Санкт-Петербургский политехнический университет имени Петра Великого

Высшая школа прикладной математики и физики

Математическая статистика
Отчёт по лабораторной работе №9

Выполнил:

Студент: Теплов Андрей Сергеевич Группа: 5030102/10201

Принял:

к. ф.-м. н., доцент
Баженов Александр Николаевич

Санкт-Петербург
2024г..

Содержание

Список иллюстраций

1. Постановка задачи

Исследование из области солнечной энергетики [1]. На рис 1 показана схема установки для исследования фотоэлектрических характеристик.

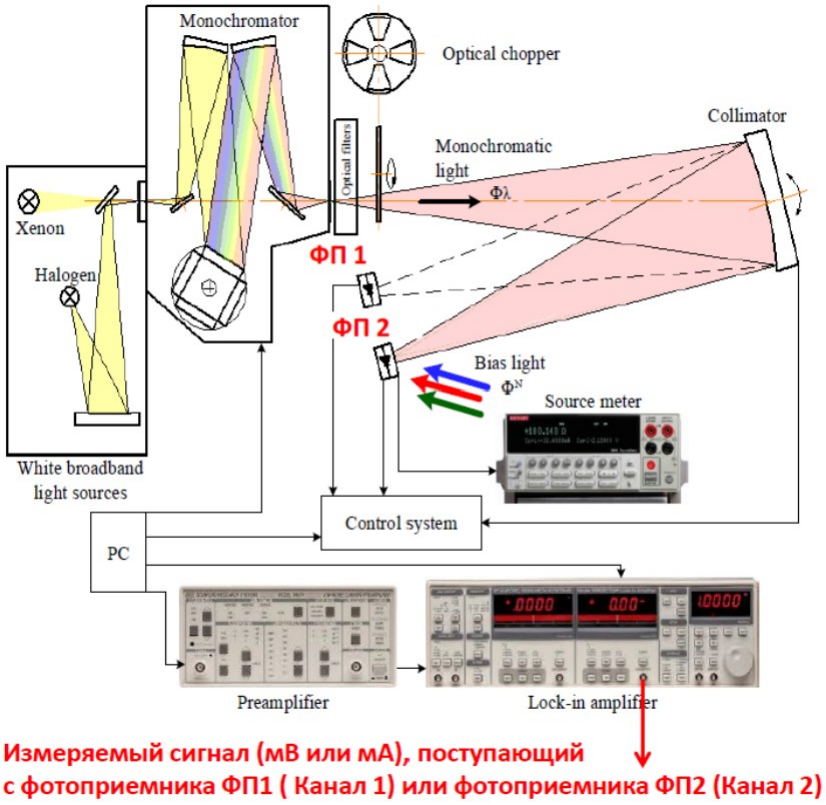

Рис. 1. Схема установки для исследования фотоэлектрических характеристик.

Калибровка датчика ФП1 производится по эталону ФП2. Зависимость между квантовыми эффективностями датчиков предполагается одинаковой для каждой пары измерений

$$QE_2 = \frac{I_2}{I_1} * QE_1 \tag{1}$$

QE_2, QE_1 - эталонная эффективность эталонного и исследумого датчика, а I_2, I_1 - измеренные точки. Данные с датчиков находятся в файлах Ch2_800nm_2m.csv и Ch1_800nm_2m.csv.

Треубется определить коэффициент калибровки

$$R_{21} = \frac{I_2}{I_1} \tag{2}$$

при помощи линейной регрессии на множестве интервальных данных и коэффициента Жаккара.

2. Теория

2.1. Представление данных

В первую очередь представим данные таким образом, чтобы применить понятия статистики данных с интервальной неопределенностью.

Один из распространённых способов получения интервальных результатов в первичных измерениях - это "обинтерваливание"точечных значений, когда к точечному базовому зачению $\dot{x}$, которое считывается по показаниям измерительного прибора, прибавляется *интервал погрешности* ϵ:

$$\mathbf{x} = \dot{x} + \epsilon \tag{3}$$

Интервал погрешности зададим как

$$\epsilon = [-\epsilon; \epsilon]$$

В конкретных измерениях примем $\epsilon = 10^{-4}$ мВ.

Согласно терминологии интервального анализа, рассматриваемая выборка - это вектор интервалов, или интервальный вектор $x = (x_1, x_2, ..., x_n)$.

2.2. Линейная регрессия

2.2.1. Описание модели

Линейная регрессия - регрессионная модель зависимости одной переменной от другой с линейной функцией зависимости:

$$y_i = X_i\beta_i + \epsilon_i$$

где X - заданные значения, y - параметры отклика, ϵ - случайная ошибка модели. В случае, если у нас y_i зависит от одного параметра x_i, то модель выглядит следующим образом:

$$y_i = \beta_0 + \beta_1 * x_i + \epsilon_i \tag{4}$$

В данной можели мы пренебрегаем прогрешностью и считаем, что она получается при измерении y_i.

2.2.2. Метод наименьших модулей

Для наиболее точного приближения входных с фотоприемников данных y_i линейной регрессией $f(x_i)$ используется метод наименьших квадратов. Этот метот основывается на минимизации нормы разности последовательности:

$$\|f(x_i) - y_i\|_{l^1} \to min \tag{5}$$

В данном случае ставится задача линейного программирования, решение которой дает нам коэффициенты β_0 и β_1, а также вектор множителей коррекции данных w. По итогу получается следующая задача линейного программирования

$$\sum_{i=1}^{n} |w_i| \to min \tag{6}$$

$$\beta_0 + \beta_1 * x_i - w_i * \epsilon \leq y_i, i = 1..n \tag{7}$$

$$\beta_0 + \beta_1 * x_i + w_i * \epsilon \leq y_i, i = 1..n \tag{8}$$

$$1 \leq w_i, i = 1..n \tag{9}$$

2.3. Предварительная обработка данных

Для оценки постоянной, как можно будет увидет далее, необходима предварительная обработка данных. Займемся линейной моделью дейфа.

$$Lin_i(n) = A_i + B_i * n, n = 1, 2, ...N \tag{10}$$

Поставив и решив задачу линейного программирования, найдем коэффициенты A_i, B_i и вектор w_i множителей коррекции данных(где i = 1 соответствует данным с ФП1, а i = 2 соответственно ФП2). В последствии множитель коррекции данных необходимо применить к погрешностям выборки, чтобы получить данные, которые согласовывались с линейной моделью дрейфа:

$$I_i^f(n) = \dot{x}(n) + \epsilon * w_i(n), n = 1, 2, ...N \tag{11}$$

По итоге необходимо построить "спрямленные"данные выборки: получить их можно путем вычитания из исходных данных линейную компоненту:

$$I_i^c(n) = I_i^f(n) - B_i * n, n = 1, 2, ...N \tag{12}$$

2.4. Коэффициент Жаккара

Коэффициент Жаккара - мера сходства множеств. В интервальных данных рассматривается некоторая модификация этого коэффициента: в качестве меры множества (в данном случае интервала) рассматривается его длина, а в качестве пересечения и оъединения - взятие минимума и максимума в интервальной арифметике Каухера соответственно. Можно заметить, что в силу возможности минимума по включению быть неправильным инервалом, коэффициент Жаккара может достишать значения только в интервале [-1; 1].

$$JK(x) = \frac{width(\wedge x_i)}{width(\vee x_i)} \tag{13}$$

2.5. Процедура оптимизации

Чтоб найти оптимальный параметр калиброфки R_{21} необходимо поставить и решить задачу максимизации коэффициента Жаккара, зависящего от параметра калибровки:

$$JK(I_1^c(n) * R \cup I_2^c(n)) =\rightarrow max \tag{14}$$

где I_1^c и I_2^c - полученные спрямленные выборки, а R - параметр калибровки. Найденный таким образом R и будет искомым оптимальным R_{21} в силу наибольшего совпадения, оцененного коэффицентом Жаккара.

3. Реализация

Лабораторная работа выполнена на языке программирования Python(3.7) с использованием следующих библиотек: Numpy, Scipy, Tabulate, Statsmodels, Matplotlib.
Отчет написан в онлайн редакторе LaTeX - Overleaf.

4. Результаты

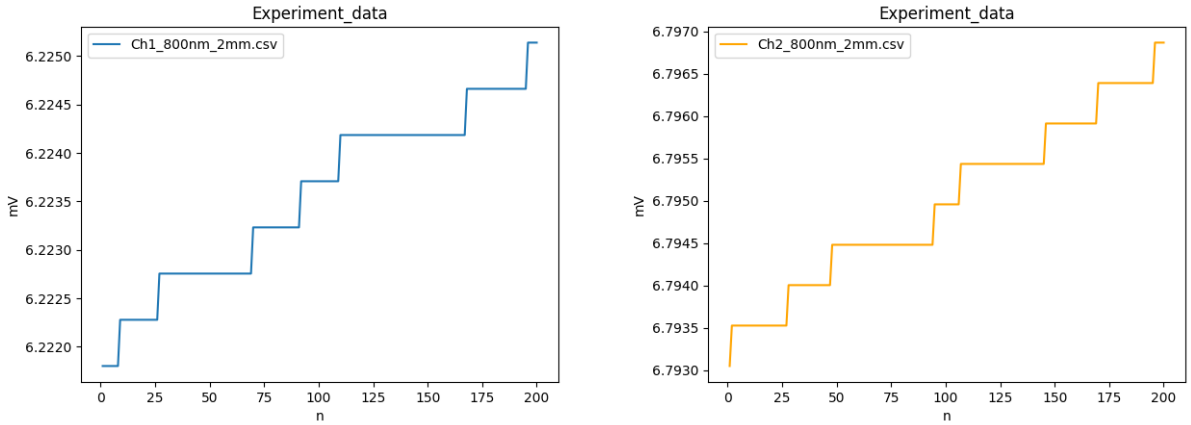

Рис. 2. Исходные данные из экспериментов

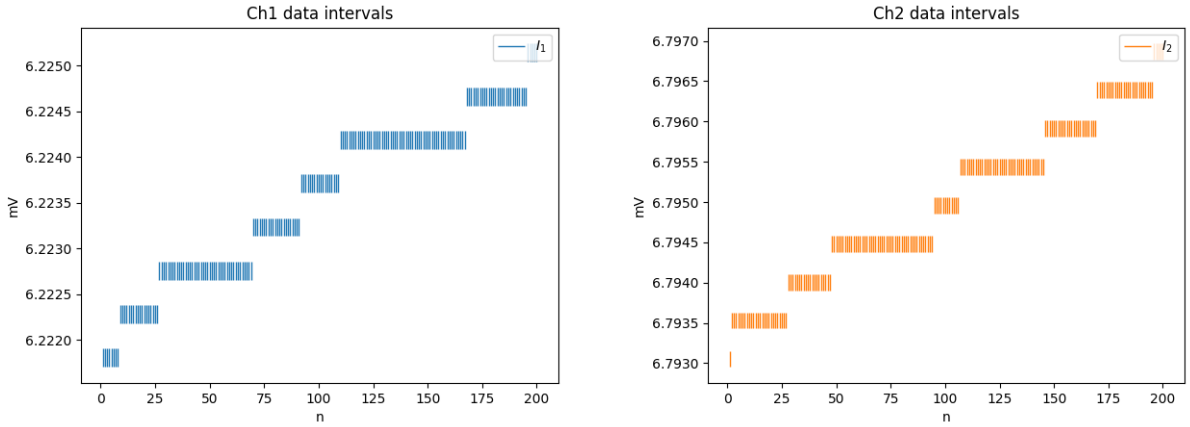

Рис. 3. Интервальное представление исходных данных

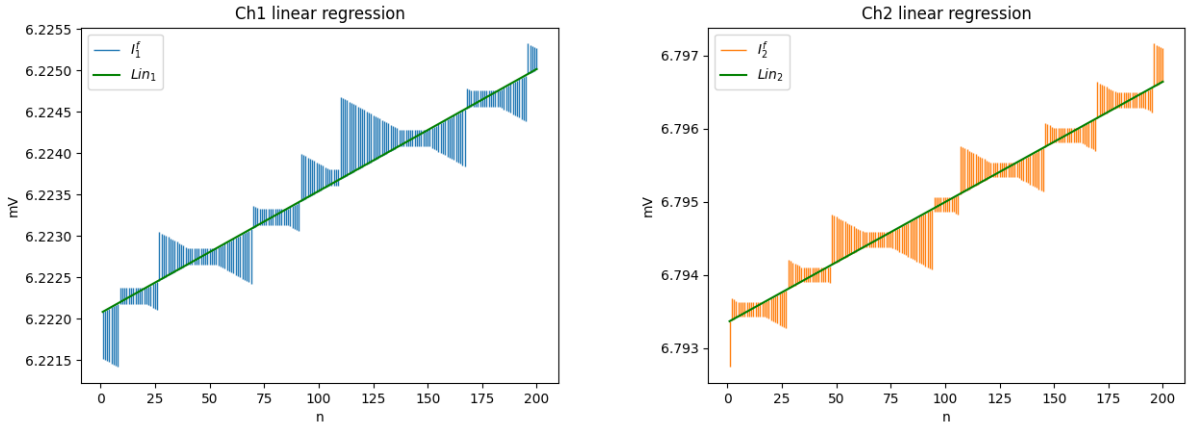

Рис. 4. Линейная модель дрейфа данных

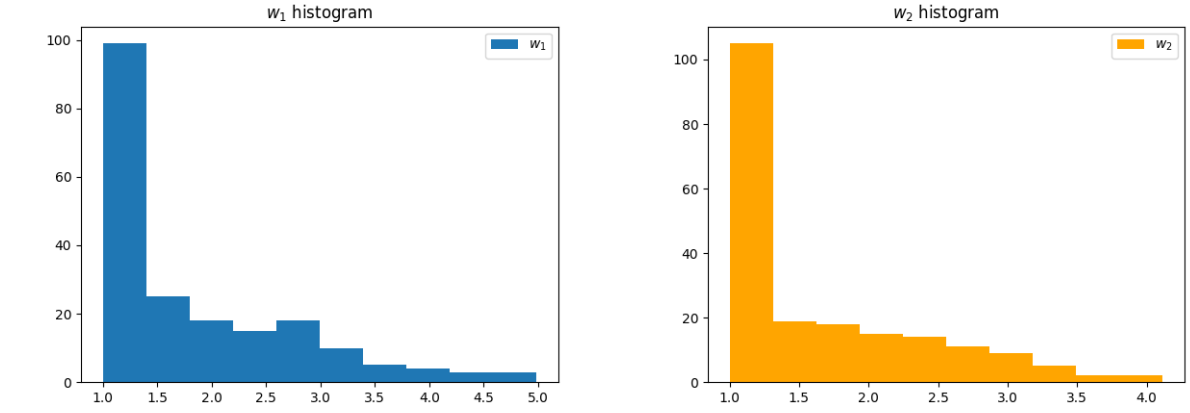

Рис. 5. Гистограммы значений множителей коррекции w

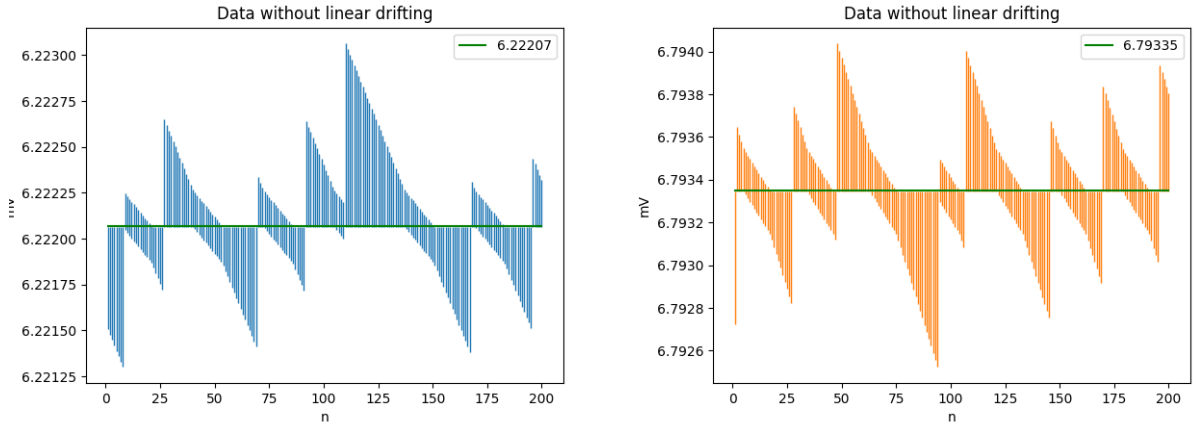

Рис. 6. Скорректированные модели данных

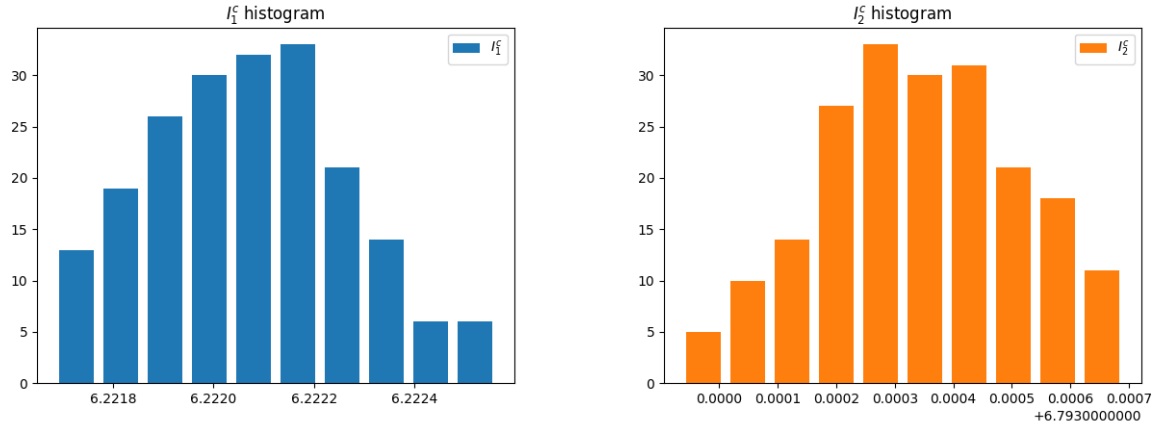

Рис. 7. Гистограммы скорректированных данных

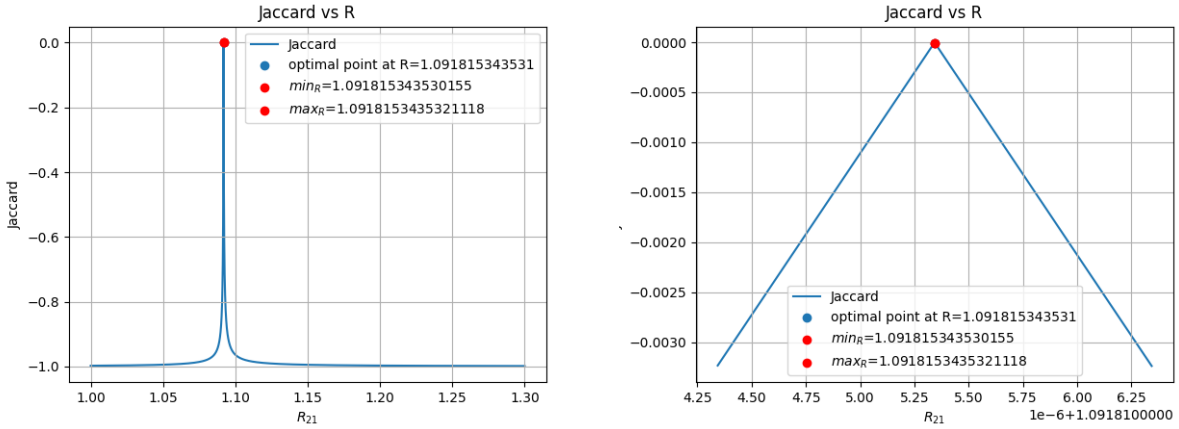

Рис. 8. Значение коэффициента Жаккара от калибровочного множителя от R_{21}

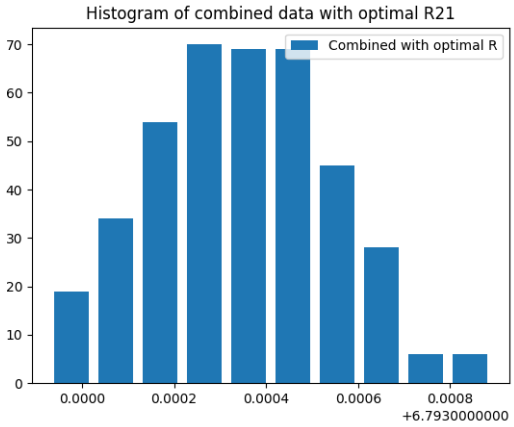

Рис. 9. Гистограмма объединённых данных при оптимальном значении R_{21}

5. Обсуждение

Множители коррекции w. На гистограммах значений множителей коррекции (Рис.5), видно, что половина (для эталлоного фотопередатчика даже больше) не требует коррекции. Этот факт свидетельстует о том, что линейная модель дрейфа данных является разумным приближением.

Коэффициент Жаккара На рис.8 видно, что оптимальным множителем R_{21} является число, равное 1.091815343531. Однако видно, что коэффициент Жаккара при оптимальном значении едва-едва превышает 0, а интервал, при котором JK $>= 0$, соизмерим с точкой (длина интервала оценивается $10^{-9} - 10^{-10}$). Это показывает на то, что исходные данные имеют ряд неточностей, которые сложно устранить.

Гистограмма объединённых данных при оптимальном значении R. Рассматривая гистограммы скорректированных данных(Рис. 7), можно заметить, что выборки I_1^f и I_2^f имеют характерные "пики"около центра, что имеет отражение на гисстограмме объединённых данных(Рис. 9).

6. Ссылка на репозиторий

https://github.com/PopovIV/MathematicalStatistics

Список литературы

[1] А.Н.Баженов Введение в анализ данных с интервальной неопределенностью. 2022.