Обзор некоторых статистических методов скорейшего обнаружения

Алексей Артемов

8 мая 2014

Общая постановка задачи (дискретное время)

- Предполагается, что наблюдаемые данные описываются числовой последовательностью

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

— это результаты наблюдений над случайными величинами

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

- Распределение процесса $\xi = (\xi_t)_{t\geqslant 0)}$:
  - до момента θ — P_∞
  - после момента θ — P_0
- Момент разладки θ заранее неизвестен

Общая постановка задачи (дискретное время)

- Варианты постановок задач:
  - offline проверка гипотезы о наличии изменения
  - offline обнаружение изменения (сегментация)
  - online обнаружение изменения
    - finite horizon: $0 \leqslant t \leqslant T < \infty$
    - infinite horizon: $0 \leqslant t \leqslant \infty$
- Момент разладки θ может быть:
  - случайной величиной с некоторым распределением
  - просто неизвестным параметром
- Существует множество различных критериев оптимальности методов скорейшего обнаружения (= постановок задач)

Общая постановка задачи (дискретное время)

- Можно использовать свойства наблюдений
  - зависимость и независимость
  - скалярность и векторность
  - распределение
- Бывает необходимо учитывать, что
  - разладка может появляться более одного раза в течение наблюдений
  - параметры процесса после разладки, как правило, неизвестны
- ...и всё это только для стационарных процессов

Из этого доклада вы не узнаете

Ввиду объема материала, сложности изложения, релевантности поставленным задачам, а также личного опыта докладчика опущены:

- случай finite horizon
- сегментация временного ряда освещается ограниченно
- оценивание значения временного ряда
- случай зависимых наблюдений
- векторный случай

Понятие разладки

- Свойства данных всюду предполагаются либо постоянными, либо медленно меняющимися
- Многие адаптивные алгоритмы оценивания хорошо работают для медленных изменений
- Разладка — это **быстрое** по сравнению с шагом дискретизации изменение

Общие обозначения

- $\theta \in \{0, 1, \ldots, \infty\}$ — момент разладки
- P_θ — распределение процесса ξ в предположении, что разладка происходит в момент θ
- P_0 — распределение процесса ξ в предположении, что разладка произошла с самого начала $(\theta = 0)$
- P_∞ — распределение процесса ξ в предположении, что разладка не происходит никогда $(\theta = \infty)$
- $f_\theta(\cdot), f_0(\cdot), f_\infty(\cdot)$ — плотности мер $\mathrm{P}_\theta, \mathrm{P}_0, \mathrm{P}_\infty$
- $\mathbf{E}_\theta, \mathbf{E}_0, \mathbf{E}_\infty$ — средние по мерам $\mathrm{P}_\theta, \mathrm{P}_0, \mathrm{P}_\infty$

Метод Неймана-Пирсона

- Различение двух гипотез по фиксированному числу наблюдений
- Требуется принять одну из гипотез $H_0(\theta = 0)$ или $H_\infty(\theta = \infty)$
- Решающее правило $d = d(X_1, \ldots, X_n)$ характеризуется вероятностями ошибок I и II рода $\alpha(d) = \mathrm{P}_\infty(H_0)$, $\beta(d) = \mathrm{P}_0(H_\infty)$,
- Возможные постановки задач:
  - $\alpha(d) + \beta(d) \sim \inf\limits_d$
  - $\beta(d) \sim \inf\limits_{d \in D_\alpha}$, $\quad D_\alpha = \{d : \alpha(d) \leqslant \alpha\}$
- Статистика — отношение правдоподобия

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

Метод Неймана-Пирсона

- Если $L_n > h_\alpha$, то принимается H_0. Иначе — H_∞.
- Удобно также использовать статистику $Z_n = \log L_n$
- **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

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

Оптимальное решающее правило:

$$d(X_1, \ldots, X_n) = \begin{cases} H_0 & \text{если } Z_n \geqslant h \\ H_\infty & \text{если } Z_n < h. \end{cases}$$

- Величина h выбирается исходя из заданных величин ошибок I и II рода

Контрольные карты Шухарта

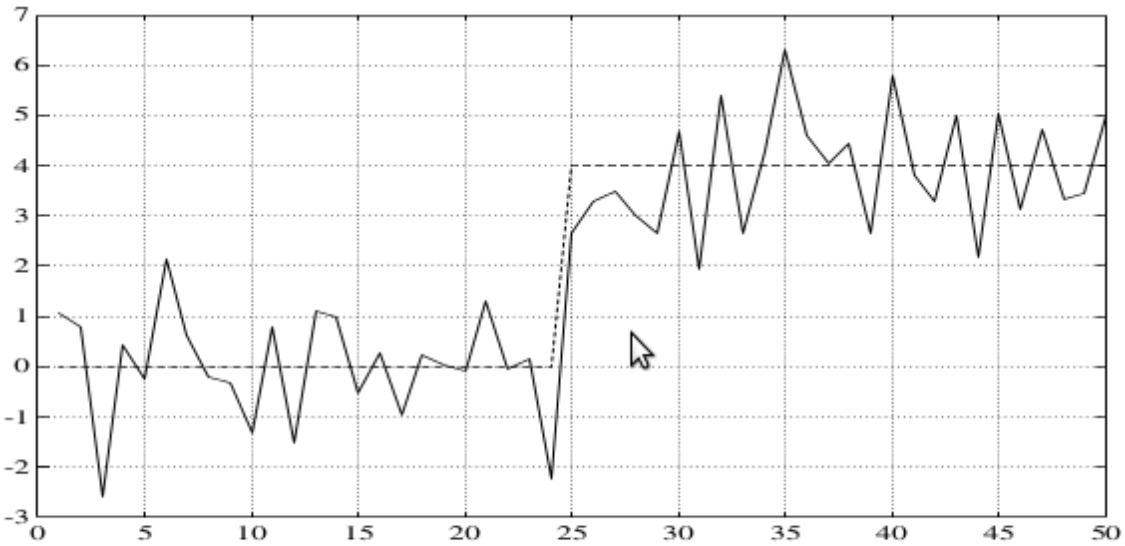
- Наблюдения $X_1, X_2, \ldots$ разбиваются на группы размера N
- Для каждой группы $X_K = (X_{(K-1)N+1}, \ldots X_{KN})$, $K = 1, 2, \ldots$ ставится задача различения двух гипотез:

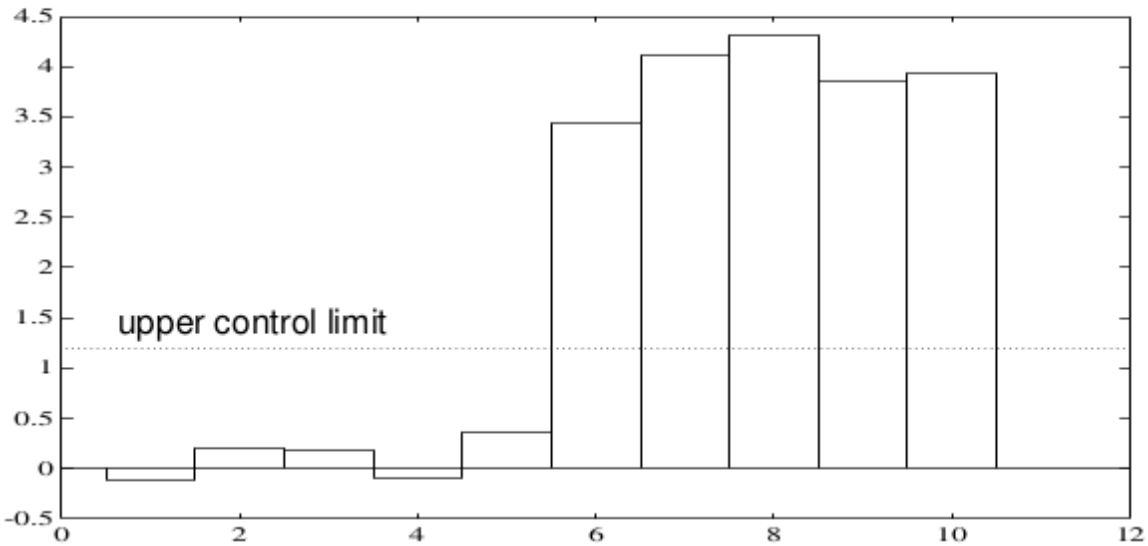
$$H_0 : \theta \leqslant (j-1)N + 1 \quad \text{и} \quad H_\infty : \theta \leqslant (j-1)N + 1$$

- Решающее правило $d_K = d(X_K)$ следует из леммы Неймана-Пирсона
- Момент остановки — первый момент принятия гипотезы H_0

$$\tau = N \cdot \min\{K : d_K = H_0\}$$

Пример [Basseville et al., 1993]





Последовательное различение гипотез

- Последовательный тест — пара (τ, φ):
  - $\tau = \tau(X) \in \{0, 1, \ldots, \infty\}$ — момент остановки относительно $\{\mathcal{F}_n, n \geqslant 1\}$
  - $\varphi = \varphi(X) \in [0, 1]$ — решающая функция
- Останавливаем наблюдения в момент τ и принимаем гипотезу H_0 с вероятностью $\varphi(X_1, \ldots, X_\tau)$
- Характеристики — средние длительности наблюдений $\mathbf{E}_0\tau$ и $\mathbf{E}_\infty\tau$, а также вероятности ошибок I и II рода

$$\alpha(\varphi) = \mathbf{E}_\infty\varphi \qquad \text{и} \qquad \beta(\varphi) = \mathbf{E}_0(1 - \varphi)$$

- Пусть заданы два числа α и β и класс тестов

$$\Delta(\alpha, \beta) = \left\{\delta = (\tau, \varphi) : \alpha(\varphi) \leqslant \alpha, \beta(\varphi) \leqslant \beta, \mathbf{E}_0\tau \leqslant 0, \mathbf{E}_\infty\tau \leqslant \infty\right\}$$

Критерий оптимальности теста $\delta^* = (\tau^*, \varphi^*) \in \Delta(\alpha, \beta)$:

$$\forall(\tau, \varphi) \in \Delta(\alpha, \beta) \quad \mathbf{E}_0\tau^* \leqslant \mathbf{E}_0\tau, \quad \mathbf{E}_\infty\tau^* \leqslant \mathbf{E}_\infty\tau$$

Метод Вальда

- Выбираются две константы $A < 0$ и $B > 0$ и момент

$$\tau_{A,B} = \inf\{n \geqslant 1 : Z_n \notin (A, B)\}$$

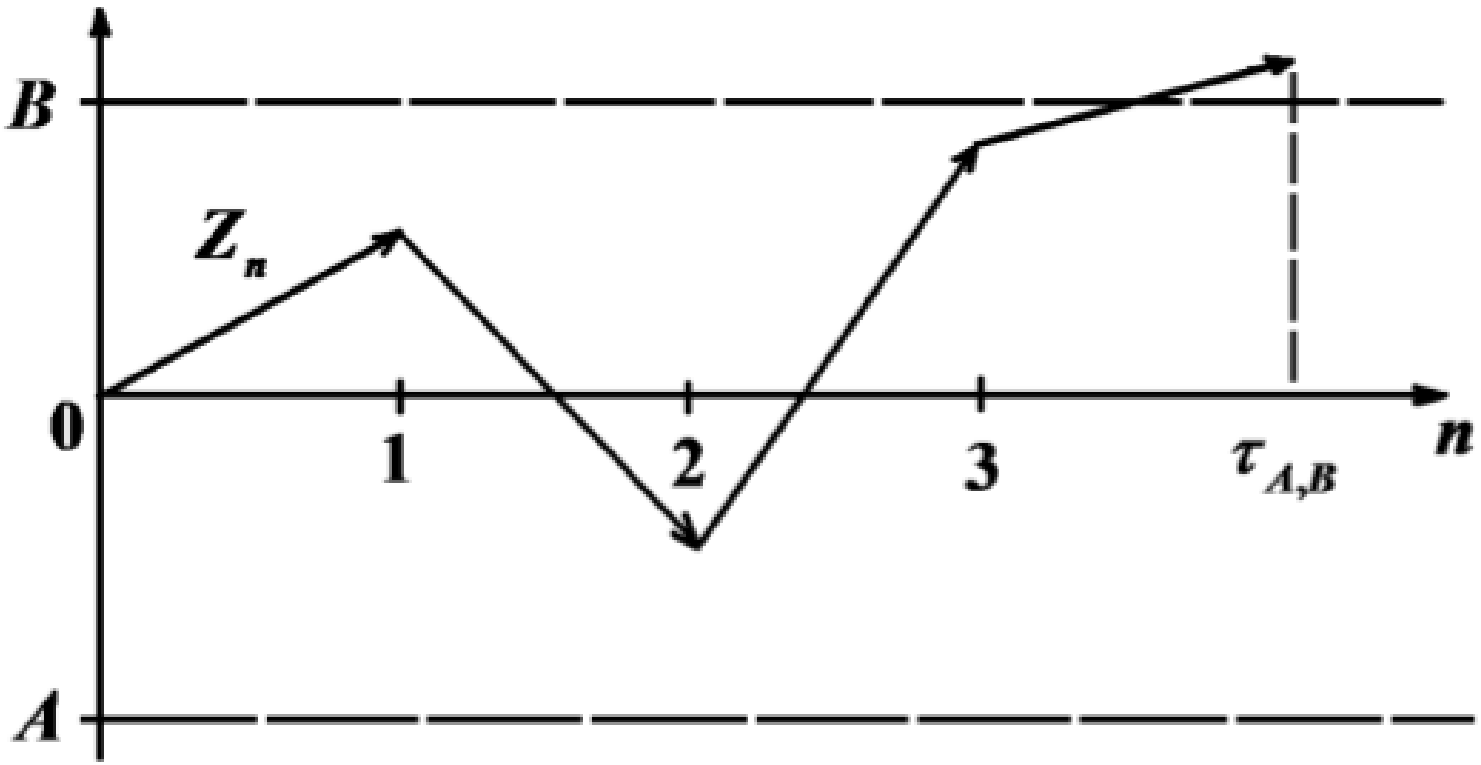
- Решающее правило

$$\varphi_{A,B} = \begin{cases} 1 & \text{если } Z_{\tau_{A,B}} \geqslant B, \\ 0 & \text{если } Z_{\tau_{A,B}} \leqslant A. \end{cases}$$

- Пороги A, B связаны с заданными вероятностями ошибок I и II рода α, β соотношениями

$$A = \log \frac{\beta}{1-\alpha}, \qquad B = \log \frac{1-\beta}{\alpha}$$

Пример [Ширяев, 2011]



Геометрическое среднее (EWMA)

- Подходит для обнаружения изменения среднего значения или дисперсии временного ряда
- Идея в рекурсивной оценке среднего вида

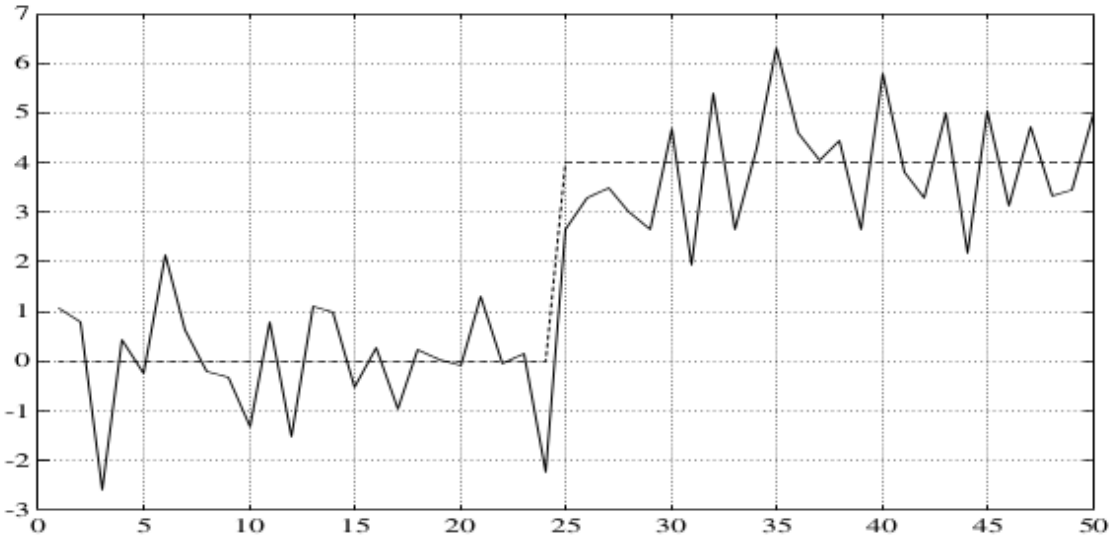
$$\hat{m}_k = (1-\lambda)\hat{m}_{k-1} + \lambda X_k, \qquad k = 1, 2, \ldots,$$

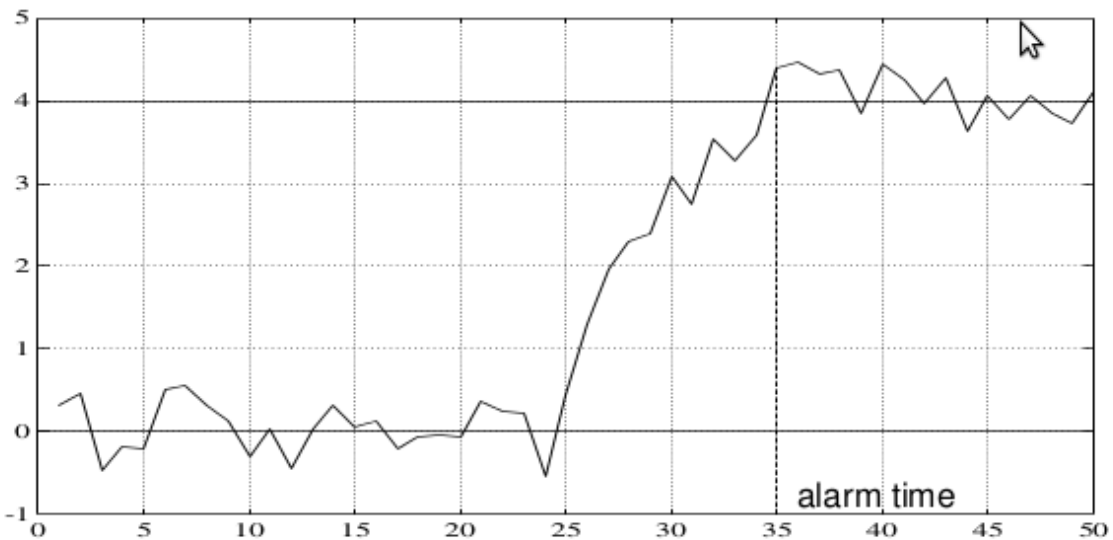
где λ («вес» новых данных) — параметр алгоритма.

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

$$\tau = \min\{k \geqslant 1 : \hat{m}_k \geqslant h\}$$

Пример [Basseville et al., 1993]





Оптимальные алгоритмы

- Рассмотрим некоторые постановки задач скорейшего обнаружения в режиме online
- Обычно ставятся для детектирования появления сноса у броуновского движения (в непрерывном времени):

$$\xi_t = \begin{cases} W_t & \text{если } 0 \leqslant t < \theta \\ \mu t + W_t & \text{если } t \geqslant \theta. \end{cases}$$

- Для дискретного времени возможна постановка:

$$\xi_t = \begin{cases} Z_t & \text{если } 0 \leqslant t < \theta \\ \mu + Z_t & \text{если } t \geqslant \theta, \end{cases}$$

где $Z_t, t = 1, 2, \ldots$ — нормальные i.i.d.r.v.

Кумулятивные суммы

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

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

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

- Качество алгоритма задается величиной

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

Кумулятивные суммы

- Вводятся статистики

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

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

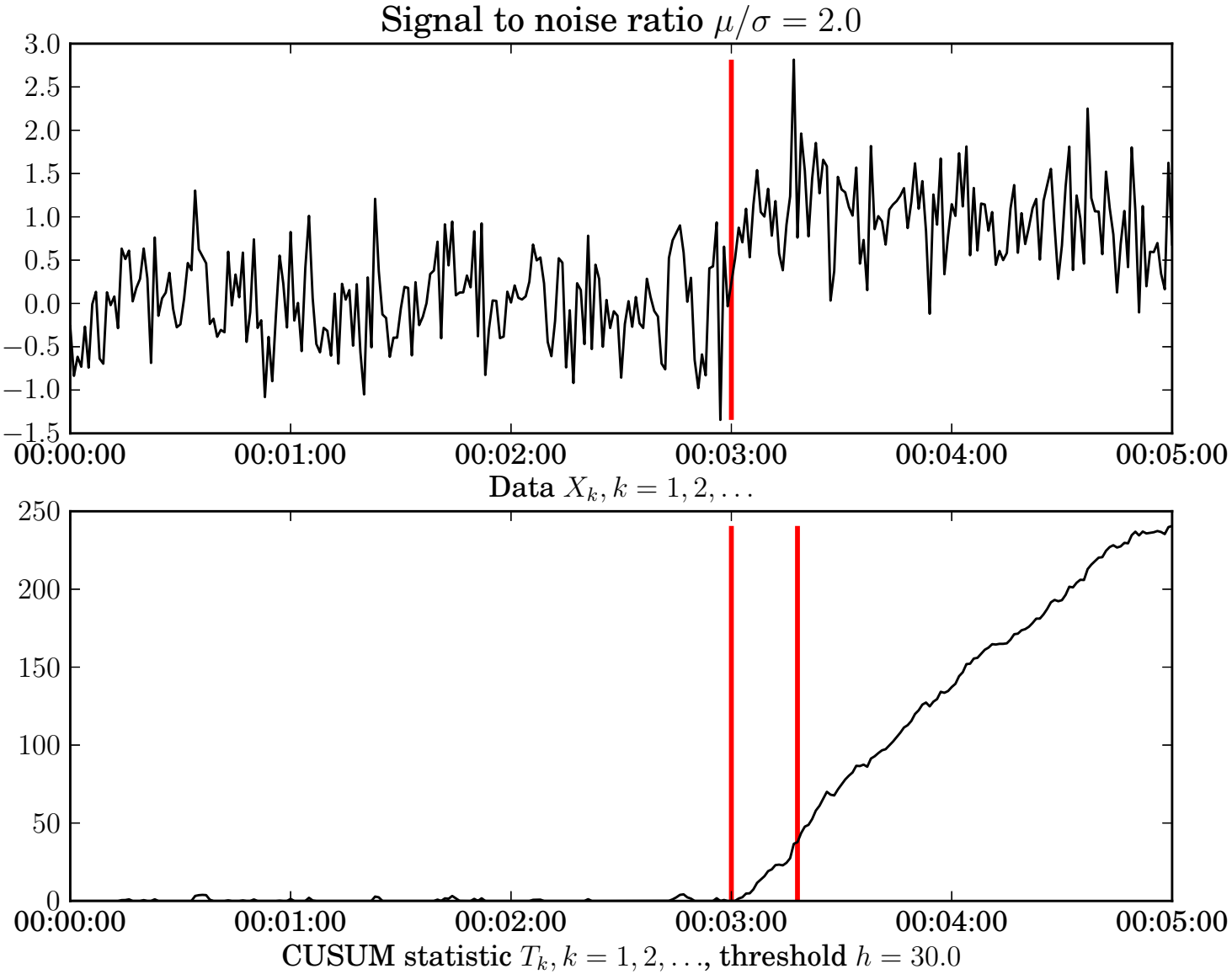
$$\gamma_n = \max\left\{1, \max_{1 \leqslant \theta \leqslant n} \prod_{k=\theta}^{n} \frac{f_0(X_k)}{f_\infty(X_k)}\right\},$$

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

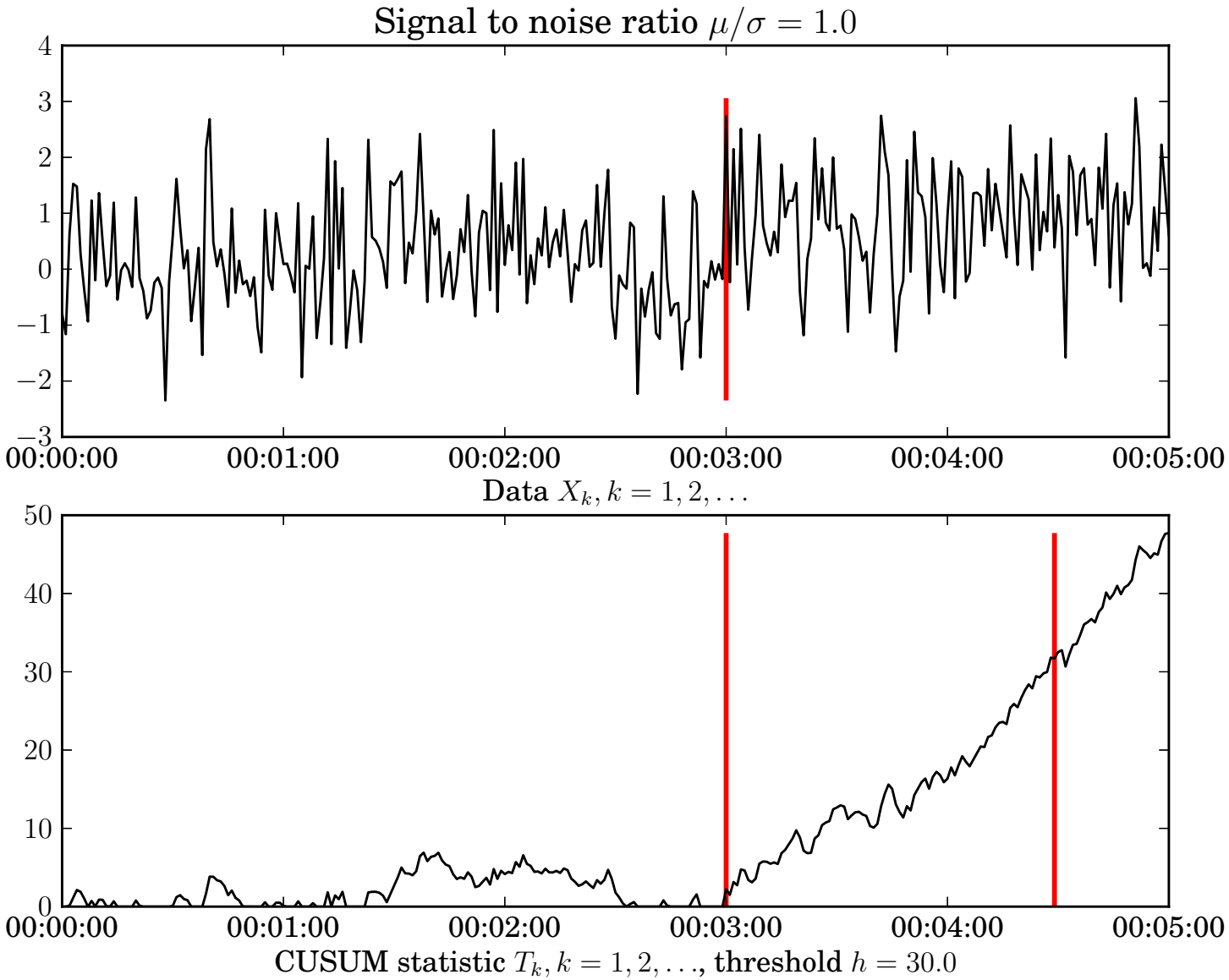
- Статистика T_n обладает свойством $T_n = \max(0, T_{n-1} + \zeta_n)$ и называется статистикой кумулятивных сумм (CUmulative SUMs, CUSUM).
- Остановка в момент τ_{CUSUM} минимизирует величину $\mathbf{D}(T)$

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

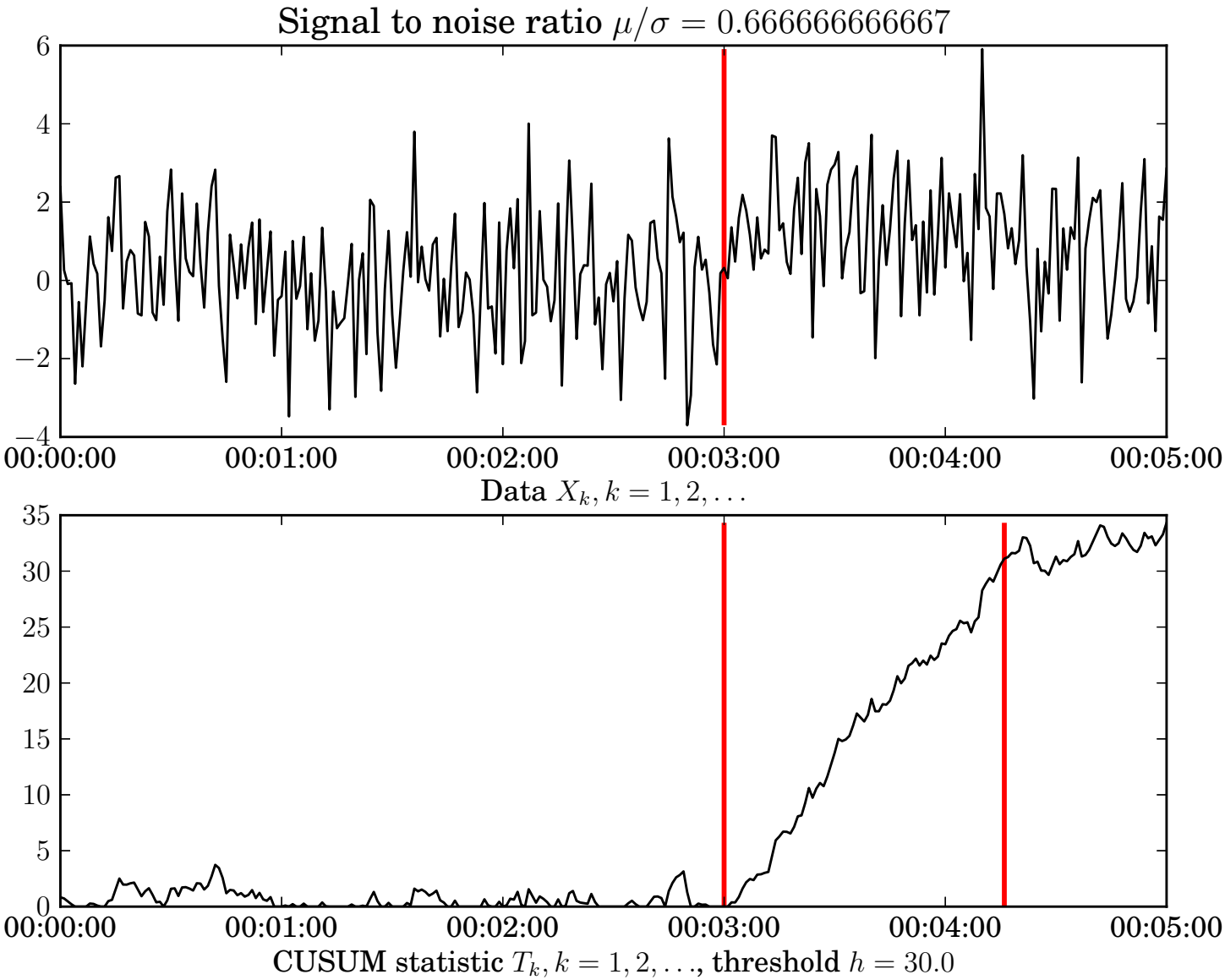
Пример 1 [Razladki]



Пример 2 [Razladki]



Пример 3 [Razladki]



Статистика Ширяева-Робертса

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

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

Статистика Ширяева-Робертса

- Вводится статистика

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

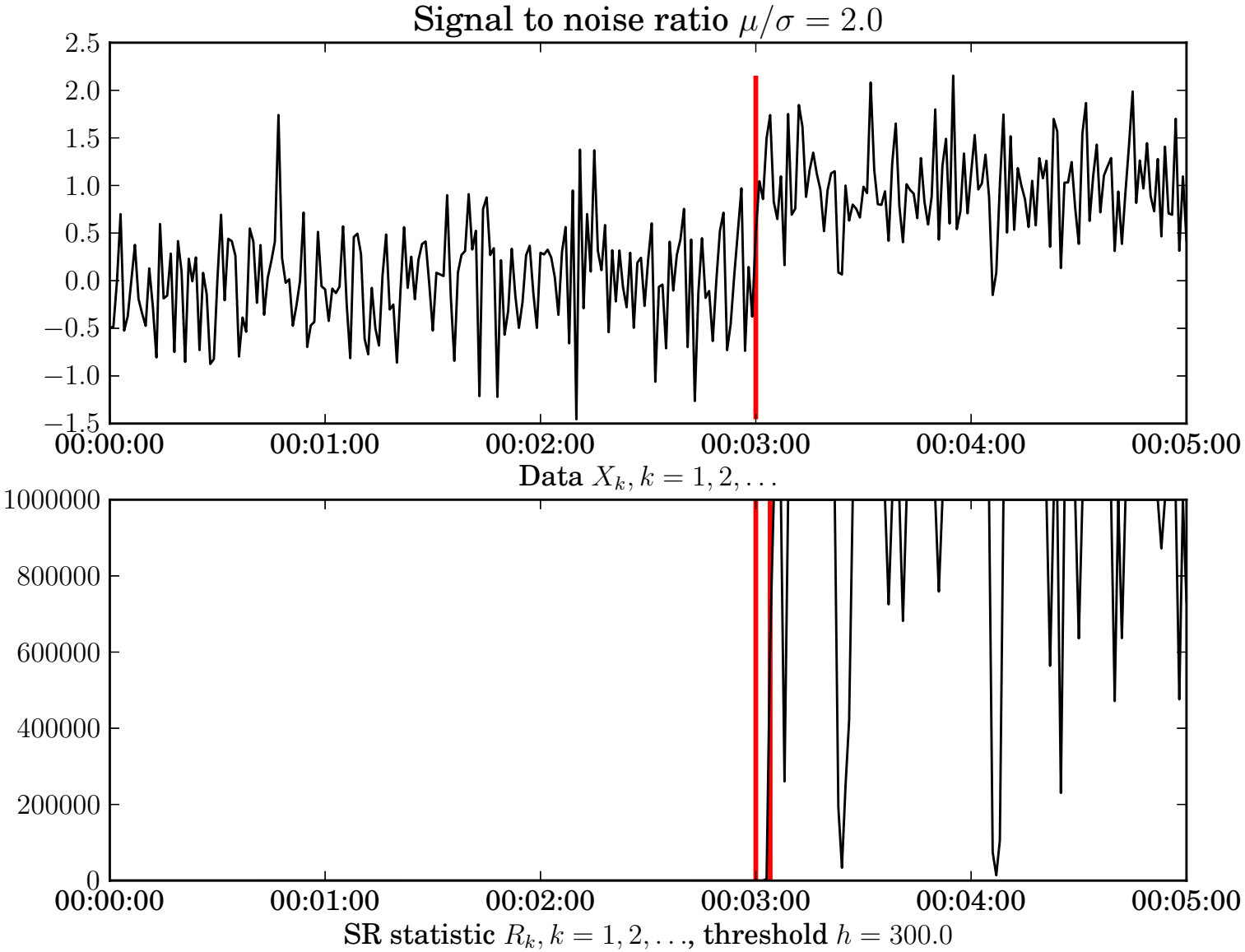
- Если случайные величины $\xi_1, \ldots, \xi_n$ независимы, то

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

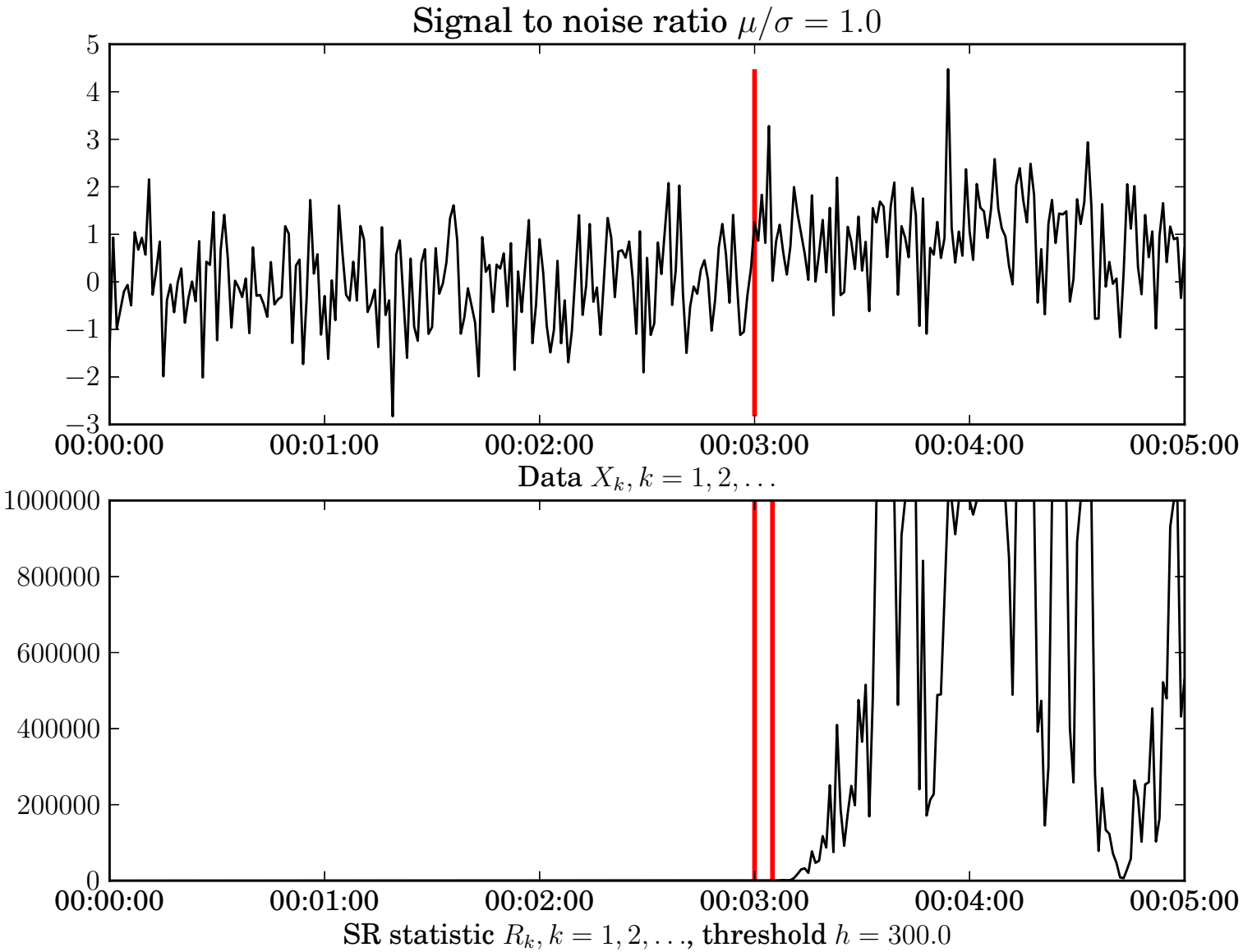
- Статистика R_n обладает свойством $R_n = (1 + R_{n-1})l_k$ и называется статистикой Ширяева-Робертса (Shiryaev-Roberts, SR).
- Остановка в момент τ_{SR} минимизирует величину $\mathbf{C}(T)$

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

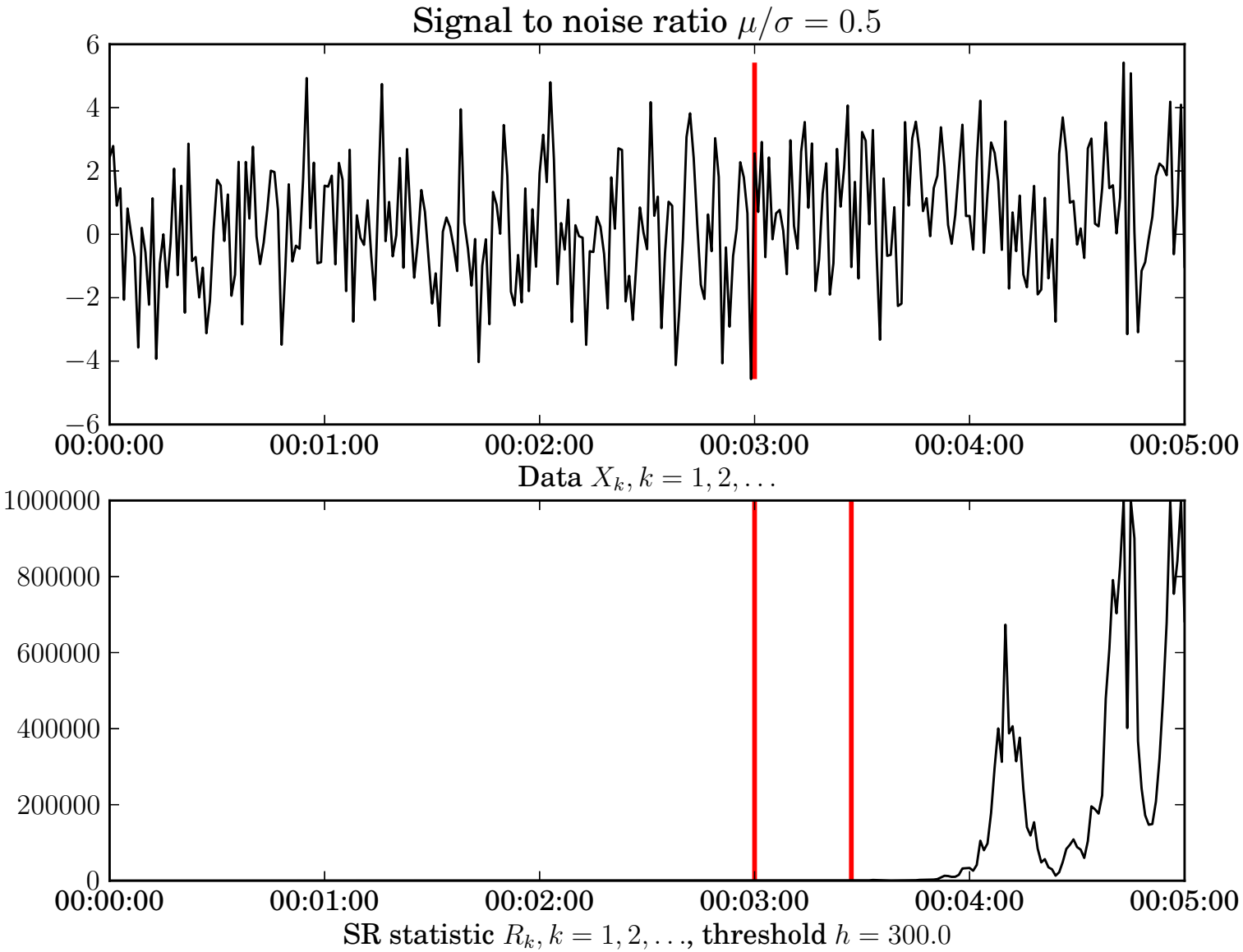
Пример 1 [Razladki]



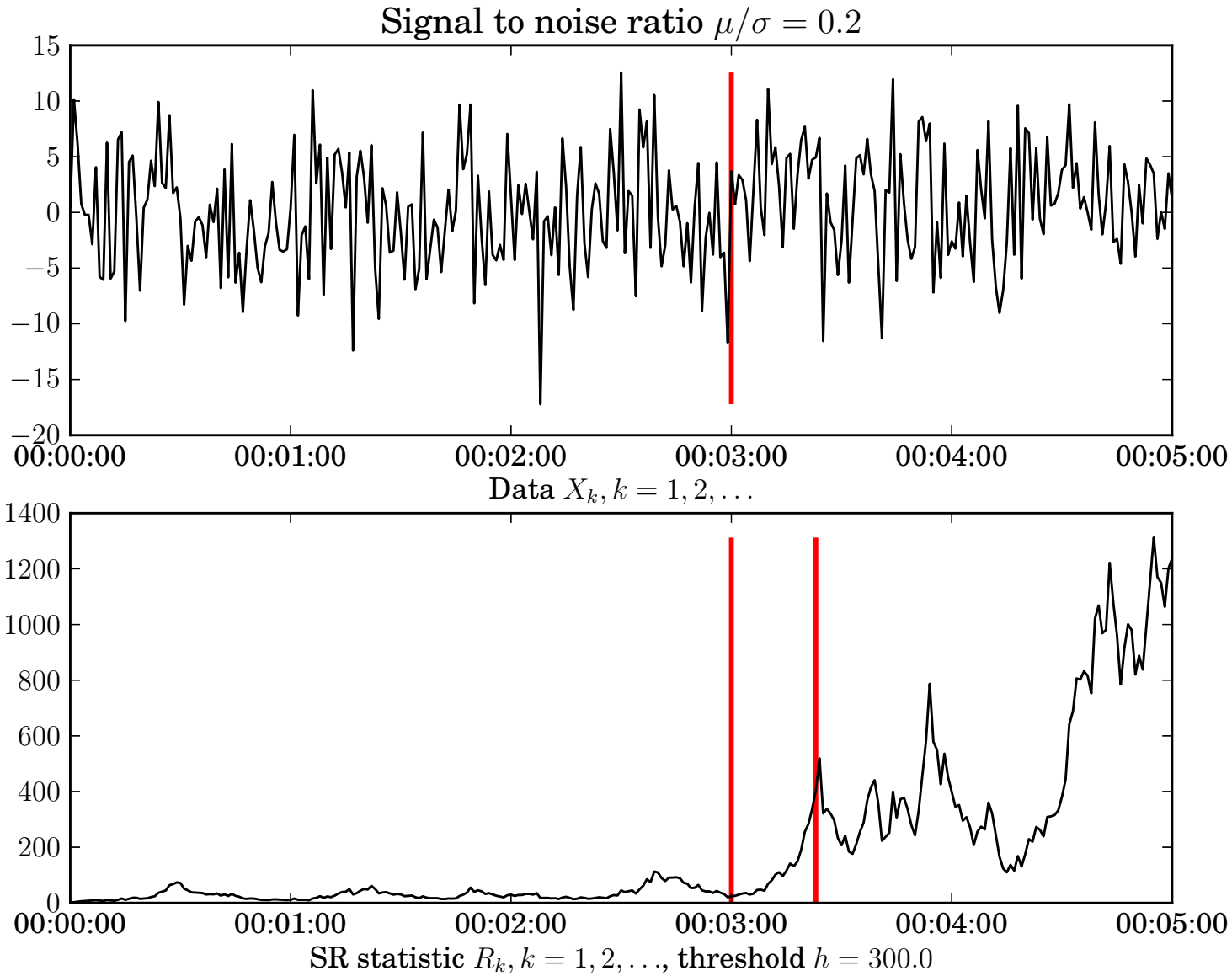
Пример 2 [Razladki]



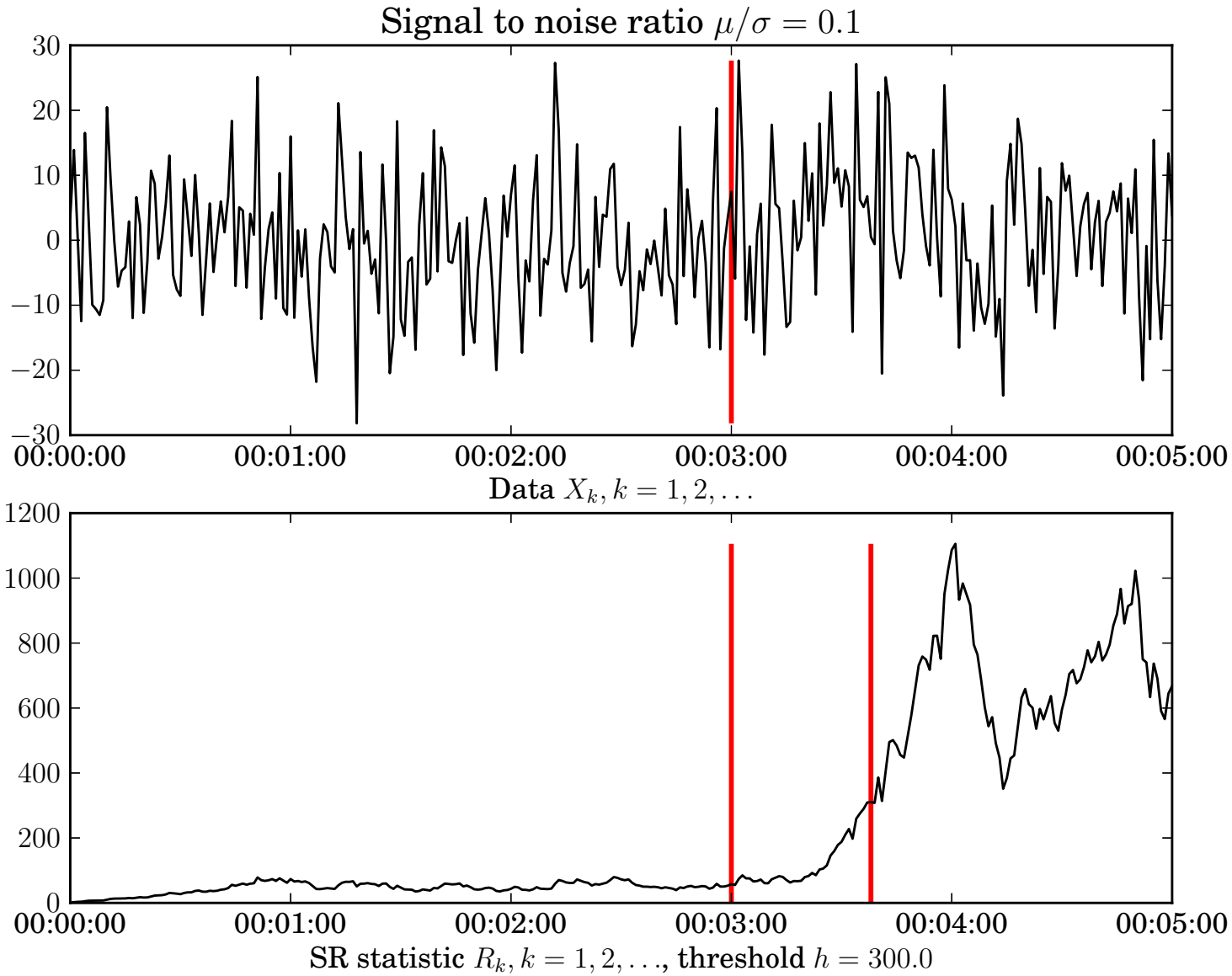
Пример 3 [Razladki]



Пример 4 [Razladki]



Пример 5 [Razladki]



Статистика π_t

- 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, $q = 1 - p$.
- Фиксируется некоторое число $\alpha \in (0, 1]$ и задается класс

$$\mathcal{M}_\alpha = \{\tau : \mathrm{P}(\tau < \theta) \leqslant \alpha\}$$

  тех моментов остановки, для которых вероятность ложной тревоги не выше α.

Статистика апостериорной вероятности

- Качество алгоритма задается величиной

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

- Этот критерий эквивалентен условно-байесовскому критерию

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

Статистика апостериорной вероятности

- Вводится статистика

$$\pi_n = \frac{\varphi_n}{1 + \varphi_n},$$

где для φ_n справедливо рекуррентное соотношение

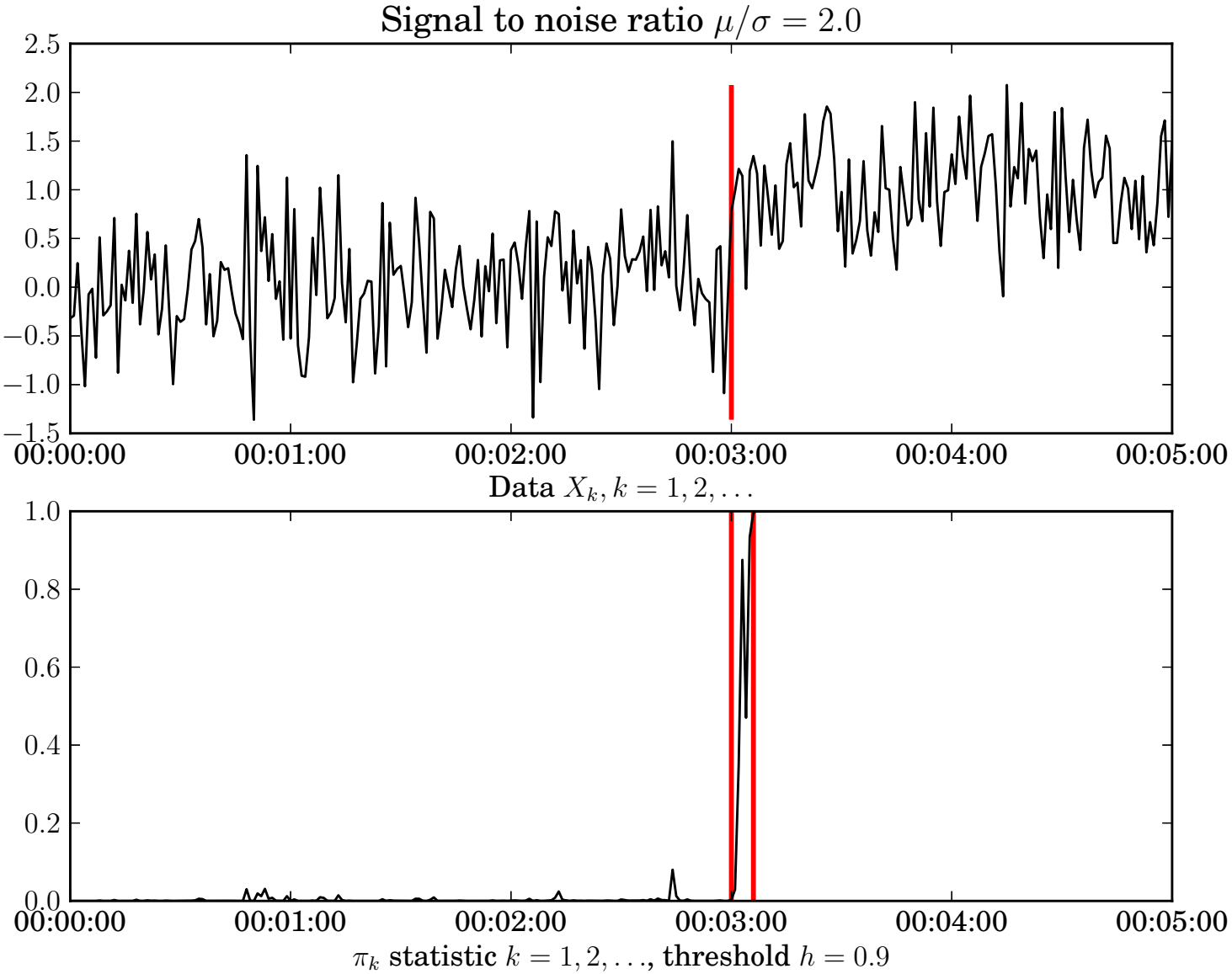
$$\varphi_{n+1} = (p + \varphi_n)\frac{f_0(X_n)}{q f_\infty(X_n)}.$$

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

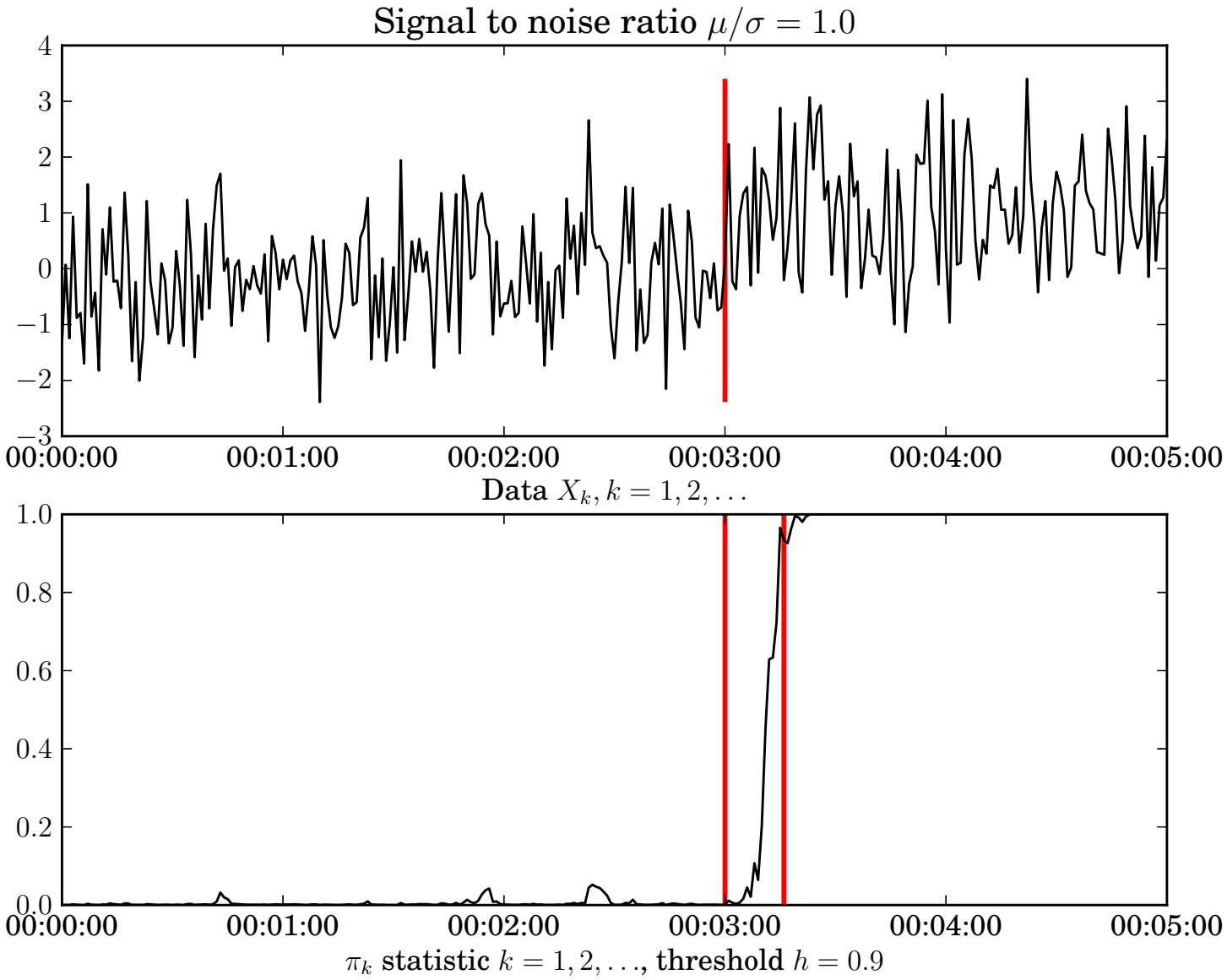
$$\tau_\pi = \inf\{n \geqslant 0 : \pi_n \geqslant h\}$$

- Значение π_n — апостериорная вероятность появления разладки до момента времени n в предположении, что получены наблюдения $\{X_k, k = \overline{1, n}\}$.

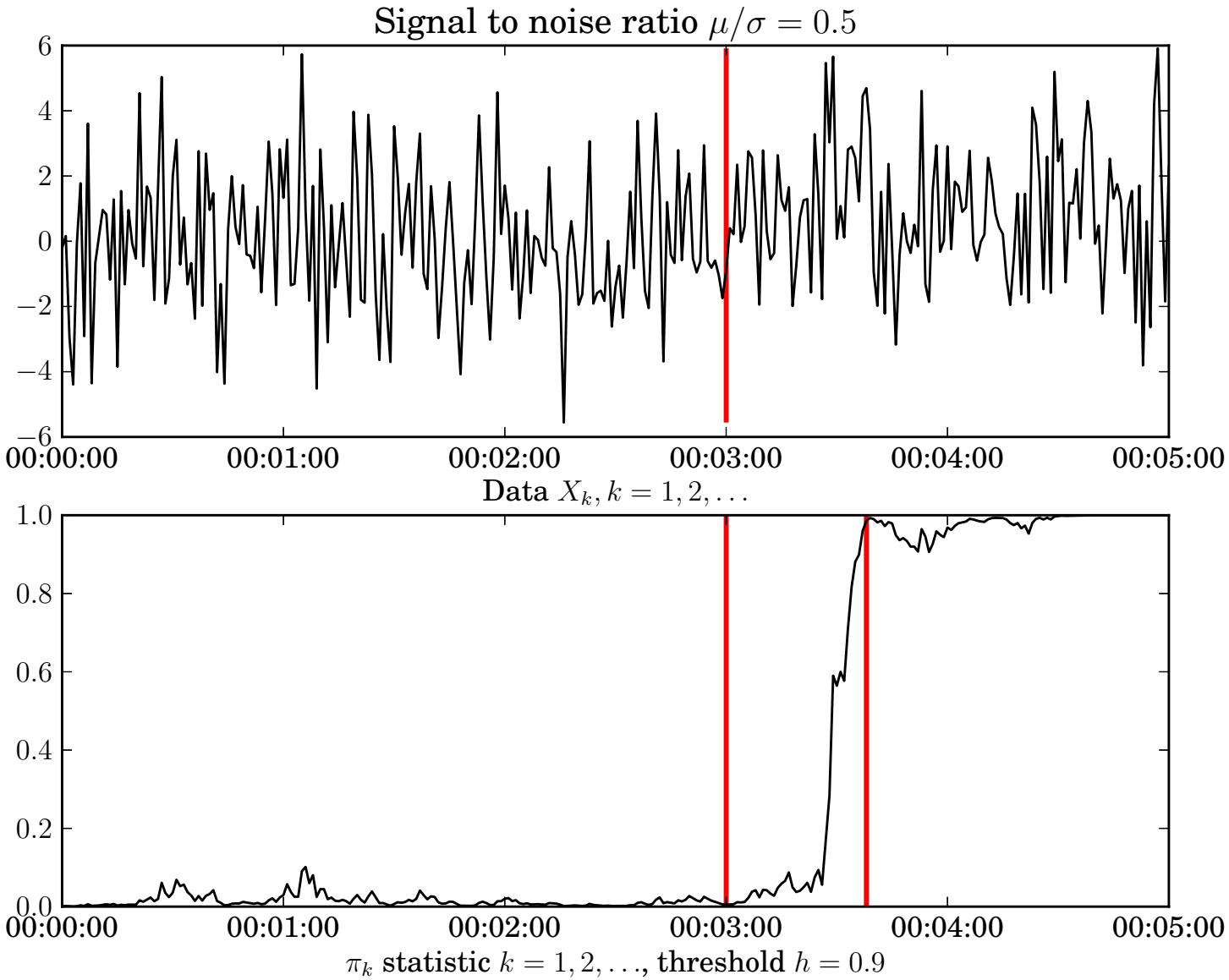
Пример 1 [Razladki]



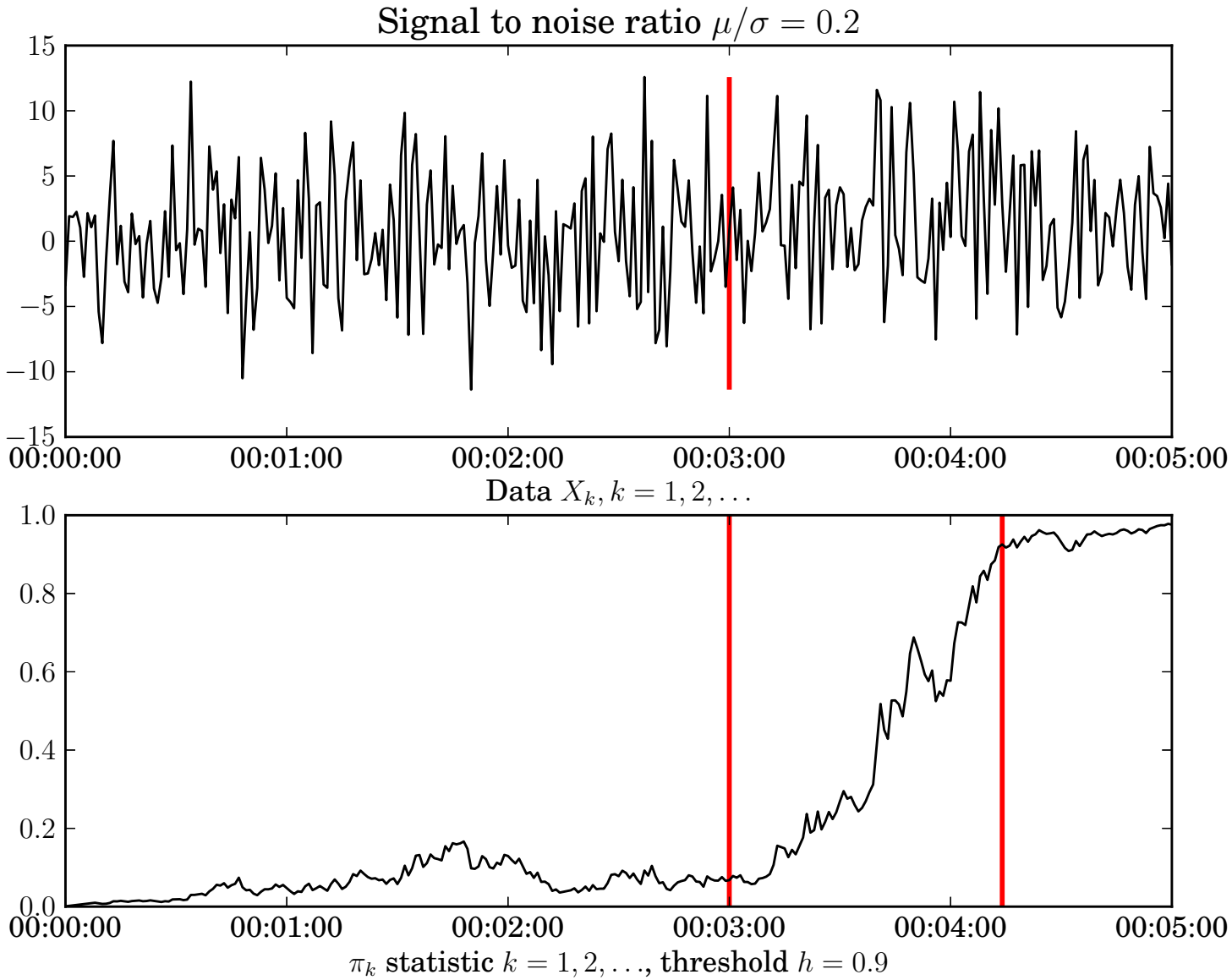
Пример 2 [Razladki]



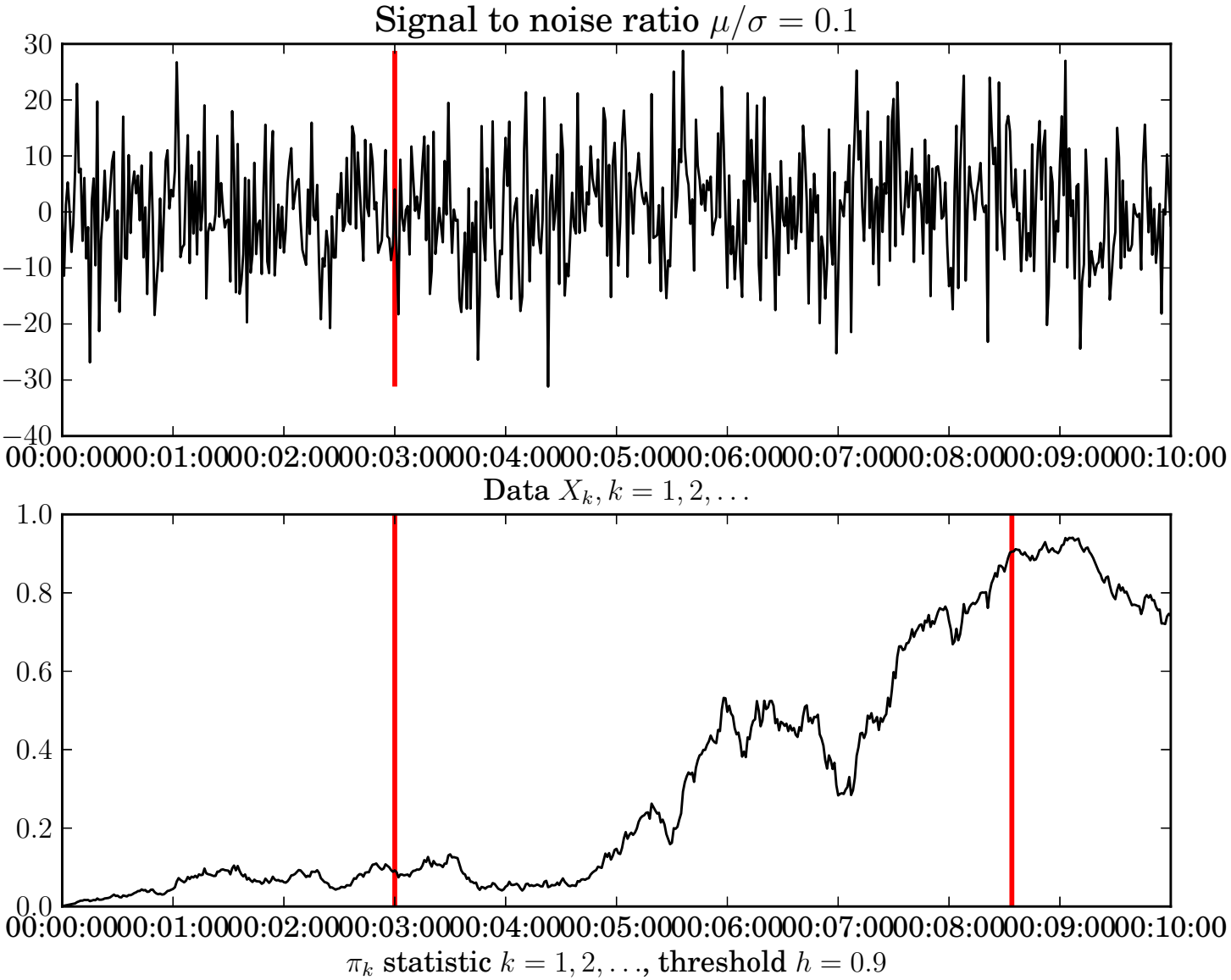
Пример 3 [Razladki]



Пример 4 [Razladki]



Пример 5 [Razladki]



Общая литература

Базовая литература по теории оптимальной остановки

- Обзор методов неоптимальной и оптимальной остановки с примерами и обсуждением [Basseville et al., 1993]
- Обзор методов оптимальной остановки с доказательствами и выводом формул [Ширяев, 2011]

Литература по оптимальным методам

Кумулятивные суммы

- Введены в [Page, 1954]
- Критерий оптимальности — [Lorden et al., 1971]
- Оптимальность для дискретного случая — [Moustakides et al., 1986], [Ritov, 1990]
- Оптимальность для непрерывного случая — [Ширяев, 1996]

Статистика Ширяева-Робертса

- Введена независимо в [Ширяев, 1961] и [Roberts, 1966]

Статистика апостериорных вероятностей

- Введена в [Ширяев, 1969]

Литература I

Michèle Basseville, Igor V Nikiforov, et al. *Detection of abrupt changes: theory and application*, volume 104. Prentice Hall Englewood Cliffs, 1993.

Gary Lorden et al. Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, 42(6): 1897–1908, 1971.

George V Moustakides et al. Optimal stopping times for detecting changes in distributions. *The Annals of Statistics*, 14(4): 1379–1387, 1986.

ES Page. Continuous inspection schemes. *Biometrika*, pages 100–115, 1954.

Ya'acov Ritov. Decision theoretic optimality of the cusum procedure. *The Annals of Statistics*, pages 1464–1469, 1990.

Литература II

SW Roberts. A comparison of some control chart procedures. *Technometrics*, 8(3):411–430, 1966.

Альберт Николаевич Ширяев. Задача скорейшего обнаружения нарушения стационарного режима. In *Докл. АН СССР*, volume 138, pages 1039–1042, 1961.

Альберт Николаевич Ширяев. *Статистический последовательный анализ: Оптималъные правила остановки*. Наука, 1969.

Альберт Николаевич Ширяев. Минимаксная оптимальность метода кумулятивных сумм (cusum) в случае непрерывного времени. *Успехи математических наук*, 51(4 (310)):173–174, 1996.

Альберт Николаевич Ширяев. *Вероятностно-статистические методы в теории принятия решений*, volume 144. МЦНМО, 2011.