Санкт-Петербургский государственный университет

Прикладная математика и информатика

Статистическое моделирование

Алиева Наталия Дмитриевна

Исследования по задаче “размазанной” разреженности

Выпускная квалификационная работа

Научный руководитель:

д. ф.-м. н., профессор М. С. Ермаков

Рецензент:

д. ф.-м. н., профессор Г. Л. Шевляков

Санкт-Петербург

2017

Saint Petersburg State University
Applied Mathematics and Computer Science
Statistical Modelling

Alieva Nataliia

Investigation on fused sparsity problem

Graduation Project

Scientific Supervisor:
Doctor of Physics and Mathematics,
Professor M. S. Ermakov

Reviewer:
Doctor of Physics and Mathematics,
Professor G. L. Shevlyakov

Saint Petersburg
2017

Оглавление

Введение

Целью данной выпускной работы является изучение вопроса обнаружения скачков и оценивания скачкообразных функций, когда имеется много моментов разладки.

Такие задачи встречаются во многих сферах деятельности, в том числе экономике, медицине и естественных науках, мобильной связи, обработке изображений и т.д. Этот вопрос изучался в большом числе работ. Например, обнаружение вариаций числа копий генов в ДНК (CNV - Copy Number Variations) сводится к задаче нахождения моментов разладки в исходных данных. В статье [1] был рассмотрен метод для решения такой задачи. Вопрос обнаружения моментов разладки в зашумленных данных был также изучен в работах [2], [3]. При этом одним из ключевых вопросов являлась минимальная ширина скачка, которую можно обнаружить. Этот вопрос был досконально изучен для одномерной регрессии.

Нами изучается данный вопрос для скачков в пуассоновских процессах. Мы находим точную (с точности до константы) асимптотическую ширину скачка (разрыва) для пуассоновских процессов, которую можно обнаружить в контексте одного из первых алгоритмов обнаружения моментов раздалки, представленных в работе [5]. Существует много методов оценивания регрессионной модели со скачками, но наиболее распространенными являются работы [4] и [1], однако результаты этих работ не сравнивались ранее. Таким образом, основной задачей было сравнить работу алгоритмов в работах [4] и [1], а также рассмотреть один из первых алгоритмов [5], реализовать его и сравнить с методами [4] и [1].

В главе (2) приведен подробный обзор существующих методов решений такого типа задачи.

Основные результаты, полученные в ходе исследованияя представлены в главе (1), а именно:

- Теоретические оценки для исходного и модифицированного алгоритмов;
- Практические результаты и сравнительный анализ алгоритмов.

Подробный анализ и доказательства приведены в главе (3).

Глава 1

Основные результаты

1.1. Теоретические результаты

В статье [5] был предложен адаптивный метод обнаружения скачков для нормального распределения с различными оценками. В этом разделе будет представлены оценки для пуассновских случайных процессов, а также оценки для исходного алгоритма, которые еще не были получены. Все предложенные оценки были найдены для одномерного и двумерного случаев.

Подробные доказательства представлены в главе (3).

1.1.1. Модель с гауссовским шумом

Рассмотрим область $[0,1]^d$, где $d = 1, 2$, и модель регрессии для одномерного случая (для $d = 2$ подробная постановка задачи введена в разделе (3.1.2)):

$$Y_i = f(X_i) + \xi_i, i = 1, ..n, \tag{1.1}$$

где $i = 1, ..n$, $X_i, Y_i \in \mathbb{R}$, ξ_i — независимые случайные величины нормального распределения $(\mathcal{N}(0, \sigma^2))$, где σ^2 — дисперсия. Оценка нижней границы ширины интервала была приведена в статье [5], поэтому для одномерного случая приведена только оценка верхней границы ширины интервала скачка.

Для двумерного случая приведены оценки для нижней и верхней границ стороны квадрата со скачком. Подробные доказательства представлены в разделе (3.1).

Полученные результаты:

- Одномерный случай:

$$\mathrm{h}(n) = Cn^{-1} \ln n, \tag{1.2}$$

  где $C \asymp 2 + \varepsilon$.

- Двумерный случай:

  - Нижняя граница:

$$\mathrm{h}(n) = Cn^{-1} \ln n, \tag{1.3}$$

    где $C \asymp \sqrt{2 - \varepsilon}$.

- Верхняя граница:

$$\mathrm{h}(n) = Cn^{-1}\ln n, \tag{1.4}$$

где $C \asymp \sqrt{2+\varepsilon}$.

1.1.2. Пуассоновский случайный процесс

Рассмотрим постановку задачу для одномерного случая. Постановка задачи для двумерного случая вводится аналогичным образом в разделе (3.2.2).

Пусть наблюдается неоднородный пуассоновский процесс $X(t)$, при $t \in [0,1]$, постоянной интенсивности $\lambda(t) = n$ за исключением малого интервала $[a_0, a_1] \in [0,1]$, где $0 < a_0 < a_1 < 1$, а интенсивность скачка $\lambda(t) = an$, $(a > 1)$. Обозначим ширину этого интервала за $\mathrm{h}(n)$.

Рассмотрим задачу о проверке гипотезы H_0 о том, что интенсивность процесса Пуассона постоянна и равна n, против альтернативы H_1 о том, что интенсивность процесса Пуассона имеет вид:

$$\lambda(t) = \begin{cases} na, & t \in [a_0, a_1] \\ n, & t \notin [a_0, a_1]. \end{cases} \tag{1.5}$$

Требуется получить состоятельную оценку для ширины интервала со скачком (в одномерном случае) и состоятельную оценку для стороны квадрата со скачком (в двумерном случае). Подробные доказательства представлены в разделе (3.2).

Полученные результаты:

- Одномерный случай:

  - Нижняя граница:

$$\mathrm{h}(n) = Cn^{-1}\ln n, \tag{1.6}$$

  где $C \asymp \frac{\ln(1+a)}{(2\ln 2-1)(a-1)}$.

  - Верхняя граница:

$$\mathrm{h}(n) = Cn^{-1}\ln n, \tag{1.7}$$

  где $C \asymp \frac{\ln(1+a)}{(2\ln 2-1)(a-1)}$.

- Двумерный случай:

- Нижняя граница:

$$\mathrm{h}(n) = Cn^{-1} \ln n, \tag{1.8}$$

где $C \asymp \sqrt{\frac{\ln(1+a)}{(2\ln 2-1)(a-1)}}$.

- Верхняя граница:

$$\mathrm{h}(n) = Cn^{-1} \ln n, \tag{1.9}$$

где $C \asymp \sqrt{\frac{\ln(1+a)}{(2\ln 2-1)(a-1)}}$.

1.2. Результаты моделирования

В этом разделе приведен сравнительный анализ алгоритма В. Г. Спокойного [5] с другими алгоритмами.

Для сравнения были выбраны следующие алгоритмы:

- Алгоритм обнаружения скачков А. Мунка [4].
- Алгоритм отсеивания и ранжирования (SaRa) [1].

Необходимо провести сравнительный анализ алгоритмов для гауссовского шума и пуассоновского процесса.

В этом разделе представлены примеры обнаружения скачков для разных объемов выборки, оценки полученных параметров и, кроме того, приведена визуализация времени работы алгоритмов. В приложении А представлена реализация всех алгоритмов, в том числе реализация расчета времени работы алгоритмов.

1.3. Гауссовский шум

1.3.1. Обнаружение и оценка моментов разладки

Случай 1

Рассмотрим ситуацию, когда скачок находится справа, то есть имеется всего 1 точка разладки.

Ниже приведены примеры обнаружения такого скачка для различных объемов выборки для SaRa (рис. ??), алгоритма Мунка (рис. ??) и алгоритма В. Г. Спокойного (рис. ??). Красным цветом выделена оценка высоты скачка.

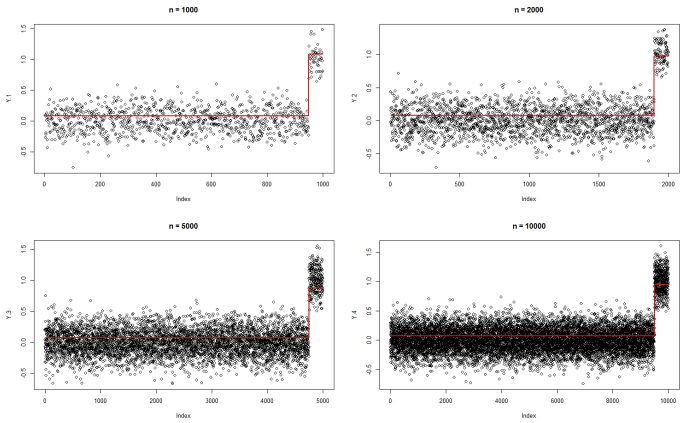

Рис. 1.1. Пример обнаружения одного скачка на границе с помощью SaRa для гауссовского шума

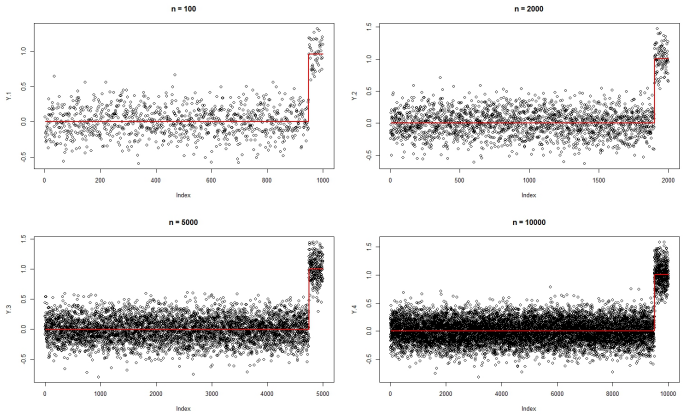

Рис. 1.2. Пример обнаружения одного скачка на границе с помощью алгоритма Мунка для гауссовского шума

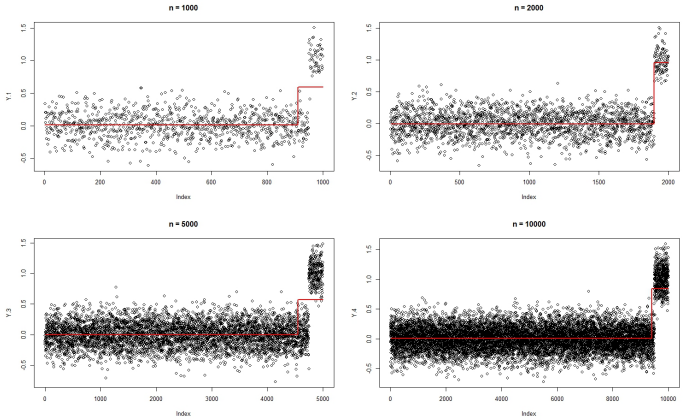

Рис. 1.3. Пример обнаружения одного скачка на границе с помощью алгоритма В. Г. Спокойного для гауссовского шума

Случай 2

Рассмотрим ситуацию, когда скачок находится посередине, то есть имеются 2 точки разладки.

Ниже приведены примеры обнаружения такого скачка для различных объемов выборки для SaRa (рис. ??), алгоритма Мунка (рис. ??) и алгоритма В. Г. Спокойного (рис. ??). Красным цветом выделена оценка высоты скачка.

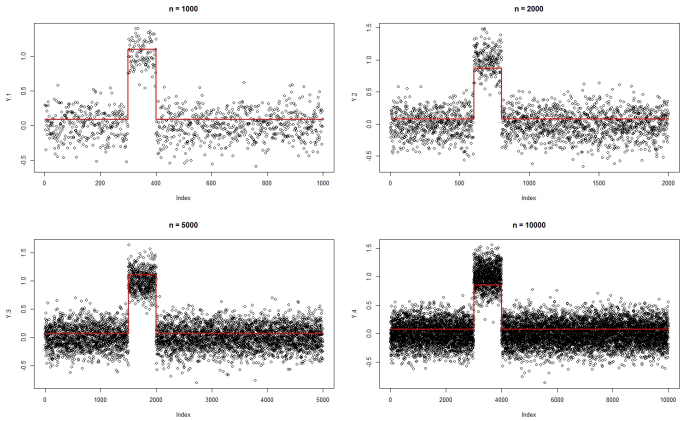

Рис. 1.4. Пример обнаружения одного скачка внутри интервала с помощью SaRa для гауссовского шума

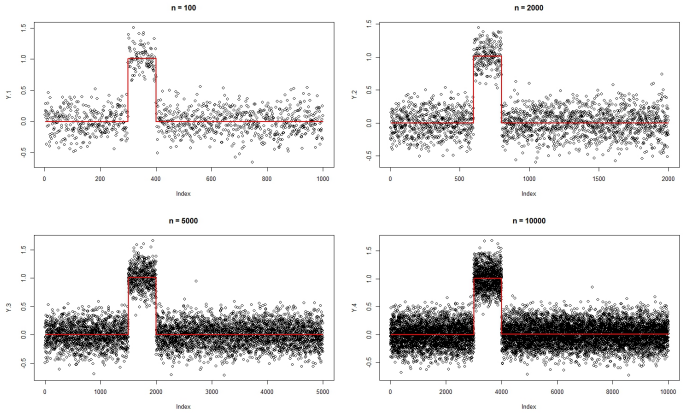

Рис. 1.5. Пример обнаружения одного скачка внутри интервала с помощью алгоритма Мунка для гауссовского шума

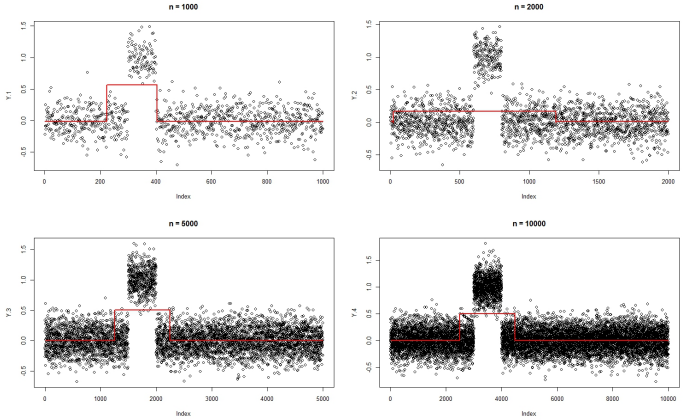

Рис. 1.6. Пример обнаружения одного скачка внутри интервала с помощью алгоритма В. Г. Спокойного для гауссовского шума

Сравнительные характеристики алгоритмов в зависимости от параметров

1. **Параметры нормального распределения:** $\mu = 0, \sigma = 0.2$**.**

   Исходные нормированные моменты разладки: $(0.3, 0.4)^{\mathrm{T}}$.
   Исходная ширина интервала со скачком (нормированная под отрезок $[0, 1]$): 0.1.
   Исходная высота скачка: 1.

   Были получены оценки смещения, дисперсии и стандартного отклонения на 500 итерациях для различных объемов выборки ($n = 100, n = 500, n = 1000$).

Таблица 1.1. Оценки смещения, дисперсии, RMSE для моментов разладки

Алгоритм	Математическое ожидание	Дисперсия	RMSE
Алгоритм Мунка [4], $n = 100$	(0.29,0.4)	(0.012,0.01)	(0.5,0.01)
Алгоритм Мунка [4], $n = 500$	(0.29,0.39)	(0.016,0.005)	(1.1,0.003)
Алгоритм Мунка [4], $n = 1000$	(0.3,0.39)	(0.016,0.01)	(1.5, 0.03)
SaRa [1], $n = 100$	(0.25,0.35)	(0.6,7)	(0.9,0.5)
SaRa [1], $n = 500$	(0.29,0.38)	(10.4,130)	(1.3,0.5)
SaRa [1], $n = 1000$	(0.295,0.39)	(19.6,545)	(1.7,0.8)
Алгоритм Спокойного [5], $n = 100$	(0.21,0.56)	(119,560)	(1.7,0.28)
Алгоритм Спокойного [5], $n = 500$	(0.17,0.64)	(3467,16597)	(4.6,7.9)
Алгоритм Спокойного [5], $n = 1000$	(0.19,0.6)	(13588,64046)	(6.1,10.4)

В таблице 1.1 представлены оценки результатов обнаружения моментов разладки для всех 3–х алгоритмов, а именно: математическое ожидание, дисперсия и RMSE.

Заметим, что в среднем наиболее точным из всех алгоритмов является алгоритм Мунка [4] и результаты также обладают наименьшей лисперсией с увеличением объема выборки. Оценки SaRa [1] улучшаются с увеличением объема выборки, однако также увеличивается и RMSE. Худший результат показал алгоритм Спокойного [5], в среднем оценки колеблятся около первого приближения, и с увели-

чением объема выборки результат не приближается к истинному. Кроме того, для его результатов получили очень большую дисперсию.

Таблица 1.2. Оценки смещения, дисперсии, RMSE для ширины интервала

Алгоритм	Математическое ожидание	Дисперсия	RMSE
Алгоритм Мунка [4], $n = 100$	0.1	0.02	0.5
Алгоритм Мунка [4], $n = 500$	0.1	0.02	1.1
Алгоритм Мунка [4], $n = 1000$	0.1	0.02	1.5
SaRa [1], $n = 100$	0.101	7.4	0.58
SaRa [1], $n = 500$	0.095	139.7	1.1
SaRa [1], $n = 1000$	0.095	562.2	1.6
Алгоритм Спокойного [5], $n = 100$	0.34	1147	4.5
Алгоритм Спокойного [5], $n = 500$	0.46	33813.13	12.4
Алгоритм Спокойного [5], $n = 1000$	0.42	131491.8	16.39

В таблице 1.2 представлены оценки ширины скачка для всех трех алгоритмов. Заметим, что алгоритм Мунка [4] оценивает точно ширину, начиная с объема выборки $n = 100$, кроме того, сохраняется небольшая дисперсия по сравнению с остальными алгоритмами.

Оценки для SaRa [1] колеблются около истинного значения, однако нет тенденции к улучшению с увеличением объема выборки.

Оценки для алгоритма Спокойного [5] получились хуже остальных алгоритмов, как и предыдущем случае.

Таблица 1.3. Оценки смещения, дисперсии, RMSE для высоты скачка

Алгоритм	Математическое ожидание	Дисперсия	RMSE
Алгоритм Мунка [4], $n = 100$	1	0.004	0.0064
Алгоритм Мунка [4], $n = 500$	0.99	0.0008	0.00125
Алгоритм Мунка [4], $n = 1000$	0.99	0.0004	6.95E-04
SaRa [1], $n = 100$	0.82	0.014	0.02
SaRa [1], $n = 500$	0.8824	0.016	0.07
SaRa [1], $n = 1000$	0.8828	0.0164	0.05
Алгоритм Спокойного [5], $n = 100$	0.4	0.1	0.065
Алгоритм Спокойного [5], $n = 500$	0.47	0.1	0.02
Алгоритм Спокойного [5], $n = 1000$	0.51	0.1	0.018

В таблице 1.3 представлены оценки высоты скачка для всех трех алгоритмов. Заметим, что алгоритм Мунка [4] оценивает точно высоту, начиная с объема выборки

$n = 100$, однако с увеличением объема выборки результат в среднем немного ухудшается, тем не менее колеблется в окрестности истинного значения. Кроме того, с увеличением обхема выборки уменбшаются дисперсия и RMSE.

Оценки для SaRa [1] колеблются около истинного значения, однако они хуже, чем самая плохая оценка алгоритма Мунка [4].

Оценки для алгоритма Спокойного [5] получились хуже остальных алгоритмов, как и предыдущем случае.

2. **Параметры нормального распределения:** $\mu = 0, \sigma = 0.5$**.**

Исходные нормированные моменты разладки: $(0.3, 0.4)^{\mathrm{T}}$.

Исходная ширина интервала со скачком (нормированная под отрезок $[0, 1]$): 0.1.

Исходная высота скачка: 1.

Были получены оценки смещения, дисперсии и стандартного отклонения на 500 итерациях для различных объемов выборки ($n = 100, n = 500, n = 1000$).

Таблица 1.4. Оценки смещения, дисперсии, RMSE для моментов разладки

Алгоритм	Математическое ожидание	Дисперсия	RMSE
Алгоритм Мунка [4], $n = 100$	(0.24,0.42)	(135.39,155)	(1.57,1.26)
Алгоритм Мунка [4], $n = 500$	(0.29,0.39)	(1.75,6.25)	(1.1,0.1)
Алгоритм Мунка [4], $n = 1000$	(0.3,0.4)	(1.21,1.13)	(1.5, 0.03)
SaRa [1], $n = 100$	(0.26,0.36)	(40.1,165.9)	(1.09,1.34)
SaRa [1], $n = 500$	(0.29,0.47)	(2287.6,9126.9)	(2.45,4.59)
SaRa [1], $n = 1000$	(0.3,0.5)	(17033.1,43895.8)	(4.3,7.4)
Алгоритм Спокойного [5], $n = 100$	(0.22,0.5)	(108.2,502.1)	(1.65,2.6)
Алгоритм Спокойного [5], $n = 500$	(0.18,0.6)	(3442.8,16685.1)	(4.5,7.8)

Алгоритм Спокойного [5], $n = 1000$	(0.19,0.6)	(13805,63966.2)	(6.24,10.4)

В таблице 1.4 представлены оценки моментов разладки для всех трех алгоритмов. Заметим, что алгоритм Мунка [4] оценивает точно, начиная при объеме выборки $n = 1000$, при этом с $n = 100$ до $n = 1000$ замечено улучшение оценок. Кроме того, с увеличением объема выборки уменьшается дисперсия оценки.

Оценки для SaRa [1] улучшаются с увеличением объема выборки и так же, как и в алгоритме Мунка [4] приходят к истинному значению при обхеме выборки $n = 1000$. Однако с увеличением обхема выборки происходит увеличение дисперсии оценки.

Оценки для алгоритма Спокойного [5] получились хуже остальных алгоритмов, как и предыдущем случае.

Таблица 1.5. Оценки смещения, дисперсии, RMSE для ширины интервала

Алгоритм	Математическое ожидание	Дисперсия	RMSE
Алгоритм Мунка [4], $n = 100$	0.18	275.5	2.1
Алгоритм Мунка [4], $n = 500$	0.09	8.29	1.1
Алгоритм Мунка [4], $n = 1000$	0.1	2.19	1.5
SaRa [1], $n = 100$	0.09	139.2	1.28
SaRa [1], $n = 500$	0.09	7329.9	4.8
SaRa [1], $n = 1000$	0.2	34059.8	7.7

Алгоритм Спокойного [5], $n = 100$	0.3	1033.3	4.2
Алгоритм Спокойного [5], $n = 500$	0.4	33577.7	12.25
Алгоритм Спокойного [5], $n = 1000$	0.4	132496.4	16.4

В таблице 1.5 представлены оценки ширины скачка для всех трех алгоритмов. Заметим, что алгоритм Мунка [4] оценивает точно ширину, начиная с объема выборки $n = 1000$, при этом с $n = 100$ до $n = 1000$ замечено улучшение оценок. Кроме того, с увеличением объема выборки уменьшается дисперсия оценки.

Оценки для SaRa [1] колеблются в окрестности истинного значения, однако для объема выборки $n = 1000$ результат превысил истинный на 0.1.

Оценки для алгоритма Спокойного [5] получились хуже остальных алгоритмов, как и предыдущем случае. Кроме того, они ухудшаются с увеличением объема выборки.

Таблица 1.6. Оценки смещения, дисперсии, RMSE для высоты скачка

Алгоритм	Математическое ожидание	Дисперсия	RMSE
Алгоритм Мунка [4], $n = 100$	1	0.02	0.02
Алгоритм Мунка [4], $n = 500$	1	0.01	0.005
Алгоритм Мунка [4], $n = 1000$	1	0.0028	0.001
SaRa [1], $n = 100$	0.82	0.048	0.027

SaRa [1], $n = 500$	0.92	0.042	0.009
SaRa [1], $n = 1000$	0.94	0.03	0.006
Алгоритм Спокойного [5], $n = 100$	0.47	0.1	0.06
Алгоритм Спокойного [5], $n = 500$	0.51	0.1	0.02
Алгоритм Спокойного [5], $n = 1000$	0.53	0.1	0.018

В таблице 1.6 представлены оценки высоты скачка для всех трех алгоритмов. Заметим, что алгоритм Мунка [4] оценивает точно высоту, начиная с объема выборки $n = 100$. Кроме того, с увеличением объема выборки дисперсия и RMSE уменьшаются.

Оценки для SaRa [1] улучшаются с увеличением объема выборки, однако даже при $n = 1000$ не достигают истинного значения. Кроме того, дисперсия и RMSE уменьшаются с увеличением объема выборки.

Оценки для алгоритма Спокойного [5] получились хуже остальных алгоритмов, как и предыдущем случае.

1.4. Пуассоновский процесс

1.4.1. Обнаружение и оценка моментов разладки

Рассмотрим ситуацию, когда скачок находится внутри интервала.

1. **Исходные нормированные моменты разладки:** $(0.3, 0.4)^{\mathrm{T}}$.
   Исходная ширина интервала со скачком (нормированная под отрезок $[0, 1]$): 0.1.
   Исходная высота скачка: 2.

Были получены оценки смещения, дисперсии и стандартного отклонения на 500 итерациях для различных объемов выборки ($n = 100, n = 500, n = 1000$).

Таблица 1.7. Оценки смещения, дисперсии, RMSE для моментов разладки

Алгоритм	Математическое ожидание	Дисперсия	RMSE
Алгоритм Мунка [4], $n = 100$	(0.3,0.4)	(0,0)	(0.5,0)
Алгоритм Мунка [4], $n = 500$	(0.3,0.4)	(0,0)	(1.1,0)
Алгоритм Мунка [4], $n = 1000$	(0.3,0.4)	(0,0)	(1.5, 0)
SaRa [1], $n = 100$	(0.25,0.36)	(0,0)	(1,0.4)
SaRa [1], $n = 500$	(0.29,0.392)	(0,0)	(1.3,0.17)
SaRa [1], $n = 1000$	(0.295,0.396)	(0,0)	(1.7,0.12)
Алгоритм Спокойного [5], $n = 100$	(0.279,0.4)	(1.5,0.24)	(0.7,0.065)
Алгоритм Спокойного [5], $n = 500$	(0.28,0.4)	(0.39,0.16)	(1.5,0.02)

Алгоритм Спокойного [5], $n = 1000$	(0.28,0.4)	(0.66,1.78)	(2.2,0.04)

В таблице 1.7 представлены оценки моментов разладки для всех трех алгоритмов. Заметим, что алгоритм Мунка [4] оценивает точно моменты разладки, начиная с объема выборки $n = 100$.

Оценки для SaRa [1] улучшаются с увеличением объема выборки, однако даже при $n = 1000$ не достигают истинного значения.

Оценки для алгоритма Спокойного [5] получились хуже остальных алгоритмов, однако с увеличением объема выборки приближаются к истинному значению, но дисперсия увеличивается.

Таблица 1.8. Оценки смещения, дисперсии, RMSE для ширины интервала

Алгоритм	Математическое ожидание	Дисперсия	RMSE
Алгоритм Мунка [4], $n = 100$	0.1	0	0.5
Алгоритм Мунка [4], $n = 500$	0.1	0	1.1
Алгоритм Мунка [4], $n = 1000$	0.1	0	1.5
SaRa [1], $n = 100$	0.11	0	0.6
SaRa [1], $n = 500$	0.1	0	1.1
SaRa [1], $n = 1000$	0.1	0	1.5
Алгоритм Спокойного [5], $n = 100$	0.12	1.84	0.76

Алгоритм Спокойного [5], $n = 500$	0.12	0.59	1.5
Алгоритм Спокойного [5], $n = 1000$	0.11	0.35	2.1

В таблице 1.8 представлены оценки ширины скачка для всех трех алгоритмов. Заметим, что алгоритм Мунка [4] оценивает точно ширину, начиная с объема выборки $n = 100$.

Оценки для SaRa [1] улучшаются с увеличением объема выборки и достигают истинного значения, начиная с объема выборки $n = 500$.

Оценки для алгоритма Спокойного [5] получились хуже остальных алгоритмов, однако с увеличением объема выборки приближаются к истинному значению.

Таблица 1.9. Оценки смещения, дисперсии, RMSE для высоты скачка

Алгоритм	Математическое ожидание	Дисперсия	RMSE
Алгоритм Мунка [4], $n = 100$	1.8	0.001	0.08
Алгоритм Мунка [4], $n = 500$	1.8	0.0002	0.03
Алгоритм Мунка [4], $n = 1000$	1.8	0.0001	0.02
SaRa [1], $n = 100$	1.8	0.0003	0.082
SaRa [1], $n = 500$	1.96	1.42E-05	0.04
SaRa [1], $n = 1000$	1.98	3.14E-06	0.03

Алгоритм Спокойного [5], $n = 100$	1.44	0.008	0.045
Алгоритм Спокойного [5], $n = 500$	1.47	0.0006	0.02
Алгоритм Спокойного [5], $n = 1000$	1.48	0.0004	0.014

В таблице 1.9 представлены оценки высоты скачка для всех трех алгоритмов. Заметим, что алгоритм Мунка [4] оценивает точно высоту, начиная с объема выборки $n = 100$.

Оценки для SaRa [1] улучшаются с увеличением объема выборки и достигают истинного значения, начиная с объема выборки $n = 500$.

Оценки для алгоритма Спокойного [5] получились хуже остальных алгоритмов, однако с увеличением объема выборки приближаются к истинному значению.

2. **Исходные нормированные моменты разладки:** $(0.3, 0.4)^{\mathrm{T}}$.
   Исходная ширина интервала со скачком (нормированная под отрезок $[0, 1]$): 0.1.
   Исходная высота скачка: 10.

Были получены оценки смещения, дисперсии и стандартного отклонения на 500 итерациях для различных объемов выборки ($n = 100, n = 500, n = 1000$).

Таблица 1.10. Оценки смещения, дисперсии, RMSE для моментов разладки

Алгоритм	Математическое ожидание	Дисперсия	RMSE
Алгоритм Мунка [4], $n = 100$	(0.3,0.4)	(0,0)	(0.5,0)
Алгоритм Мунка [4], $n = 500$	(0.3,0.4)	(0,0)	(1.1,0)

Алгоритм Мунка [4], $n = 1000$	(0.3,0.4)	(0,0)	(1.5, 0)
SaRa [1], $n = 100$	(0.25,0.36)	(0,0)	(1,0.4)
SaRa [1], $n = 500$	(0.29,0.392)	(0,0)	(1.34,0.17)
SaRa [1], $n = 1000$	(0.295,0.396)	(0,0)	(1.7,0.12)
Алгоритм Спокойного [5], $n = 100$	(0.279,0.4)	(1.5,0.2)	(0.7,0.08)
Алгоритм Спокойного [5], $n = 500$	(0.28,0.4)	(0.39,0.24)	(1.56,0.02)
Алгоритм Спокойного [5], $n = 1000$	(0.28,0.4)	(0.66,0.04)	(2.2,0.06)

В таблице 1.10 представлены оценки моментов скачка для всех трех алгоритмов. Заметим, что алгоритм Мунка [4] оценивает точно расположение моментов разладки, начиная с объема выборки $n = 100$.

Оценки для SaRa [1] улучшаются с увеличением объема выборки, но не достигают истинного значения.

Оценки для алгоритма Спокойного [5] получились хуже остальных алгоритмов, однако с увеличением объема выборки приближаются к истинному значению.

Таблица 1.11. Оценки смещения, дисперсии, RMSE для ширины интервала

Алгоритм	Математическое ожидание	Дисперсия	RMSE
Алгоритм Мунка [4], $n = 100$	0.1	0	0.5

Алгоритм Мунка [4], $n = 500$	0.1	0	1.1
Алгоритм Мунка [4], $n = 1000$	0.1	0	1.5
SaRa [1], $n = 100$	0.11	0	0.6
SaRa [1], $n = 500$	0.1	0	1.1
SaRa [1], $n = 1000$	0.1	0	1.5
Алгоритм Спокойного [5], $n = 100$	0.12	1.84	0.7
Алгоритм Спокойного [5], $n = 500$	0.12	0.59	1.58
Алгоритм Спокойного [5], $n = 1000$	0.11	0.35	2.1

В таблице 1.11 представлены оценки ширины скачка для всех трех алгоритмов. Заметим, что алгоритм Мунка [4] оценивает точно ширину скачка, начиная с объема выборки $n = 100$.

Оценки для SaRa [1] улучшаются с увеличением объема выборки и достигают истинного значения, начиная с $n = 500$.

Оценки для алгоритма Спокойного [5] получились хуже остальных алгоритмов, однако с увеличением объема выборки приближаются к истинному значению.

Таблица 1.12. Оценки смещения, дисперсии, RMSE для высоты скачка

Алгоритм	Математическое ожидание	Дисперсия	RMSE

Алгоритм Мунка [4], $n = 100$	9.8	0.001	0.08
Алгоритм Мунка [4], $n = 500$	9.8	0.0002	0.03
Алгоритм Мунка [4], $n = 1000$	9.8	0.0001	0.02
SaRa [1], $n = 100$	9.1	0.0003	0.082
SaRa [1], $n = 500$	9.8	1.42E-05	0.04
SaRa [1], $n = 1000$	9.9	3.03E-06	0.03
Алгоритм Спокойного [5], $n = 100$	7.8	0.089	0.6
Алгоритм Спокойного [5], $n = 500$	8.09	0.08	0.3
Алгоритм Спокойного [5], $n = 1000$	8.16	0.08	0.22

В таблице 1.12 представлены оценки высоты скачка для всех трех алгоритмов. Оценки для алгоритма Мунка [4] близи к истинному значению. Кроме того, с увеличением обхема выборки, дисперсия и RMSE уменьшаются.

Оценки для SaRa [1] улучшаются с увеличением объема выборки, но не достигают истинного значения. Кроме того, с увеличением объема выборки также уменьшается дисперсия и RMSE.

Оценки для алгоритма Спокойного [5] получились хуже остальных алгоритмов, однако с увеличением объема выборки приближаются к истинному значению.

1.5. Анализ времени работы алгоритмов

Исходя из выводов предыдущей секции, реализация алгоритма В. Г. Спокойного в R обнаруживает скачки хуже, чем алгоритм Мунка и SaRa.

Рассмотрим время, за которое каждый из алгоритмов обнаруживает скачок.

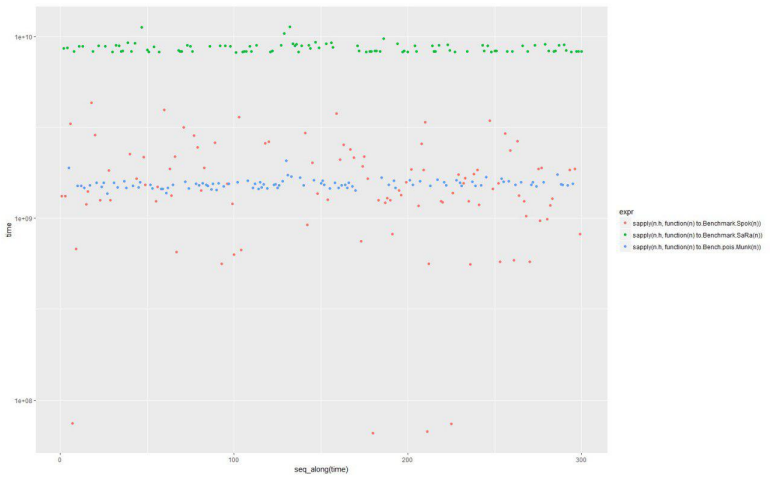

Рис. 1.7. График времени работы алгоритмов Мунка, В. Г. Спокойного и SaRa в зависимости от объема выборки

Голубым цветом обозначен алгоритм Мунка, красным обозначен алгоритм Спокойного, зеленым обозначен алгоритм SaRa (см. рис. **??**).

Заметим, что в среднем SaRa [1] работает дольше, чем алгоритмы Мунка [4] и Спокойного [5].

Глава 2

Методы обнаружения разладки

Для задачи множественной разладки основная цель заключается в оценке числа моментов разладки и их расположения. Существует несколько подходов определения моментов разладки [6]. Обозначим за J^* *истинное число моментов разладки*, а за $\tau^* = (\tau_1^*, \ldots, \tau_J^*)^{\mathrm{T}}$ — их *вектор локации* конкретных данных порождающего процесса.

2.1. Исчерпывающий поиск (exhaustive search)

Несмотря на вычислительную сложность,алгоритм исчерпывающего поиска может быть применим для всех возможных $0 \leq J \leq n-1$ и $0 < \tau_1 < \cdots < \tau_J < n$. Определим $\mathbf{\Gamma}_J = \{\tau = (\tau_1, \ldots, \tau_n) : 0 < \tau_1 < \cdots < \tau_J < n\}$ — множество всевозможных упорядоченных J–мерных векторов, которые характеризуют местоположения моментов разладки.

Для любого J определим:

$$\hat{\sigma}_J^2 = \min_{\tau \in \mathbf{\Gamma}_J} \hat{\sigma_\tau}^2,$$

где $\hat{\sigma}_\tau^2$ — оценка максимального правдоподобия для условной дисперсии на позиции τ момента разладки. Свойство состоятельности оценщика $\hat{J}$ определяется Байесовским информационным критерием (BIC):

$$\hat{J} = \operatorname{argmin}_J \frac{n}{2} \log \hat{\sigma}_J^2 + J \log n. \tag{2.1}$$

Также была показана скорость сходимости:

$$\hat{\tau}_\kappa - \tau_\kappa^* = O_P(1), 1 \leq \kappa \leq J^*,$$

где $\hat{\tau} = (\hat{\tau}_1, \ldots, \hat{\tau}_{J^*})^{\mathrm{T}} = \operatorname{argmin}_{\tau \in \mathbf{\Gamma}_{J^*}} \hat{\sigma}_\tau^2$ и J^* — истинное число моментов разладки, которое может быть оценено с помощью формулы (2.1)

Метод exhaustive search не является эффективным с вычислительной точки зрения и слишком трудоемкий при больших n. Пользуясь последовательностью структуры, можно применить технику динамического программирования, что может уменьшить вычислительную сложность до $O(N^2)$. Возможно также уменьшить ее до $O(N)$, однако в таком предположении, которое не применимо в приложениях.

2.2. Пошаговый отбор (stepwise selection)

Пошаговым отбором часто заменяют обыкновенный перебор ввиду понижения сложности и простоты реализации. Прямой и обратный методы отбора являются хорошо известными пошаговыми процедурами и оба могут быть применены к задаче разладки. Однако есть и свои тонкости, которые нужно учитывать при применении этих алгоритмов.

В частности основной недостаток *прямого метода отбора (BS — binary segmentation)* — плохо ищет короткие сегменты внутри длинных. Для увеличения мощности алгоритма было предложено использовать алгоритм циклической бинарной сегментации (CBS — Circular Binary Segmentation). Главным отличием от стандартного BS является то, что CBS тестирует альтернативу на каждом сегменте. Однако для больших n этот алгоритм является достаточно медленным, так как он считает все пары точек при расчетет тестовых статистик.

Кроме того, был предложен метод дикой бинарной сегмендации (WBS — Wild Binary Segmentation), который улучшает работу стандартного BS. В отличие от BS вместо глобального теста WBS считает тест на случайном сегменте, при этом значения тестовых статистик и истинного числа моментов разладки будут совпадать с BS, затем находит первый момент разладки на случайном интервале. Далее алгоритм повторяет тут же процедуру, пока не выполнятся условия остановки.

Обратный метод отбора (BWD — Backward selection approach) похож на кластерный анализ bottom–up, когда небольшие кластеры на нижних уровнях объединяются в более крупные на верхних уровнях. Алгоритм начинается с n групп, каждая из которых содержит по одному моменту разладки, а затем группы сливаются на основе некоторого критерия до момента остановки. Алгоритм BWD легко реализовать, и он имеет вычислительную сложность $O(n \log n)$.

2.3. Алгоритм отсеивания и ранжирования (SaRa)

Этот алгоритм представлен в статье [1]. Для модели нормального среднего считаем локально определенные статистики на каждой позиции $h \leq j \leq n - h$,

$$D_h(j) = \frac{\Big(\sum\limits_{k=j+1}^{j+h} Y_k - \sum\limits_{k=j-h+1}^{j} Y_k \Big)}{h},$$

где h — фиксированная ширина. По сути $|D_h(\cdot)|$ является тестовой статистикой отношения правдоподобия для задачи локального тестирования. Значит, последовательность $|D_h(\cdot)|$ в каком–то смысле измеряет относительную вероятность быть моментом разладки для каждой позиции.

Алгоритм *SaRa* может быть описан, как модифицированный метод бинарной сегментации (раздел 2.2). На первом этапе рассчитывается $|D_h(\cdot)|$, и ее глобальный максимизатор делит последовательность на два сегмента. Затем в каждом сегменте мы повторяем ту же процедуру, пока максимум $|D_h(\cdot)|$ ниже некоторого значения для каждого сегмента. Таким образом, разницей между SaRa и BS является использование локальной или глобальной тестовой статистики отношения правдоподобия в качестве основы оптимизации. SaRa имеет несколько преимуществ:

- тестовые статистики не нужно пересчитывать на каждом шаге,
- единое пороговое значение можно применять одновременно
- в рамках SaRa можно решить проблему разладки с помощью множественной проверки гипотез,
- вычислительная сложность — $O(N)$, так как последовательность $|D_h(\cdot)|$ может быть вычислена по рекуррентной формуле.

2.4. Алгоритм одновременного многомасштабного оценивания скачков (SMUCE)

Этот алгоритм прелставлен в источнике [4]. Данный алгоритм $SMUCE$ — *Simultaneous MUltiscale Change–point Estimator* позволяет оценить число скачков, их расположение, значение функции в этих точках, а также доверительные интервалы для моментов разладки и самой функции.

Рассмотрим независимые случайные величины $Y = (Y_1, \ldots, Y_n)$, подчиняющиеся закону распределения:

$$Y_i \sim F_{\nu(i/n)}, i = 1, \ldots, n, \tag{2.2}$$

где $\{F_\theta\}_{\theta\in\Theta}$ — одномерное экспоненциальное семейство с плотностьями f_θ и $\nu : [0,1) \longrightarrow \Theta \subseteq \mathbb{R}$ ступенчатая непрерывная справа функция с неизвестным числом K скачков.

Определим множество $\mathcal{S}$, как пространство всех непрервыных справа ступенчатых функций с произвольным, но конечным числом скачков на единичном интервале $[0,1)$ со значениями в некоем множестве Θ. Для $\nu \in \mathcal{S}$ определим за $J(\nu)$ упорядоченный вектор моментов разладки и за $\#J(\nu)$ — его длину, то есть количество скачков.

На первом шаге алгоритму $SMUCE$ требуется решить (невыпуклую) оптимизационную задачу:

$$\inf_{\nu\in\mathcal{S}} \#J(\nu), \tag{2.3}$$

так, что $T_n(Y,\nu) \leq q$, где q — пороговое значение, которое будет определено позднее. $T_n(Y,\nu)$ — некоторая многомасштабная статистика для функции–"кандидата" $\nu \in \mathcal{S}$. Статистика T_n оценивает максимум над локальной статистикой отношения правдоподобия на всех дискретных интервалах $[i/n, j/n]$, таких, что ν на этих интервалах — константа со значением $\theta = \theta_{ij}$, то есть:

$$T_n(Y,\nu) = \max_{\substack{1\leq i<j\leq n \\ \nu(t)=\theta, t\in[i/n,j/n]}} \Big(\sqrt{2T_i^j(Y,\theta)} - \sqrt{2\log\frac{en}{j-i+1}}\Big), \tag{2.4}$$

где локальная статистика отношения правдоподобия T_i^j для проверки гипотезы $H_0 : \theta = \theta_0$ против альтернативы $H_1 : \theta \neq \theta_0$ на интервале $[i/n, j/n]$ определяется как:

$$T_i^j(Y,\nu) = \log\Big(\frac{\sup_{\theta\in\Theta}\prod_{l=i}^j f_\theta(Y_l)}{\prod_{l=i}^j f_{\theta_0}(Y_l)}\Big). \tag{2.5}$$

Эта статистика измеряет, как хорошо данные могут быть описаны с помощью константы θ_0 на интервале $[i/n, j/n]$.

Алгоритм $SMUCE$ проводит многомасштабное тестирование следующим образом: выбор модели (оценивание K), затем оценивание ν (зная K). Минимальное значение $\#J$ дает оценку числа скачков. Далее определим множество всех решений в соответствии с условиями и выбором трешхолда q:

$$\mathcal{C}(q) = \{\nu \in \mathcal{S} : \#J(\nu) = \hat{K}(q), T_n(Y,\nu) \leq q\}. \tag{2.6}$$

Затем $SMUCE$ определяет $\hat{\nu}(q)$ таким образом:

$$\hat{\nu}(q) = \underset{\nu \in \mathcal{C}(q)}{\operatorname{argmax}} \sum_{i=1}^{n} \log(f_{\nu(i/n)}(Y_i)). \tag{2.7}$$

2.5. Обнаружение скачков с помощью диадических интервалов

Задача обнаружения скачков функции может быть переформулирована в задачу нахождения геометрических объектов в зашумленных данных. На вход подаются данные, необходимо найти множества, на которых математическое ожидание будет повышенным (места, где возможно отделить какой-либо сегмент), на этих множествах и будут скачки.

В статье [2] рассмотрен класс методов, предназначенных для обнаружения различных геометрических объектов с помощью быстрых алгоритмов, подробно разобран одномерный случай, а также даны предпосылки для многомерного случая.

2.5.1. Метрическое пространство интервалов

Пусть n бинарное целочисленное $n = 2^J$ и пусть $\mathbb{J}_n$ — коллекция всевозможных диадических подынтервалов

$$I_{j,k} = \{k2^j, ...(k+1)2^j - 1\},$$

где $0 \le j \le \log_2(n)$ и $0 \le k < n/2^J$. Обозначим мощность интервала I за $|I|$.

Любой интервал тесно связан с его максимально возможным диадическим интервалом. То есть на самом деле можем ввести некоторую меру зависимости между интервалами (*афинная мера*), которая будет иметь вид:

$$\rho(I, J) = \frac{|I \cap J|}{\sqrt{|I||J|}}. \tag{2.8}$$

Лемма 1 (доказательство и пояснения к лемме представлены в статье [2])**.** *Диадические интервалы $\mathbb{J}_n$ образуют ε-сеть для коллекции $\mathbb{I}_n$ всех интервалов, причем $\varepsilon = \sqrt{1/2}$.*

Опишем, как происходит поиск “перспективных” диадических интервалов для размерности $d = 1$ (ясно, что этот поиск будет аналогичным и для многомерного случая).

Конечно, значение $\varepsilon = \sqrt{1/2}$ не является малым. Вызывает вопрос тот факт, что мы допускает $\varepsilon < 1$ независимо от n. Чтобы получить приближения с малыми ε, мы используем диадические интервалы в качестве “базы”и формируем смеси интервалов, добавляя дополнительные бинарные интервалы на концы уже существующего. Формально говорим, что интервал J_l — расширение уровня l, если оно было получено следующим образом:

1) Начинаем с базового интервала J_0, который является либо диадическим интервалом $I_{j,k}$, либо объединением двух соседних интервалов $I_{j,k}$ и $I_{j,k+1}$, где k — нечетное.

2) На стадиях $g = 1, ..., l$ продлеваем J_{g-1}, чтобы получить J_g, прикрепляя к одному или обоим концам интервала J_{g-1} диадические интервалы длины $2^{-g}|I_{j,k}|$, или ничего не делая (то есть $J_{g-1} = J_g$).

Коллекцию всех расширений уровня l для диадического интервала I обозначим за $\mathbb{J}_l[I]$. Множество всех расширений уровня l обозначим $\mathbb{J}_{n,l}$.

Лемма 2. $\#\mathbb{J}_n \leq n4^{l+1}$
$\rho^*_{n,l} = \min\limits_{I\in\mathbb{I}_n} \max\limits_{J\in\mathbb{J}_n} \rho(I, J) \geq 1/\sqrt{1 + 2 \cdot 2^{-l}}$.

2.5.2. Одномерный случай

Пусть $X = (x(i) : 0 \leq i < n)$ — массив случайных величин, содержащих белый шум, кроме, возможно, интервала $S = [a; b]$, на котором математическое ожидание повышенное. Обозначим $\mathbb{S}_n$ за множество всевозможных интервалов с максимальной правой границей n. Случайные величины $x(i)$ имеют вид:

$$x(i) = \mu \cdot \mathbb{I}_{\{a\leq i<b\}} + z(i), i = 0, ..., n - 1. \tag{2.9}$$

- a, b — границы интервала, удовлетворяющие условию: $0 \leq a < b \leq n$, но заранее неинзвестны,
- μ — амплитуда сигнала. Однако принято работать с *нормализованной* амплитудой: $A = \mu/\sqrt{b - a}$
- $z(i)$ — независимые одинаково распределенные случайные величины (гауссовский белый шум)

Наша задача определить, можно ли отделить в конкретном случае сигнал от шума. Для этого мы проверяем соответствующую гипотезу: $H_0 : A = 0$, против альтернативы: $H_{1,n} : A > 0, S \in \mathbb{S}_n$.

Пусть величина $\xi_S(i) = \xi_{a,b}(i) = \mathbb{I}_{\{a \le i < b\}}/\sqrt{b-a}$ — это l^2-нормализованный прототип интервала. Статистика критерия:

$$X[S] = \langle \xi_S, X \rangle = \sum_i \xi_S(i) x(i),$$

позволяет нам применить тест Неймана–Пирсона для H_0 против $H_{1,S,A}$, проверяя неравенство:

$$X[S] > t,$$

для некоторого threshold'a t. По сути, это переформулировка критерия отношения правдоподобия. Для смеси альтернативных гипотез, где $A > 0$ и $S \in \mathbb{S}_n$ (и оба неизвестны), нам необходимо провести следующий тест:

$$X_n^* \equiv \max_{S \in \mathbb{S}_n} X[S] > t_{n,\alpha}^*,$$

где $t_{n,\alpha}^*$ — подходящее пороговое значение, рассчитанное, чтобы сохранить общий $\alpha - level$ полученного теста при выполнении гипотезы H_0. Такой подход будем называть обобщенным критерием отношения правдоподобия (GLRT — Generalized Likelihood Ratio Test). Отдельной задачей будет определение подходящего threshold'a. Теоретические факты, предложенные для решения этой проблемы, представлены в статье [2].

Пусть множество $\mathbb{I}_n$ определяет коллекцию всех подынтервалов $\{0, ..., n-1\}$. Рассматриваем случайное поле $X[I]$, определенное на $I \in \mathbb{I}_n$. Статистика X_n^* — максимум по всей коллекции X. Предложенный подход обращает внимание на *диадические интервалы*, как особенные подмножества коллекции всех интервалов. Нас интересуют бинарные интервалы, так как их мощность равна $2n$, а не $\approx n^2/2$, тогда построим ε-сеть в пространстве интервалов в специальной "метрике обнаружения", которую мы определим ниже при $\varepsilon < 1$.

В данной статье предложен алгоритм, основанный на расширении "перспективных" бинарных интервалов, определенных в разделе 2.5.1. Зафиксируем $\eta > 0$ и, возвращаясь к определению (2) $\rho_{n,l}^*$, зафиксируем l так, что

$$\rho_{n,l}^* > \sqrt{\frac{1+\eta/2}{1+\eta}}.$$

1) *Поиск "перспективных" диадических интервалов.* Определим все бинарные интервалы $I \in \mathbb{I}_n$ с помощью

$$X[I] > \frac{1}{3}\sqrt{\log(n)}.$$

Если имеется более $n^{19/20}$ таких интервалов, чем $2n$ диадических, тогда отвергаем $H_{0,n}$.

2) *Расширение "перспективных" интервалов.* Для каждого интервала I, найденного на стадии 1, пронумеруем все расширения уровней l

$$X_l^*[I] = \max\{X[I'] : I' \in \mathbb{J}_l[I]\}.$$

3) *Принятие решения.* Если максимум $X[I]$ на стадии 1 превышает $\sqrt{2\log(n)}$, или максимум $X_l^*[I]$ на стадии 2 превышает $\sqrt{2(1+\eta/3)\log(n)}$, отвергаем $H_{0,n}$.

Данный алгоритм является эффективным для вычисления по построению. Подробное разъяснение и доказательство этого представлены в статье [2].

2.6. Проверка гипотез: false discovery

В некоторых из перечисленных методов, чтобы найти точку скачка, проверяется множество гипотез.Таким образом, необходимо контролировать среднюю долю множественных отклонений нулевой гипотезы, а именно *FDR (False Discovery Rate)*. В предложенных методах FDR определяется с помощью поправки Бонферрони. В этом исследовании при дальнейшем рассмотрении методов планируется рассмотреть более тонкий подход нахождения FDR. Рассмотрим некоторые способы определения *false discovery*, и каким образом его контролировать [7].

Для данного $m \geq 1$, допустим, что $\Theta_i \subseteq \mathfrak{R}^{d_i}, 1 \leq d_i < \infty$, и $A_i \subset \Theta_i, i = 1, \ldots, m$. Пусть $X \sim \mathbb{P}_{\theta_1,\ldots,\theta_m}$, $\theta_i \in \Theta_i, i = 1, \ldots, m$ и $X \in \mathfrak{R}^d, d \geq 1$.

Определим нулевую гипотезу H_{0i}: $\theta_i \in A_i, i = 1, \ldots, m$.

Определим тестовую статистику $T_i = T_i(X), i = 1, \ldots, m$ и будем предполагать также, что гипотеза H_{0i} отвергается, если $T_i > c_i$, то есть нулевая гипотеза отвергается при больших значениях тестовой статистики. Пусть с учетом наших данных p_i будет означать p–value, основанные на статистике T_i. Здесь гипотеза H_{0i} такая, что все p–значения определены и равномерно распределены под соответствующими нулями. Для некоторого m_0, $0 \leq m_0 \leq m$, предположим, что m_0 нулевых гипотез не отверглись. Обозначим через R общее количество отказов из m тестов в общем применении данной последовательности тестов и определим:

- $W = m - R$,
- $V = \sum_{i=1}^{m} \mathbb{I}_{T_i > c_i \cap \{H_{0i} istrue\}}$ — количество отвергнутых правдивых гипотез (false discovery),
- $S = R - V$ и $U = m_0 - V$ — соответствующие частоты “правильных действий”,
- $T = m - m_0 - S$ — число неотвергнутых неверных гипотез.

Разумно было бы сохранять U и S большими, а V и T малыми. Однако обычно критические области статистических тестов являются вложенными, поэтому желание сохранить и V, и T малыми противоречиво. Можно пытаться сохранять V малым или найти баланс, чтобы V и T были минимальными. Ниже приведены несколько способов контроля ложных открытий:

- **Familywise error rate (FWER)** $\mathbb{P}_{\theta_1,\ldots,\theta_m}(V \geq 1)$

- **False-discovery rate (FDR)** $\mathbb{E}_{\theta_1,\ldots,\theta_m}(Q)$, $Q = \frac{V}{R}\mathbb{I}_{R>0} + 0\mathbb{I}_{R=0} = \frac{V}{R \vee 1}$,
- **Marginal false-discovery rate (MFDR)** $\frac{\mathbb{E}_{\theta_1,\ldots,\theta_m}(V)}{\mathbb{E}_{\theta_1,\ldots,\theta_m}(R)}$,
- **Marginal realized false-discovery rate (MRFDR)** $\frac{\mathbb{E}_{\theta_1,\ldots,\theta_m}(V)}{r}\mathbb{I}_{r>0} + 0\mathbb{I}_{r=0}$, где r — реализованное значение R,
- **False-discovery proportion distribution (FDPD)** $\mathbb{P}_{\theta_1,\ldots,\theta_m}(Q > q)$,
- **Positive false-discovery proportion distribution (PFDPD)** $\mathbb{P}_{\theta_1,\ldots,\theta_m}(Q > q|R > 0)$,
- **Positive false-discovery rate (PFDR)** $\mathbb{E}_{\theta_1,\ldots,\theta_m}(Q|R > 0)$.

Известно, что вычислять false discovery (FD) при фиксированном объеме выборок достаточно просто. Интересно посмотреть, можно ли вычислять FD при растущем объеме наблюдений, то есть в процессе получения данных на каждом шаге.

Глава 3

Обнаружение и оценивание скачков для различного шума

В статье [5] был предложен адаптивный метод обнаружения скачков для нормального распределения с различными оценками. В этом разделе будет предложена модификация алгоритма [5] для пуассновских случайных процессов, продолжение исходного алгоритма для двумерного случая и асимптотические оценки для верхней и нижней границ ширины интервала (области) со скачком для всех случаев.

3.1. Гауссовский шум

3.1.1. Одномерный случай

Верхняя граница ширины интервала

Рассмотрим модель регрессии:

$$Y_i = f(X_i) + \xi_i, i = 1, ..n, \tag{3.1}$$

где $i = 1, ..n$, $X_i, Y_i \in \mathbb{R}$, ξ_i — независимые случайные величины нормального распределения ($\mathcal{N}(0, \sigma^2)$), где σ^2 — дисперсия. Оценка нижней границы ширины интервала была приведена в статье [5], поэтому в этом разделе для одномерного случая будет приведена только оценка верхней границы ширины интервала скачка.

Так как рассматриваем оценку верхней границы ширины скачка, необходимо действовать следующим образом: рассматривается сетка с шагом ε, т.е. разобьем отрезок $[0, 1]$ на множество подынтервалов с шагом ε. При этом рассматриваться будет не одно разбиение, как в случае с нижней границей, а множество различных решеток с шагом ε, которые будут покрывать отрезок $[0, 1]$.

На рис. 3.1 приведен пример разбиения интервала на множество подынтервалов с шагом ε, когда имеется только одна решетка. Красным выделен искомый интервал. Заметим, что при таком разбиении ни одно окно разбиения не покрывает искомый интервал. Поэтому необходимо рассматривать множество решеток.

Рис. 3.1. Пример случая, когда искомый интервал не покрывается окном шириной ε

Таким образом, искомый интервал (как на рис. 3.1) не покрывается окном шириной ε, поэтому необходимо рассмотреть множество решеток. Рассмотрим ε^{-1} таких решеток. Обозначим множество всех решеток за $\mathcal{W}$.

Таким образом, для верхней границы необходимо показать следующее:

- Если скачок есть, то вероятность того, что мы его обнаружим $\longrightarrow 1$;
- Если скачка нет то вероятность того, что мы его найдем $\longrightarrow 0$.

Рассмотрим ширину скачка вида: $Cn^{-1}\ln n$, и рассмотрим константу $C = (2+\varepsilon)$. Такой вид ширины скачка (без учета константы) был приведен в статье [5].

Скачок отсутствует

В этом случае необходимо показать, чтот если скачок отсутствует, то вероятность обнаружить его будет стремиться к 0.

Рассмотрим все имеющиеся наблюдения $\xi_i \sim \mathcal{N}(0, \sigma^2)$ (рассматривать Y_i нет смысла, так как нам известно, что скачок отсутствует, соответственно, все наблюдения подчиняются гауссовскому закону распределения). Возьмем максимум из них, и также максимум по всем решеткам (как было определено ранее).

Аналогично статье [5] формализуем вышеописанное событие, и в итоге получим следующее: взятый выше максимум больше, чем некоторое заранее выбранное пороговое значение X_{bound}.

$$\max_{\mathcal{W}} \max_{i} \xi_i > X_{bound}. \tag{3.2}$$

Обозначим за $X_{bound} = (2 + 2\varepsilon/3)\ln n$. Так как мы считаем, что скачка нет, то $\forall i\ \xi_i = \sum\limits_{j=1}^{(2+\varepsilon)\ln n} \zeta_j \sim \mathcal{N}(0, (2+\varepsilon)\ln n)$, где $(2+\varepsilon)\ln n$ — дисперсия σ^2. Число точек в интервале со скачком также равно $(2+\varepsilon)\ln n$.

Таким образом, необходимо показать:

$$\mathrm{P}\left(\max_{\mathcal{W}}\max_{i}\xi_i > (2+2\varepsilon/3)\ln n\right) \stackrel{?}{\longrightarrow} 0. \tag{3.3}$$

Отнормируем ξ_i: $\eta_i = (\sqrt{(2+\varepsilon)\ln n})^{-1}\xi_i \sim \mathcal{N}(0,1)$

Подставим η_i в выражение:

$$\mathrm{P}\left(\max_{\mathcal{W}}\max_{i}\xi_i > (2+2\varepsilon/3)\ln n\right) = \mathrm{P}\left(\max_{\mathcal{W}}\max_{i}\eta_i > \frac{(2+2\varepsilon/3)\ln n}{\sqrt{(2+\varepsilon)\ln n}}\right). \tag{3.4}$$

Оценим получившееся выражение в правой части неравенства:

$$\frac{(2+2\varepsilon/3)\ln n}{\sqrt{(2+\varepsilon)\ln n}} \asymp \sqrt{(2+\varepsilon/3)\ln n}. \tag{3.5}$$

По формуле (3.5) получаем:

$$\begin{aligned}
&\mathrm{P}\left(\max_{\mathcal{W}}\max_{i}\eta_i > \frac{(2+2\varepsilon/3)\ln n}{\sqrt{(2+\varepsilon)\ln n}}\right) = \mathrm{P}\left(\max_{\mathcal{W}}\max_{i}\eta_i > \sqrt{(2+\varepsilon/3)\ln n}\right) \leq \\
&\leq \sum_{W}\mathrm{P}\left(\max_{i}\eta_i > \sqrt{(2+\varepsilon/3)\ln n}\right) \leq \sum_{\mathcal{W}}\sum_{i}\mathrm{P}\left(\eta_i > \sqrt{(2+\varepsilon/3)\ln n}\right) \leq \\
&\leq \varepsilon^{-2}\cdot n\cdot\frac{e^{-\frac{(\sqrt{(2+\varepsilon/3)\ln n})^2}{2}}}{2\pi\sqrt{(2+\varepsilon/3)\ln n}}.
\end{aligned} \tag{3.6}$$

В последнем переходе суммирование по решеткам превращается в коэффициент ε^{-2} (число решеток), а суммирование по всем i превращается в n, то есть число всех наблюдений.

После этого преобразуем полученное выражение к виду:

$$\varepsilon^{-2}\cdot n\cdot\frac{e^{-\frac{(\sqrt{(2+\varepsilon/3)\ln n})^2}{2}}}{2\pi\sqrt{(2+\varepsilon/3)\ln n}} = \frac{\varepsilon^{-2}n}{2\pi\sqrt{(2+\varepsilon/3)\ln n}}e^{-(1+\varepsilon/6)\ln n} = \frac{\varepsilon^{-2}n^{-\varepsilon/6}}{2\pi\sqrt{(2+\varepsilon/3)\ln n}} \tag{3.7}$$

Таким образом:

$$\frac{\varepsilon^{-2}n^{-\varepsilon/6}}{2\pi\sqrt{(2+\varepsilon/3)\ln n}} \underset{n\longrightarrow\infty}{\longrightarrow} 0. \tag{3.8}$$

В итоге показали, что и требовалось:

$$\mathrm{P}\left(\max_{\mathcal{W}}\max_{i}\xi_i > (2+2\varepsilon/3)\ln n\right) \longrightarrow 0. \tag{3.9}$$

Скачок есть

Теперь необходимо показать, что при наличии скачка вероятность его обнаружить стремится к 1.

Аналогично предыдущему пункту (3.1.1) и доказательству теоремы в статье В.Г. Спокойного, рассматривается событие:

$$\max_{\mathcal{W}} \max_{i} \xi_i > (2 + 2\varepsilon/3) \ln n. \tag{3.10}$$

Покажем, что:

$$\mathrm{P}\left(\max_{\mathcal{W}} \max_{i} \xi_i > (2 + 2\varepsilon/3) \ln n\right) \xrightarrow{?} 1. \tag{3.11}$$

Так как скачок существует (по предположению), рассмотрим только те величины, которые попали в интервал скачка (I_{jump}):

$$\mathrm{P}\left(\max_{\mathcal{W}} \max_{i} \xi_i > (2 + 2\varepsilon/3) \ln n\right) \geq \mathrm{P}\left(\max_{\mathcal{W}} \max_{j \in I_{jump}} \xi_j > (2 + 2\varepsilon/3) \ln n\right). \tag{3.12}$$

Заметим, что величины, попавшие в интервал скачка имеют такое распределение: $\xi_j \sim \mathcal{N}((2 + \varepsilon) \ln n, (2 + \varepsilon) \ln n)$.

Приведем полученное выражение (предыдущая формула), отнормировав ξ_j:

$$\mathrm{P}\left(\max_{\mathcal{W}} \max_{j \in I_{jump}} \xi_j > (2 + 2\varepsilon/3) \ln n\right) = \mathrm{P}\left(\max_{\mathcal{W}} \max_{j \in I_{jump}} \zeta_j + \sqrt{(2 + \varepsilon) \ln n} > \frac{(2 + 2\varepsilon/3) \ln n}{\sqrt{(2 + \varepsilon) \ln n}}\right). \tag{3.13}$$

Оценим $\frac{(2+2\varepsilon/3) \ln n}{\sqrt{(2+\varepsilon) \ln n}}$. Аналогично предыдущему пункту получаем:

$$\frac{(2 + 2\varepsilon/3) \ln n}{\sqrt{(2 + \varepsilon) \ln n}} = \sqrt{(2 + \varepsilon/3) \ln n}. \tag{3.14}$$

Подставим его в формулу:

$$\begin{aligned} &\mathrm{P}\left(\max_{\mathcal{W}} \max_{j \in I_{jump}} \zeta_j + \sqrt{(2 + \varepsilon) \ln n} > \sqrt{(2 + \varepsilon/3) \ln n}\right) = \\ &= \mathrm{P}\left(\max_{\mathcal{W}} \max_{j \in I_{jump}} \zeta_j > \sqrt{(2 + \varepsilon/3) \ln n} - \sqrt{(2 + \varepsilon) \ln n}\right) \end{aligned} \tag{3.15}$$

Подставим полученное выражение:

$$\begin{aligned}
&\mathrm{P}\left(\max_{\mathcal{W}}\max_{j\in I_{jump}}\zeta_j > (\sqrt{(2+\varepsilon/3)}-\sqrt{(2+\varepsilon)})\sqrt{\ln n}\right) = \\
&= \mathrm{P}\left(\max_{\mathcal{W}}\max_{j\in I_{jump}}\zeta_j < (\sqrt{(2+\varepsilon)}-\sqrt{(2+\varepsilon/3)})\sqrt{\ln n}\right) = \\
&= 1 - \mathrm{P}\left(\max_{\mathcal{W}}\max_{j\in I_{jump}}\zeta_j > (\sqrt{(2+\varepsilon)}-\sqrt{(2+\varepsilon/3)})\sqrt{\ln n}\right).
\end{aligned} \tag{3.16}$$

Теперь необходимо показать:

$$\mathrm{P}\left(\max_{\mathcal{W}}\max_{j\in I_{jump}}\zeta_j > (\sqrt{(2+\varepsilon)}-\sqrt{(2+\varepsilon/3)})\sqrt{\ln n}\right) \overset{?}{\longrightarrow} 0. \tag{3.17}$$

Аналогично предыдущему пункту (3.1.1):

$$\begin{aligned}
&\mathrm{P}\left(\max_{\mathcal{W}}\max_{j\in I_{jump}}\zeta_j > (\sqrt{(2+\varepsilon)}-\sqrt{(2+\varepsilon/3)})\sqrt{\ln n}\right) \leq \\
&\leq \varepsilon^{-2}(2+\varepsilon)\ln n\,\mathrm{P}\left(\zeta_j > (\sqrt{(2+\varepsilon)}-\sqrt{(2+\varepsilon/3)})\sqrt{\ln n}\right) \leq \\
&\leq \varepsilon^{-2}(2+\varepsilon)\ln n\frac{e^{-\frac{(\sqrt{(2+\varepsilon)}-\sqrt{(2+\varepsilon/3)})^2\ln n}{2}}}{2\pi\cdot(\sqrt{(2+\varepsilon)}-\sqrt{(2+\varepsilon/3)})\sqrt{\ln n}}.
\end{aligned} \tag{3.18}$$

Преобразуем полученное выражение:

$$\begin{aligned}
&\varepsilon^{-2}(2+\varepsilon)\ln n\frac{e^{-\frac{(\sqrt{(2+\varepsilon)}-\sqrt{(2+\varepsilon/3)})^2\ln n}{2}}}{2\pi\cdot(\sqrt{(2+\varepsilon)}-\sqrt{(2+\varepsilon/3)})\sqrt{\ln n}} \asymp \\
&\asymp \varepsilon^{-2}(2+\varepsilon)\sqrt{\ln n}\frac{e^{-(2+2\varepsilon/3-2\sqrt{1+2\varepsilon/3})\ln n}}{2\pi\cdot(\sqrt{(2+\varepsilon)}-\sqrt{(2+\varepsilon/3)})} \asymp \\
&\asymp \varepsilon^{-2}(2+\varepsilon)\sqrt{\ln n}\frac{n^{-(2+2\varepsilon/3-2\sqrt{1+2\varepsilon/3})}}{2\pi\cdot(\sqrt{(2+\varepsilon)}-\sqrt{(2+\varepsilon/3)})}.
\end{aligned} \tag{3.19}$$

В итоге получаем, что:

$$\varepsilon^{-2}(2+\varepsilon)\sqrt{\ln n}\frac{n^{-(2+2\varepsilon/3-2\sqrt{1+2\varepsilon/3})}}{2\pi\cdot(\sqrt{(2+\varepsilon)}-\sqrt{(2+\varepsilon/3)})} \underset{n\longrightarrow\infty}{\longrightarrow} 0. \tag{3.20}$$

Таким образом, вероятность исполнения исходно рассматриваемого события стремится к 1, что и требовалось показать:

$$\mathrm{P}\left(\max_{\mathcal{W}}\max_{i}\xi_i > (2+2\varepsilon/3)\ln n\right) \longrightarrow 1. \tag{3.21}$$

3.1.2. Двумерный случай

Нижняя граница ширины области

Рассмотрим “квадрат” $[0;1] \times [0;1]$ и модель регрессии для двумерного случая:

$$Y_{i,j} = f(\mathbb{X}_{i,j}) + \xi_{i,j}, \tag{3.22}$$

где $i, j = 1, \ldots, n$. Разделим нашу область на M^2 “квадратов” (т.е. каждый отрезок на M интервалов, каждый из который шириной $\mathrm{h}(n) = \mathcal{C} n^{-1} \log n$, $M = 1/\mathrm{h}(n)$). Обозначим такое разбиение за $\mathcal{F}$. Рассмотрим константу $\mathcal{C} = \sqrt{2 - \varepsilon}$

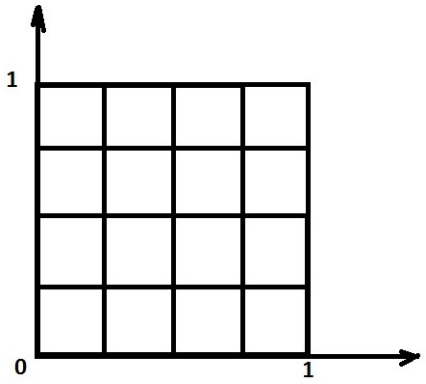

Рис. 3.2. Простейший пример разбиения квадрата

При этом каждая полученная область из разбиения $\mathcal{F}$ содержит ровно N точек плана, $N = n\,\mathrm{h}(n) + 1)$. Теперь предположим, что наша функция f имеет такой вид:

$$f_I(\mathbb{X}_{i,j}) = \begin{cases} 1, \mathbb{X}_{i,j} \in I \\ 0, \mathbb{X}_{i,j} \notin I, \end{cases}$$

где $\mathbb{X}_{i,j}$ соответствует паре точек (x_i, x_j).

Теперь исходная задача сводится к задаче нахождения оценки размера области I (как элемента конечного множества $\mathcal{F}$) на основе данных.

Пусть $Z_{I,n}$ — логарифм отношения правдоподобия:

$$Z_{I,n} = \log\Big(\frac{d\mathbb{P}_{f_I}}{d\mathbb{P}_0}\Big), \tag{3.23}$$

где $\mathbb{P}_0$ соответствует функции $f \equiv 0$.

Напоминаем, что мы рассматриваем центрированное нормальное распределение, то есть $\mathcal{N}(0, \sigma^2)$.

Таким образом, в нашем случае $Z_{I,n}$ выглядит так:

$$Z_{I,n} = \frac{1}{2} \sum_{(\frac{i}{n};\frac{j}{n})\in I} [Y_{i,j}^2 - (Y_{i,j} - 1)^2] = \sum_{(\frac{i}{n};\frac{j}{n})\in I} Y_{i,j} - N/2. \tag{3.24}$$

Замечание 1. *Композиция символов* $\sum\limits_{(\frac{i}{n};\frac{j}{n})\in I}$ *означает двойное суммирование по индексам* i *и* j*, но так, что пара* (i, j) *являлась парой индексов для точки плана* (x_i, x_j)*. То есть на самом деле, много сочетаний различных* i *и* j *приэтом суммировании должно обнулиться.*

В итоге получаем, что байесовская оценка $\hat{I}$ выглядит как:

$$\hat{I} = \underset{I}{\operatorname{argmax}} Z_I.$$

Зафиксируем теперь произвольно область $I_0 \in \mathcal{F}$, и рассмотрим теперь вероятность $\mathbb{P}_{I_0}(\hat{I} \neq I_0)$, где $\mathbb{P}_{I_0}$ — мера, относящаяся к f_{I_0}.

Отметим, что с мерой $\mathbb{P}_{I_0}$ с вероятностью 1 выполняется следующее:

$$\sum_{I_0} Y_{i,j} = \sqrt{N} \cdot \zeta_{I_0} + N,$$

$$\sum_{I} Y_{i,j} = \sqrt{N} \cdot \zeta_I, I \neq I_0.$$

При этом $\zeta_I = \frac{1}{\sqrt{N}} \sum\limits_I \xi_{i,j} \sim \mathcal{N}(0, 1)$ (согласно ЦПТ).

Теперь:

$$\mathbb{P}_{I_0}(\hat{I} \neq I_0) = \mathbb{P}(\max_{I\neq I_0} \zeta_I - \sqrt{N}/2 > \zeta_{I_0} + \sqrt{N}/2) = \mathbb{P}(\max_{I\in\mathcal{F}} \zeta_I > \sqrt{N}). \tag{3.25}$$

Согласно ссылке в статье Спокойного можем привести полученное выражение при $\alpha < 2$:

$$\mathbb{P}(\max_{I\in\mathcal{F}} \zeta_I > \sqrt{N}) = \mathbb{P}(\max_{I\in\mathcal{F}} \zeta_I > \sqrt{\alpha \log M^2}). \tag{3.26}$$

Введем замену: $\sqrt{\alpha \log M^2} = \mathcal{C}_{\mathcal{M}}$. Подставим ее в формулу (3.26):

$$\mathbb{P}(\max_{I\in\mathcal{F}} \zeta_I > \mathcal{C}_M) = 1 - (\mathbb{P}(\xi_i < \mathcal{C}_M))^{M^2} = 1 - (1 - \mathbb{P}(\xi_i > \mathcal{C}_M))^{M^2} = 1 - e^{M^2 \log(1-\mathbb{P}(\xi_i > \mathcal{C}_M))}. \tag{3.27}$$

Заменим логарифм:

$$1 - e^{M^2 \log(1-\mathbb{P}(\xi_i > \mathcal{C}_M))} = 1 - e^{-M^2 \mathbb{P}(\xi_i > \mathcal{C}_M)}. \tag{3.28}$$

Напомним, что хвост нормального распределения выглядит так:

$$\mathbb{P}(\xi_i > \mathcal{C}_M) = \frac{1}{\sqrt{2\pi}\mathcal{C}_M} e^{-\mathcal{C}_M^2/2}. \tag{3.29}$$

Подставим (3.29) в получившееся выражение (3.28):

$$1 - e^{-M^2 \mathbb{P}(\xi_i > \mathcal{C}_M)} = 1 - e^{-\frac{M^2}{\sqrt{2\pi\alpha \log(M^2)}} e^{-\frac{\alpha \log M^2}{2}}} = 1 - e^{-\frac{M^2}{\sqrt{2\pi\alpha \log M^2}} M^{-\alpha}}. \tag{3.30}$$

Обозначим $\tilde{M} = M^2$:

$$1 - e^{-\frac{M^2}{\sqrt{2\pi\alpha \log M^2}} M^{-\alpha}} = 1 - e^{-\frac{\tilde{M}}{\sqrt{2\pi\alpha \log \tilde{M}}} \tilde{M}^{-\alpha/2}} \longrightarrow 1. \tag{3.31}$$

Верхняя граница ширины области

Аналогично одномерному случаю приводятся доказательства для квадратной сетки.

Разобьем область $[0,1] \times [0,1]$ на множество подобластей (“квадратов”) со стороной ε. При этом рассматриваться будет не одно разбиение, как в случае с нижней границей, а множество различных решеток с шагом ε. Так как в данном случае рассматривается сетка на плоскости, число таких решеток будет ε^{-2}. Обозначим множество решеток аналогично предыдущему пункту: $\mathcal{W}$.

Таким образом, для верхней границы необходимо показать следующее:

- Если скачок есть, то вероятность того, что мы его обнаружим $\longrightarrow 1$;
- Если скачка нет то вероятность того, что мы его найдем $\longrightarrow 0$.

Напомним, что сторона “квадрата” (области, где обнаружен скачок) имеет вид: $Cn^{-1} \ln n$, и рассмотрим $C = \sqrt{2+\varepsilon}$.

Доказательство для верхней оценки скачка проводится аналогично одномерому случаю с точностью до числа решеток.

3.2. Пуассоновский случайный процесс

3.2.1. Одномерный случай

Пусть наблюдается неоднородный пуассоновский процесс $X(t)$, при $t \in [0,1]$, постоянной интенсивности $\lambda(t) = n$ за исключением малого интервала $[a_0, a_1] \in [0,1]$, где $0 < a_0 < a_1 < 1$, а интенсивность скачка $\lambda(t) = an$, $(a > 1)$. Обозначим ширину этого интервала за $\mathrm{h}(n)$.

Рассмотрим задачу о проверке гипотезы H_0 о том, что интенсивность процесса Пуассона постоянна и равна n, против альтернативы H_1 о том, что интенсивность процесса Пуассона имеет вид:

$$\lambda(t) = \begin{cases} na, & t \in [a_0, a_1] \\ n, & t \notin [a_0, a_1]. \end{cases} \tag{3.32}$$

Изучим вопрос: при каких $\mathrm{h}(n)$ существует состоятельный критерий проверки гипотезы. Обозначим для критерия K_n вероятности ошибок I и II рода $\alpha(K_n)$ и $\beta(K_n)$, соответственно. Последовательность критериев K_n называется состоятельной, если:

$$\lim_{n \longrightarrow \infty} \alpha(K_n) + \beta(K_n) = 0. \tag{3.33}$$

В этом случае говорим, что существует состоятельное обнаружение скачка.

В противном случае скажем, что обнаружение скачка невозможно.

Вернемся к исходной постановке задачи. Наблюдается неоднородный пуассоновский процесс с постоянной интенсивностью $\lambda(t)$ за исключением малого (a_0, a_1) внутри отрезка $[0,1]$.

Разделим отрезок $[0,1]$ на M одинаковых подынтервалов, где ширина каждого имеет вид $\mathrm{h}(n) = \frac{C \log n}{n}$. При этом $M = \frac{1}{\mathrm{h}(n)}$. Теперь задача сводится к оцениванию интервала со скачком.

Рассмотрим логарифм функции отношения правдоподобия. воспользовавшись леммой [8]:

$$\mathcal{Z}_I = -C(a-1) \cdot \ln n + \ln(1+a) \cdot (X(a_1) - X(a_0)). \tag{3.34}$$

Оценим интервал, на котором произошел скачок:

$$\hat{I} = \text{argmax}(-C(a-1)\cdot \ln n + \ln(1+a)\cdot(X(a_1) - X(a_0))). \tag{3.35}$$

Обозначим интервал со скачком за I_0:

- Для I_0 случайная величина $X(a_1) - X(a_0) \sim \text{Poisson}((a+1)\cdot C\ln n)$;
- Для $I \neq I_0$ случайная величина $X(a_1) - X(a_0) \sim \text{Poisson}(C\ln n)$.

Нижняя граница ширины интервала

Необходимо показать, что:

$$\text{P}\left(\max_I(-(a-1)n\,\text{h}(n) + \ln(1+a)\xi_I) > -(a-1)n\,\text{h}(n)\ + \ln(1+a)\xi_{I_0}\right) \xrightarrow{?} 0. \tag{3.36}$$

Таким образом, необходимо показать, что вероятность обнаружить скачок вне интервала I_0 стремится к 0.

$$\begin{aligned}&\text{P}\left(\max_I(-(a-1)n\,\text{h}(n) + \ln(1+a)\xi_I) > -(a-1)n\,\text{h}(n) + \ln(1+a)\xi_{I_0}\right) = \\ &= \text{P}\left(\max_I(\ln(1+a)\xi_I - (a-1)n\,\text{h}(n)) > (a-1)n\,\text{h}(n) + \ln(1+a)\xi_{I_0} - 2(a-1)n\,\text{h}(n)\right).\end{aligned} \tag{3.37}$$

Покажем, что $\ln(1+a)\xi_{I_0} - 2(a-1)n\,\text{h}(n)$ мало, и этим слагаемым можно пренебречь. Для этого воспользуемся неравенствами Чернова для $X \sim \text{Poisson}(\lambda)$:

$$\text{P}\left(X \leq x\right) \leq \frac{e^{-\lambda}(e\lambda)^x}{x^x}, x < \lambda$$
$$\text{P}\left(X \geq x\right) \leq \frac{e^{-\lambda}(e\lambda)^x}{x^x}, x > \lambda$$

В нашем случае $\lambda = C(a+1)\ln n$, $x = \frac{2C(a-1)\ln n}{\ln(1+a)}$

Подставим компоненты неравенства:

$$\text{P}\left(\xi_{I_0} \leq \frac{2C(a-1)\ln n}{\ln(1+a)}\right) \leq \frac{e^{-C(a+1)\ln n}(e(C(a+1)\ln n))^{\frac{2C(a-1)\ln n}{\ln(1+a)}}}{(\frac{2C(a-1)\ln n}{\ln(1+a)})^{\frac{2C(a-1)\ln n}{\ln(1+a)}}} \tag{3.38}$$

$$\text{P}\left(\xi_{I_0} \geq \frac{2C(a-1)\ln n}{\ln(1+a)}\right) \leq \frac{e^{-C(a+1)\ln n}(e(C(a+1)\ln n))^{\frac{2C(a-1)\ln n}{\ln(1+a)}}}{(\frac{2C(a-1)\ln n}{\ln(1+a)})^{\frac{2C(a-1)\ln n}{\ln(1+a)}}} \tag{3.39}$$

Таким образом, необходимо показать, что правая часть неравенства (в обоих случаях) стремится к 0.

Покажем это:

$$\frac{e^{-C(a+1)\ln n}\big(e(C(a+1)\ln n)\big)^{\frac{2C(a-1)\ln n}{\ln(1+a)}}}{\big(\frac{2C(a-1)\ln n}{\ln(1+a)}\big)^{\frac{2C(a-1)\ln n}{\ln(1+a)}}} = \\ = e^{-C(a+1)\ln n+\frac{2C(a-1)\ln n}{\ln(1+a)}}\cdot\Big(\frac{(a+1)\ln(1+a)}{2(a-1)}\Big)^{\frac{2C(a-1)\ln n}{\ln(1+a)}} = \\ = n^{-C(a+1)+\frac{2C(a-1)}{\ln(1+a)}}\cdot\Big(\frac{(a+1)\ln(1+a)}{2(a-1)}\Big)^{\frac{2C(a-1)\ln n}{2(a-1)}} \tag{3.40}$$

Заметим, что при $a>0$ выражение $-C(a+1)+\frac{2C(a-1)}{\ln(1+a)}<0$. Преобразуем получившееся выражение:

$$n^{-C(a+1)+\frac{2C(a-1)}{\ln(1+a)}}\cdot\Big(\frac{(a+1)\ln(1+a)}{2(a-1)}\Big)^{\frac{2C(a-1)\ln n}{2(a-1)}} = \\ = \frac{\Big(\frac{(a+1)\ln(1+a)}{(a-1)e}\Big)^{\ln n\,\frac{2C(a-1)}{\ln(1+a)}}}{n^{-C(a+1)}} \tag{3.41}$$

Таким образом, получается:

$$\frac{\Big(\frac{(a+1)\ln(1+a)}{(a-1)}\Big)^{\ln n\,\frac{2C(a-1)}{\ln(1+a)}}}{n^{C(a+1)-\frac{2C(a-1)}{\ln(1+a)}}} \underset{n\longrightarrow\infty}{\longrightarrow} 0. \tag{3.42}$$

В итоге по доказанному 3.42 получаем, что :

$$\mathrm{P}\Big(\max_I\big(\ln(1+a)\xi_I-(a-1)n\,\mathrm{h}(n)\big)>(a-1)n\,\mathrm{h}(n)+\ln(1+a)\xi_{I_0}-2(a-1)n\,\mathrm{h}(n)\Big) = \\ = \mathrm{P}\Big(\max_I\big(\ln(1+a)\xi_I-(a-1)n\,\mathrm{h}(n)\big)>(a-1)n\,\mathrm{h}(n)\cdot(1-\delta)\Big). \tag{3.43}$$

Приведем полученное выражение к виду:

$$\mathrm{P}\Big(\max_I\Big(\xi_I-\frac{(a-1)n\,\mathrm{h}(n)}{\ln(1+a)}\Big)>\frac{(a-1)n\,\mathrm{h}(n)\cdot(1-\delta)}{\ln(1+a)}\Big) = \\ = 1-\Big(1-\mathrm{P}\Big(\xi_i-\frac{(a-1)n\,\mathrm{h}(n)}{\ln(1+a)}>\frac{(a-1)n\,\mathrm{h}(n)\cdot(1-\delta)}{\ln(1+a)}\Big)\Big)^{\frac{1}{\mathrm{h}(n)}} = \\ = 1-\exp\Big\{-\frac{1}{\mathrm{h}(n)}\,\mathrm{P}\Big(\xi_i-\frac{(a-1)n\,\mathrm{h}(n)}{\ln(1+a)}>\frac{(a-1)n\,\mathrm{h}(n)\cdot(1-\delta)}{\ln(1+a)}\Big)\Big\}. \tag{3.44}$$

По результатам источника [9] верно следующее:

$$\mathrm{P}\left(|\xi_I - n\lambda| > nx\right) \leq \exp\Big\{ - n(\lambda + x)\ln\left(1 + \frac{x}{\lambda}\right) + nx\Big\}, \tag{3.45}$$

где $\lambda \approx x$.

Применим этот факт [9] и подставим $\mathrm{h}(n) = \frac{C\ln n}{n}$ (пренебрегая δ):

$$\begin{aligned} \mathrm{P}\left(\xi_i - \frac{(a-1)n\,\mathrm{h}(n)}{\ln(1+a)} > \frac{(a-1)n\,\mathrm{h}(n)\cdot(1-\delta)}{\ln(1+a)}\right) \leq \\ \leq \exp\Big\{ - n\cdot\Big(\frac{C(a-1)\ln n}{n\ln(1+a)} + \frac{C(a-1)\ln n}{n\ln(1+a)}\Big)\ln\Big(1+1-\delta\Big) + \frac{(a-1)n\,\mathrm{h}(n)}{\ln(1+a)}\Big\} = \\ = \exp\Big\{ - n\cdot\Big(\frac{C(a-1)\ln n}{n\ln(1+a)} + \frac{C(a-1)\ln n}{n\ln(1+a)}\Big)\ln 2 + \frac{(a-1)n\,\mathrm{h}(n)}{\ln(1+a)}\Big\} = \\ = \exp\Big\{ - \frac{C(a-1)}{\ln(1+a)}\ln n(2\ln 2 - 1)\Big\} \end{aligned} \tag{3.46}$$

Подставим полученное выражение в 3.77:

$$\begin{aligned} 1 - \exp\Big\{ - \frac{1}{\mathrm{h}(n)}\,\mathrm{P}\left(\xi_i - n\,\mathrm{h}(n) > n\,\mathrm{h}(n)(1-\delta)\right)\Big\} \geq \\ \geq 1 - \exp\Big\{ - \frac{1}{\mathrm{h}(n)}\exp\Big\{ - \frac{C(a-1)}{\ln(1+a)}\ln n(2\ln 2 - 1)\Big\}\Big\} = \\ = 1 - \exp\Big\{ - \frac{n}{C\ln n}\exp\Big\{ - \frac{C(a-1)}{\ln(1+a)}\ln n(2\ln 2 - 1)\Big\}\Big\} \overset{?}{\longrightarrow} 0. \end{aligned} \tag{3.47}$$

Полученное выражение будет стремиться к 0 в случае, если $\frac{C(a-1)}{\ln(1+a)}(2\ln 2 - 1) = 1$. Проверим это:

$$1 - \exp\Big\{ - \frac{n}{C\ln n}\exp\Big\{ - \frac{C(a-1)}{\ln(1+a)}\ln n(2\ln 2 - 1)\Big\}\Big\} = 1 - \exp\Big\{ - \frac{1}{C\ln n}\Big\} \underset{n\longrightarrow\infty}{\longrightarrow} 0. \tag{3.48}$$

Таким образом, была найдена константа:

$$C = \frac{\ln(1+a)}{(2\ln 2 - 1)(a-1)}. \tag{3.49}$$

Верхняя граница ширины интервала

Аналогично случаю с гауссовским шумом (3.1.1) рассмотрим сетку с шагом ε, т.е. разобьем отрезок $[0, 1]$ на множество подынтервалов с шагом ε. При этом рассматриваться будет не одно разбиение, как в случае с нижней границей, а множество различных решеток с шагом ε.

Искомый интервал может не покрываться окном шириной ε, поэтому необходимо рассмотреть множество решеток. Таких решеток всего ε^{-1}.

Допустим, ширина скачка имеет вид: $Cn^{-1}\ln n$, как и в случае с нижней границей интервала (3.2.1).

При такой же постановке задачи и тех же условиях, как в разделе (3.2.1), проведем аналогичные действия для верхней границы интервала со скачком.

Таким образом, для оценки верхней границы необходимо показать следующее:

Необходимо показать, что:

$$\mathrm{P}\left(\max_{\mathcal{W}}\max_{I}(-(a-1)n\,\mathrm{h}(n)+\ln(1+a)\xi_I) > -(a-1)n\,\mathrm{h}(n)\ +\ln(1+a)\xi_{I_0}\right)\xrightarrow{?}0. \quad (3.50)$$

Таким образом, необходимо показать, что вероятность обнаружить скачок вне интервала I_0 стремится к 0 по всем решеткам.

Кроме того, необходимо оценить константу, при которой это условие выполняется.

$$\mathrm{P}\left(\max_{\mathcal{W}}\max_{I}\left(-(a-1)n\,\mathrm{h}(n)+\ln(1+a)\xi_I\right) > -(a-1)n\,\mathrm{h}(n)+\ln(1+a)\xi_{I_0}\right) =$$
$$= \mathrm{P}\left(\max_{\mathcal{W}}\max_{I}\left(\ln(1+a)\xi_I-(a-1)n\,\mathrm{h}(n)\right) > (a-1)n\,\mathrm{h}(n)+\ln(1+a)\xi_{I_0}-2\cdot(a-1)n\,\mathrm{h}(n)\right). \quad (3.51)$$

Аналогично доказательству в разделе 3.2.1 пренебрегаем слагаемым $\ln(1+a)\xi_{I_0} - 2\cdot(a-1)n\,\mathrm{h}(n)$ и получаем, что :

$$\mathrm{P}\left(\max_{\mathcal{W}}\max_{I}\left(\ln(1+a)\xi_I-(a-1)n\,\mathrm{h}(n)\right) > (a-1)n\,\mathrm{h}(n)+\ln(1+a)\xi_{I_0}-2\cdot(a-1)n\,\mathrm{h}(n)\right) =$$
$$= \mathrm{P}\left(\max_{\mathcal{W}}\max_{I}\left(\ln(1+a)\xi_I-(a-1)n\,\mathrm{h}(n)\right) > (a-1)n\,\mathrm{h}(n)\cdot(1-\delta)\right). \quad (3.52)$$

Приведем полученное выражение к виду:

$$\begin{aligned}
&\mathrm{P}\left(\max_{\mathcal{W}}\max_{I}\ \left(\xi_I-\frac{(a-1)n\,\mathrm{h}(n)}{\ln(1+a)}\right) > \frac{(a-1)n\,\mathrm{h}(n)\cdot(1-\delta)}{\ln(1+a)}\right) \le \\
&\le \varepsilon^{-2}\,\mathrm{P}\left(\max_{I}\ \left(\xi_I-\frac{(a-1)n\,\mathrm{h}(n)}{\ln(1+a)}\right) > \frac{(a-1)n\,\mathrm{h}(n)\cdot(1-\delta)}{\ln(1+a)}\right) = \\
&= \varepsilon^{-2}\left(1-\left(1-\mathrm{P}\left(\xi_i-\frac{(a-1)n\,\mathrm{h}(n)}{\ln(1+a)} > \frac{(a-1)n\,\mathrm{h}(n)\cdot(1-\delta)}{\ln(1+a)}\right)\right)^{\frac{1}{\mathrm{h}(n)}}\right) = \\
&= \varepsilon^{-2}\left(1-\exp\left\{-\frac{1}{\mathrm{h}(n)}\,\mathrm{P}\left(\xi_i-\frac{(a-1)n\,\mathrm{h}(n)}{\ln(1+a)} > \frac{(a-1)n\,\mathrm{h}(n)\cdot(1-\delta)}{\ln(1+a)}\right)\right\}\right).
\end{aligned} \quad (3.53)$$

Осталось показать, что:

$$1-\exp\Big\{-\frac{1}{\mathrm{h}(n)}\,\mathrm{P}\left(\xi_i-\frac{(a-1)n\,\mathrm{h}(n)}{\ln(1+a)}>\frac{(a-1)n\,\mathrm{h}(n)\cdot(1-\delta)}{\ln(1+a)}\right)\Big\}\overset{?}{\longrightarrow}0. \tag{3.54}$$

Чтобы это показать, необходимо оценить $\mathrm{P}\left(\xi_i-\frac{(a-1)n\,\mathrm{h}(n)}{\ln(1+a)}>\frac{(a-1)n\,\mathrm{h}(n)\cdot(1-\delta)}{\ln(1+a)}\right)$.

По результатам источника [9] верно следующее:

$$\mathrm{P}\Big(|\xi_I-n\lambda|>nx\Big)\le\exp\Big\{-n(\lambda+x)\ln\Big(1+\frac{x}{\lambda}\Big)+nx\Big\}, \tag{3.55}$$

где $\lambda\approx x$.

Применим этот факт [9] и подставим $\mathrm{h}(n)=\frac{C\ln n}{n}$ (пренебрегая δ):

$$\begin{aligned}
&\mathrm{P}\left(\xi_i-\frac{(a-1)n\,\mathrm{h}(n)}{\ln(1+a)}>\frac{(a-1)n\,\mathrm{h}(n)\cdot(1-\delta)}{\ln(1+a)}\right)\le\\
\le\exp\Big\{-n\cdot\Big(\frac{C\cdot(a-1)\ln n}{n\ln(1+a)}+\frac{C\cdot(a-1)\ln n}{n\ln(1+a)}\Big)\ln\Big(1+1-\delta\Big)+\frac{(a-1)n\,\mathrm{h}(n)}{\ln(1+a)}\Big\}&=\\
=\exp\Big\{-n\cdot\Big(\frac{C\cdot(a-1)\ln n}{n\ln(1+a)}+\frac{C\cdot(a-1)\ln n}{n\ln(1+a)}\Big)\ln 2+\frac{(a-1)n\,\mathrm{h}(n)}{\ln(1+a)}\Big\}&=\\
=\exp\Big\{-\frac{C\cdot(a-1)}{\ln(1+a)}\ln n(2\ln 2-1)\Big\}&
\end{aligned} \tag{3.56}$$

Подставим полученное выражение в 3.77:

$$\begin{aligned}
1-\exp\Big\{-\frac{1}{\mathrm{h}(n)}\,\mathrm{P}\left(\xi_i-(a-1)n\,\mathrm{h}(n)>(a-1)n\,\mathrm{h}(n)(1-\delta)\right)\Big\}&\ge\\
\ge1-\exp\Big\{-\frac{1}{\mathrm{h}(n)}\exp\Big\{-\frac{C\cdot(a-1)}{\ln(1+a)}\ln n(2\ln 2-1)\Big\}\Big\}&=\\
=1-\exp\Big\{-\frac{n}{C\ln n}\exp\Big\{-\frac{C\cdot(a-1)}{\ln(1+a)}\ln n(2\ln 2-1)\Big\}\Big\}&\overset{?}{\longrightarrow}0.
\end{aligned} \tag{3.57}$$

Полученное выражение будет стремиться к 0 в случае, если $\frac{C\cdot(a-1)}{\ln(1+a)}(2\ln 2-1)=1$. Проверим это:

$$1-\exp\Big\{-\frac{n}{C\ln n}\exp\Big\{-\frac{C\cdot(a-1)}{\ln(1+a)}\ln n(2\ln 2-1)\Big\}\Big\}=1-\exp\Big\{-\frac{1}{C\ln n}\Big\}\underset{n\longrightarrow\infty}{\longrightarrow}0. \tag{3.58}$$

Таким образом, была найдена константа:

$$C=\frac{\ln(1+a)}{(2\ln 2-1)(a-1)}. \tag{3.59}$$

3.2.2. Двумерный случай

Пусть наблюдается неоднородный пуассоновский процесс $X(t)$, при $t \in [0,1]\times[0,1]$, постоянной интенсивности $\lambda(t) = n$ за исключением малой области (квадрата) $[x_0, x_1] \times [y_0, y_1] \in [0,1] \times [0,1]$, где $0 < x_0 < x_1 < 1$ и $0 < y_0 < y_1 < 1$, а интенсивность скачка $\lambda(t) = an$, $(a > 1)$. Обозначим сторону этого квадрата за $\mathrm{h}(n)$.

Рассмотрим задачу о проверке гипотезы H_0 о том, что интенсивность процесса Пуассона постоянна и равна n, против альтернативы H_1 о том, что интенсивность процесса Пуассона имеет вид:

$$\lambda(t) = \begin{cases} na, & t \in [x_0, x_1] \times [y_0, y_1] \\ n, & t \notin [x_0, x_1] \times [y_0, y_1]. \end{cases} \tag{3.60}$$

Изучим вопрос: при каких $\mathrm{h}(n)$ существует состоятельный критерий проверки гипотезы. Обозначим для критерия K_n вероятности ошибок I и II рода $\alpha(K_n)$ и $\beta(K_n)$, соответственно. Последовательность критериев K_n называется состоятельной, если:

$$\lim_{n\longrightarrow\infty} \alpha(K_n) + \beta(K_n) = 0. \tag{3.61}$$

В этом случае говорим, что существует состоятельное обнаружение скачка.

В противном случае скажем, что обнаружение скачка невозможно.

Вернемся к исходной постановке задачи. Наблюдается неоднородный пуассоновский процесс с постоянной интенсивностью $\lambda(t)$ за исключением малой области $[x_0, x_1]\times[y_0, y_1]$ внутри квадрата $[0,1] \times [0,1]$.

Разделим квадрат $[0,1] \times [0,1]$ на M^2 одинаковых малых квадратов (см. рис. 3.3), где сторона каждого имеет вид $\mathrm{h}(n) = \frac{C \log n}{n}$. При этом $M = \frac{1}{\mathrm{h}(n)}$. Теперь задача сводится к оцениванию малого квадрата со скачком.

Рассмотрим логарифм функции отношения правдоподобия. воспользовавшись леммой [8]:

$$\mathcal{Z}_I = -(a-1) \cdot (C \ln n)^2 + \ln(1+a) \cdot (X(x_1, y_1) - X(x_0, y_0)). \tag{3.62}$$

Оценим интервал, на котором произошел скачок:

$$\hat{I} = \mathrm{argmax}(-(a-1) \cdot (C \ln n)^2 + \ln(1+a) \cdot (X(x_1, y_1) - X(x_0, y_0))). \tag{3.63}$$

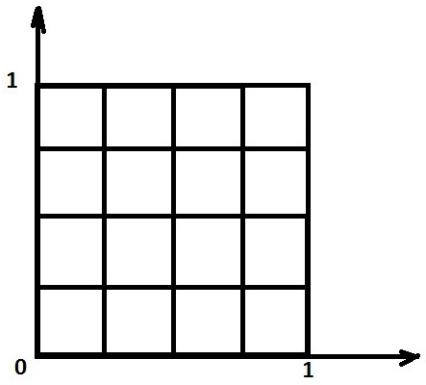

Рис. 3.3. Простейший пример разбиения квадрата

Обозначим интервал со скачком за I_0:

- Для I_0 случайная величина $X(a_1) - X(a_0) \sim \text{Poisson}((a+1) \cdot (C \ln n)^2)$;
- Для $I \neq I_0$ случайная величина $X(a_1) - X(a_0) \sim \text{Poisson}((C \ln n)^2)$.

Нижняя граница стороны квадрата

Необходимо показать, что:

$$\text{P}\left(\max_I (-(a-1)(n\,\text{h}(n))^2 + \ln(1+a)\xi_I) > -(a-1)(n\,\text{h}(n))^2 + \ln(1+a)\xi_{I_0} \right) \xrightarrow{?} 0. \quad (3.64)$$

Таким образом, необходимо показать, что вероятность обнаружить скачок вне интервала I_0 стремится к 0.

$$\text{P}\left(\max_I (-(a-1)(n\,\text{h}(n))^2 + \ln(1+a)\xi_I) > -(a-1)(n\,\text{h}(n))^2 + \ln(1+a)\xi_{I_0} \right) =$$
$$= \text{P}\left(\max_I (\ln(1+a)\xi_I - (a-1)(n\,\text{h}(n))^2) > (a-1)(n\,\text{h}(n))^2 + \ln(1+a)\xi_{I_0} - 2(a-1)(n\,\text{h}(n))^2 \right). \quad (3.65)$$

Аналогично доказательству в разделе 3.2.1 пренебрегаем слагаемым $\ln(1+a)\xi_{I_0} - 2 \cdot (a-1)(n\,\text{h}(n))^2$ и получаем, что :

$$\text{P}\left(\max_I (\ln(1+a)\xi_I - (a-1)(n\,\text{h}(n))^2) > (a-1)(n\,\text{h}(n))^2 + \ln(1+a)\xi_{I_0} - 2(a-1)(n\,\text{h}(n))^2 \right) =$$
$$= \text{P}\left(\max_I (\ln(1+a)\xi_I - (a-1)(n\,\text{h}(n))^2) > (a-1)(n\,\text{h}(n))^2 \cdot (1-\delta) \right). \quad (3.66)$$

Приведем полученное выражение к виду:

$$\begin{aligned}
&\mathrm{P}\Big(\max_I\ \Big(\xi_I - \frac{(a-1)(n\,\mathrm{h}(n))^2}{\ln(1+a)}) > \frac{(a-1)(n\,\mathrm{h}(n))^2\cdot(1-\delta)}{\ln(1+a)}\Big) = \\
&= 1 - \Big(1 - \mathrm{P}\Big(\xi_i - \frac{(a-1)(n\,\mathrm{h}(n))^2}{\ln(1+a)} > \frac{(a-1)(n\,\mathrm{h}(n))^2\cdot(1-\delta)}{\ln(1+a)}\Big)\Big)^{\frac{1}{(\mathrm{h}(n))^2}} = \\
&= 1 - \exp\Big\{-\frac{1}{(\mathrm{h}(n))^2}\,\mathrm{P}\Big(\xi_i - \frac{(a-1)(n\,\mathrm{h}(n))^2}{\ln(1+a)} > \frac{(a-1)(n\,\mathrm{h}(n))^2\cdot(1-\delta)}{\ln(1+a)}\Big)\Big\}.
\end{aligned} \tag{3.67}$$

По результатам источника [9] верно следующее:

$$\mathrm{P}\Big(|\xi_I - n\lambda| > nx\Big) \le \exp\Big\{-n(\lambda+x)\ln\Big(1+\frac{x}{\lambda}\Big) + nx\Big\}, \tag{3.68}$$

где $\lambda \approx x$.

Применим этот факт [9] и подставим $\mathrm{h}(n) = \frac{C\ln n}{n}$ (пренебрегая δ):

$$\begin{aligned}
&\mathrm{P}\Big(\xi_i - \frac{(a-1)(n\,\mathrm{h}(n))^2}{\ln(1+a)} > \frac{(a-1)(n\,\mathrm{h}(n))^2\cdot(1-\delta)}{\ln(1+a)}\Big) \le \\
&\le \exp\Big\{-n^2\cdot\Big(\frac{(a-1)(C\ln n)^2}{n^2\ln(1+a)} + \frac{(a-1)(C\ln n)^2}{n^2\ln(1+a)}\Big)\ln\Big(1+1-\delta\Big) + \frac{(a-1)(n\,\mathrm{h}(n))^2}{\ln(1+a)}\Big\} = \\
&= \exp\Big\{-n^2\cdot\Big(\frac{(a-1)(C\ln n)^2}{n^2\ln(1+a)} + \frac{(a-1)(C\ln n)^2}{n^2\ln(1+a)}\Big)\ln 2 + \frac{(a-1)(n\,\mathrm{h}(n))^2}{\ln(1+a)}\Big\} = \\
&= \exp\Big\{-\frac{(a-1)C^2}{\ln(1+a)}(\ln n)^2(2\ln 2 - 1)\Big\}
\end{aligned} \tag{3.69}$$

Подставим полученное выражение в 3.77:

$$\begin{aligned}
&1 - \exp\Big\{-\frac{1}{(\mathrm{h}(n))^2}\,\mathrm{P}\Big(\xi_i - \frac{(a-1)(n\,\mathrm{h}(n))^2}{\ln(1+a)} > \frac{(a-1)(n\,\mathrm{h}(n))^2}{\ln(1+a)}(1-\delta)\Big)\Big\} \ge \\
&\ge 1 - \exp\Big\{-\frac{1}{(\mathrm{h}(n))^2}\exp\Big\{-\frac{C^2(a-1)}{\ln(1+a)}(\ln n)^2(2\ln 2-1)\Big\}\Big\} = \\
&= 1 - \exp\Big\{-\frac{n^2}{(C\ln n)^2}\exp\Big\{-\frac{C^2(a-1)}{\ln(1+a)}(\ln n)^2(2\ln 2-1)\Big\}\Big\} \overset{?}{\longrightarrow} 0.
\end{aligned} \tag{3.70}$$

Полученное выражение будет стремиться к 0 в случае, если $\frac{C^2(a-1)}{\ln(1+a)}(2\ln 2 - 1) = 1$. Проверим это:

$$1 - \exp\Big\{-\frac{n}{C\ln n}\exp\Big\{-\frac{C^2(a-1)}{\ln(1+a)}(\ln n)^2(2\ln 2-1)\Big\}\Big\} = 1 - \exp\Big\{-\frac{n^{2-\ln n}}{(C\ln n)^2}\Big\} \underset{n\longrightarrow\infty}{\longrightarrow} 0. \tag{3.71}$$

Таким образом, была найдена константа:

$$C^2 = \frac{\ln(1+a)}{(2\ln 2 - 1)(a-1)}, \tag{3.72}$$

соответственно, извлекая корень, получаем:

$$C = \sqrt{\frac{\ln(1+a)}{(2\ln 2 - 1)(a-1)}}. \tag{3.73}$$

Верхняя граница ширины области

Аналогично случаю с гауссовским шумом (3.1.2) рассмотрим сетку с шагом ε, т.е. разобьем квадрат $[0,1] \times [0,1]$ на множество малых квадратов с шагом ε по обеим осям OX и OY. При этом рассматриваться будет не одно разбиение, как в случае с нижней границей, а множество различных решеток с шагом ε.

Искомый интервал может не покрываться окном шириной ε, поэтому необходимо рассмотреть множество решеток. Таких решеток всего ε^{-2}.

Допустим, сторона малого квадрата со скачком имеет вид: $Cn^{-1}\ln n$, как и в случае с нижней границей интервала (3.2.2).

При такой же постановке задачи и тех же условиях, как в разделе (3.2.2), проведем аналогичные действия для верхней границы интервала со скачком.

Таким образом, для оценки верхней границы необходимо показать следующее:

Необходимо показать, что:

$$\mathrm{P}\left(\max_I \left(-(a-1)(n\,\mathrm{h}(n))^2 + \ln(1+a)\xi_I\right) > -(a-1)(n\,\mathrm{h}(n))^2 + \ln(1+a)\xi_{I_0} \right) \xrightarrow{?} 0. \tag{3.74}$$

Таким образом, необходимо показать, что вероятность обнаружить скачок вне интервала I_0 стремится к 0.

$$\mathrm{P}\left(\max_{\mathcal{W}} \max_I \left(-(a-1)(n\,\mathrm{h}(n))^2 + \ln(1+a)\xi_I\right) > -(a-1)(n\,\mathrm{h}(n))^2 + \ln(1+a)\xi_{I_0} \right)$$
$$= \mathrm{P}\left(\max_{\mathcal{W}} \max_I \left(\ln(1+a)\xi_I - (a-1)(n\,\mathrm{h}(n))^2\right) > (a-1)(n\,\mathrm{h}(n))^2 + \ln(1+a)\xi_{I_0} - 2(a-1)(n\,\mathrm{h}(n))^2 \right) \tag{3.75}$$

Аналогично доказательству в разделе 3.2.1 пренебрегаем слагаемым $\ln(1+a)\xi_{I_0} - 2\cdot(a-1)(n\,\mathrm{h}(n))^2$ и получаем, что :‘

$$\mathrm{P}\Big(\max_{\mathcal{W}}\max_{I}\left(\ln(1+a)\xi_I-(a-1)(n\,\mathrm{h}(n))^2\right)>(a-1)(n\,\mathrm{h}(n))^2+\ln(1+a)\xi_{I_0}-2(a-1)(n\,\mathrm{h}(n))^2\Big)$$
$$=\mathrm{P}\Big(\max_{\mathcal{W}}\max_{I}\left(\ln(1+a)\xi_I-(a-1)(n\,\mathrm{h}(n))^2\right)>(a-1)(n\,\mathrm{h}(n))^2\cdot(1-\delta)\Big) \tag{3.76}$$

Приведем полученное выражение к виду:

$$\mathrm{P}\Big(\max_{\mathcal{W}}\max_{I}\Big(\xi_I-\frac{(a-1)(n\,\mathrm{h}(n))^2}{\ln(1+a)}\Big)>\frac{(a-1)(n\,\mathrm{h}(n))^2\cdot(1-\delta)}{\ln(1+a)}\Big)\le$$
$$\le\varepsilon^{-4}\,\mathrm{P}\Big(\max_{I}\Big(\xi_I-\frac{(a-1)(n\,\mathrm{h}(n))^2}{\ln(1+a)}\Big)>\frac{(a-1)(n\,\mathrm{h}(n))^2\cdot(1-\delta)}{\ln(1+a)}\Big)=$$
$$=\varepsilon^{-4}\Big(1-\Big(1-\mathrm{P}\Big(\xi_i-\frac{(a-1)(n\,\mathrm{h}(n))^2}{\ln(1+a)}>\frac{(a-1)(n\,\mathrm{h}(n))^2\cdot(1-\delta)}{\ln(1+a)}\Big)\Big)^{\frac{1}{(\mathrm{h}(n))^2}}\Big)=$$
$$=\varepsilon^{-4}\Big(1-\exp\Big\{-\frac{1}{(\mathrm{h}(n))^2}\,\mathrm{P}\Big(\xi_i-\frac{(a-1)(n\,\mathrm{h}(n))^2}{\ln(1+a)}>\frac{(a-1)(n\,\mathrm{h}(n))^2\cdot(1-\delta)}{\ln(1+a)}\Big)\Big\}\Big). \tag{3.77}$$

Осталось показать, что:

$$1-\exp\Big\{-\frac{1}{\mathrm{h}(n)}\,\mathrm{P}\Big(\xi_i-\frac{(a-1)n\,\mathrm{h}(n)}{\ln(1+a)}>\frac{(a-1)n\,\mathrm{h}(n)\cdot(1-\delta)}{\ln(1+a)}\Big)\Big\}\overset{?}{\longrightarrow}0. \tag{3.78}$$

Чтобы это показать, необходимо оценить $\mathrm{P}\Big(\xi_i-\frac{(a-1)(n\,\mathrm{h}(n))^2}{\ln(1+a)}>\frac{(a-1)(n\,\mathrm{h}(n))^2\cdot(1-\delta)}{\ln(1+a)}\Big)$.

По результатам источника [9] верно следующее:

$$\mathrm{P}\Big(|\xi_I-n\lambda|>nx\Big)\le\exp\Big\{-n(\lambda+x)\ln\Big(1+\frac{x}{\lambda}\Big)+nx\Big\}, \tag{3.79}$$

где $\lambda\approx x$.

Применим этот факт [9] и подставим $\mathrm{h}(n)=\frac{C\ln n}{n}$ (пренебрегая δ):

$$\mathrm{P}\Big(\xi_i-\frac{(a-1)(n\,\mathrm{h}(n))^2}{\ln(1+a)}>\frac{(a-1)(n\,\mathrm{h}(n))^2\cdot(1-\delta)}{\ln(1+a)}\Big)\le$$
$$\le\exp\Big\{-n^2\cdot\Big(\frac{(a-1)(C\ln n)^2}{n^2\ln(1+a)}+\frac{(a-1)(C\ln n)^2}{n^2\ln(1+a)}\Big)\ln\Big(1+1-\delta\Big)+\frac{(a-1)(n\,\mathrm{h}(n))^2}{\ln(1+a)}\Big\}=$$
$$=\exp\Big\{-n^2\cdot\Big(\frac{(a-1)(C\ln n)^2}{n^2\ln(1+a)}+\frac{(a-1)(C\ln n)^2}{n^2\ln(1+a)}\Big)\ln 2+\frac{(a-1)(n\,\mathrm{h}(n))^2}{\ln(1+a)}\Big\}=$$
$$=\exp\Big\{-\frac{(a-1)C^2}{\ln(1+a)}(\ln n)^2(2\ln 2-1)\Big\} \tag{3.80}$$

Подставим полученное выражение в 3.77:

$$1-\exp\Big\{-\frac{1}{(\mathrm{h}(n))^2}\,\mathrm{P}\Big(\xi_i-\frac{(a-1)(n\,\mathrm{h}(n))^2}{\ln(1+a)}>\frac{(a-1)(n\,\mathrm{h}(n))^2}{\ln(1+a)}(1-\delta)\Big)\Big\}\geq$$
$$\geq 1-\exp\Big\{-\frac{1}{(\mathrm{h}(n))^2}\exp\Big\{-\frac{C^2(a-1)}{\ln(1+a)}(\ln n)^2(2\ln 2-1)\Big\}\Big\}= \tag{3.81}$$
$$=1-\exp\Big\{-\frac{n^2}{(C\ln n)^2}\exp\Big\{-\frac{C^2(a-1)}{\ln(1+a)}(\ln n)^2(2\ln 2-1)\Big\}\Big\}\overset{?}{\longrightarrow}0.$$

Полученное выражение будет стремиться к 0 в случае, если $\frac{C^2(a-1)}{\ln(1+a)}(2\ln 2-1)=1$. Проверим это:

$$1-\exp\Big\{-\frac{n}{C\ln n}\exp\Big\{-\frac{C^2(a-1)}{\ln(1+a)}(\ln n)^2(2\ln 2-1)\Big\}\Big\}=1-\exp\Big\{-\frac{n^{2-\ln n}}{(C\ln n)^2}\Big\}\underset{n\longrightarrow\infty}{\longrightarrow}0. \tag{3.82}$$

Таким образом, была найдена константа:

$$C^2=\frac{\ln(1+a)}{(2\ln 2-1)(a-1)}, \tag{3.83}$$

соответственно, извлекая корень, получаем:

$$C=\sqrt{\frac{\ln(1+a)}{(2\ln 2-1)(a-1)}}. \tag{3.84}$$

Заключение

В данной выпускной работе была изучена проблема оценивания разреженных функций и также был подробно изучен вопрос обнаружения скачков в зашумленных данных.

В ходе работы было выполнено следующее:

- Найдена нижняя и верхняя границы ширины скачка для обнаружения его в пуассоновских процессах. Кроме того, было показано, что эти границы достигаются (см. раздел 3.2). Также нами был рассмотрен двумерный вариант скачка для исходной постановки задачи (регрессионной модели с гауссовским шумом) и для пуассоновских случайных процессов (см. разделы 3.1.2 и 3.2.2).

- Дан обзор методов, оценивающих модель регрессии, которая имеет скачкообразный характер.

- Осуществлено сравнение 3–х алгоритмов (Алгоритм Мунка [4], Спокойного [5] и SaRa [1]), которые ранее не сравнивались между собой (см. раздел 1.2). Реализация алгоритмов выполнена в среде программирования R A.

Сравнив все три алгоритма ([4], [1], [5]), можно сделать такие выводы:

- Для модели регрессии с гауссовским шумом точнее всего оценивает моменты разладки, ширину и высоту скачка алгоритм Мунка [4]. Оценки SaRa [1] прближаются к истинному значению, но гораздо хуже оценок алгоритма Мунка. Алгоритм Спокойного [5] в текущей реализации оценивает моменты разладки, высоту и ширину скачка хуже остальных алгоритмов.

- Для пуассоновских процессов самым точным также оказался алгоритм Мунка [4]. Оценки алгоритмов SaRa [1] и Спокойного [5] также приближаются к истинным значениям, однако хуже, чем алгоритм Мунка.

Полученные в данной выпускной работе результаты могут послужить основой для продолжения исследования по задаче “размазанной” разреженности. Дальнейшими перспективами являются изучение задачи асимптотического обнаружения скачка, когда его размер зависиит от мощности скачка, а также изучение сложной геометрии многомерных разрывов.

Список литературы

1. Hao N., Niu Y. S., Zhang H. Multiple change–point detection via screening and ranking algorithm // Statistica Sinica. — 2013. — Vol. 23. — P. 20.
2. Arias-Castro E., Donoho D. L. Near-optimal detection of geometric objects by fast multiscale methods. — 2005. — July. — Vol. 51, no. 7.
3. Arias-Castro E., Wang M. Distribution-free tests for sparse heterogeneous mixtures // Department of Mathematics, University of California, San Diego. — 2013. — November. — Vol. 15.
4. Frick K., Munk A., Sieling H. Multiscale change-point inference // Mathematics Subject Classification. — 2010.
5. Spokoiny V. G. Estimation of a function with discontinuities via local polynomial fit with an adaptive window choice // The Annals of Statistics. — 1998. — Vol. 26, no. 4. — P. 1356 – 1378.
6. Niu Y. S., Hao N., Zhang H. Multiple change-point detection: a selective overview.
7. Asymptotic Theory of Statistics and Probability / Ed. by I. Olkin G. Casella, S. Fienberg. — Purdue University 150 North University Street West Lafayette, IN 47907 : Anirban DasGupta, 2008.
8. Ingster Y. I., Kutoyants Y. A. Nonparametric hypothesis testing for an intensity of Poisson process. — St. Petersburg State Electrotechnical University and Laboratoire de Statistique et Processus, Universit du Maine.
9. M. A. The large deviation principle for stochastic processes. ii // Theory.Probab.Appl. — 2004. — Vol. 48. — P. 19–44.
10. Tibshirani R., Wang P. Spatial smoothing and hot spot detection for CGH data using the fused lasso // Biostatistics. — 2007. — May.
11. Reconstructing DNA copy number by penalized estimation and imputation / Zhongyang Zhang, Kenneth Lange, Roel Ophoff, Chiara Sabatti // The Annals of Applied Statistics. — 2010. — Vol. 4, no. 4. — P. 1749–1773.
12. Detection of DNA copy number variations using penalized least absolute deviations regression : Rep. : 386 / The University of Iowa ; Executor: Xiaoli Gao, Jian Huang : 2007. — December.
13. Detection of DNA copy number alterations using penalized least squares regression /

Tao Huang, Baolin Wu, Paul Lizardi, Hongyu Zhao // Bioinformatics. — 2005. — August. — Vol. 21, no. 20. — P. 3811–3817.

14. Хандыго Т. Применение метода имитации отжига в задаче аппроксимации функции регрессии. Дипломная работа. — 2015.
15. Mitzenmacher M., Upfal E. Probability and Computing. Randomized Algorithms and Probabilistic Analysis. — Press Syndicate of the University of Cambrige, 2005.

Приложение А

Анализ алгоритмов для одномерного случая

А.1. Реализация алгоритмов

А.1.1. Реализация алгоритма Мунка

При реализации алгоритма был использован пакет **stepR**, в котором присутствует необходимая функция по обнаружению скачков.

```
library(stepR)
######################## Munk algorithm ############

Munk.alg <- function(n, Y.input)
{
  result <- numeric(5)
  res.temp <- smuceR(Y.input,
  family="poisson", confband = TRUE)
  result[2] <- res.temp$rightEnd[
  res.temp$value==max(res.temp$value)]
  result[1] <- res.temp$leftEnd[
  res.temp$value==max(res.temp$value)]
  result[5] <- result[2] - result[1]
  height.max <- mean(Y.input[result[1]:result[2]])
  height.min <- mean(Y.input[-c(result[1]:result[2])])
  result[3:4] <- c(height.min,height.max)
  return(result)

}
########## estimates ##########################
prefix = "a_20/Poiss_pr"
prefix_res = "a_20/munk_res_20_Poiss_pr"
```

```
res.100 <- read.csv(
paste(prefix_res,"_100.csv",sep = ""),
sep=",",dec=".", header=T)
res.100 <- res.100[,-1]

res.500 <- read.csv(
paste(prefix_res,"_500.csv",sep = ""),
sep=",",dec=".", header=T)
res.500 <- res.500[,-1]

res.1000 <- read.csv(
paste(prefix_res,"_1000.csv",sep = ""),
sep=",",dec=".", header=T)
res.1000 <- res.1000[,-1]

res.100.mod <- res.100
res.100.mod[3,] <- res.100[4,]-res.100[3,]
res.100.mod[4,] <- res.100[5,]
res.100.mod <- res.100.mod[-5,]

mean.100 <- sapply(1:4,function(i){
mean(as.numeric(res.100.mod[i,]))})
var.100 <- sapply(1:4,function(i)
{var(as.numeric(res.100.mod[i,]))})
rmse.100 <- sapply(1:4,function(i){
rmse.f(as.numeric(res.100.mod[i,]),sim.100[i,])})

res.500.mod <- res.500
res.500.mod[3,] <- res.500[4,]-res.500[3,]
res.500.mod[4,] <- res.500[5,]
res.500.mod <- res.500.mod[-5,]
```

```
mean.500 <- sapply(1:4,function(i){
mean(as.numeric(res.500.mod[i,]))})
var.500 <- sapply(1:4,function(i){
var(as.numeric(res.500.mod[i,]))})
rmse.500 <- sapply(1:4,function(i){
rmse.f(as.numeric(res.500.mod[i,]),sim.500[i,])})

res.1000.mod <- res.1000
res.1000.mod[3,] <- res.1000[4,]-res.1000[3,]
res.1000.mod[4,] <- res.1000[5,]
res.1000.mod <- res.1000.mod[-5,]

mean.1000 <- sapply(1:4,function(i){
mean(as.numeric(res.1000.mod[i,]))})
var.1000 <- sapply(1:4,function(i){
var(as.numeric(res.1000.mod[i,]))})
rmse.1000 <- sapply(1:4,function(i){
rmse.f(as.numeric(res.1000.mod[i,]),sim.1000[i])})
```

A.1.2. Реализация SaRa

Алгоритм реализован на основании статьи Yue S. Niu и Heping Zhang.

```
######################## D SaRa ######################################
D.SaRa.1.new <- function(j,Y,h)
{
  s <- (sum(sapply((j+1):(j+h),
  function(k){Y[k]})) - sum(sapply((j-h+1):j,function(k){Y[k]})))/h
  s
}

######################### maximum if D ############
if.local.max.1 <- function(x, D.ans)
{

  if (all(D.ans[x] >= D.ans[(x-h.1):(x+h.1)]))
  {
    return(x)
  } else return(NA)
}
######################### SaRa algorithm ###

Sara.alg <- function(n,Y.input)
{
  result <- numeric(5)
  D.ans <- abs(sapply(h:(n-h),function(k){
  D.SaRa.1.new(k,Y.input,h)}))
  n.D <- length(D.ans)
  ind.max.1 <- sapply((1+h.1):(n.D-h.1),
  function(z){if.local.max.1(z, D.ans)})
  ind.max.1 <- ind.max.1[!is.na(ind.max.1)]
  D.est1.1 <- D.ans[ind.max.1]
  ind.est.1 <- ind.max.1[D.ans[ind.max.1] > 0.5]
```

```
  if (length(ind.est.1) > 2)
  {
    ind.temp.1 <- ind.est.1[
    D.ans[ind.est.1] == max(D.ans[ind.est.1])]
    if (length(ind.temp.1) != 2) {
      ind.temp <- ind.est.1[-which(ind.temp.1 == ind.est.1)]
      ind.temp.2 <- ind.est.1[D.ans[ind.temp] == max(D.ans[ind.temp])]
      if (ind.temp.1 > ind.temp.2)
      {
        result[1] <- ind.temp.2
        result[2] <- ind.temp.1
      } else {
        result[1] <- ind.temp.1
        result[2] <- ind.temp.2
      }
    }

  } else {
    result[1:2] <- ind.est.1
  }
  height.min <- mean(D.ans[-result[1:2]])
  height.max <- mean(D.ans[result[1]],D.ans[result[2]])
  result[3:4] <- c(height.min,height.max)
  result[5] <- result[2] - result[1]
  return(result)
}
#####################################################################
########## estimates ########################
prefix = "a_20/Poiss_pr"
prefix_res = "a_20/sara_res_20_Poiss_pr"

res.100 <- read.csv(
```

```
paste(prefix_res,"_100.csv",sep = ""),
sep=",",dec=".", header=T)
res.100 <- res.100[,-1]

res.500 <- read.csv(
paste(prefix_res,"_500.csv",sep = ""),
sep=",",dec=".", header=T)
res.500 <- res.500[,-1]

res.1000 <- read.csv(
paste(prefix_res,"_1000.csv",sep = ""),
sep=",",dec=".", header=T)
res.1000 <- res.1000[,-1]

res.100.mod <- res.100
res.100.mod[3,] <- res.100[4,]-res.100[3,]
res.100.mod[4,] <- res.100[5,]
res.100.mod <- res.100.mod[-5,]

mean.100 <- sapply(1:4,function(i){
mean(as.numeric(res.100.mod[i,]))})
var.100 <- sapply(1:4,function(i){
var(as.numeric(res.100.mod[i,]))})
rmse.100 <- sapply(1:4,function(i){
rmse.f(as.numeric(res.100.mod[i,]),sim.100[i,])})

res.500.mod <- res.500
res.500.mod[3,] <- res.500[4,]-res.500[3,]
res.500.mod[4,] <- res.500[5,]
res.500.mod <- res.500.mod[-5,]

mean.500 <- sapply(1:4,function(i){
```

```
mean(as.numeric(res.500.mod[i,]))})
var.500 <- sapply(1:4,function(i){
var(as.numeric(res.500.mod[i,]))})
rmse.500 <- sapply(1:4,function(i){
rmse.f(as.numeric(res.500.mod[i,]),sim.500[i,])})

res.1000.mod <- res.1000
res.1000.mod[3,] <- res.1000[4,]-res.1000[3,]
res.1000.mod[4,] <- res.1000[5,]
res.1000.mod <- res.1000.mod[-5,]

mean.1000 <- sapply(1:4,function(i){
mean(as.numeric(res.1000.mod[i,]))})
var.1000 <- sapply(1:4,function(i){
var(as.numeric(res.1000.mod[i,]))})
rmse.1000 <- sapply(1:4,function(i){
rmse.f(as.numeric(res.1000.mod[i,]),sim.1000[i])})
```

А.1.3. Реализация алгоритма В. Г. Спокойного

Алгоритм реализован на основании статьи В. Г. Спокойного, а также выведенного алгоритма для шума с независимыми приращениями пуассоновского процесса.

```
#################### Z loglikelihood NORM ####


Z.loglik.norm <- function(x.par,Y)
{
  n <- length(Y)
  N.int <- est.const*log(n) + 1 # number of design points in interval
  s <- -sum(sapply(1:n,function(i) {
  ifelse(is.between(1.0*i/n,
  x.par[1],x.par[2]),Y[i],0)})) + N.int/2.0
  return(s)
}

#################### Z loglikelihood POISSON ###


Z.loglik.poiss <- function(x.par,Y)
{
  n <- length(Y)
  s <- -(Y[x.par[2]*n]-Y[x.par[1]*n])
  *log(1+a) + est.const*log(n)
  return(s)
}

######################## Spok algorithm ############


Spok.alg <- function(n, Y.input, Z.loglik)
```

```
{
  result <- numeric(5)



  res.temp <- optim(par.init,Z.loglik,
  Y = Y.input,method = "L-BFGS-B",
  lower=c(0.01,0.01),upper=c(1.0,1.0))
  result[1] <- min(
  round(res.temp$par[1]*n),round(res.temp$par[2]*n))
  result[2] <- max(
  round(res.temp$par[1]*n),round(res.temp$par[2]*n))
  result[5] <- abs(result[2] - result[1])
  result[3] <- mean(Y.input[-c(result[1]:result[2])])
  result[4] <- mean(Y.input[c(result[1]:result[2])])

  return(result)

}

########## estimates ##########################
prefix = "st_dev_2_h_20/Norm"
prefix_res = "st_dev_2_h_20/spok_res_2_h_20_Norm"

res.100 <- read.csv(
paste(prefix_res,"_100.csv",sep = ""),
sep=",",dec=".", header=T)
res.100 <- res.100[,-1]

res.500 <- read.csv(
paste(prefix_res,"_500.csv",sep = ""),
sep=",",dec=".", header=T)
res.500 <- res.500[,-1]
```

```
res.1000 <- read.csv(
paste(prefix_res,"_1000.csv",sep = ""),
sep=",",dec=".", header=T)
res.1000 <- res.1000[,-1]

res.100.mod <- res.100
res.100.mod[3,] <- res.100[4,]-res.100[3,]
res.100.mod[4,] <- res.100[5,]
res.100.mod <- res.100.mod[-5,]

mean.100 <- sapply(1:4,function(i){
mean(as.numeric(res.100.mod[i,]))})
var.100 <- sapply(1:4,function(i){
var(as.numeric(res.100.mod[i,]))})
rmse.100 <- sapply(1:4,function(i){
rmse.f(as.numeric(res.100.mod[i,]),sim.100[i,])})

res.500.mod <- res.500
res.500.mod[3,] <- res.500[4,]-res.500[3,]
res.500.mod[4,] <- res.500[5,]
res.500.mod <- res.500.mod[-5,]

mean.500 <- sapply(1:4,function(i){
mean(as.numeric(res.500.mod[i,]))})
var.500 <- sapply(1:4,function(i){
var(as.numeric(res.500.mod[i,]))})
rmse.500 <- sapply(1:4,function(i){
rmse.f(as.numeric(res.500.mod[i,]),sim.500[i,])})

res.1000.mod <- res.1000
res.1000.mod[3,] <- res.1000[4,]-res.1000[3,]
```

```
res.1000.mod[4,] <- res.1000[5,]
res.1000.mod <- res.1000.mod[-5,]

mean.1000 <- sapply(1:4,function(i){
mean(as.numeric(res.1000.mod[i,]))})
var.1000 <- sapply(1:4,function(i){
var(as.numeric(res.1000.mod[i,]))})
rmse.1000 <- sapply(1:4,function(i){
rmse.f(as.numeric(res.1000.mod[i,]),sim.1000[i])})
```