Стохастические задачи о разладке для случайных процессов. Основные статистики в задачах скорейшего обнаружения разладки

А. В. Артёмов

ФТиАД ФКН ВШЭ,

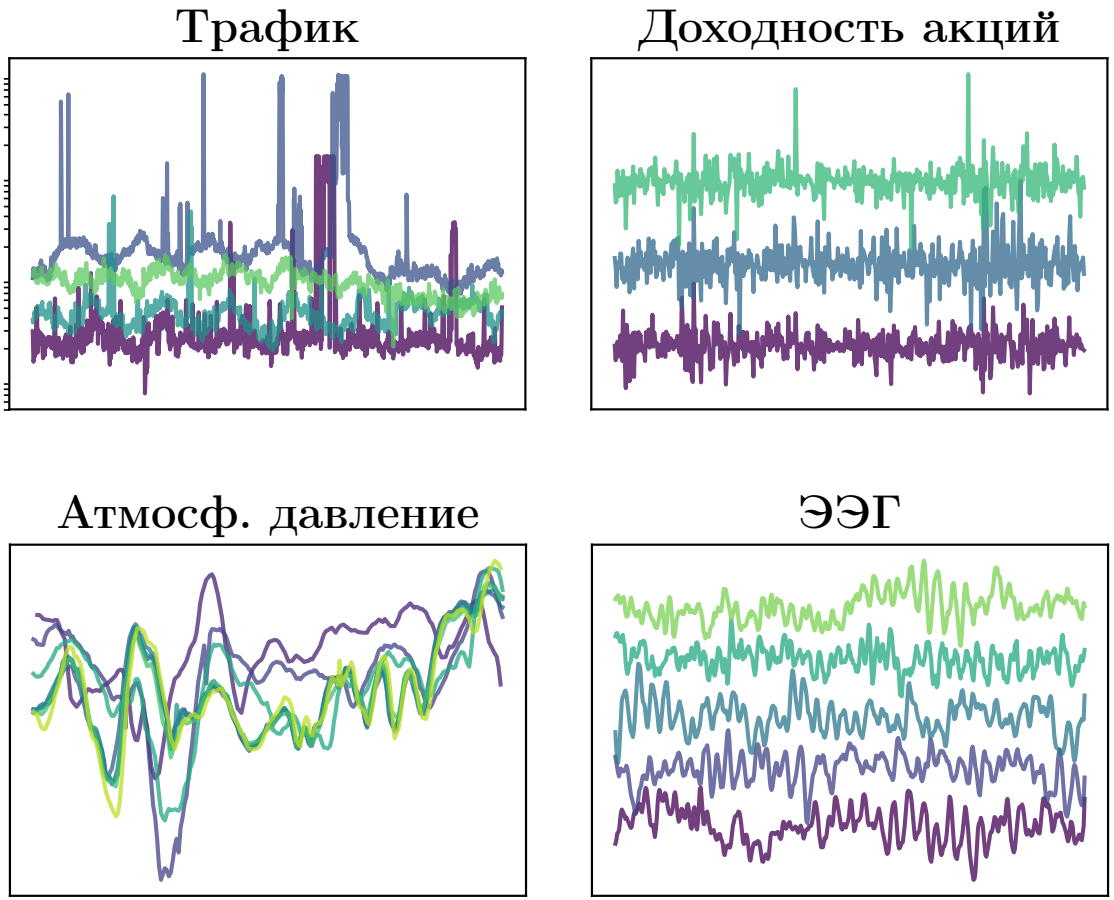
Вероятностные модели и прикладная статистика в финансовой математике, весна 2018

Содержание

Содержание

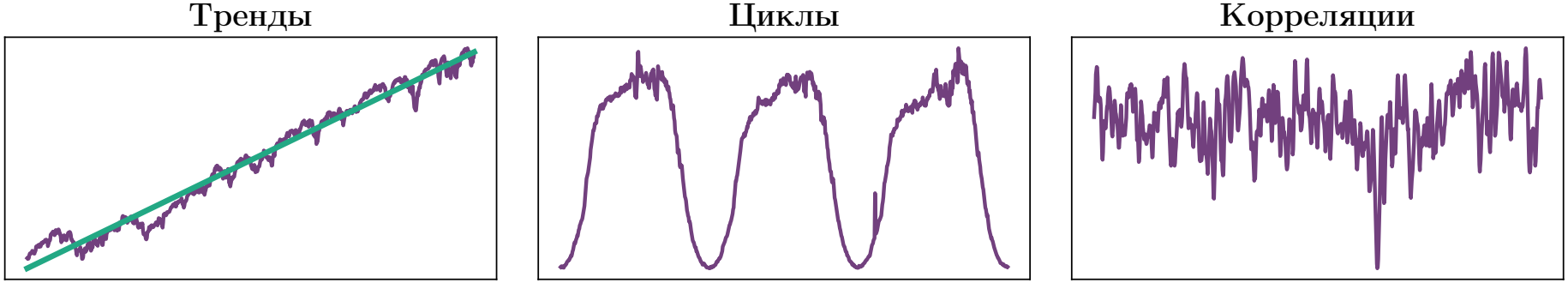
Что такое разладки?

- Многие реальные процессы описываются (многомерными) временными рядами



Что такое разладки?

- В этих временных рядах выделяют компоненты (статистические характеристики)

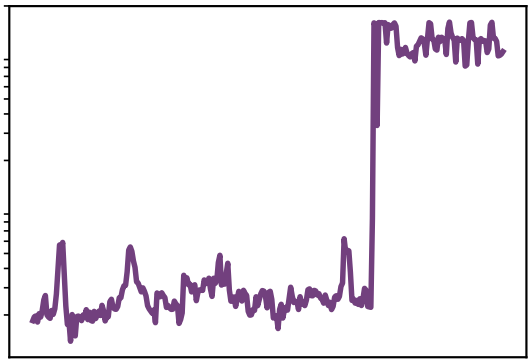


- **Описание, оценивание, выводы** о наблюдаемых временных рядах: теория и статистика случайных процессов
  - фильтрация, сегментация, шумоподавление, анализ трендов, корреляционный, дисперсионный, регрессионный, морфологический анализ, ...

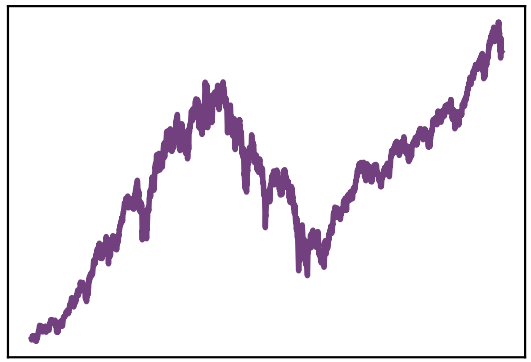
Что такое разладки?

- **Разладка:** изменение статистических свойств ряда

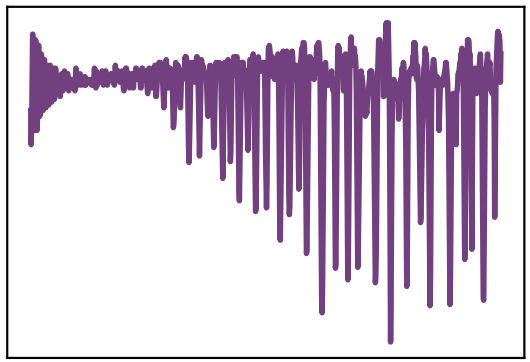
Разрывы



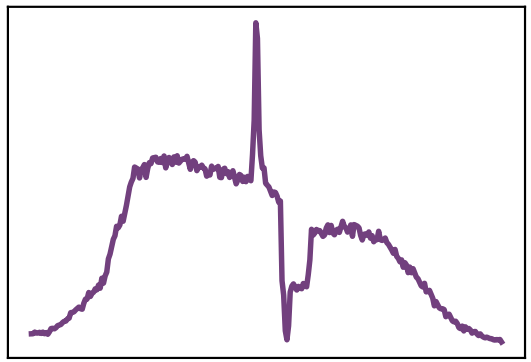
Изломы



Рост разброса

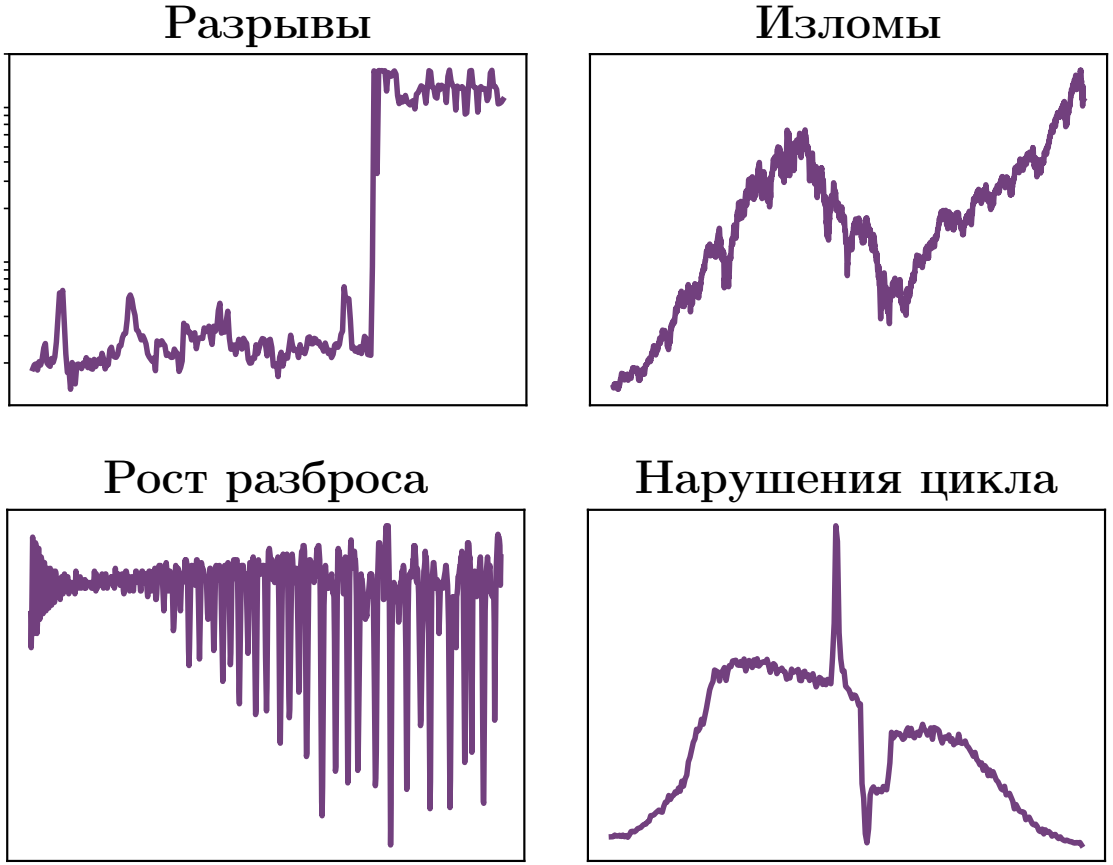


Нарушения цикла



Что такое разладки?

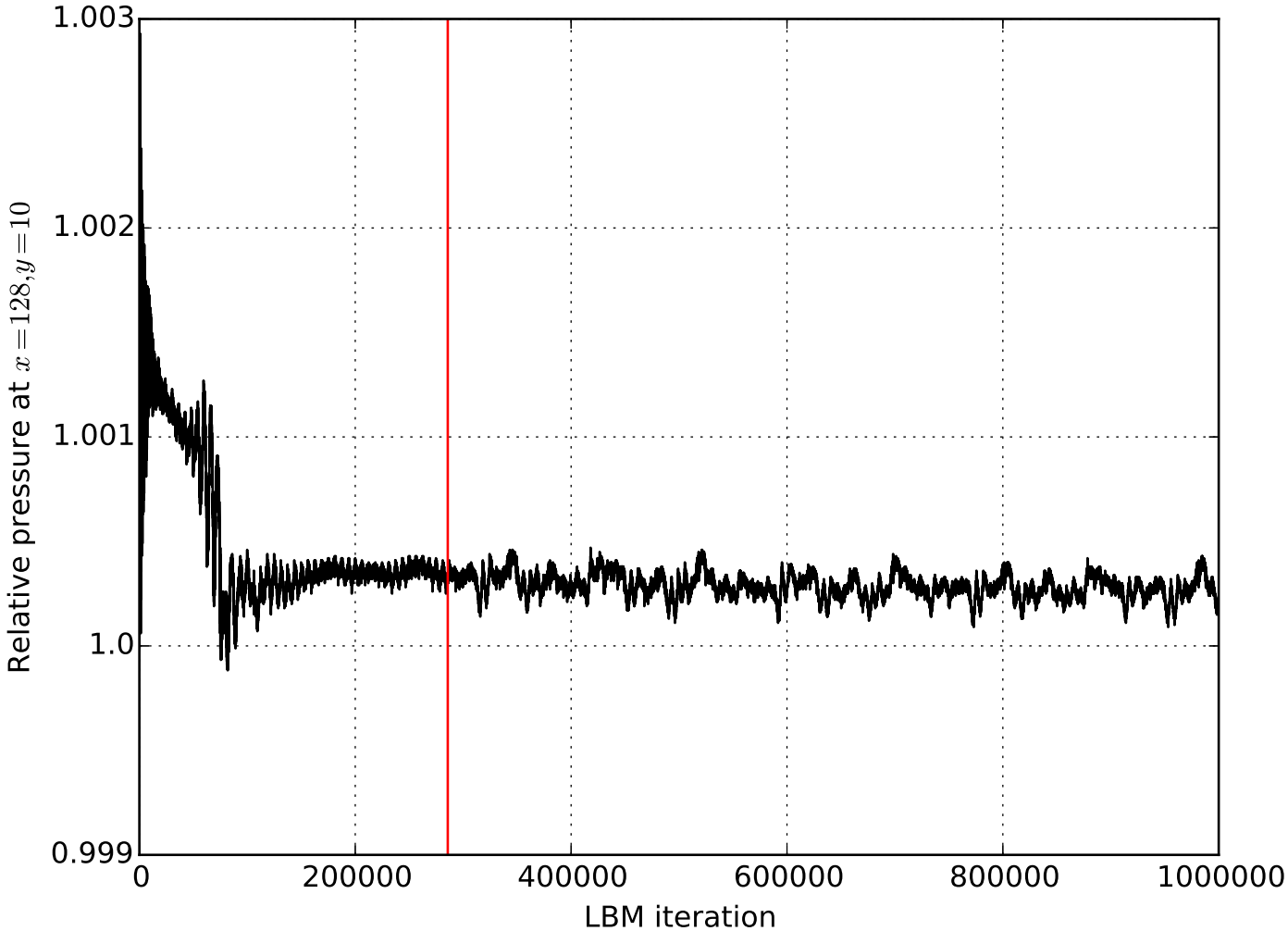
- **Разладка:** изменение статистических свойств ряда



- **Задача «о разладке»:** выявить возникающее изменение

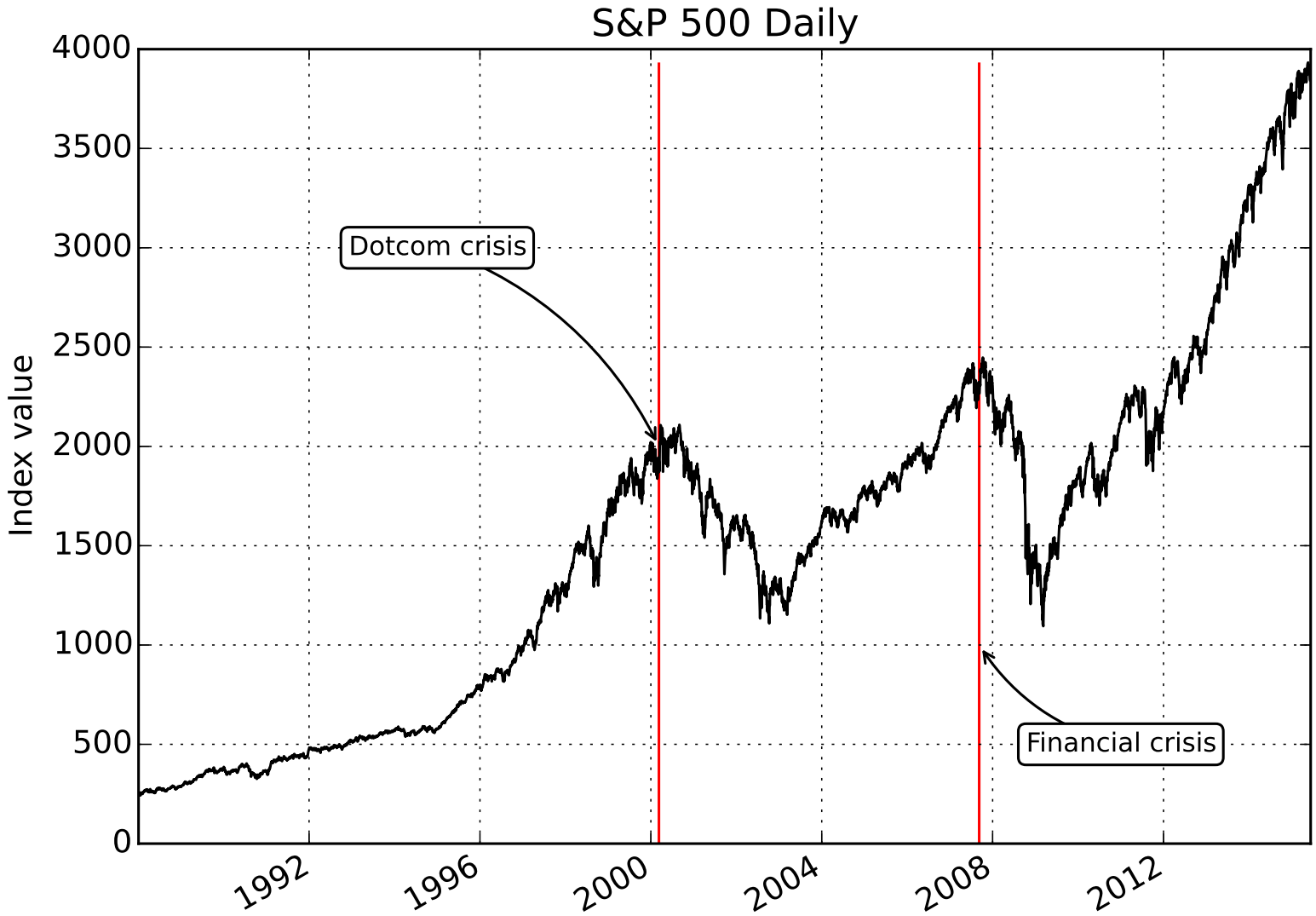
Пример из реального мира (модель гидродинамики)

Пример: давление жидкости в гидродинамической системе (модель на основе метода Больцмана).



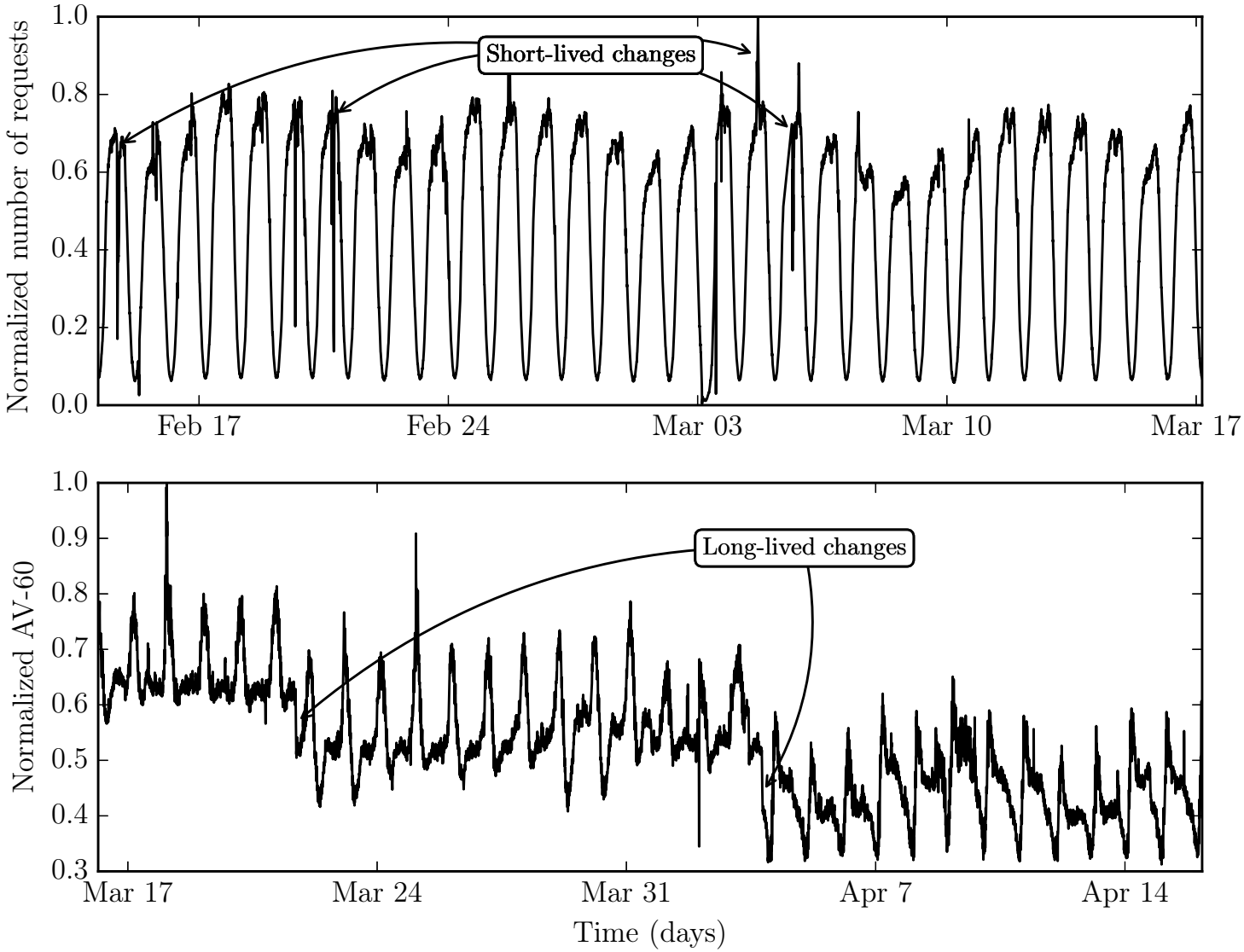
Пример из реального мира (финансовые данные)

Пример: динамика индекса S&P 500 за 27 лет.



Пример из реального мира (интернет-данные)

Пример: посещаемость интернет-сервиса.



Насколько это важно? Примеры приложений I

- Обнаружение внедрений в компьютерные сети (атак, ведущих к изменению объема передаваемого трафика) [Kim и др. 2004; Alexander G Tartakovsky 2003; Alexander G. Tartakovsky и др. 2006]
- Обнаружение аномалий в сетях передачи данных (видеопотоки в системах видеонаблюдения, сетевой трафик и др.) [Casas и др. 2010; Lakhina, Crovella и Christiphe Diot 2004; Lakhina, Crovella и Christophe Diot 2004; Pham и др. 2014; A. Tartakovsky 2013]
- Обнаружение и изоляция отказов узлов систем управления транспортными средствами [Malladi и др. 1999; Willsky 1976]
- Мониторинг целостности системы геопозиционирования [Basseville и др. 2002]
- Обнаружение изменений структуры породы при бурении скважин [Adams и др. 2007]

Насколько это важно? Примеры приложений II

- Обнаружение начала рецессии или экономического роста [Andersson и др. 2002]
- Обнаружение изменений волатильности индекса Dow Jones [Adams и др. 2007]
- Обнаружение сигнала при наблюдении подводных целей [Streit и др. 1999]
- Автоматическое обнаружение аномального человеческого поведения при видеонаблюдении [Pham и др. 2014]
- Автоматический контроль качества выпускаемой продукции [Ben-Gal и др. 2003; Girshick, Meyer A and Rubin 1952; Shewhart 1931]
- Мониторинг и анализ смертности и заболеваемости раком легких [Dass 2009; Dass и др. 2011; Taweab и др. 2015]
- Обнаружение возникновения эпидемий [MacNeill и др. 1995]

Насколько это важно? Примеры приложений III

- Обнаружение аритмии (внезапных изменений ритма биения) сердца [Willsky 1976]
- Предсказание транзиторных ишемических атак (преходящих нарушений мозгового кровообращения) [Cerutti S. и др. 1993]
- Диагностика задержки внутриутробного роста [Petzold и др. 2004]
- Анализ несчастных случаев на угольных шахтах [Adams и др. 2007]
- Мониторинг уровня хлора в питьевой воде [Guépié и др. 2012]

Насколько это важно? Объем публикуемой литературы

Поиск в системе индексации Google Scholar выдает, **начиная с 2000 года:**

- `change point detection` — **10 200** статей
- `anomaly detection` — **53 900** статей
- `break detection` — **3 980** статей
- `обнаружение разладок, обнаружение аномалий, обнаружение изменений` — **765** статей

Первые работы по разладкам: 1931 год, W. A. Shewhart (цель — контроль качества выпускаемой продукции).

Содержание

Структура наблюдаемых данных

- Наблюдаемый случайный процесс $\xi = (\xi_t)_{t \geqslant 0}$ задан на пространстве $(\Omega, \mathcal{F}, \mathrm{P})$ и имеет структуру

$$\xi_1, \xi_2, \ldots, \xi_{\theta-1}, \xi_\theta, \xi_{\theta+1}, \ldots$$

и описывается данными $X = (X_t)_{t \geqslant 0}$

$$X_1, X_2, \ldots, X_{\theta-1}, X_\theta, X_{\theta+1}, \ldots$$

Структура наблюдаемых данных

- Наблюдаемый случайный процесс $\xi = (\xi_t)_{t \geqslant 0}$ задан на пространстве $(\Omega, \mathcal{F}, \mathrm{P})$ и имеет структуру

$$\underbrace{\xi_1, \xi_2, \ldots, \xi_{\theta-1},}_{\text{до разладки } \mathrm{P}=\mathrm{P}_\infty} \quad \underbrace{\xi_\theta, \xi_{\theta+1}, \ldots}_{\text{разладка: } \mathrm{P}=\mathrm{P}_0}$$

  и описывается данными $X = (X_t)_{t \geqslant 0}$

$$X_1, X_2, \ldots, X_{\theta-1}, \quad X_\theta, X_{\theta+1}, \ldots$$

Структура наблюдаемых данных

- Наблюдаемый случайный процесс $\xi = (\xi_t)_{t \geqslant 0}$ задан на пространстве $(\Omega, \mathcal{F}, \mathrm{P})$ и имеет структуру

$$\underbrace{\xi_1, \xi_2, \ldots, \xi_{\theta-1},}_{\text{до разладки } \mathrm{P}=\mathrm{P}_\infty} \quad \underbrace{\xi_\theta, \xi_{\theta+1}, \ldots}_{\text{разладка: } \mathrm{P}=\mathrm{P}_0}$$

  и описывается данными $X = (X_t)_{t \geqslant 0}$

$$X_1, X_2, \ldots, X_{\theta-1}, \quad X_\theta, X_{\theta+1}, \ldots$$

Структура наблюдаемых данных

- Наблюдаемый случайный процесс $\xi = (\xi_t)_{t \geqslant 0}$ задан на пространстве $(\Omega, \mathcal{F}, \mathrm{P})$ и имеет структуру

$$\underbrace{\xi_1, \xi_2, \ldots, \xi_{\theta-1},}_{\text{до разладки } \mathrm{P}=\mathrm{P}_\infty} \quad \underbrace{\xi_\theta, \xi_{\theta+1}, \ldots}_{\text{разладка: } \mathrm{P}=\mathrm{P}_0}$$

  и описывается данными $X = (X_t)_{t \geqslant 0}$

$$X_1, X_2, \ldots, X_{\theta-1}, \quad X_\theta, X_{\theta+1}, \ldots$$

- θ — момент появления **разладки,** который требуется оценить по данным

Структура наблюдаемых данных

- Наблюдаемый случайный процесс $\xi = (\xi_t)_{t\geqslant 0}$ задан на пространстве $(\Omega, \mathcal{F}, \mathrm{P})$ и имеет структуру

$$\underbrace{\xi_1, \xi_2, \ldots, \xi_{\theta-1},}_{\text{до разладки } \mathrm{P}=\mathrm{P}_\infty} \quad \underbrace{\xi_\theta, \xi_{\theta+1}, \ldots}_{\text{разладка: } \mathrm{P}=\mathrm{P}_0}$$

  и описывается данными $X = (X_t)_{t\geqslant 0}$

$$X_1, X_2, \ldots, X_{\theta-1}, \quad X_\theta, X_{\theta+1}, \ldots$$

- θ — момент появления **разладки,** который требуется оценить по данным
- Классическая модель в непрерывном времени:

$$\xi_t = \mu \mathbb{1}_{\{t\geqslant\theta\}}(t) + W_t, \quad W = (W_t)_{t\geqslant 0} - \text{БД}$$

Роль распределений процесса P_∞, P_0 и P_θ

- Наблюдаемый процесс $\xi = (\xi_t)_{t \geqslant 0}$ имеет структуру

$$\xi_1, \xi_2, \ldots, \xi_{\theta-1}, \quad \xi_\theta, \xi_{\theta+1}, \ldots$$

Роль распределений процесса P_∞, P_0 и P_θ

- Наблюдаемый процесс $\xi = (\xi_t)_{t \geqslant 0}$ имеет структуру

$$\xi_1, \xi_2, \ldots, \xi_{\theta-1}, \quad \xi_\theta, \xi_{\theta+1}, \ldots$$

- P_∞: распределение ξ в предположении, что разладка не появляется никогда ($\theta = \infty$)

$$\xi_1, \xi_2, \ldots, \xi_{\theta-1}, \xi_\theta, \xi_{\theta+1}, \ldots \quad \sim \quad \mathrm{P}_\infty$$

Роль распределений процесса P_∞, P_0 и P_θ

- Наблюдаемый процесс $\xi = (\xi_t)_{t \geqslant 0}$ имеет структуру

$$\xi_1, \xi_2, \ldots, \xi_{\theta-1}, \quad \xi_\theta, \xi_{\theta+1}, \ldots$$

- P_∞: распределение ξ в предположении, что разладка не появляется никогда ($\theta = \infty$)

$$\xi_1, \xi_2, \ldots, \xi_{\theta-1}, \xi_\theta, \xi_{\theta+1}, \ldots \quad \sim \quad \mathrm{P}_\infty$$

- P_0: распределение ξ в предположении, что разладка произошла в момент старта наблюдений ($\theta = 0$)

$$\xi_1, \xi_2, \ldots, \xi_{\theta-1}, \xi_\theta, \xi_{\theta+1}, \ldots \quad \sim \quad \mathrm{P}_0$$

Роль распределений процесса P_∞, P_0 и P_θ

- Наблюдаемый процесс $\xi = (\xi_t)_{t \geqslant 0}$ имеет структуру

$$\xi_1, \xi_2, \ldots, \xi_{\theta-1}, \quad \xi_\theta, \xi_{\theta+1}, \ldots$$

- P_∞: распределение ξ в предположении, что разладка не появляется никогда ($\theta = \infty$)

$$\xi_1, \xi_2, \ldots, \xi_{\theta-1}, \xi_\theta, \xi_{\theta+1}, \ldots \quad \sim \quad \mathrm{P}_\infty$$

- P_0: распределение ξ в предположении, что разладка произошла в момент старта наблюдений ($\theta = 0$)

$$\xi_1, \xi_2, \ldots, \xi_{\theta-1}, \xi_\theta, \xi_{\theta+1}, \ldots \quad \sim \quad \mathrm{P}_0$$

- У мер P_∞ и P_0 есть плотности $f_\infty(\cdot)$ и $f_0(\cdot)$ и соответствующие матожидания E_∞ и E_0

Роль распределений процесса P_∞, P_0 и P_θ

- Наблюдаемый процесс $\xi = (\xi_t)_{t \geqslant 0}$ имеет структуру

$$\xi_1, \xi_2, \ldots, \xi_{\theta-1}, \quad \xi_\theta, \xi_{\theta+1}, \ldots$$

- P_θ: распределение ξ в предположении, что разладка произошла в момент θ
- Плотность $f_\theta^n(x), x \in \mathbb{R}^n$ меры P_θ имеет специальный вид

$$f_\theta^n(X_1, \ldots, X_n) = f_\infty(X_1, \ldots, X_{\theta-1}) \cdot f_0(X_\theta, \ldots, X_n)$$

Пример: бракованные изделия

- $\xi_1, \xi_2, \ldots$: длина выпускаемых изделий, $\xi_i \in \mathbb{R}_+$

Пример: бракованные изделия

- $\xi_1, \xi_2, \ldots$: длина выпускаемых изделий, $\xi_i \in \mathbb{R}_+$
- *Нормальный* ход индустриального процесса:
  $\xi_1, \xi_2, \ldots$ — i.i.d., $\xi_i \sim \mathcal{N}(\mu_\infty, \sigma^2)$, $(\theta = \infty)$

Пример: бракованные изделия

- $\xi_1, \xi_2, \ldots$: длина выпускаемых изделий, $\xi_i \in \mathbb{R}_+$
- *Нормальный* ход индустриального процесса:
  $\xi_1, \xi_2, \ldots$ — i.i.d., $\xi_i \sim \mathcal{N}(\mu_\infty, \sigma^2)$, $(\theta = \infty)$
- Изначально производятся *бракованные* изделия:
  $\xi_1, \xi_2, \ldots$ — i.i.d., $\xi_i \sim \mathcal{N}(\mu_0, \sigma^2)$, $(\theta = 0)$

Пример: бракованные изделия

- $\xi_1, \xi_2, \ldots$: длина выпускаемых изделий, $\xi_i \in \mathbb{R}_+$
- *Нормальный* ход индустриального процесса:
  $\xi_1, \xi_2, \ldots$ — i.i.d., $\xi_i \sim \mathcal{N}(\mu_\infty, \sigma^2)$, $(\theta = \infty)$
- Изначально производятся *бракованные* изделия:
  $\xi_1, \xi_2, \ldots$ — i.i.d., $\xi_i \sim \mathcal{N}(\mu_0, \sigma^2)$, $(\theta = 0)$
- Типичный случай: сначала имеет место нормальный ход, но в момент θ наступает «сбой» («разладка»):

$$\underbrace{\xi_1, \xi_2, \ldots, \xi_{\theta-1},}_{\mathcal{N}(\mu_\infty,\sigma^2)} \quad \underbrace{\xi_\theta, \xi_{\theta+1}, \cdots}_{\mathcal{N}(\mu_0,\sigma^2)}$$

Обнаружение разладки: момент подачи *сигнала тревоги*

- Пусть до момента времени n доступны наблюдения

$$\mathbf{X}_n = (X_1, \ldots, X_n)$$

Обнаружение разладки: момент подачи *сигнала тревоги*

- Пусть до момента времени n доступны наблюдения

$$\mathbf{X}_n = (X_1, \ldots, X_n)$$

- Требуется подать *сигнал тревоги* в момент $\tau = n$, если есть доказательства появления разладки

Обнаружение разладки: момент подачи *сигнала тревоги*

- Пусть до момента времени n доступны наблюдения

$$\mathbf{X}_n = (X_1, \ldots, X_n)$$

- Требуется подать *сигнал тревоги* в момент $\tau = n$, если есть доказательства появления разладки
- **Момент остановки:** статистика $\tau = \tau(\mathbf{X}_n)$

$$\tau \in \{0, 1 \ldots, \infty\}, \qquad \{\mathbf{X}_n : \tau(\mathbf{X}_n) = n\} \in \sigma(\mathbf{X}_n)$$

Обнаружение разладки: момент подачи *сигнала тревоги*

- Пусть до момента времени n доступны наблюдения

$$\mathbf{X}_n = (X_1, \ldots, X_n)$$

- Требуется подать *сигнал тревоги* в момент $\tau = n$, если есть доказательства появления разладки
- **Момент остановки:** статистика $\tau = \tau(\mathbf{X}_n)$

$$\tau \in \{0, 1 \ldots, \infty\}, \qquad \{\mathbf{X}_n : \tau(\mathbf{X}_n) = n\} \in \sigma(\mathbf{X}_n)$$

- Используется лишь накопленная к *настоящему времени* информация (не используется будущее)

Обнаружение разладки: момент подачи *сигнала тревоги*

- Пусть до момента времени n доступны наблюдения

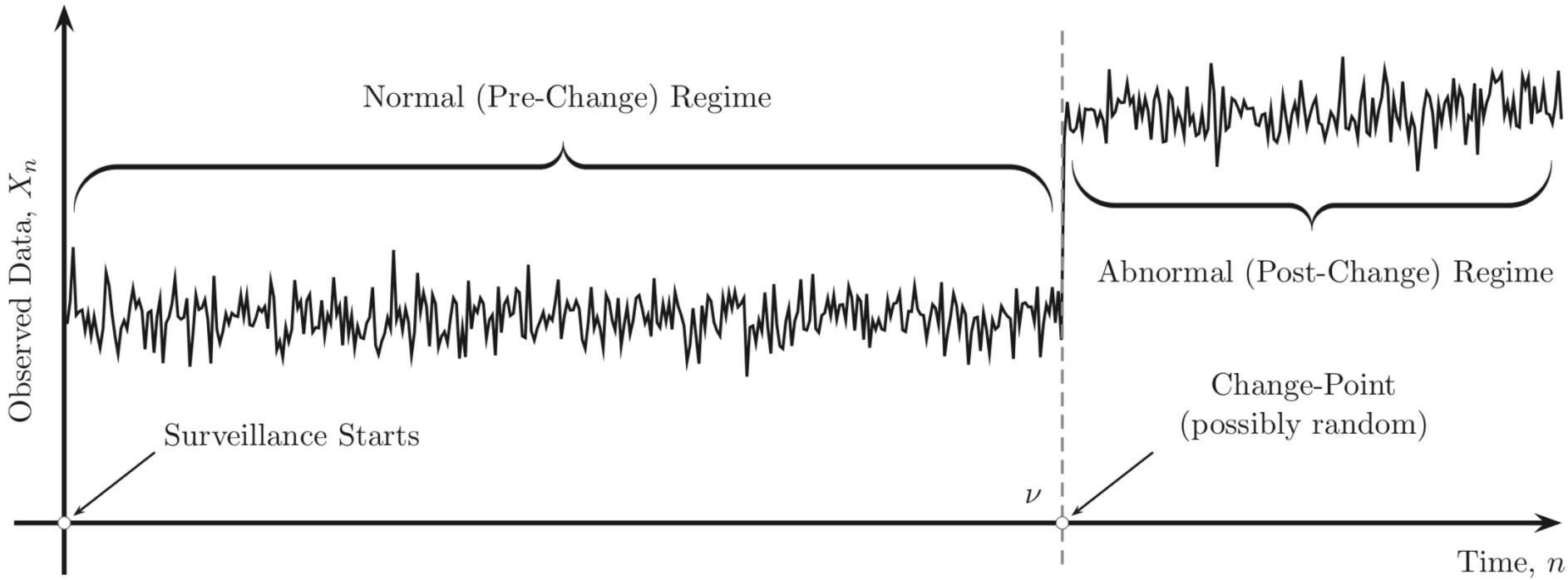
$$\mathbf{X}_n = (X_1, \ldots, X_n)$$

- Требуется подать *сигнал тревоги* в момент $\tau = n$, если есть доказательства появления разладки
- **Момент остановки:** статистика $\tau = \tau(\mathbf{X}_n)$

$$\tau \in \{0, 1 \ldots, \infty\}, \qquad \{\mathbf{X}_n : \tau(\mathbf{X}_n) = n\} \in \sigma(\mathbf{X}_n)$$

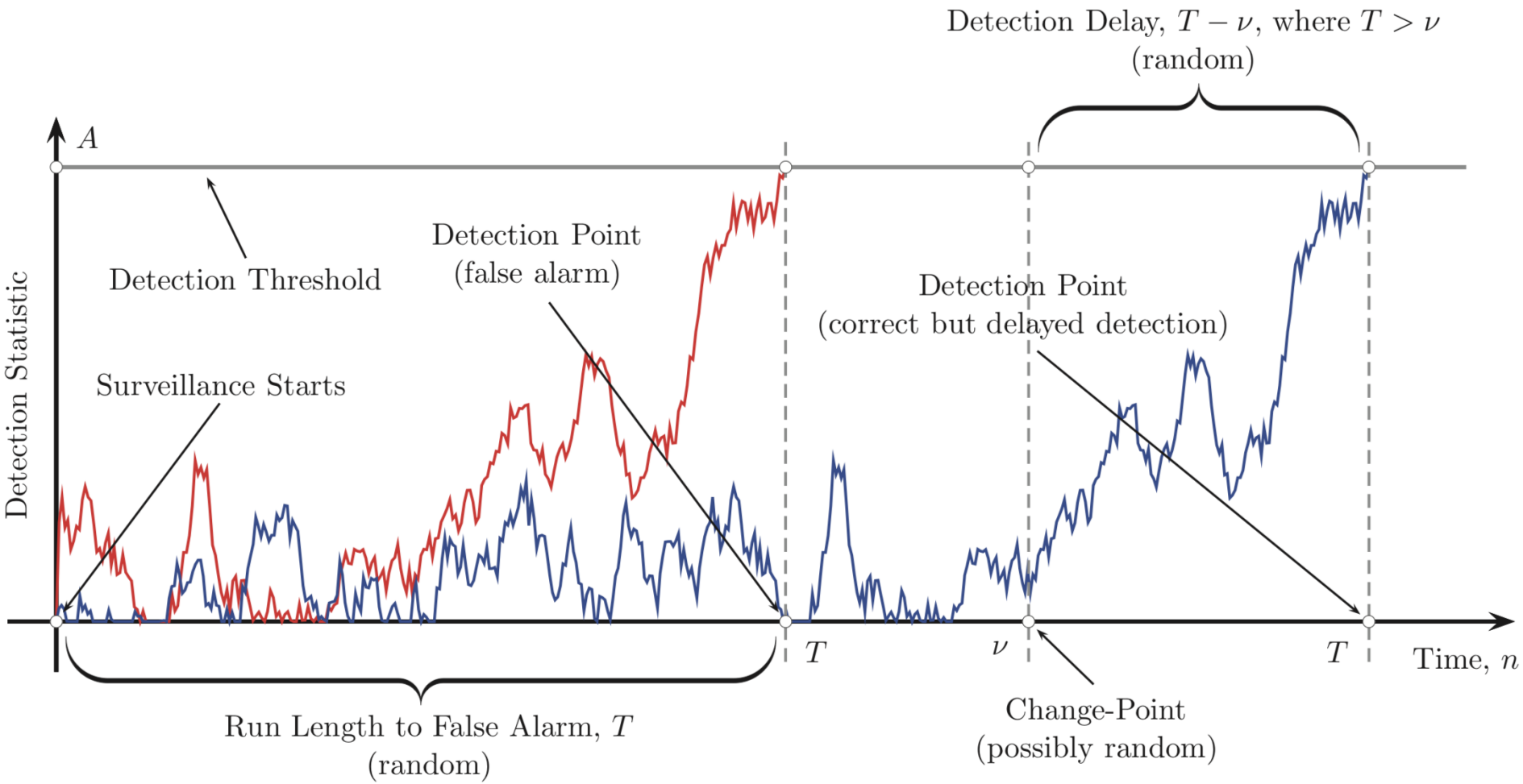
- Используется лишь накопленная к *настоящему времени* информация (не используется будущее)
- Момент остановки τ строится на основе некоторой статистики наблюдений $\gamma = (\gamma_t)_{t \geqslant 0}, \gamma_n = \gamma_n(\mathbf{X}_n)$

Типичный сценарий обнаружения разладки



Изображение: *Polunchenko, Aleksey S., and Alexander G. Tartakovsky. "State-of-the-art in sequential change-point detection."Methodology and computing in applied probability 14.3 (2012): 649-684.*

Типичный сценарий обнаружения разладки, $\nu = \theta, T = \tau$



Изображение: *Polunchenko, Aleksey S., and Alexander G. Tartakovsky. "State-of-the-art in sequential change-point detection."Methodology and computing in applied probability 14.3 (2012): 649-684.*

Высококачественный момент $\tau(\mathbf{X}_n)$

- Момент остановки: статистика $\tau = \tau(\mathbf{X}_n)$

Высококачественный момент $\tau(\mathbf{X}_n)$

- Момент остановки: статистика $\tau = \tau(\mathbf{X}_n)$
- $\mathrm{E}_\infty \tau$: среднее время до ложной тревоги (false detection delay, $\mathrm{FDD}(\tau)$)

Высококачественный момент $\tau(\mathbf{X}_n)$

- Момент остановки: статистика $\tau = \tau(\mathbf{X}_n)$
- $\mathrm{E}_\infty \tau$: среднее время до ложной тревоги (false detection delay, $\mathrm{FDD}(\tau)$)
- **Хорошо:** $\mathrm{E}_\infty \tau \to \infty$ (редкие ложные тревоги)

Высококачественный момент $\tau(\mathbf{X}_n)$

- Момент остановки: статистика $\tau = \tau(\mathbf{X}_n)$
- $\mathrm{E}_\infty \tau$: среднее время до ложной тревоги (false detection delay, $\mathrm{FDD}(\tau)$)
- **Хорошо:** $\mathrm{E}_\infty \tau \to \infty$ (редкие ложные тревоги)
- $\mathrm{E}_0 \tau$ или $\mathrm{E}_\theta[\tau - \theta | \tau > \theta]$: средняя задержка в обнаружении (average detection delay, $\mathrm{ADD}(\tau)$)

Высококачественный момент $\tau(\mathbf{X}_n)$

- Момент остановки: статистика $\tau = \tau(\mathbf{X}_n)$
- $\mathrm{E}_\infty \tau$: среднее время до ложной тревоги (false detection delay, $\mathrm{FDD}(\tau)$)
- **Хорошо:** $\mathrm{E}_\infty \tau \to \infty$ (редкие ложные тревоги)
- $\mathrm{E}_0 \tau$ или $\mathrm{E}_\theta[\tau - \theta | \tau > \theta]$: средняя задержка в обнаружении (average detection delay, $\mathrm{ADD}(\tau)$)
- **Хорошо:** $\mathrm{E}_0 \tau \to 0$ (быстрое обнаружение)

Некачественный момент $\tau(\mathbf{X}_n)$

- Момент остановки: статистика $\tau = \tau(\mathbf{X}_n)$

Некачественный момент $\tau(\mathbf{X}_n)$

- Момент остановки: статистика $\tau = \tau(\mathbf{X}_n)$
- $\mathrm{P}_\theta(\tau < \theta)$: вероятность ложной тревоги (probability of false alarm, $\mathrm{PFA}(\tau)$)

Некачественный момент $\tau(\mathbf{X}_n)$

- Момент остановки: статистика $\tau = \tau(\mathbf{X}_n)$
- $\mathrm{P}_\theta(\tau < \theta)$: вероятность ложной тревоги (probability of false alarm, $\mathrm{PFA}(\tau)$)
- **Плохо:** $\mathrm{P}_\theta(\tau < \theta) \to 1$ (частые ложные тревоги)

Некачественный момент $\tau(\mathbf{X}_n)$

- Момент остановки: статистика $\tau = \tau(\mathbf{X}_n)$
- $\mathrm{P}_\theta(\tau < \theta)$: вероятность ложной тревоги (probability of false alarm, $\mathrm{PFA}(\tau)$)
- **Плохо:** $\mathrm{P}_\theta(\tau < \theta) \to 1$ (частые ложные тревоги)
- $\mathrm{E}_0\,\tau$ или $\mathrm{E}_\theta[\tau - \theta | \tau > \theta]$: средняя задержка в обнаружении (average detection delay, $\mathrm{ADD}(\tau)$)

Некачественный момент $\tau(\mathbf{X}_n)$

- Момент остановки: статистика $\tau = \tau(\mathbf{X}_n)$
- $\mathrm{P}_\theta(\tau < \theta)$: вероятность ложной тревоги (probability of false alarm, $\mathrm{PFA}(\tau)$)
- **Плохо:** $\mathrm{P}_\theta(\tau < \theta) \to 1$ (частые ложные тревоги)
- $\mathrm{E}_0\,\tau$ или $\mathrm{E}_\theta[\tau - \theta | \tau > \theta]$: средняя задержка в обнаружении (average detection delay, $\mathrm{ADD}(\tau)$)
- **Плохо:** $\mathrm{E}_\theta[\tau - \theta | \tau > \theta] \to \infty$ (медленное обнаружение)

Некачественный момент $\tau(\mathbf{X}_n)$

- Момент остановки: статистика $\tau = \tau(\mathbf{X}_n)$
- $\mathrm{P}_\theta(\tau < \theta)$: вероятность ложной тревоги (probability of false alarm, $\mathrm{PFA}(\tau)$)
- **Плохо:** $\mathrm{P}_\theta(\tau < \theta) \to 1$ (частые ложные тревоги)
- $\mathrm{E}_0\,\tau$ или $\mathrm{E}_\theta[\tau - \theta | \tau > \theta]$: средняя задержка в обнаружении (average detection delay, $\mathrm{ADD}(\tau)$)
- **Плохо:** $\mathrm{E}_\theta[\tau - \theta | \tau > \theta] \to \infty$ (медленное обнаружение)
- **В практике:** задержка срабатывания $\mathrm{ADD}(\tau)$ и время без ложных тревог $\mathrm{FDD}(\tau)$ — **конфликтующие критерии** (строится зависимость одного от другого)

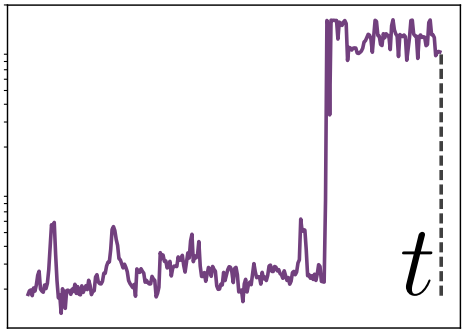
Содержание

Подходы к обнаружению разладки временного ряда

Некоторые процедуры обнаружения разладки, использующие частные особенности данных:

- Контрольные карты [Shewhart 1931]
- Алгоритм кумулятивных сумм (CUSUM) [Page 1954]
- Экспоненциально взвешенное скользящее среднее [Roberts 1959]
- Фильтр Калмана [Kalman 1960]
- Байесовские методы [Girshick, Meyer A and Rubin 1952; А. Ширяев 1961]
- Процедура Ширяева-Робертса [Roberts 1966; Альберт Николаевич Ширяев 1961]
- Метод обобщенного отношения правдоподобия [Willsky 1976]
- Методы на основе контекстных деревьев [Ben-Gal и др. 2003]
- Ядерные методы [Desobry и др. 2005]
- Энтропийный подход [Дарховский 2013]

Построение процедур обнаружения разладки



Детектор
$\Pi(\{X_s, s \leqslant t\})$

$\begin{cases} \text{ОК}, \\ \text{Тревога} \end{cases}$
в момент t

- Π: детектор, процедура обнаружения (должен учитывать некоторые предположения о модели сигнала и разладки)
- Π: момент тревоги $\tau = \inf\{t \geqslant 0 : \gamma_t \geqslant h\}$

Общие обозначения

- Пусть $X_1, \ldots, X_n$ — наблюдения, доступные до момента времени n
- Основная статистика — отношение правдоподобия

$$L_n = \frac{f_0(X_1, \ldots, X_n)}{f_\infty(X_1, \ldots, X_n)}$$

- Удобно также использовать статистику $Z_n = \log L_n$
- Если наблюдения $X_1, \ldots, X_n$ независимы:

$$L_n = \prod_{k=1}^{n} \frac{f_0(X_k)}{f_\infty(X_k)} = \prod_{k=1}^{n} l_k, \quad Z_n = \sum_{k=1}^{n} \log \frac{f_0(X_k)}{f_\infty(X_k)} = \sum_{k=1}^{n} \zeta_k$$

Простой пример

- Пусть $\xi_1, \ldots, \xi_n$ — нормальные i.i.d.r.v., причем

$$f_0(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-r_0)^2}{2\sigma^2}}, \qquad f_\infty(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-r_\infty)^2}{2\sigma^2}}$$

Тогда правдоподобие выборки $X_1, \ldots, X_n$

$$L_n = \prod_{k=1}^{n} \exp\left\{ \frac{r_\infty - r_0}{\sigma^2} \left[X_k - \frac{r_0 + r_\infty}{2} \right] \right\},$$

а его логарифм —

$$Z_n = \frac{r_\infty - r_0}{\sigma^2} \left[\overline{X_n} - \frac{r_0 + r_\infty}{2} n \right]$$

Контрольные карты Шухарта

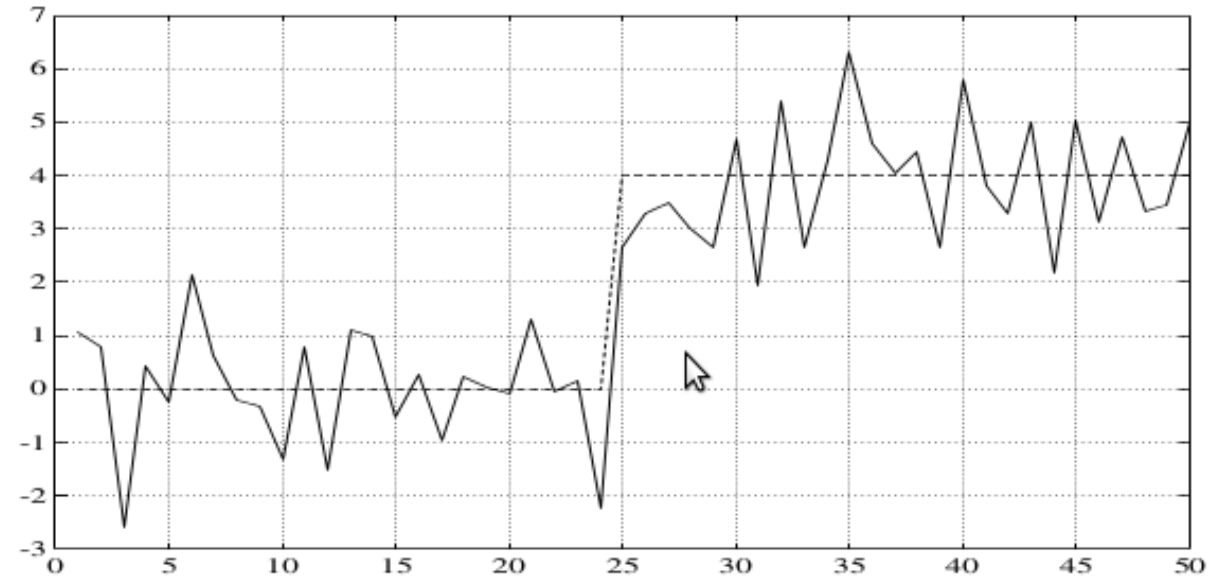
- Наблюдения $X_1, X_2, \ldots$ разбиваются на группы (батчи) размера N (N — параметр алгоритма)
- Для каждой группы $\mathbf{X}_K = (X_{N\cdot(K-1)+1}, \ldots, X_{N\cdot K})$, $K = 1, 2, \ldots$ подсчитывается логарифм правдоподобия:

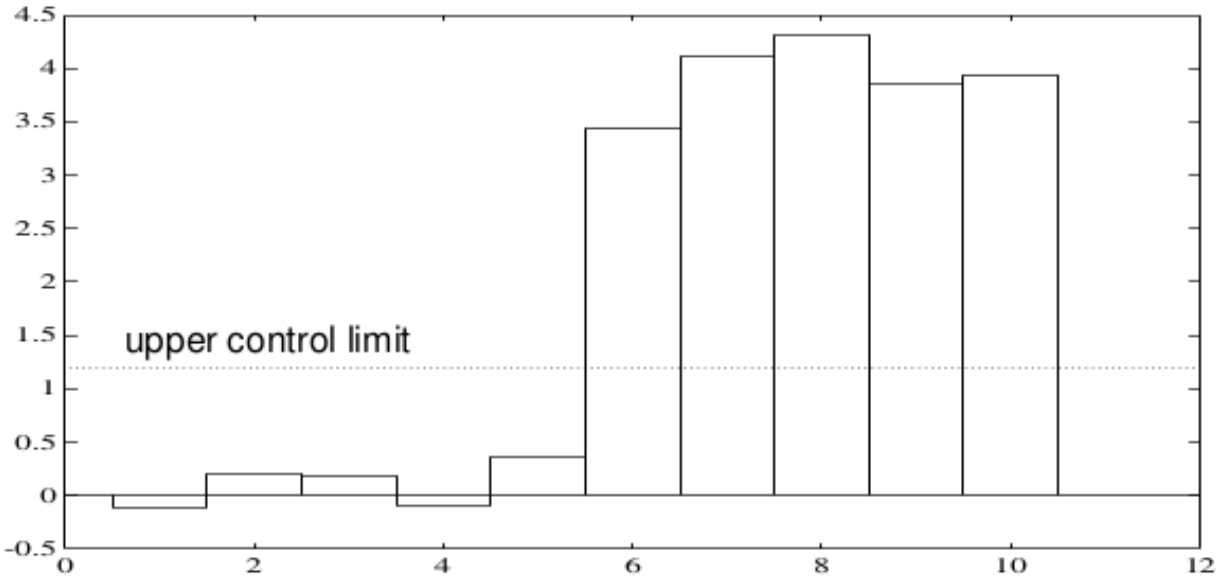
$$S_i^k = \sum_{j=i}^{k} \zeta_j$$

- Момент остановки — первый момент выхода статистики $S_{N\cdot(K-1)+1}^{N\cdot K}$ на заданный уровень h:

$$\tau_{\text{SH}} = N \cdot \min\{K : S_{N\cdot(K-1)+1}^{N\cdot K} \geqslant h\}$$

Пример [Michèle Basseville и др. 1993]





Экспоненциально взвешенное скользящее среднее

- Задается рекурсивная оценка среднего значения

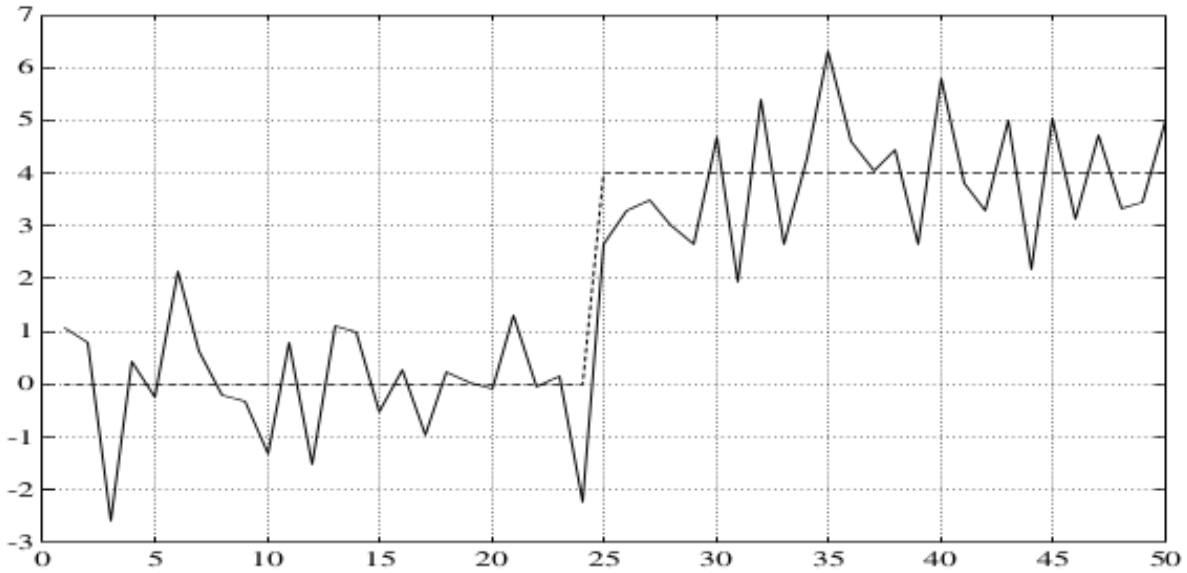
$$\hat{m}_k = (1-\lambda)\hat{m}_{k-1} + \lambda X_k, \qquad k = 1, 2, \ldots,$$

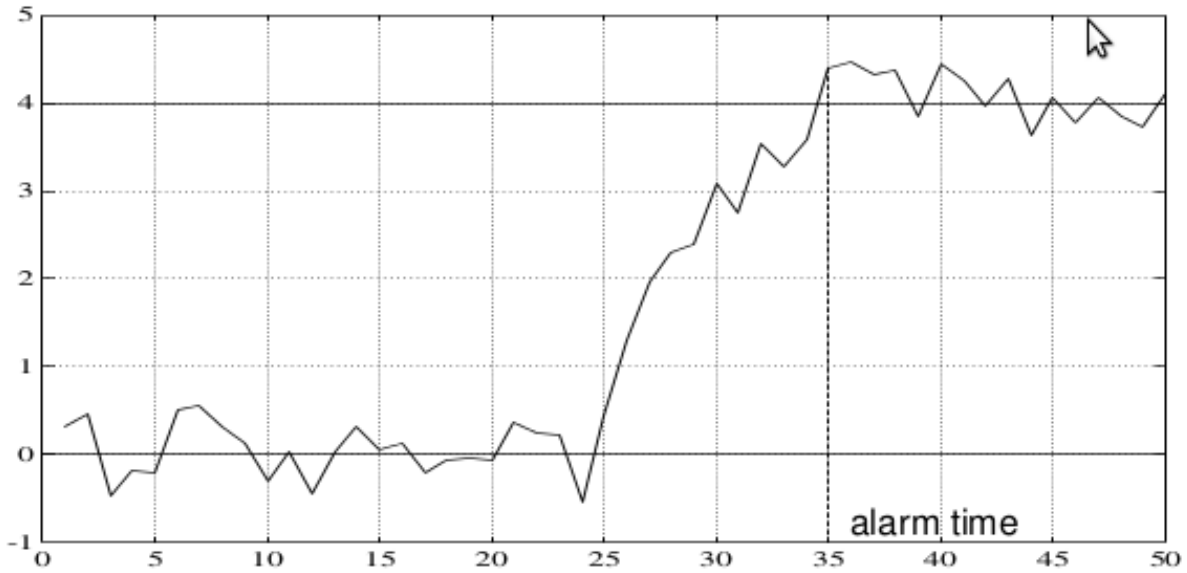
где λ («вес» новых данных) — параметр алгоритма

- Момент остановки — момент первого выхода статистики $\hat{m}_k$ на заданный уровень h:

$$\tau_{\mathrm{EWMA}} = \min\{k \geqslant 1 : \hat{m}_k \geqslant h\}$$

Пример [Michèle Basseville и др. 1993]





Содержание

Кумулятивные суммы

- О параметре θ не делается никаких предположений

Кумулятивные суммы

- О параметре θ не делается никаких предположений
- Фиксируется $T > 0$ и задается класс

$$\mathcal{M}_T = \{\tau : \mathrm{E}_\infty \tau \geqslant T\}$$

тех моментов остановки, для которых среднее время до ложной тревоги не меньше T

Кумулятивные суммы

- О параметре θ не делается никаких предположений
- Фиксируется $T > 0$ и задается класс

$$\mathcal{M}_T = \{\tau : \mathrm{E}_\infty \tau \geqslant T\}$$

тех моментов остановки, для которых среднее время до ложной тревоги не меньше T

- Качество алгоритма задается величиной

$$\mathbf{D}(T) = \sup_{\theta \geqslant 0} \operatorname{ess\,sup}_{\omega} \mathrm{E}_\theta\Big((\tau - \theta)^+ \Big| \mathcal{F}_\theta\Big)(\omega) \quad \sim \quad \inf_{\tau \in \mathcal{M}_T}$$

Кумулятивные суммы

- О параметре θ не делается никаких предположений
- Фиксируется $T > 0$ и задается класс

$$\mathcal{M}_T = \{\tau : \mathrm{E}_\infty \tau \geqslant T\}$$

тех моментов остановки, для которых среднее время до ложной тревоги не меньше T

- Качество алгоритма задается величиной

$$\mathbf{D}(T) = \sup_{\theta \geqslant 0} \operatorname{ess\,sup}_{\omega} \underbrace{\mathrm{E}_\theta\big((\tau - \theta)^+ \big| \mathcal{F}_\theta\big)(\omega)}_{\substack{\text{среднее время} \\ \text{обнаружения разладки}}} \quad \sim \quad \inf_{\tau \in \mathcal{M}_T}$$

Кумулятивные суммы

- О параметре θ не делается никаких предположений
- Фиксируется $T > 0$ и задается класс

$$\mathcal{M}_T = \{\tau : \mathrm{E}_\infty \tau \geqslant T\}$$

тех моментов остановки, для которых среднее время до ложной тревоги не меньше T

- Качество алгоритма задается величиной

$$\mathbf{D}(T) = \sup_{\theta \geqslant 0} \underbrace{\operatorname{ess\,sup}_{\omega} \underbrace{\mathrm{E}_\theta\big((\tau - \theta)^+ \big| \mathcal{F}_\theta\big)(\omega)}_{\text{среднее время обнаружения разладки}}}_{\text{наихудшее среди всех траекторий}} \quad \sim \quad \inf_{\tau \in \mathcal{M}_T}$$

Кумулятивные суммы

- О параметре θ не делается никаких предположений
- Фиксируется $T > 0$ и задается класс

$$\mathcal{M}_T = \{\tau : \mathrm{E}_\infty \tau \geqslant T\}$$

тех моментов остановки, для которых среднее время до ложной тревоги не меньше T

- Качество алгоритма задается величиной

$$\mathbf{D}(T) = \sup_{\theta \geqslant 0} \underbrace{\operatorname*{ess\,sup}_{\omega} \underbrace{\underbrace{\mathrm{E}_\theta\big((\tau-\theta)^+ \big| \mathcal{F}_\theta\big)(\omega)}_{\text{среднее время обнаружения разладки}}}_{\text{наихудшее среди всех траекторий}}}_{\text{и всех моментов } \theta \text{ появления разладки}} \quad \sim \quad \inf_{\tau \in \mathcal{M}_T}$$

Кумулятивные суммы

- Вводятся статистики

$$\gamma_n = \sup_{\theta \geqslant 0} \frac{f_\theta(X_1, \ldots, X_n)}{f_\infty(X_1, \ldots, X_n)} \qquad \text{и} \qquad T_n = \log \gamma_n$$

Кумулятивные суммы

- Вводятся статистики

$$\gamma_n = \sup_{\theta \geqslant 0} \frac{f_\theta(X_1, \ldots, X_n)}{f_\infty(X_1, \ldots, X_n)} \qquad \text{и} \qquad T_n = \log \gamma_n$$

- Если случайные величины $\xi_1, \ldots, \xi_n$ независимы, то

$$\gamma_n = \max\left\{1, \max_{1 \leqslant \theta \leqslant n} \prod_{k=\theta}^{n} \frac{f_0(X_k)}{f_\infty(X_k)}\right\},$$

$$T_n = \max\left\{0, \max_{1 \leqslant \theta \leqslant n} \sum_{k=\theta}^{n} \log \frac{f_0(X_k)}{f_\infty(X_k)}\right\} =$$

$$= \max\left\{0, \max_{1 \leqslant \theta \leqslant n} \sum_{k=\theta}^{n} \zeta_k\right\}$$

Кумулятивные суммы

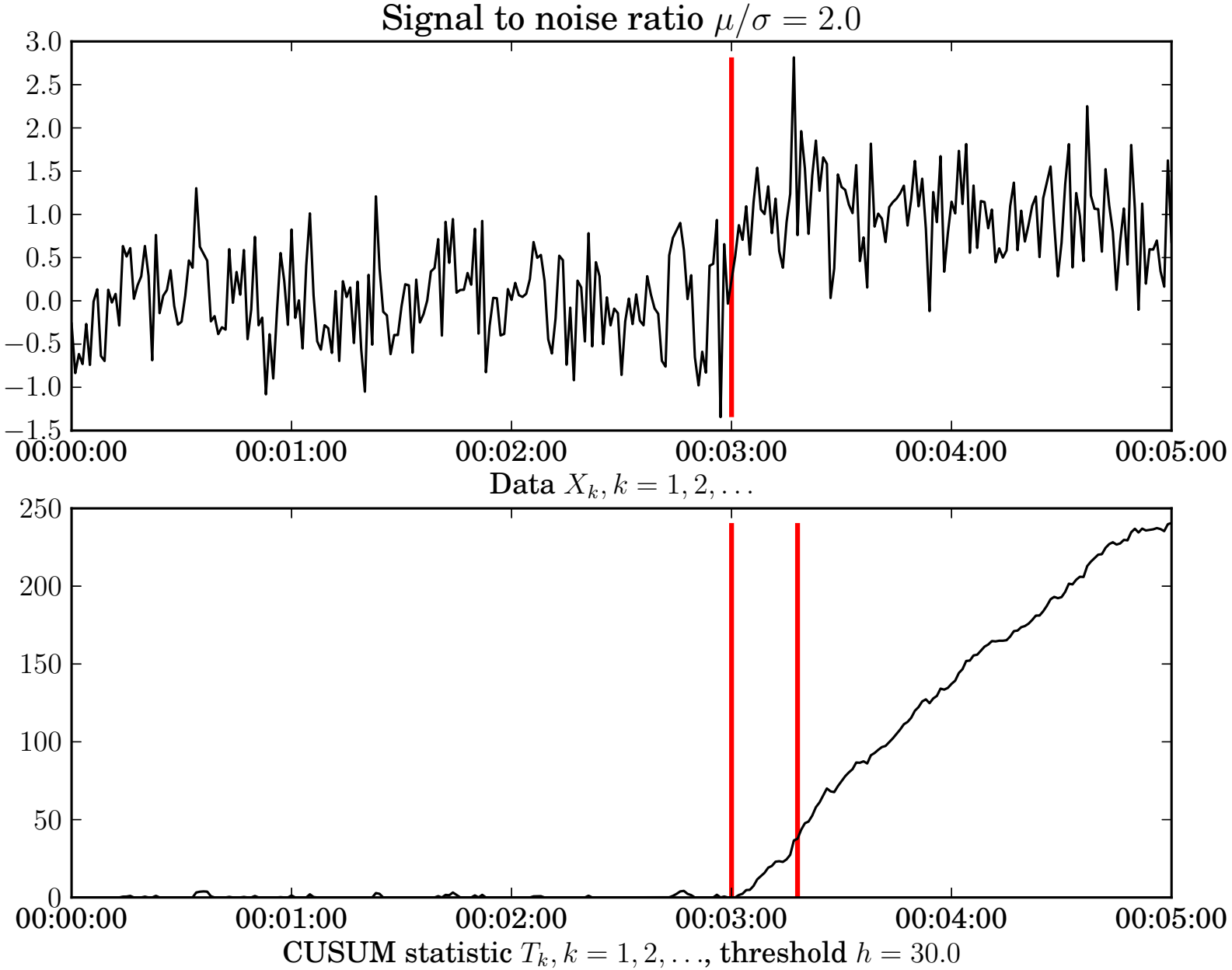
- Статистика T_n обладает свойством $T_n = \max(0, T_{n-1} + \zeta_n)$ и называется статистикой кумулятивных сумм (CUmulative SUMs, CUSUM).

Кумулятивные суммы

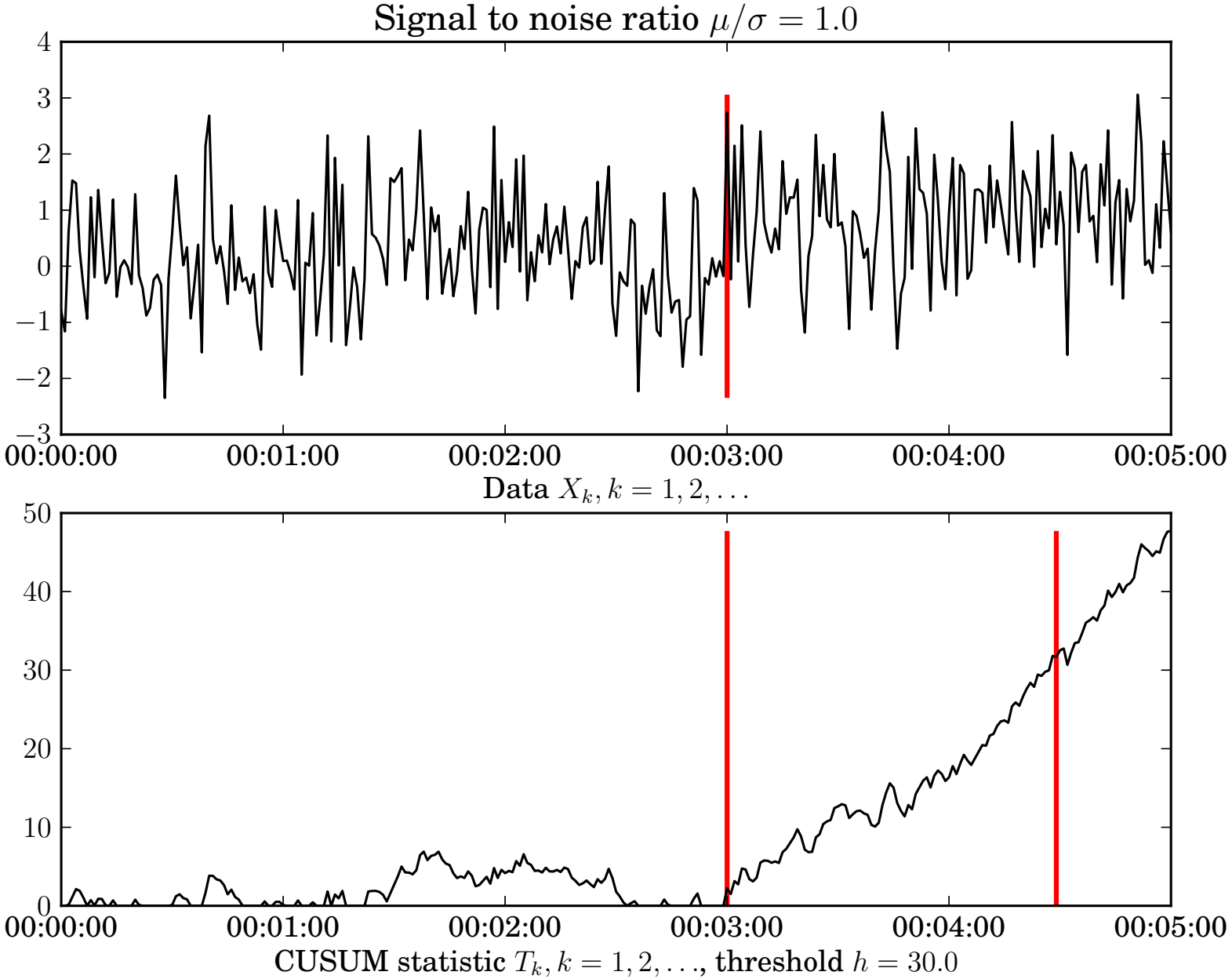
- Статистика T_n обладает свойством $T_n = \max(0, T_{n-1} + \zeta_n)$ и называется статистикой кумулятивных сумм (CUmulative SUMs, CUSUM).
- Остановка в момент τ_{CUSUM} минимизирует величину $\mathbf{D}(T)$

$$\tau_{\text{CUSUM}} = \inf\{n \geqslant 0 : T_n \geqslant h\}$$

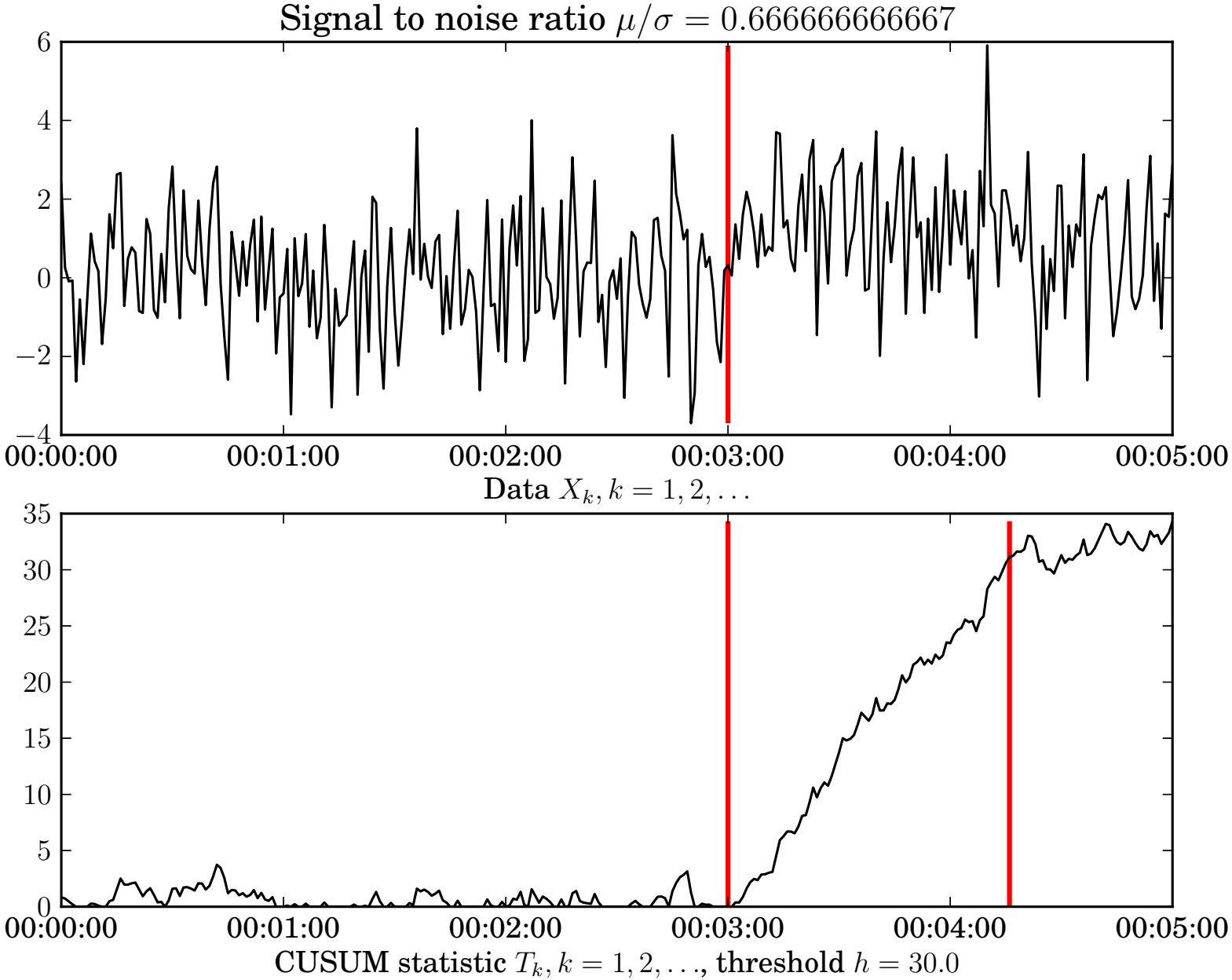
Пример 8.1 [Razladki]



Пример 8.2 [Razladki]



Пример 8.3 [Razladki]



Статистика Ширяева-Робертса: история

- Разработана независимо А. Н. Ширяевым (Альберт Николаевич Ширяев 1961, 1963) и S. W. Roberts (Roberts 1966)

Статистика Ширяева-Робертса: история

- Разработана независимо А. Н. Ширяевым (Альберт Николаевич Ширяев 1961, 1963) и S. W. Roberts (Roberts 1966)
- Появилась в эпизоде сериала Numb3rs (Episode 308: Hardball)

Episode 308:
Hardball

A professional baseball player trying to make a comeback to the major leagues dies after he is given a lethal injection of steroids. The investigation reveals that someone may have known about the victim's illegal drug use and threatened to expose him. The evidence of the player's steroid use is confirmed through an advanced system of statistical analysis created by a young man who plays fantasy baseball, and Charlie tries to determine how the stats are linked to the player's murder.

Math used:
sabermetrics, Shiryaev-Roberts change-point analysis

Статистика Ширяева-Робертса

- Условия для моментов остановки и параметра θ в точности совпадают с условиями для процедуры кумулятивных сумм

Статистика Ширяева-Робертса

- Условия для моментов остановки и параметра θ в точности совпадают с условиями для процедуры кумулятивных сумм
- Качество алгоритма задается величиной

$$\mathbf{C}(T) = \sup_{\theta \geqslant 0} \mathbb{E}_\theta \big((\tau - \theta)^+ \big| \tau \geqslant \theta \big) \quad \sim \quad \inf_{\tau \in \mathcal{M}_T}$$

Статистика Ширяева-Робертса

- Условия для моментов остановки и параметра θ в точности совпадают с условиями для процедуры кумулятивных сумм
- Качество алгоритма задается величиной

$$\mathbf{C}(T) = \sup_{\theta \geqslant 0} \underbrace{\mathbb{E}_\theta\big((\tau - \theta)^+ \big| \tau \geqslant \theta\big)}_{\substack{\text{(условное) среднее время} \\ \text{обнаружения разладки}}} \quad \sim \quad \inf_{\tau \in \mathcal{M}_T}$$

Статистика Ширяева-Робертса

- Условия для моментов остановки и параметра θ в точности совпадают с условиями для процедуры кумулятивных сумм
- Качество алгоритма задается величиной

$$\mathbf{C}(T) = \sup_{\theta \geqslant 0} \underbrace{\mathbb{E}_\theta\big((\tau-\theta)^+ \big| \tau \geqslant \theta\big)}_{\substack{\text{(условное) среднее время} \\ \text{обнаружения разладки}}} \quad \sim \quad \inf_{\tau \in \mathcal{M}_T}$$

$$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}_{\substack{\text{наихудшее среди всех} \\ \text{моментов } \theta \text{ появления разладки}}}$$

Статистика Ширяева-Робертса

- Вводится статистика $R_n = \sum_{\theta=1}^{n} \frac{f_\theta(X_1,\ldots,X_n)}{f_\infty(X_1,\ldots,X_n)}$

Статистика Ширяева-Робертса

- Вводится статистика $R_n = \sum\limits_{\theta=1}^{n} \frac{f_\theta(X_1,\ldots,X_n)}{f_\infty(X_1,\ldots,X_n)}$
- Если случайные величины $\xi_1, \ldots, \xi_n$ независимы, то

$$R_n = \sum_{\theta=1}^{n} \prod_{k=\theta}^{n} \frac{f_0(X_k)}{f_\infty(X_k)} = \sum_{\theta=1}^{n} \prod_{k=\theta}^{n} l_k.$$

Статистика Ширяева-Робертса

- Вводится статистика $R_n = \sum\limits_{\theta=1}^{n} \frac{f_\theta(X_1,\ldots,X_n)}{f_\infty(X_1,\ldots,X_n)}$
- Если случайные величины $\xi_1, \ldots, \xi_n$ независимы, то

$$R_n = \sum_{\theta=1}^{n} \prod_{k=\theta}^{n} \frac{f_0(X_k)}{f_\infty(X_k)} = \sum_{\theta=1}^{n} \prod_{k=\theta}^{n} l_k.$$

- Статистика R_n обладает свойством $R_n = (1 + R_{n-1})l_k$ и называется статистикой Ширяева-Робертса (Shiryaev-Roberts, SR).

Статистика Ширяева-Робертса

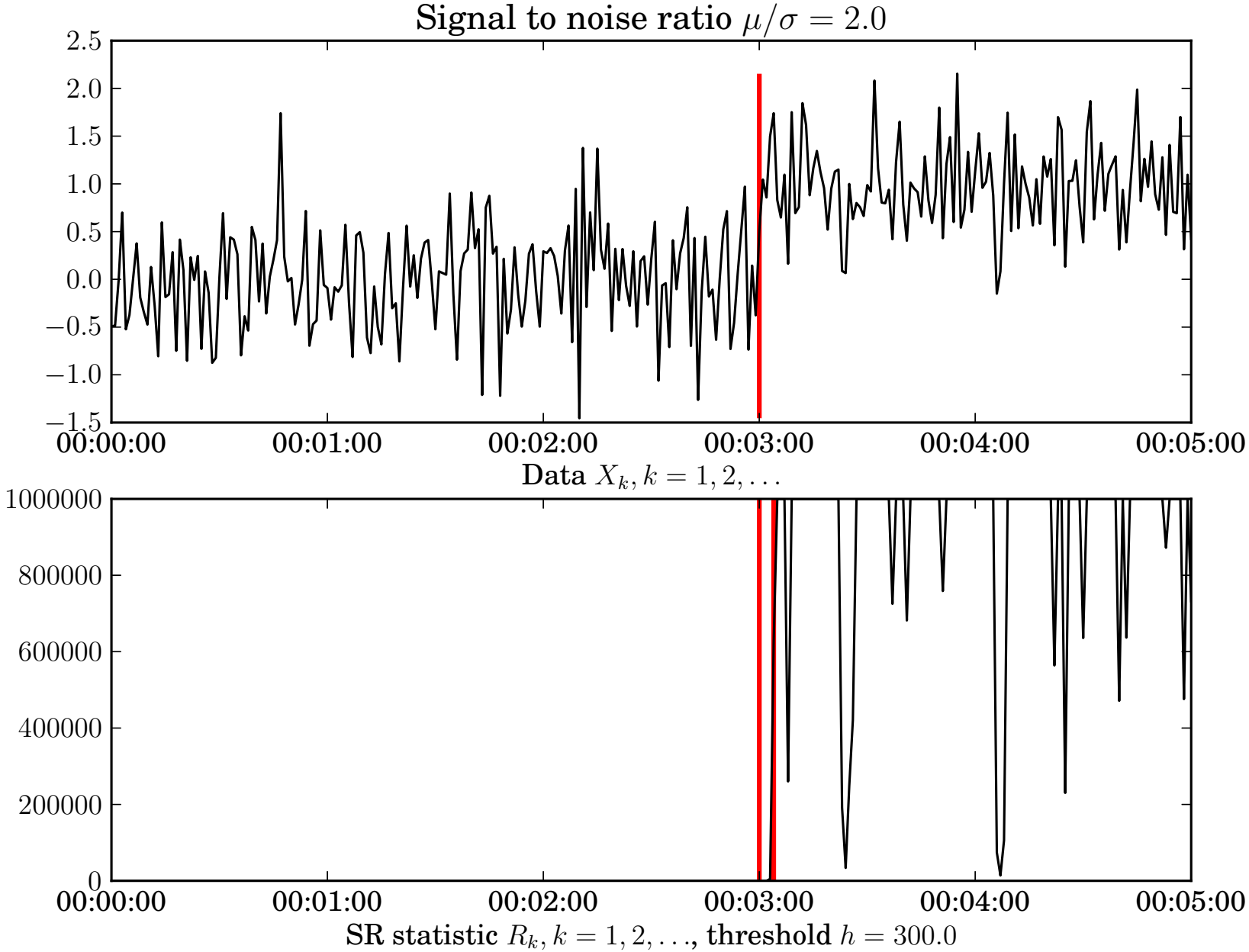
- Вводится статистика $R_n = \sum\limits_{\theta=1}^{n} \frac{f_\theta(X_1,\ldots,X_n)}{f_\infty(X_1,\ldots,X_n)}$
- Если случайные величины $\xi_1, \ldots, \xi_n$ независимы, то

$$R_n = \sum_{\theta=1}^{n} \prod_{k=\theta}^{n} \frac{f_0(X_k)}{f_\infty(X_k)} = \sum_{\theta=1}^{n} \prod_{k=\theta}^{n} l_k.$$

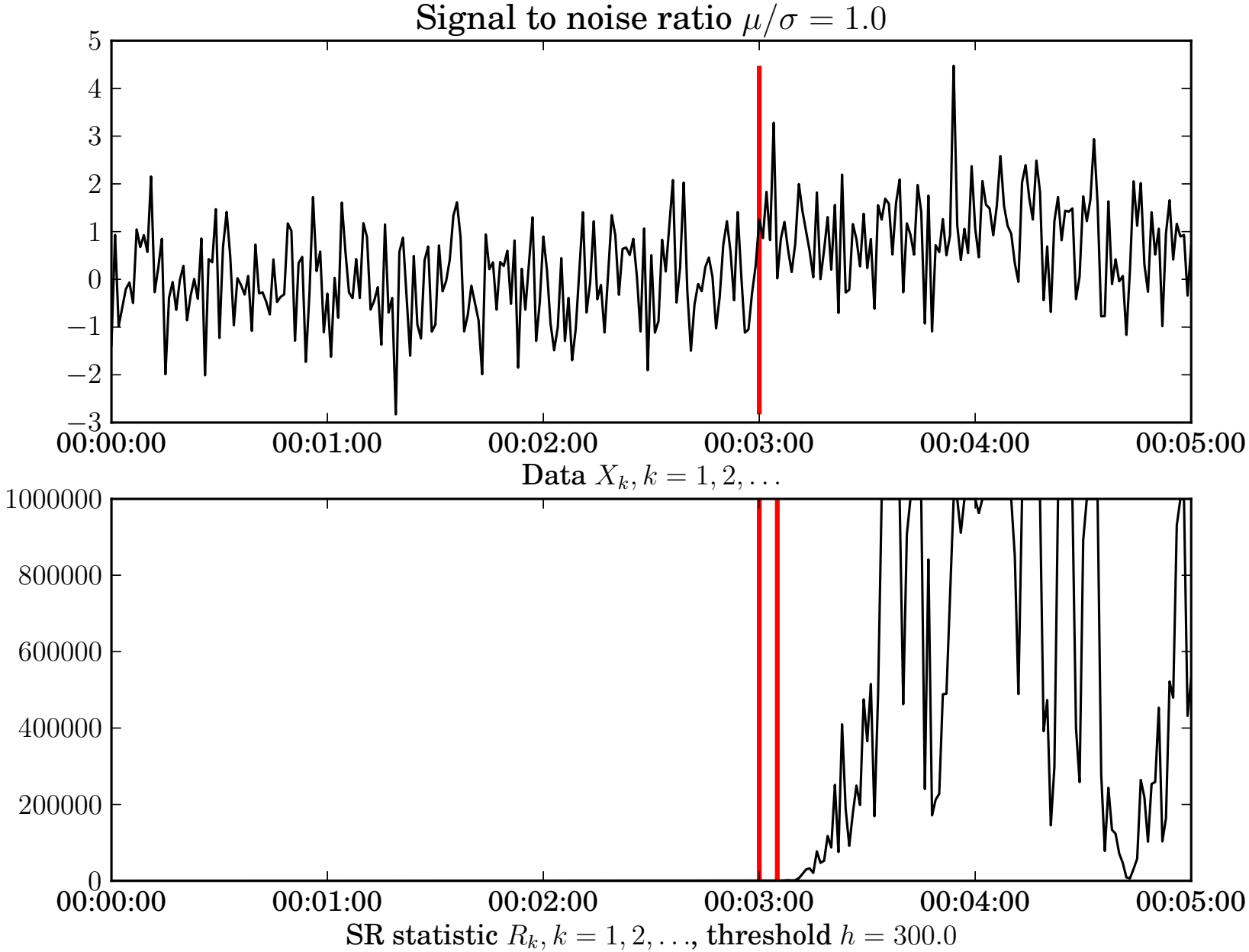
- Статистика R_n обладает свойством $R_n = (1 + R_{n-1})l_k$ и называется статистикой Ширяева-Робертса (Shiryaev-Roberts, SR).
- Остановка в момент τ_{SR} минимизирует $\mathbf{C}(T)$

$$\tau_{\mathrm{SR}} = \inf\{n \geqslant 0 : R_n \geqslant h\}$$

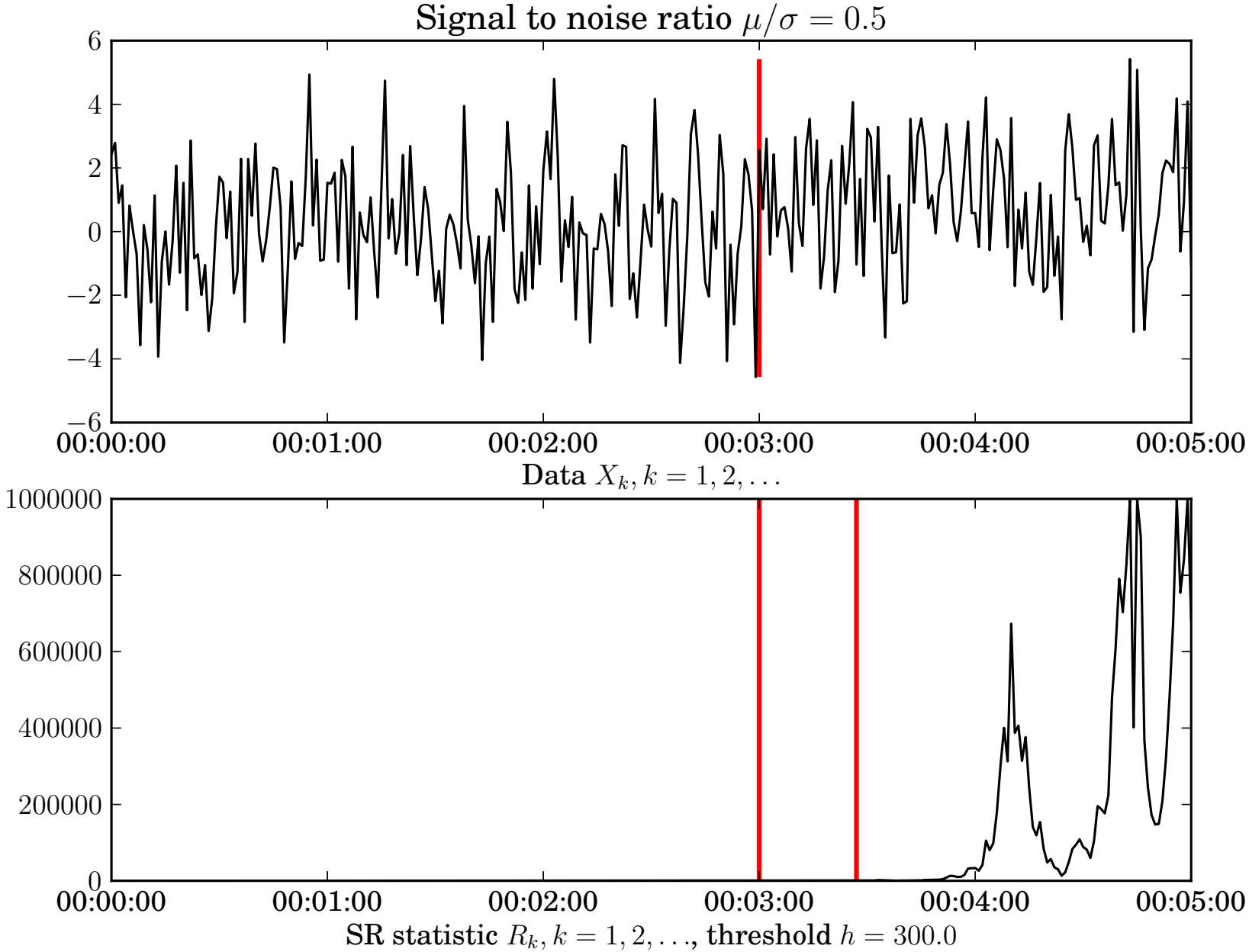
Пример 9.1 [Razladki]



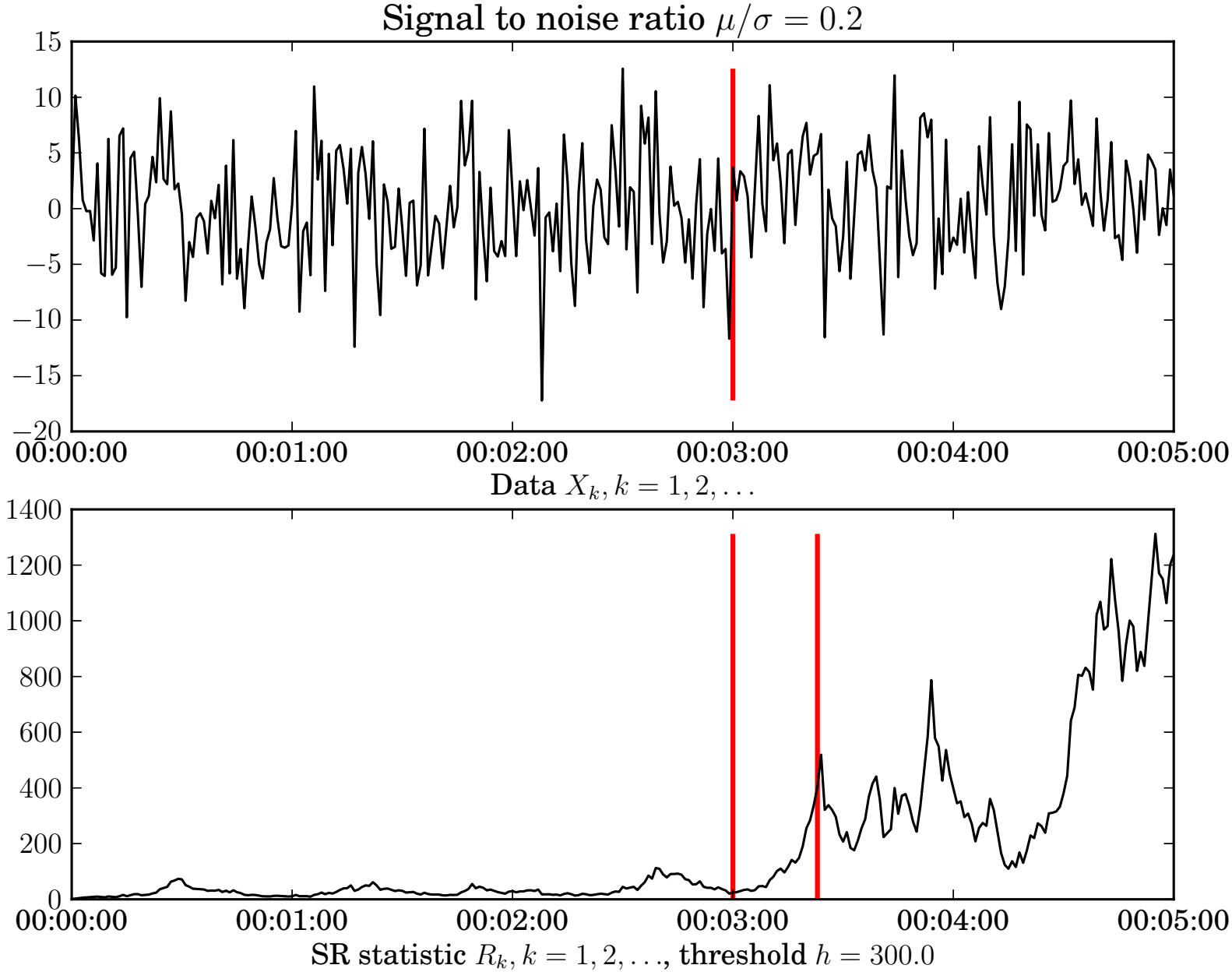
Пример 9.2 [Razladki]



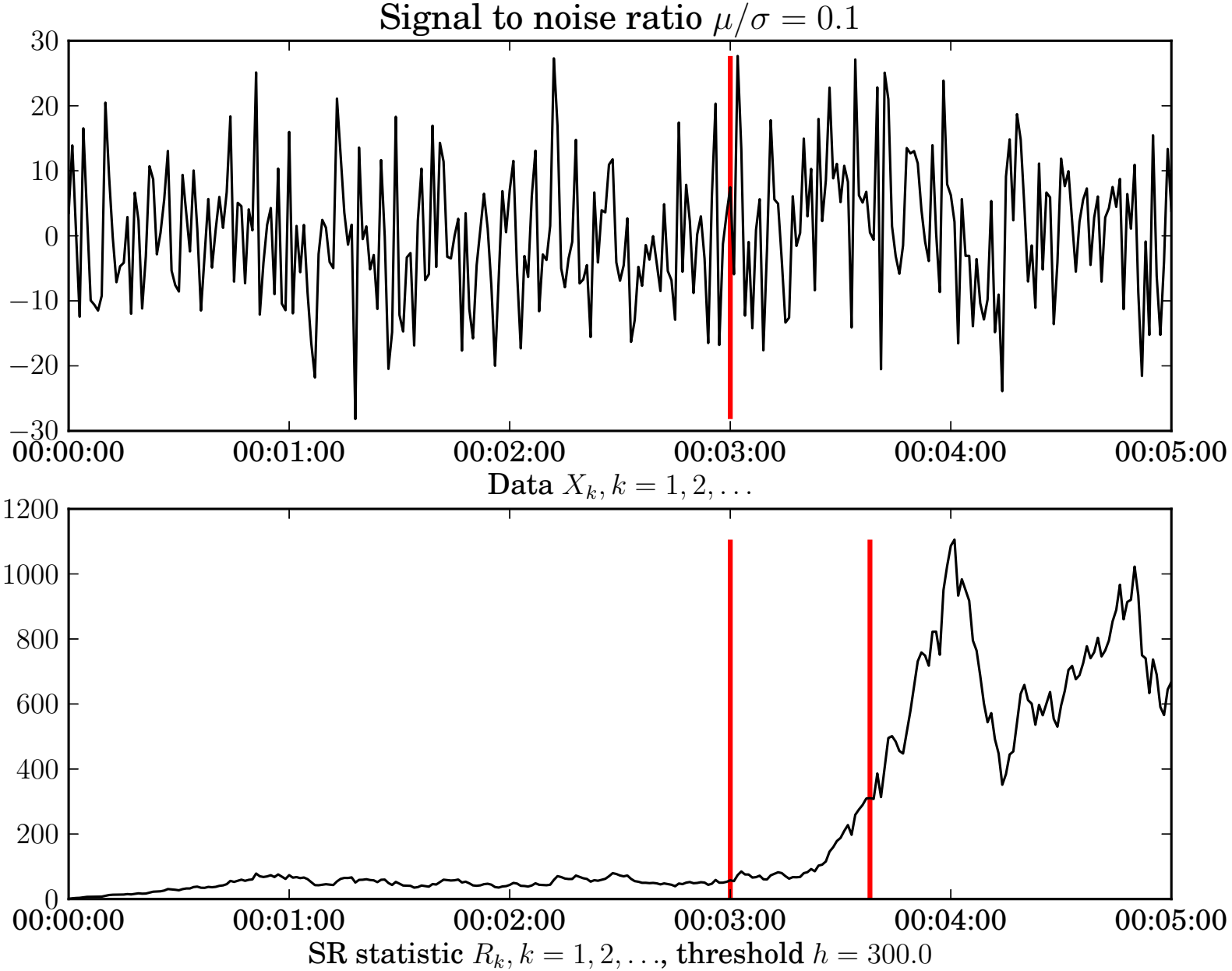
Пример 9.3 [Razladki]



Пример 9.4 [Razladki]



Пример 9.5 [Razladki]



Статистика апостериорной вероятности π_t I

- Пусть $\theta = \theta(\omega)$ — случайная величина, $\theta \perp Z_t$, имеющая (геометрическое) распределение

$$\mathrm{P}(\theta = 0) = \pi, \quad \mathrm{P}(\theta = n \big| \theta > 0) = pq^{n-1},$$

причем $\pi \in [0, 1)$ и $p \in (0, 1)$ известны, $q = 1 - p$.

Статистика апостериорной вероятности π_t II

- Качество алгоритма задается величиной

$$\mathrm{E}(\tau - \theta \big| \tau > \theta) \quad \sim \quad \inf_{\tau \in \mathcal{M}_\alpha}$$

- Этот критерий эквивалентен безусловному

$$\mathbf{A}(c) = \underbrace{\mathrm{P}(\tau < \theta)}_{\substack{\text{вероятность} \\ \text{ложной тревоги}}} + \underbrace{c\,\mathrm{E}(\tau - \theta \big| \tau > \theta)}_{\substack{\text{(условное) среднее время} \\ \text{обнаружения разладки}}} \quad \sim \quad \inf_{\tau}$$

Статистика апостериорной вероятности

- Вводится статистика

$$\pi_n = \mathrm{P}(\theta \leqslant n \Big| X_1, \ldots, X_n),$$

представимая в виде

$$\pi_n = \frac{\varphi_n}{1+\varphi_n}, \qquad \varphi_{n+1} = \frac{(p+\varphi_n)}{q} l_k$$

Статистика апостериорной вероятности

- Вводится статистика

$$\pi_n = \mathrm{P}(\theta \leqslant n \big| X_1, \ldots, X_n),$$

представимая в виде

$$\pi_n = \frac{\varphi_n}{1+\varphi_n}, \qquad \varphi_{n+1} = \frac{(p+\varphi_n)}{q} l_k$$

- Остановка в момент τ_π минимизирует $\mathbf{A}(c)$:

$$\tau_\pi = \inf\{n \geqslant 0 : \pi_n \geqslant h\}$$

Статистика апостериорной вероятности

- Вводится статистика

$$\pi_n = \mathrm{P}(\theta \leqslant n \big| X_1, \ldots, X_n),$$

представимая в виде

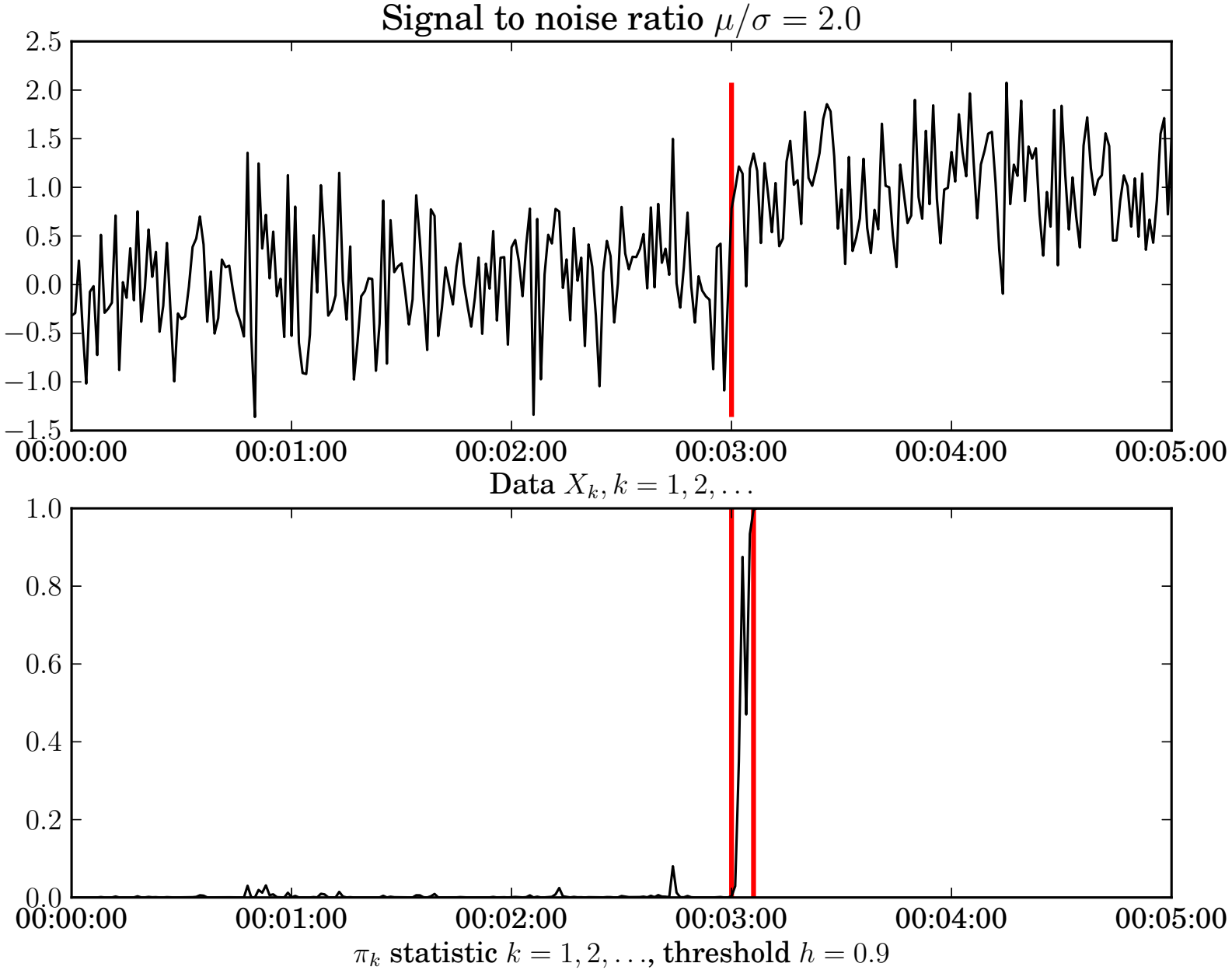
$$\pi_n = \frac{\varphi_n}{1 + \varphi_n}, \qquad \varphi_{n+1} = \frac{(p + \varphi_n)}{q} l_k$$

- Остановка в момент τ_π минимизирует $\mathbf{A}(c)$:

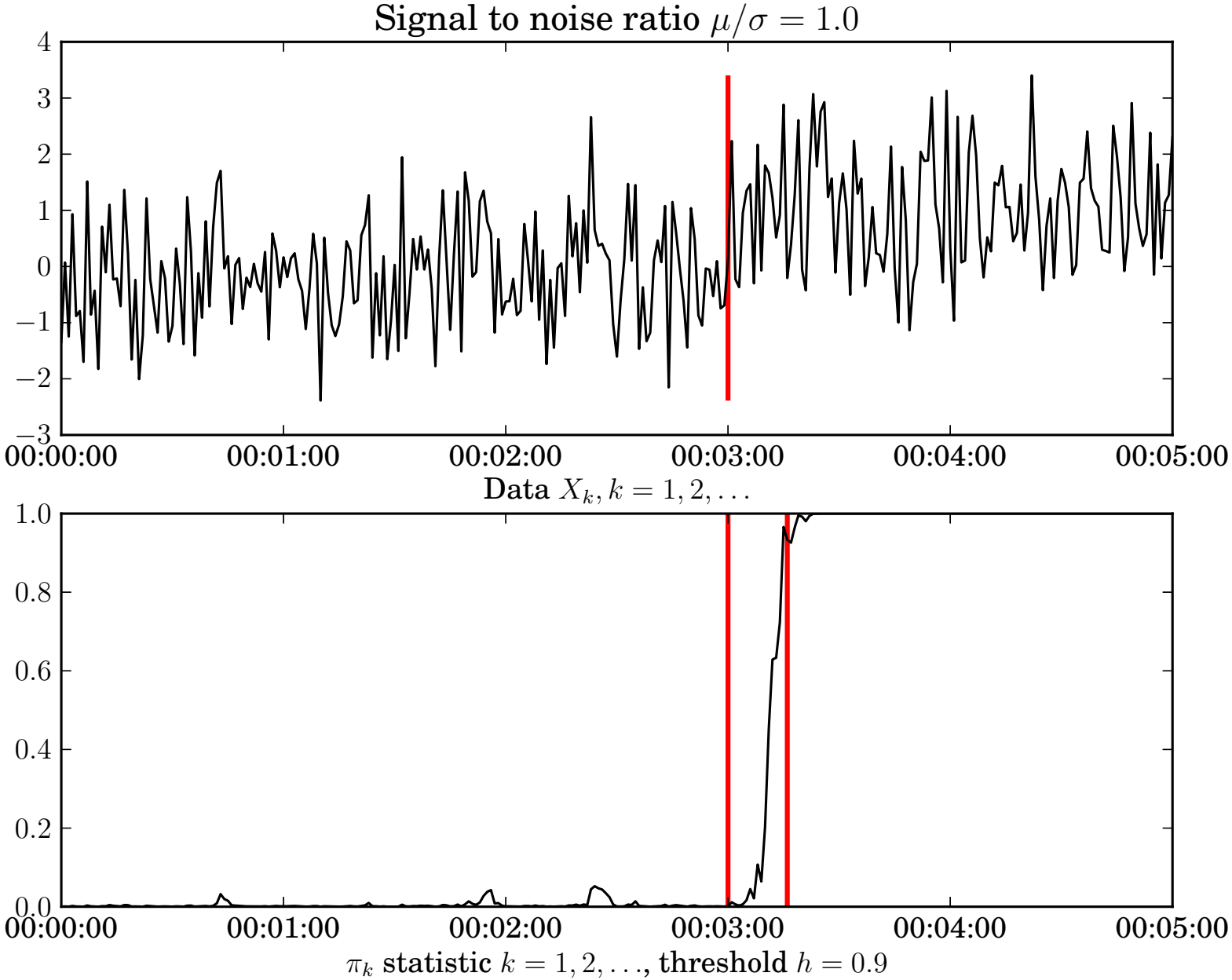
$$\tau_\pi = \inf\{n \geqslant 0 : \pi_n \geqslant h\}$$

- π_n — апостериорная вероятность появления разладки до момента времени n в предположении, что получены наблюдения $X_1, \ldots, X_n$

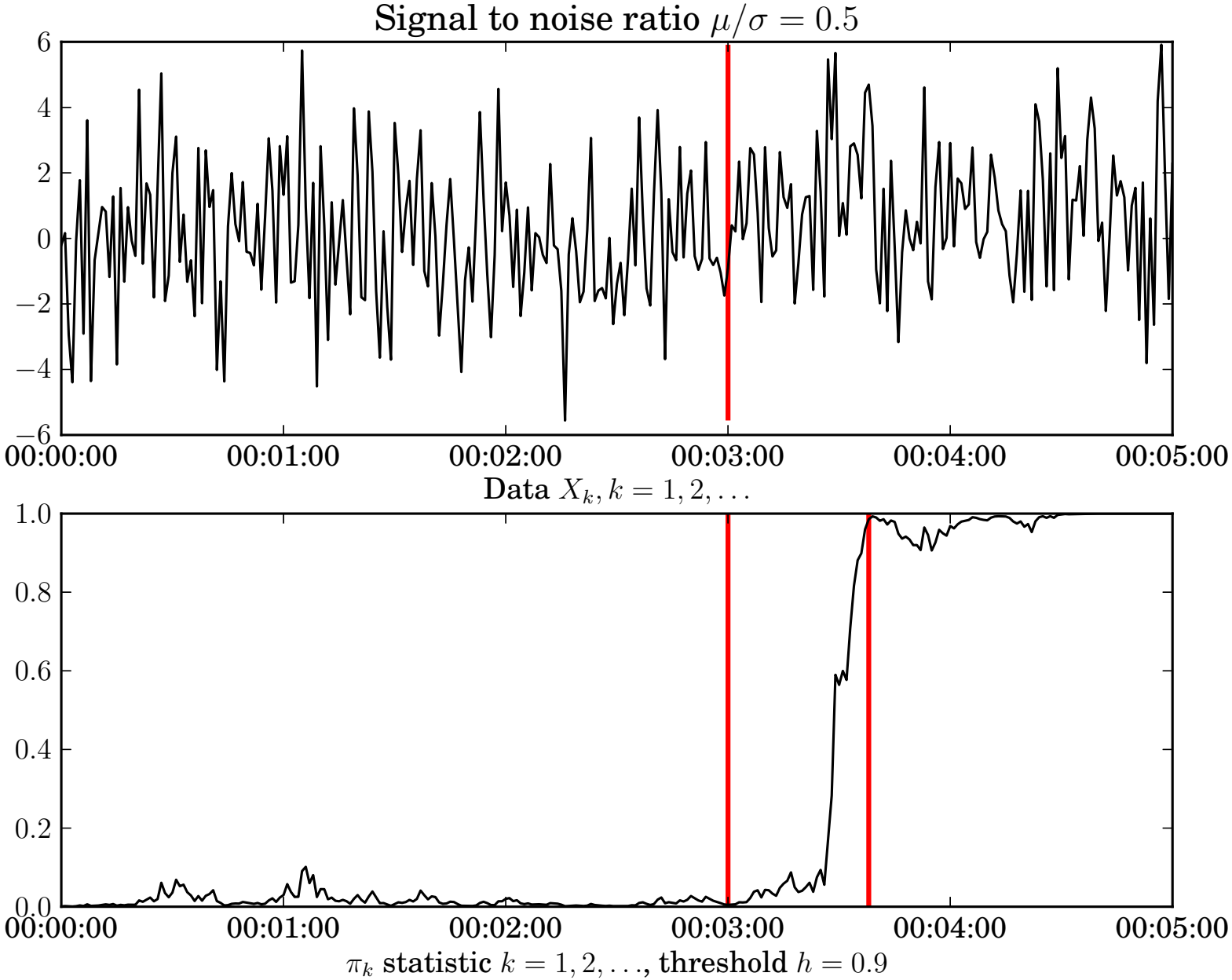
Пример 10.1 [Razladki]



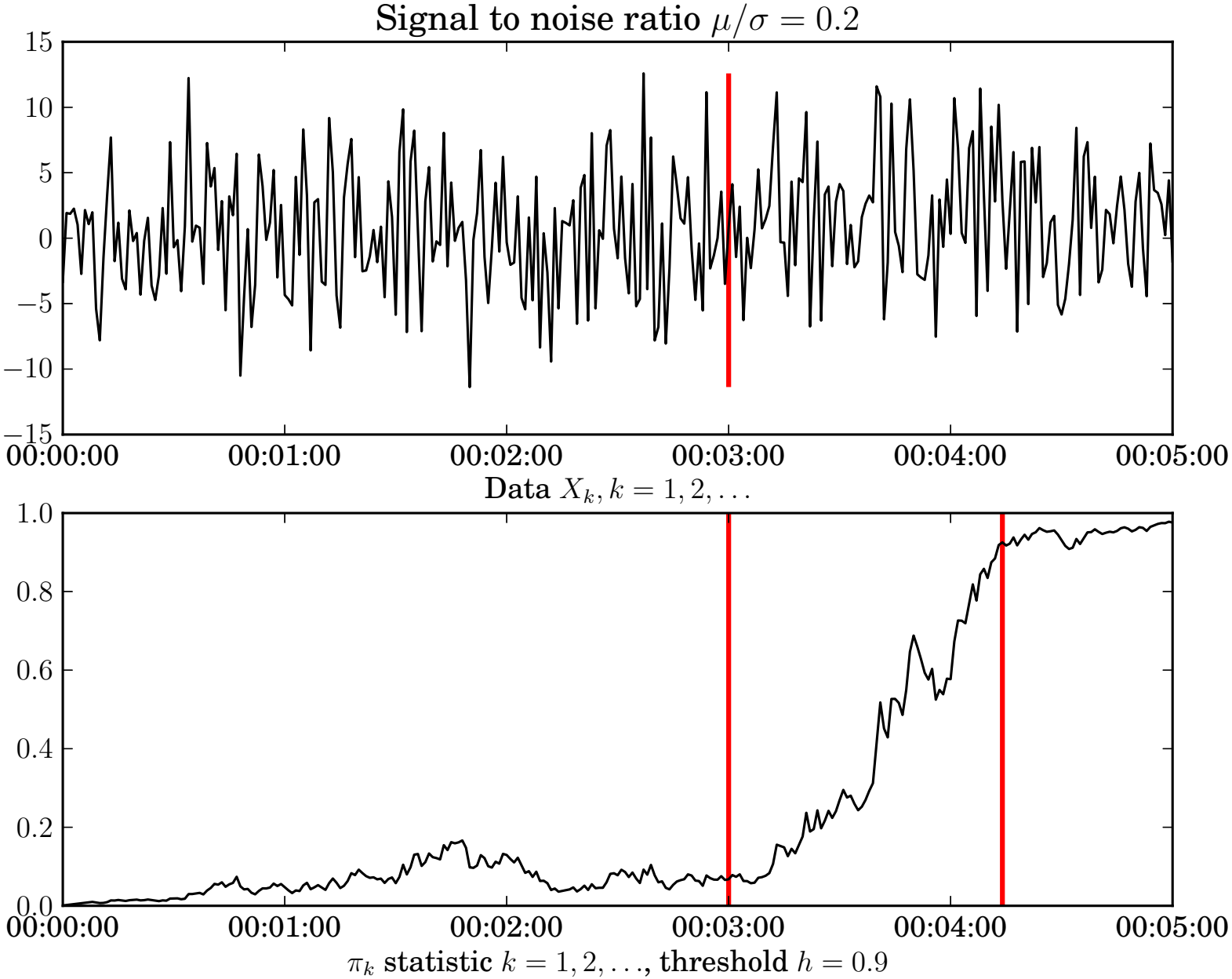
Пример 10.2 [Razladki]



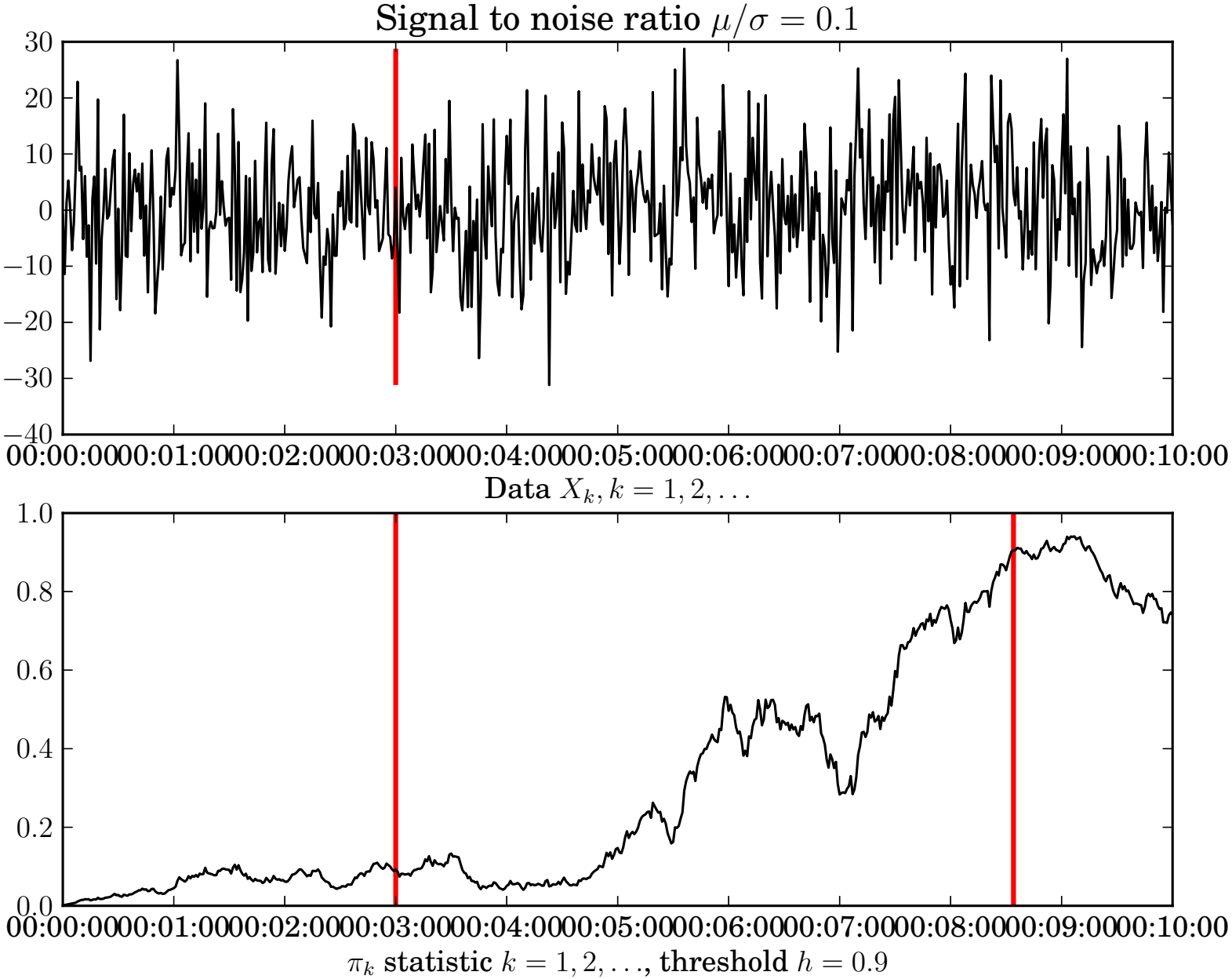
Пример 10.3 [Razladki]



Пример 10.4 [Razladki]



Пример 10.5 [Razladki]



Литература I

Adams, Ryan Prescott и David J. C. MacKay (2007). «Bayesian Online Changepoint Detection». В: *arXiv preprint arXiv:0710.3742*, с. 7. DOI: arXiv:0710.3742v1. arXiv: 0710.3742. URL: http://arxiv.org/abs/0710.3742.

Andersson, Eva, David Bock и Marianne Frisen (2002). «Department of Statistics Goteborg University Sweden with application to turns in business cycles». В:

Basseville, M и I Nikiforov (2002). «Fault isolation for diagnosis: nuisanece rejection and multiple hypothesis testing». В: *Annual Reviews in Control* 26, с. 189—202.

Basseville, Michèle и Igor V Nikiforov (1993). *Detection of abrupt changes: theory and application*. ISBN: 0-13-126780-9. DOI: 10.1016/0967-0661(94)90196-1.

Литература II

Ben-Gal, Irad, Gail Morag и Armin Shmilovici (2003). «Context-Based Statistical Process Control». В: *Technometrics* 45.4, с. 293—311. ISSN: 0040-1706. DOI: 10.1198/004017003000000122.

Casas, P. и др. (2010). «Optimal volume anomaly detection and isolation in large-scale IP networks using coarse-grained measurements». В: *Computer Networks* 54.11, с. 1750—1766. ISSN: 13891286. DOI: 10.1016/j.comnet.2010.01.013.

Cerutti S. и др. (1993). «Time variant power spectrum analysis for the detection of transient episodes in HRV signal». В: *IEEE Transactions on biomedical engineering* 40.2, с. 136—144. URL: file:///C:/Documents%20and%20Settings/avartak/Desktop/PhD%20Research/aniket%20research%20documents/HRV/Cerutti/CERUTTI%5C_power%5C_spectrum%5C_hrv%5C_ITBE%5C_1993.pdf.

Dass, Sarat C (2009). «Hierarchical Spatial Regression Models for Change Point Analysis». В: с. 3119—3133.

Литература III

Dass, Sarat C, Chae Young Lim и Tapabrata Maiti (2011). *Change Point Analysis of Cancer Mortality Rates for US States using Functional Dirichlet Processes*. Тех. отч. Technical Report RM 690, Department of Statistics и Probability, Michigan State University.

Desobry, Frédéric, Manuel Davy и Christian Doncarli (2005). «An online kernel change detection algorithm». В: *IEEE Transactions on Signal Processing* 53.8, с. 2961—2974.

Girshick, Meyer A and Rubin, Herman (1952). «A Bayes approach to a quality control model». В: *The Annals of Mathematical Statistics*, с. 114—125. ISSN: 0003-4851. DOI: 10.1214/aoms/1177733256.

Guépié, Blaise Kévin, Lionel Fillatre и Igor Nikiforov (2012). «Sequential Detection of Transient Changes». В: *Sequential Analysis* 31.4, с. 528—547. ISSN: 0747-4946. DOI: 10.1080/07474946.2012.719443.

Kalman, Rudolph Emil (1960). «A new approach to linear filtering and prediction problems». В: *Journal of Fluids Engineering* 82.1, с. 35—45.

Литература IV

Kim, Hongjoong, Boris L Rozovskii и Alexander G Tartakovsky (2004). «A Nonparametric Multichart CUSUM Test for Rapid Detection of DOS Attacks in Computer Networks». В: 2.3, с. 149—158.

Lakhina, Anukool, Mark Crovella и Christiphe Diot (2004). «Characterization of network-wide anomalies in traffic flows». В: *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement - IMC '04* 6, с. 201. ISSN: 00283940. DOI: 10.1145/1028788.1028813. URL: http://portal.acm.org/citation.cfm?doid=1028788.1028813.

Lakhina, Anukool, Mark Crovella и Christophe Diot (2004). «Diagnosing network-wide traffic anomalies». В: *ACM SIGCOMM Computer Communication Review* 34.4, с. 219. ISSN: 01464833. DOI: 10.1145/1030194.1015492.

MacNeill, I B и Y Mao (1995). «Change-point analysis for mortality and morbidity rate». В: *Applied Change Point Problems in Statistics*, с. 37—55.

Литература V

Malladi, D.P. и J.L. Speyer (1999). «A generalized Shiryayev sequential probability ratio test for change detection and isolation». В: *Automatic Control, IEEE Transactions ...* 44.8, с. 1522—1534. URL: http://ieeexplore.ieee.org/xpls/abs%5C_all.jsp?arnumber=780416.

Page, Es (1954). «Continuous inspection schemes». В: *Biometrika* 41.1, с. 100—115. ISSN: 00063444. DOI: 10.2307/2333009. URL: http://drsmorey.org/bibtex/upload/Neyman:Pearson:1928.pdf$%5Cbackslash$nhttp://www.jstor.org/stable/2333009.

Petzold, Max и др. (2004). «Surveillance in longitudinal models: Detection of intrauterine growth restriction». В: *Biometrics* 60.4, с. 1025—1033. ISSN: 0006341X. DOI: 10.1111/j.0006-341X.2004.00258.x.

Pham, Duc-Son и др. (2014). «Anomaly detection in large-scale data stream networks». В: *Data Mining and Knowledge Discovery* 28.1, с. 145—189.

Roberts, SW (1959). «Control chart tests based on geometric moving averages». В: *Technometrics* 1.3, с. 239—250.

Литература VI

Roberts, SW (1966). «A comparison of some control chart procedures». В: *Technometrics* 8.3, с. 411—430.

Shewhart, Walter Andrew (1931). *Economic control of quality of manufactured product*.

Streit, Roy L. и Peter K. Willett (1999). «Detection of random transient signals via hyperparameter estimation». В: *IEEE Transactions on Signal Processing* 47.7, с. 1823—1834. ISSN: 1053587X. DOI: 10.1109/78.771032.

Tartakovsky, Ag (2013). «Efficient computer network anomaly detection by changepoint detection methods». В: *IEEE Journal of Selected Topics in Signal Processing* 7.1, с. 7—11. DOI: 10.1109/JSTSP.2012.2233713. arXiv: arXiv:1212.1829. URL: http://ieeexplore.ieee.org/xpls/abs%5C_all.jsp?arnumber=6380529.

Tartakovsky, Alexander G (2003). «Quickest Change Detection in Distributed Sensor Systems». В: *Proceedings of the 6th International Conference on Information Fusion*, с. 756—763.

Литература VII

Tartakovsky, Alexander G. и др. (2006). «A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods». В: *IEEE Transactions on Signal Processing* 54.9, с. 3372—3381. ISSN: 1053587X. DOI: 10.1109/TSP.2006.879308.

Taweab, Fauzia, Noor Akma Ibrahim и Jayanthi Arasan (2015). «A Bounded Cumulative Hazard Model with A change- Point According to a Threshold in a covariate for Right-Censored Data». В: 74.1, с. 69—74.

Willsky, As (1976). «A survey of design methods for failure detection in dynamic systems». В: *Automatica* 12, с. 601—611. ISSN: 00051098. DOI: 10.1016/0005-1098(76)90041-8. URL: http://www.sciencedirect.com/science/article/pii/0005109876900418.

Дарховский, Борис Семенович (2013). «Обнаружение разладки случайной последовательности при минимальной априорной информации». В: *Теория вероятностей и ее применения* 58.3, с. 585—590.

Литература VIII

Ширяев, Альберт Николаевич (1961). «Задача скорейшего обнаружения нарушения стационарного режима». В: *Докл. АН СССР*. Т. 138. 5, с. 1039—1042.

– (1963). «Об оптимальных методах в задачах скорейшего обнаружения». В: *Теория вероятностей и ее применения* 8.1, с. 26—51.

Ширяев, АН (1961). «Обнаружение спонтанно возникающих эффектов». В: *Докл. АН СССР*. Т. 138. 4, с. 799—801.