Домашнее задание

Винецкая Полина

19 декабря 2022 г.

1 Задача 3

Дан набор однотипных N файлов, содержащих одномерный массив данных X и Y. Предполагая, что зависимость Y от X описывается линейной функцией вида $Y = kX + b + \eta$ (шум), определить характер и интенсивность шума, ожидаемые значения и доверительный интервал для наклона (k) и смещения (b) для отдельных реализаций и для ансамбля реализаций.

1.1 Отдельная реализация

1.1.1 Анализ параметров аппроксимации

Для примера и описания алгоритма рассмотрим подробнее данные из первого файла. С помощью модели линейной регрессии (scipy.stats.linregress) были найдены оптимальные параметры k и b, а также доверительные пределы для уровня значимости $\alpha = 5\%$.

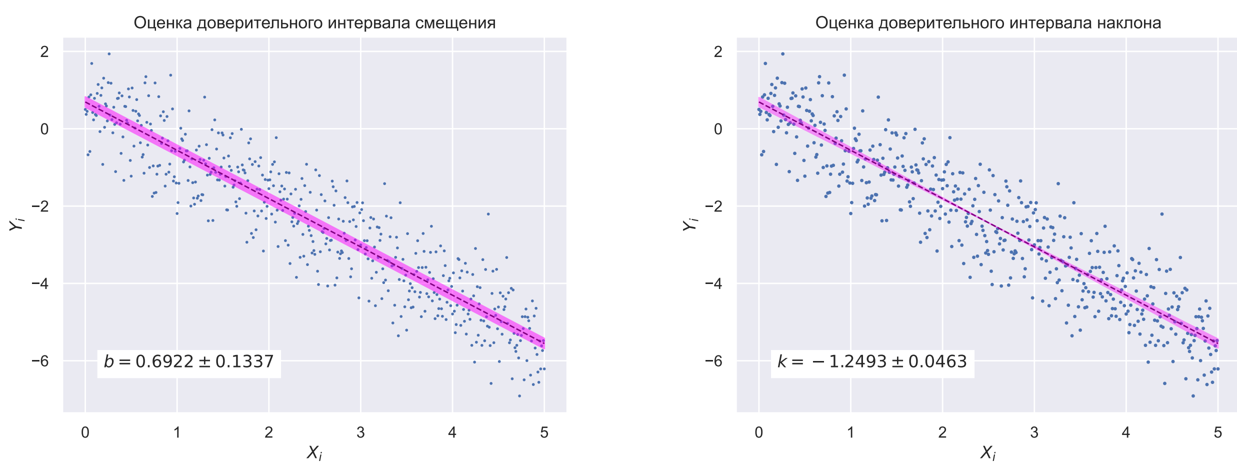


Рис. 1: График исходных данных, аппроксимирующей прямой и доверительных интервалов для параметров k и b

Поскольку уровень доверия 95% соответствует дисперсии 2σ, дисперсии наклона и смещения равны:

$$\sigma_k = 0.0231 \qquad \sigma_b = 0.0669 \tag{1}$$

Также дисперсии можно вычислить по формулам:

$$\sigma_k = \frac{\sigma_n}{\sqrt{N}\sigma_x} = 0.0236 \qquad \sigma_b = \sqrt{\frac{\sigma_n^2}{N} + \sigma_k^2 \langle x_i \rangle} = 0.0681 \tag{2}$$

Различие между дисперсиями, посчитанными аналитически и програмно составляет менее 3%.

1.1.2 Анализ шума

Теперь проанализируем шум:

$$\eta_i = Y_i - (kX + b) \tag{3}$$

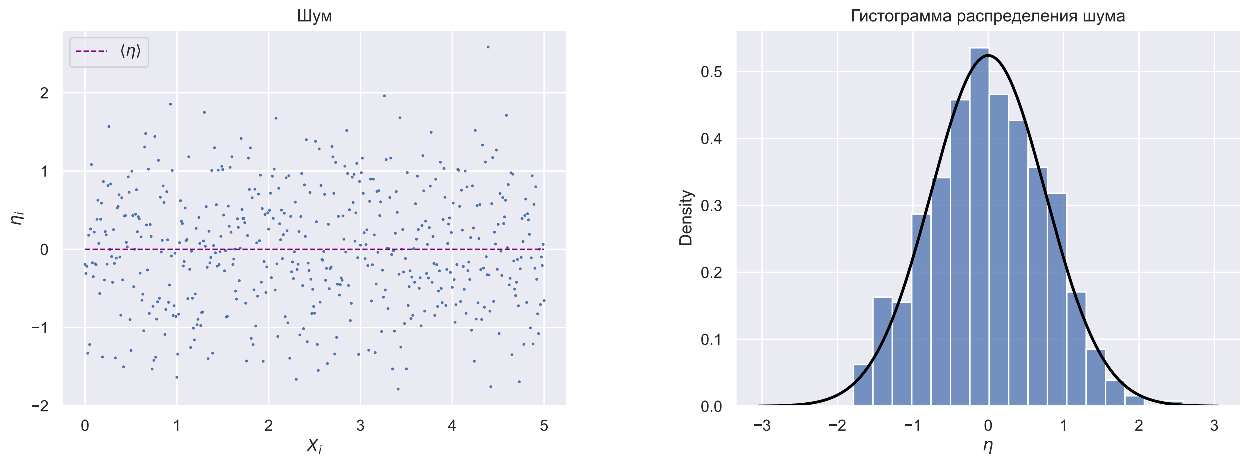


Рис. 2: График исходных данных, аппроксимирующей прямой и доверительных интервалов для параметров k и b

Легко видеть, что шум распределен по Гауссу. Его интенсивность равна корню из его дисперсии и равна в данном случае:

$$\sigma_{noise} = 0.7973 \tag{4}$$

1.2 Все реализации

Аналогичные действия были проделаны для всех файлов. Поскольку данные представляют собой реализации одного и того же эксперимента, неудивительно, что найденные значения k, b, их доверительные интервалы, а также интенсивность шума отличались слабо.

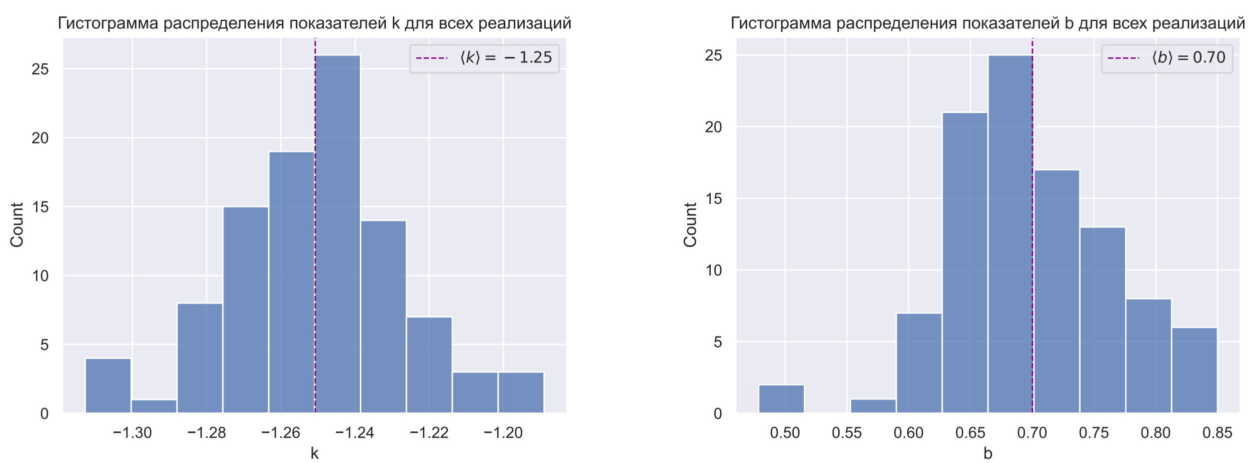


Рис. 3: Гистограммы распределения параметров k и b

	Среднее	Ст.отклонение
k	-1.25	0.02
b	0.7	0.07
σ_{noise}	0.8	0.02

Средние значения параметров k,b,σ_{noise} приведены в таблице ниже:

Также можно посчитать коррелятор между найдетнными k и b:

$$\text{Corr(k,b)} = \frac{\langle (k - \langle k \rangle)(b - \langle b \rangle) \rangle}{\sigma_k \sigma_b} = -0.85 \tag{5}$$

Такое большое по модулю отрицательное значение коррреляторa можно итрепретировать так: чем найденное k, тем меньше b. Это значит, что большая ошибка в нахождении одного параметра провоцирует и большое отклонение от истинного значения для другого. Это подтверждает график.

Рис. 4: Зависимость $b(k)$

1.3 Ансамбль реализаций

Аналогично были найдены параметры для ансамбля реализаций:

$$k = -1.2507 \pm 0.0048 \qquad b = 0.7001 \pm 0.0139 \qquad \sigma_{noise} = 0.797 \tag{6}$$

2 Задача 13

Предполагая, что зависимость Y от X описывается полиномиальной функцией $Y = a_n X^n + a_{n-1} X^{n-1} + \ldots + a_1 X + a_0 +$ шум с неизвестной степенью многочлена n, определить характер и интенсивность шума, ожидаемые значения и доверительные интервалы для коэффициентов $a_0, \ldots a_n$ для отдельных реализаций и для ансамбля реализаций

2.1 Определение степени полинома

Для начала построим все данные и найдем оптимальные значения коэффициентов для разных степеней. Видно (это также можно оценить по среднеквадратичному отклонению), что повышение степени выше 4-й уже не дает заметного улучшения, поэтому степень модельного полинома будем считать равной 4.

Значения коэффициентов и их доверительные интервалы для уровня доверия 95% следующие:

$$a_4 = 0.099 \pm 0.001 \qquad a_3 = 0.3 \pm 0.003 \tag{7}$$

$$a_2 = -0.994 \pm 0.012 \qquad a_1 = 0.994 \pm 0.021 \tag{8}$$

$$a_0 = 0.032 \pm 0.026 \tag{9}$$

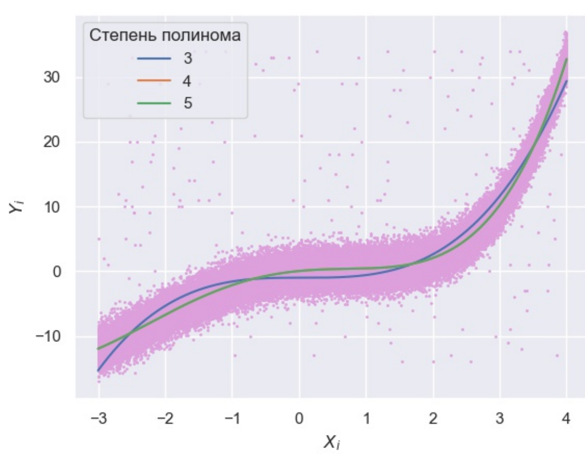


Рис. 5: Графики аппроксимирующих кривых разных нечетных степеней

Чтобы определить характер шума, построим гистограмму разности данных и значений аппроксимации. Как можно видеть из гистограммы, шум снова распределен по Гауссу. Его интенсивность равна $\sigma_{noise} = 1.86$.

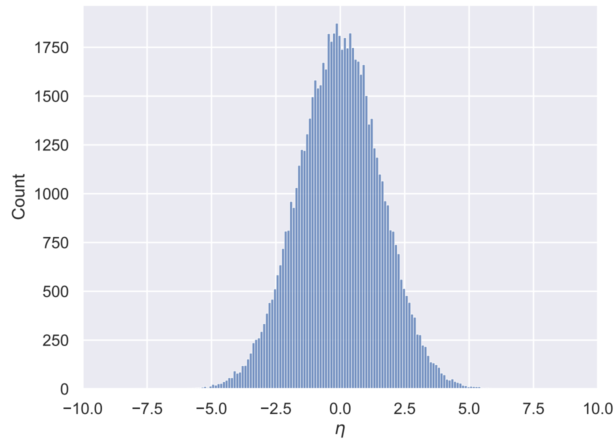
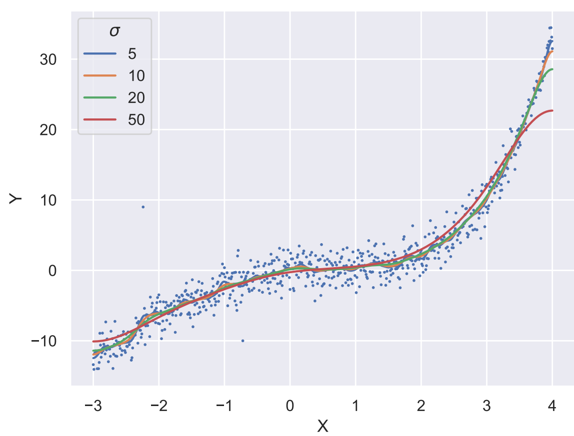


Рис. 6: Гистограмма распределения шума

2.2 Интерполяция и фильтрация

2.2.1 Отдельная реализация

Рассмотрим одну реализацию и применим в данным скользящий гауссов фильтр и построим численные производные для разных значений параметра σ. Поскольку мы уже знаем вид функции, можно сравнить результаты интерполяции с аналитической производной результата, который был получен ранее.



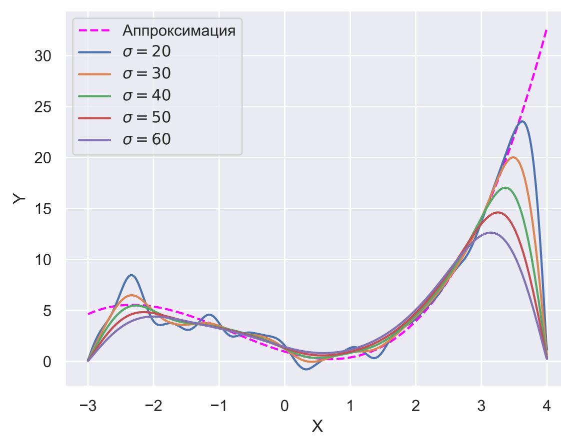


Рис. 7: Графики очищенных с помощью скользящего Гауссова фильтра данных(слева) и их численных производных(справа) для разных параметров стандартного отклонения гауссовой функции

Видно, что, во-первых, параметр должен быть не меньше 20, чтобы подавились быстрые осцилляции, а во-вторых, интерполяция верно отражает поведение данных только на премежутке порядка $[-2.5, 3.5]$. На краях же отрезка значения отклоняются и производная стремится к 0. Оптимальным можно считать график $\sigma = 40$, поскольку в нем отсутствуют быстрые осцилляциии, и при этом он не сильно отличается от графика при $\sigma = 30$. Также он обладает наименьшим среднеквадратичным отклонением от пунктирного графика аппроксимации (но я не уверена, что можно этим знанием пользоваться, поскольку знание формы функции избавляет от необходимости интерполировать данные).

Медианный фильтр в данном случае менее эффективен, поскольку тут достаточно мало выбросов и график производной получается слишком шумным.

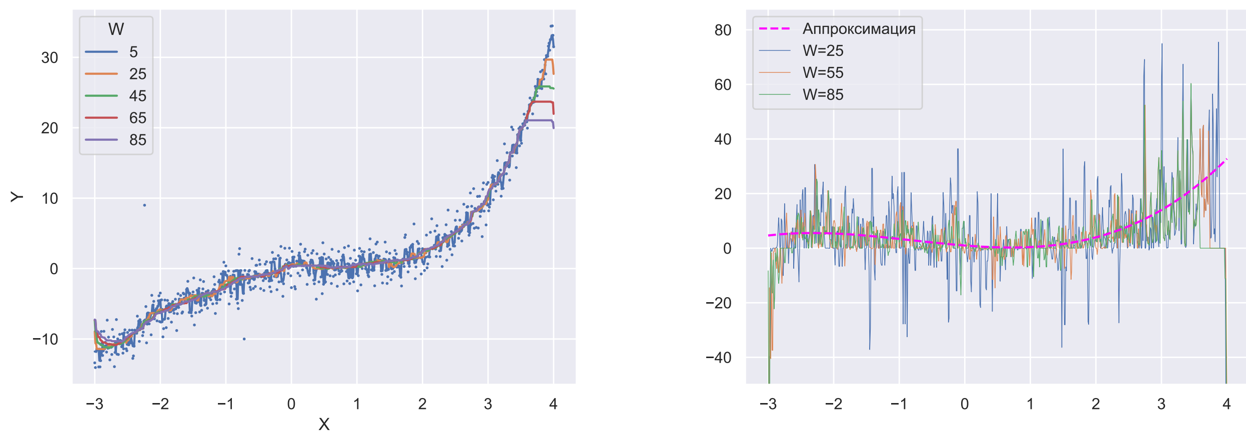


Рис. 8: Графики очищенных с помощью медианного фильтра данных(слева) и их численных производных(справа) для разных ширин окна сглаживания(в точках)

2.3 Все реализации

Поскольку данную процедуру нужно выполнить для всех файлов, выбирать параметр фильтра также надо программно. Поэтому введем меру среднеквадратичного расстояния на центральном интервале (где производная не убывает резко в 0) и будем ее минимизировать.

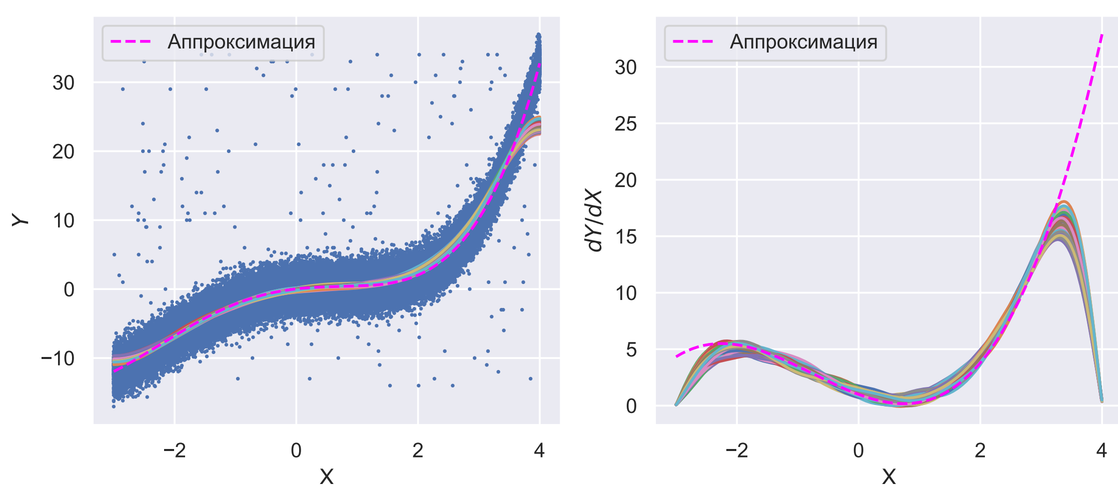


Рис. 9: Оптимальная интерполяция скользящим гауссовым фильтом каждой отдельной реализации

2.3.1 Ансамбль реализаций

Можно применить гауссов фильтр и к ансамблю реализаций:

2.3.2 Среднее по ансамблю

Уже описанную процедуру с использованием скользящего гауссова фильтра можно провести и для усредненных по реализациям данных. Как видно из графика, на исследуемом интервале [-2,3] интерполяция очень хорошо ложится на аппроксимационную кривую.

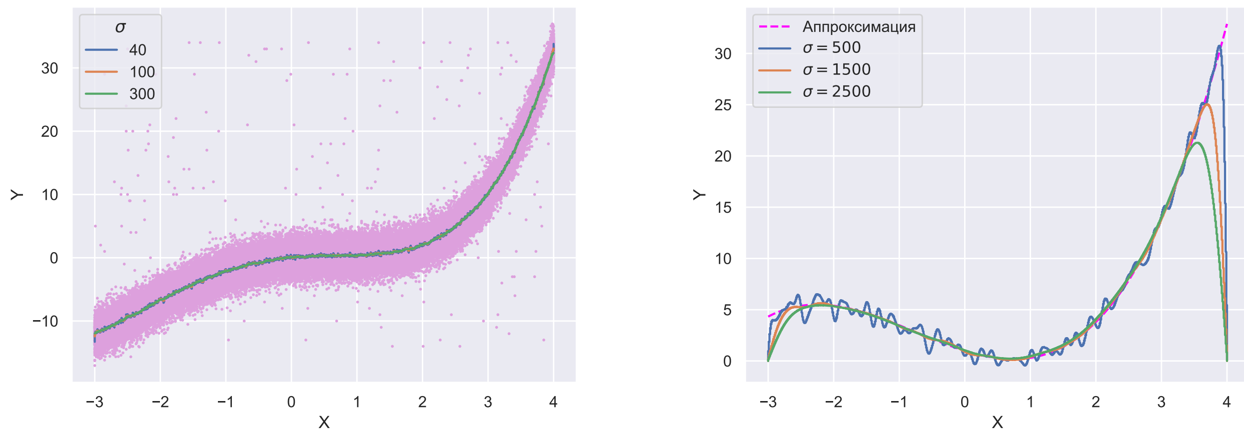


Рис. 10: Графики очищенных с помощью скользящего Гауссова фильтра данных(слева) и их численных производных(справа) для разных параметров стандартного отклонения гауссовой функции (для ансамбля реализаций)

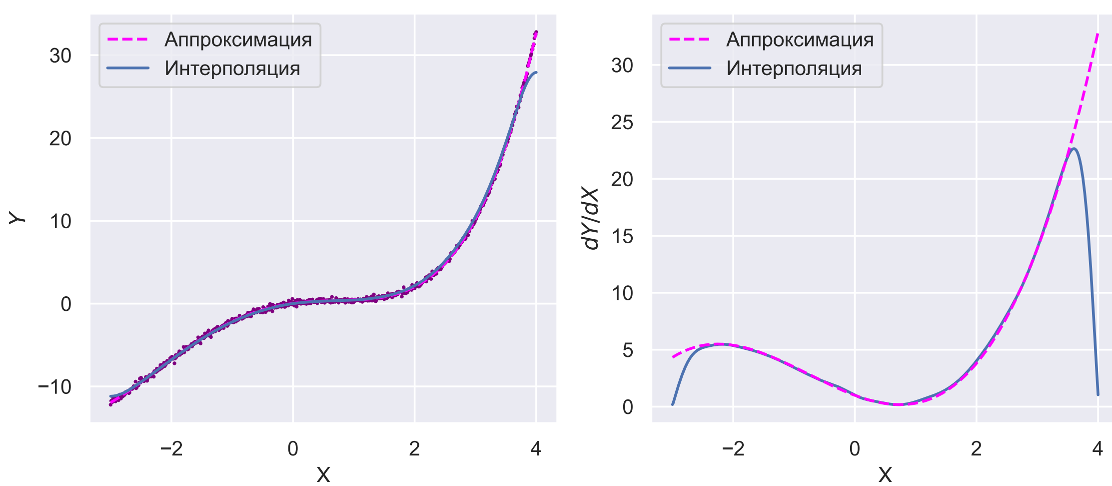


Рис. 11: Оптимальная интерполяция скользящим гауссовым фильтом усредненных данных

2.4 Исследование экстремумов

Как видно по графикам, данные могут иметь точки экстремума в области [0,2]. Однако построенные кривые производных не пересекают 0, хотя проходят довольно близко к нему. Для оценки того, может ли там быть точка экстремума, можно постоить вторую производную (рис.12). Видно, что на интересующем нас промежутке вторая производная строго положительная, значит, экстремумов нет.

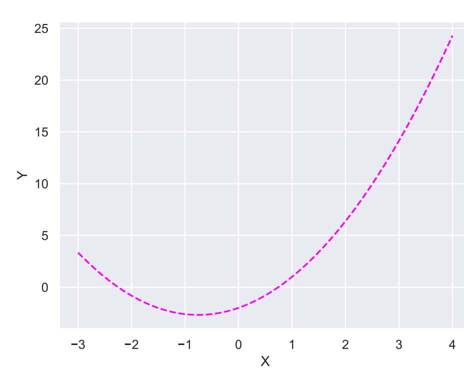


Рис. 12: Вторая производная (аналитическая для аппроксимационной кривой)

3 Задача 23

Дан набор однотипных N файлов, содержащих одномерный массив данных X и Y (первый и второй столбцы). Предполагая, что зависимость Y от X может быть представлена в виде комбинации локализованных пиков и аддитивного шума, оценить параметры пиков (средние значения и доверительные интервалы для амплитуды, положения и ширины). Определить, являются ли пики гауссовыми или лоренцевыми.

3.1 Отдельная реализация

Построив график, можно увидеть, что имеется два пика. Поскольку мы знаем, что каждый из них либо лоренцев, либо гауссов, будем использовать 3 подгоночные функции:

$$f_1 = A_1 \frac{\sigma_1^2}{(x-\mu_1)^2 + \sigma_1^2} + A_2 \frac{\sigma_2^2}{(x-\mu_2)^2 + \sigma_2^2} \tag{10}$$

$$f_2 = A_1 \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right) + A_2 \exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right) \tag{11}$$

$$f_3 = A_1 \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right) + A_2 \frac{\sigma_2^2}{(x-\mu_2)^2 + \sigma_2^2} \tag{12}$$

Даже визуально из графиков на рис. 13 видно, что пики имеют гауссову форму, однако можно также это проверить по стандартному отклонению коэффициентов подгонки:

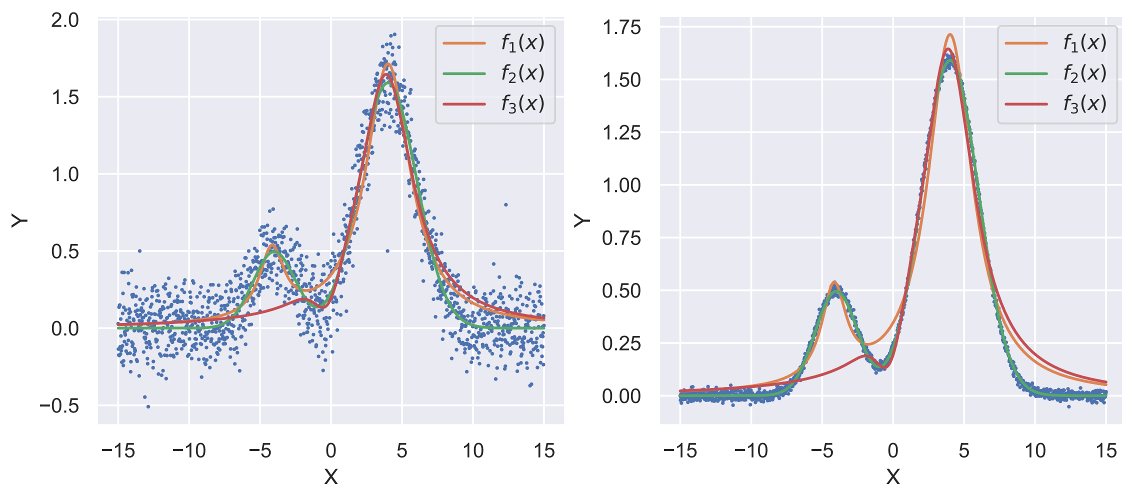


Рис. 13: Аппроксимация одной случайной реализации(слева) и среднего от всех реализаций(справа)

σ	A_1	A_2	σ_1	σ_2	μ_1	μ_1
f_1	1.712 ± 0.01	0.462 ± 0.013	1.91 ± 0.016	1.044 ± 0.044	3.992 ± 0.011	-4.153 ± 0.03
f_2	0.5 ± 0.002	1.595 ± 0.001	1.496 ± 0.006	2.001 ± 0.002	-3.999 ± 0.006	3.997 ± 0.002
f_3	-0.204 ± 0.019	1.655 ± 0.014	0.846 ± 0.104	2.219 ± 0.033	-0.134 ± 0.091	3.863 ± 0.022

Таблица 1: Таблица стандартных отклонений найденных коэффициентов аппроксимации для разных форм функций (для среднего по реализациям)

Видно, что стандартные отклонения для f_2 на порядок меньше, чем для других двух функций. Можно сделать вывод, что **сигнал имеет форму двух гауссовых пиков**.

4 Задача 33

Дан набор однотипных N файлов, содержащих одномерный массив данных X и Y (первый столбец – время в секундах, второй столбец – сигнал в произвольных единицах). Предполагая, что зависимость Y от X может быть представлена в виде комбинации периодических сигналов и аддитивного шума, определить параметры периодического сигнала (частоту, амплитуду и относительную фазу) для каждой из реализаций и для ансамбля реализаций.

4.1 Отдельная реализация

По Фурье данных видно, что периодические компоненты имеют частоты 200 и 240 Гц. Очистим полученные данные, занулив все частоты кроме выбранных и простроим обратный Фурье, для начала для каждой из двух частот по отдельности.

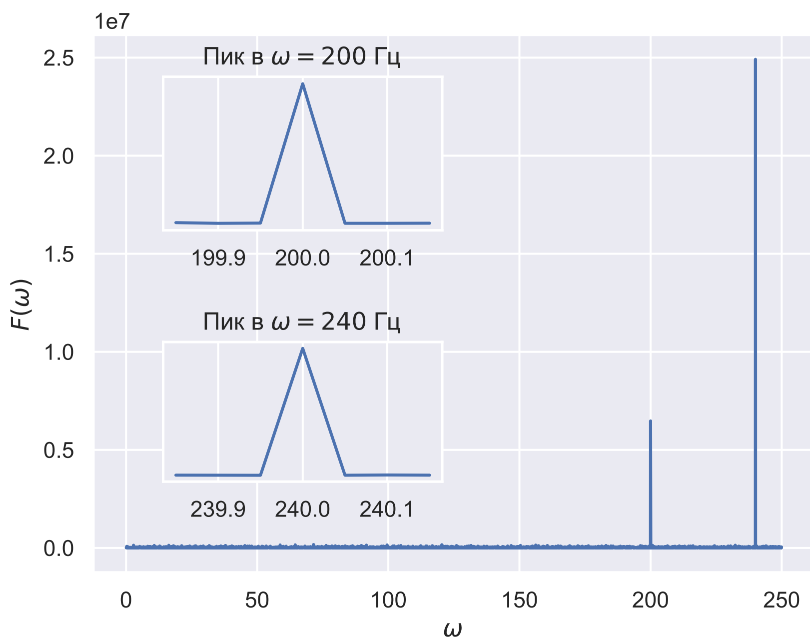


Рис. 14: Быстрое Фурье-преобразование отдельной реализации

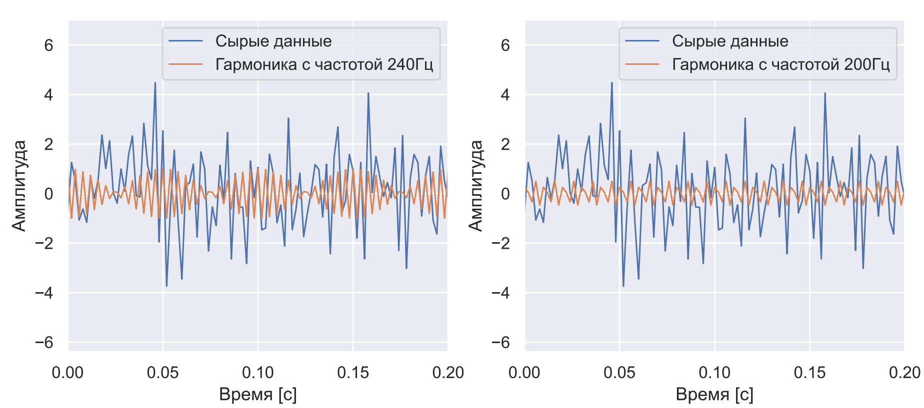


Рис. 15: Отфильтрованные данные для одной реализации

Здесь заметна проблема того, что длина данных ограничена, и для высоких частот метод Фурье дает функцию sinc, а не синус. Это можно решить путем аппроксимации синуса к очищенным данным.

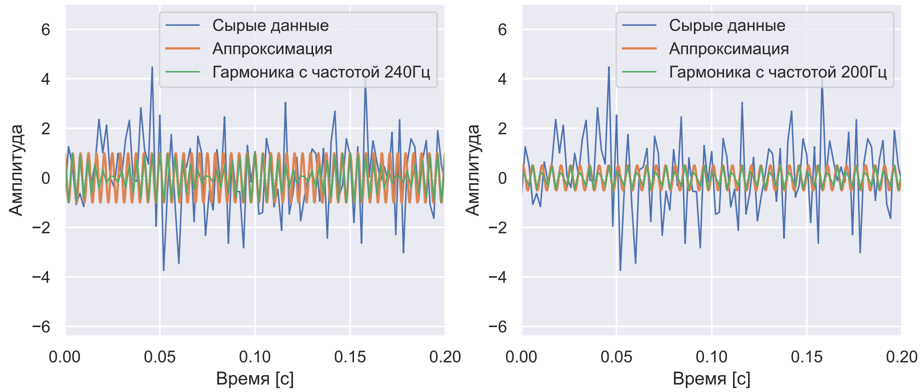


Рис. 16: Аппроксимация отфильтрованных данных для одной реализации

Частота гармоники (из спектра)	Частота гармоники (аппроксимация)	Амплитуда	Фаза
240	$240 \pm 1.2 \cdot 10^{-13}$	$0.998 \pm 4.6 \cdot 10^{-12}$	$-1 \cdot 10^{-3} \pm 6.1 \cdot 10^{-15}$
200	$200 \pm 3 \cdot 10^{-14}$	$0.5 \pm 5.6 \cdot 10^{-13}$	$-4 \cdot 10^{-4} \pm 1.7 \cdot 10^{-15}$

Таблица 2: Значения коэффициентов и их средних отклонений для одной реализации

С учетом малости стандартных отклонений найденных коэффициентов по сравнению с их величинами можно с уверенностью сказать, что исходная функция равна:

$$\sin(2\pi \cdot 240t) + 0.5\sin(2\pi \cdot 200t) \tag{13}$$

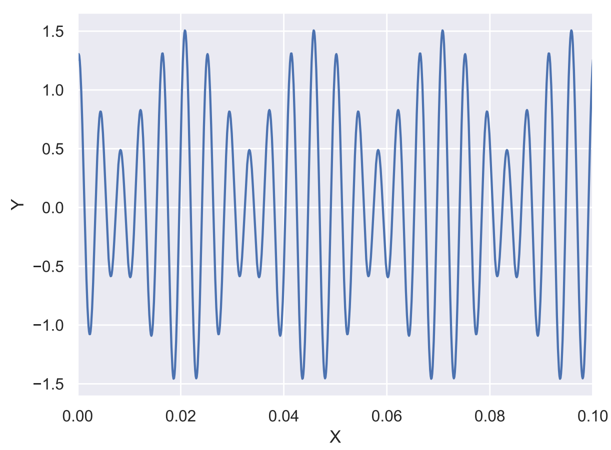


Рис. 17: Очищенный сигнал

Ситуация аналогична для всех реализаций. В качестве альтернативы (и наверное это более честно) можно использовать оконную функцию для свертки, но и метод с аппроксимацией сработал хорошо.

4.2 Ансамбль реализаций

Поскольку я буду брать Фурье, нет большого смысла брать просто объединение всех реализаций, поскольку из-за повторений появятся лишние гармоники в спектре. Поэтому будем анализировать среднее по ансамблю.

Частота гармоники (из спектра)	Частота гармоники (аппроксимация)	Амплитуда	Фаза
240	$240 \pm 9.3 \cdot 10^{-14}$	$0.99 \pm 3.4 \cdot 10^{-12}$	$-1 \cdot 10^{-3} \pm 4.5 \cdot 10^{-15}$
200	$200 \pm 2.0 \cdot 10^{-14}$	$0.5 \pm 3.7 \cdot 10^{-13}$	$-4 \cdot 10^{-4} \pm 1.2 \cdot 10^{-15}$

Таблица 3: Значения коэффициентов и их средних отклонений для среднего по ансамблю

Как видно из таблицы, значения коэффициентов не изменились, лишь уменьшились и без того малые средние отклонения. Это подтверждает предположение о верности найденной функции.