относительная с.к. погрешность оценивания тренда задается равенством:

$$RRMSE(X_t, \hat{X}_t) = \sqrt{\frac{1}{K} \sum_{k=1}^K \frac{(X_k - \hat{X}_k)^2}{X_k^2}}$$

ЗАДАЧА 3

Разработать метод обнаружения разладки
на основе ансамбля «слабых» детекторов
для повышения эффективности обнаружения разладки
в случае **нарушения стандартных предположений**
о ее модели

Характеристики искусственных данных

| Набор | Тип данных | Изменяемые параметры | Момент разрядки θ и длительность Δ | Магнитуда разрядки |
|-------------|-----------------------------|--|--|--|
| WhiteNoise | Белый гауссовский шум | математическое | | случайная, |
| Fractal | Фрактальный гауссовский шум | ожидание μ | случайные, | $\mu \sim U(0.1, 2)$ |
| Cauchy | Белый шум Коши | | $\theta \sim U(200, 800)$, | |
| ARMA-AR | Процесс ARMA(10, 3) | AR-члены φ_i | $\Delta \sim U(5, 100)$ | случайная |
| ARMA-MA | Процесс ARMA(10, 3) | MA-члены θ_j | | случайная |
| GARCH1-ARMA | GARCH(1, 1) + ARMA(10, 3) | $\alpha_1, \beta_1, \varphi_i, \theta_j$ | | случайная |
| GARCH1 | Процесс GARCH(1, 1) | α_1, β_1 | | случайная, $\alpha_1 \sim U(.4, .8)$, $\beta_1 \sim U(.1, .2)$, |

Таблица: Характеристики использованных при сравнительном анализе искусственных наборов данных

Характеристики использованных процедур

| Процедура | Статистика процедуры | Параметры процедуры | Параметры распределения |
|-------------|--------------------------------------|------------------------|---------------------------------|
| Posterior | Статистика апостериорной вероятности | $p = 0.05$ | |
| SR | Статистика Ширяева-Робертса | | $p_\infty = \mathcal{N}(0, 1),$ |
| CUSUM | Статистика кумулятивных сумм | | $p_0 = \mathcal{N}(1, 1)$ |
| Shewhart | Статистика контрольных карт | $K = 20$ | |
| Changepoint | Статистика changepoint | $K = 24$ | |
| Maj | Ансамбль голосования большинством | | $p_\infty = \mathcal{N}(0, 1),$ |
| Weight-p | Ансамбль взвешенного голосования | $p \in \{0, 1, 2, 3\}$ | $p_0 = \mathcal{N}(1, 1)$ |
| Log-p | Ансамбль-логистическая регрессия | $p \in \{0, 1, 2, 3\}$ | |

Таблица: Характеристики использованных при сравнительном анализе процедур обнаружения разладки

Анализ эффективности ансамблей

| Процедура | WhiteNoise | Fractal | Cauchy | GARCH1 | ARMA-AR | ARMA-MA | GARCH1-ARMA |
|-------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| Shewhart | 77.52 | 24.44 | 05.45 | 32.08 | 19.80 | 76.37 | 40.00 |
| CUSUM | 61.11 | 30.70 | 19.24 | 59.88 | 28.74 | 89.90 | 75.44 |
| SR | 22.22 | 6.11 | .40 | 50.06 | 24.17 | 7.15 | 72.72 |
| Changepoint | 60.62 | 45.42 | 24.18 | 21.94 | 13.03 | 57.15 | 22.98 |
| Posterior π_t | 27.38 | 7.76 | .66 | 53.60 | 29.00 | 35.69 | 74.58 |
| MAJ | 62.13 | 24.62 | 6.11 | 47.80 | 28.74 | 92.71 | 67.01 |
| WEIGHT – 0 | 71.73 | 38.94 | 25.08 | 55.62 | 24.94 | 79.48 | 67.23 |
| WEIGHT – 1 | 71.58 | 38.97 | 13.46 | 57.65 | 29.60 | 91.92 | 71.28 |
| WEIGHT – 2 | 73.89 | 39.63 | 12.29 | 56.57 | 30.45 | 91.13 | 69.28 |
| WEIGHT – 3 | 73.25 | 38.98 | 11.61 | 57.83 | 26.24 | 90.70 | 72.11 |
| LOG – 0 | 77.25 | 48.64 | 25.90 | 51.35 | 23.29 | 87.72 | 68.35 |
| LOG – 1 | 76.27 | 36.20 | 31.03 | 50.03 | 23.49 | 88.47 | 65.97 |
| LOG – 2 | 78.01 | 39.74 | 31.30 | 49.35 | 27.99 | 88.93 | 66.43 |
| LOG – 3 | 78.85 | 40.31 | 32.24 | 49.08 | 27.99 | 88.88 | 66.77 |

Таблица: Сравнительный анализ эффективности «слабых» детекторов и ансамблей для всех наборов данных в терминах площади под кривой «точность–полнота». Полужирным шрифтом выделен лучший результат для каждого набора данных.

Обнаружение разладки квазипериодического сигнала

- Модель разладки для I_t :

$$I_t = \mu \mathbb{1}_{\{t \geq \theta\}}(t) + \sigma_{\varphi(t)} \varepsilon_t,$$

- θ : неизвестный момент разладки
- μ : неизвестная величина разладки
- Рассмотрим процесс R_t :

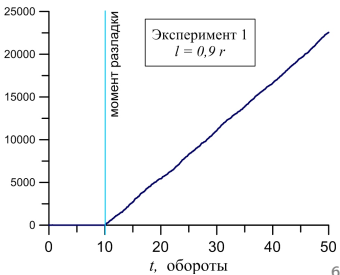
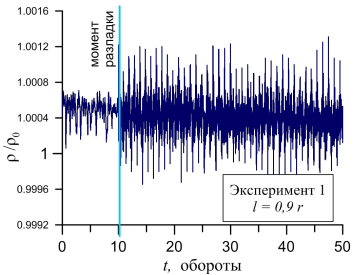
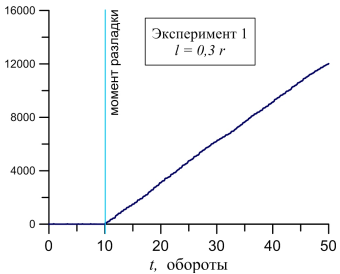
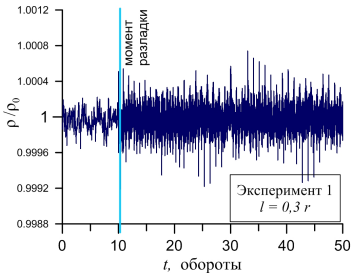
$$R_t = \frac{X_t - \hat{X}_t}{\hat{\sigma}_{\varphi(t)}}, \quad \mathbb{E}R_t \approx \begin{cases} 0, & \text{если } 0 \leq t < \theta, \\ \mu, & \text{если } t \geq \theta \end{cases}$$

- Обнаружение разладки с помощью ансамбля

Обнаружение изменения режима турбулентного течения

- Задача исследования возможности детектирования изменения режима турбулентного течения
- Моделирование течения жидкости решетчатым методом Больцмана
- Регистрируется акустический сигнал, соответствующий давлению в объеме жидкости
- Изменение типа течения при изменении геометрии объема

Приложение: обнаружение сильного изменения течения



Приложение: обнаружение слабого изменения течения

