Министерство науки и высшего образования Российской Федерации

Санкт-Петербургский политехнический университет Петра Великого

Физико-механический институт

Высшая школа прикладной математики и вычислительной физики

Дисциплина «Анализ данных с интервальной неопределенностью»

Отчет по лабораторным работам №1 и №2

Выполнил

Студент группы 5040103/90301 А. А. Северюхина

Принял

к. ф.-м. н., доцент А. Н. Баженов

Санкт-Петербург

2024

Содержание

1. Постановка задачи

Проводится исследование в области солнечной энергетики.

Чип быстрой аналоговой памяти PSI DRS4 имеет 8 каналов, каждый из которых содержит 1024 ячейки. Они содержат конденсаторы для хранения значения заряда и электронные ключи для записи сигналов и считывания напряжений через аналогово-цифровой преобразователь (АЦП). Ячейки объединяются в кольцевые буферы.

При подаче сигнала синхронизации запись напряжений на конденсаторы прекращается, а номер ячейки, в которую была сделана последняя запись, запоминается.

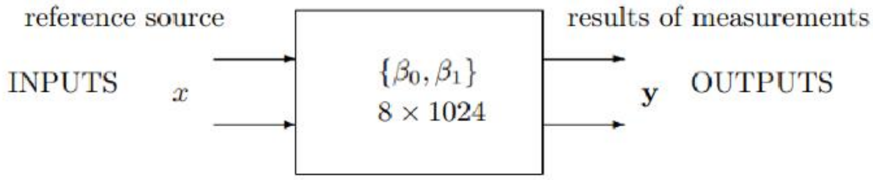

Рисунок 1. Структурная схема калибровки DRS4

Необходимо провести калибровку данного чипа.

В чип подается известное напряжение X и считываются полученные значения Y. Для каждого отдельного напряжения X эта процедура повторяется 100 раз.

Исходя из предположения, что $Y = \beta_0 * X + \beta_1$, выполняется линейная регрессия и определяются коэффициенты β_0, β_1.

2. Теория

2.1 Линейная регрессия

Пусть заданы две последовательности $X = \{x_i\}_{i=0}^{n}$, $Y = \{y_i\}_{i=0}^{n}$, $x_i, y_i \in \mathbb{R}\ \forall i = \overline{1,n}$. Линейно регрессией для этих последовательностей называется функция:

$$f(x) = \beta_0 * x + \beta_1 \tag{1}$$

Подобранная так, чтобы вектор $F = \{f(x_i)\}_{i=1}^{n}$ был максимально близок к вектору Y.

Таким образом, для решения задачи линейной регрессии необходимо найти коэффициенты β_0, β_1.

2.2 Первый подход: нахождение argmax(Tol)

Так как показания датчиков имеют погрешность, полученные данные необходимо рассматривать как интервалы, центр которых совпадает с показаниями, а радиус равен $\varepsilon = \frac{1}{2^{14}} = \frac{1}{16384}$.

Показания датчиков независимы, поэтому рассмотрим произвольную ячейку из 8*1024 ячеек. Для нее имеем 100 * 11 пар значений, где координата x соответствует напряжению и лежит в пределах [-0.5, 0.5], а координата y представляет собой интервал с шириной окна $wid = 2/16384$.

Для того, чтобы найти точечную оценку коэффициентов калибровки, воспользуемся распознающим функционалом Tol

$$\begin{gathered} Tol_i(x) = Tol(x, A, b) = \min_{1 \le i \le m} \{rad(b_i) - |(Ax)_i - mid(b_i)|\} = \\ = \min_{1 \le i \le m} \left\{ rad(b_i) - \left| \sum_{j=1}^{n} a_{ij} x_j - mid(b_i) \right| \right\} \end{gathered} \tag{2}$$

Где A – матрица вида

$$A = \begin{pmatrix} x_0 & 1 \\ \dots & \dots \\ x_m & 1 \end{pmatrix}$$

b – интервальный вектор

$$b = \begin{pmatrix} [y_0 - \ \varepsilon, y_0 + \varepsilon] \\ \dots \\ [y_m - \ \varepsilon, y_m + \varepsilon] \end{pmatrix}$$

Допустимое множество решений системы $Ax = b$ можно описать как

$$\{x \ \in \ \mathbb{R}^n | Tol(x, A, b)| \geq 0\}$$

Таким образом, если $Tol(argmax(Tol), A, b) \geq 0$, то система совместная и $argmax(Tol)$ (вектор, содержащий β_0и β_1) можно считать результатом регрессии.

Зачастую система не является совместной. В таком случае рассмотрим множество Tol_i.

$$Tol_i(x, A, b) = rad(b_i) - |\sum_{j=1}^{n} a_{ij}x_j - mid(b_i)|, 1 \leq \ i \leq \ m \#(3)$$

Заметим, что если существует i для которого выполняется $Tol_i < 0$, то $Tol < \ 0$. При этом, ля того, чтобы $Tol_i \geq 0$достаточно, чтобы $rad(b_i)$ был достаточно большим.

Таким образом, в случае отсутствия совместности, необходимо пройти по строкам матрицы A элементам b. Если для них $Tol_i < 0$, то нужно расширить интервал b_i, чтобы выполнялось $Tol_i = 0$. В таком случае $Tol(argmax(Tol), A, b) = 0$, а $argmax(Tol)$ содержит коэффициенты калибровки.

У этого подхода есть два основных недостатка.

- Расширение интервалов на практике приводит к сильной погрешности, так как интервалы расширяются не только в сторону регрессионной прямой, но и от нее.
- Результатом данного метода является точечная оценка.

2.3 Второй подход: нахождение оценки при помощи твинной арифметики

Еще один метод нахождения оценки регрессии основан на использовании твинной арифметики.

Разделим значения y_i на группы по 100 значений в зависимости от соответствующего им x_i. Тогда для каждого x_i мы получим набор значений, по которым можно построить boxplot. По boxplot определим внутреннюю и внешнюю оценки.

Для каждого x_j построим твин $[\left[\underline{y_j^{in}}, \overline{y_j^{in}}\right], [\underline{y_j^{ex}}, \overline{y_j^{ex}}]]$,

Построим распознающий функционал Tol, где

$$A = \begin{pmatrix} x_0 & 1 \\ x_0 & 1 \\ x_0 & 1 \\ x_0 & 1 \\ x_1 & 1 \\ \dots & .. \end{pmatrix}$$

$$b = \begin{pmatrix} \left[\underline{y_0^{in}}, \overline{y_0^{in}}\right] \\ \left[\underline{y_0^{ex}}, \overline{y_0^{in}}\right] \\ \left[\underline{y_0^{in}}, \overline{y_0^{ex}}\right] \\ \left[\underline{y_0^{ex}}, \overline{y_0^{ex}}\right] \\ \left[\underline{y_1^{in}}, \overline{y_1^{in}}\right] \\ \dots \end{pmatrix}$$

Если $Tol\big(argmax(Tol)\big) = 0$, решением будет $argmax(Tol)$.

В случае, если $Tol\big(argmax(Tol)\big) > 0$, можно найти множество значений (β_0, β_1) при которых $Tol > 0$.

Если $Tol\big(argmax(Tol)\big) < 0$, необходимо привести к виду, удовлетворяющему условию совместности.

Для этого рассмотрим Tol_i. Если $Tol_i < 0$, то будем удалять соответствующую строку из A и b. Так как для каждой пары (x_i, y_i) формируется 4 уравнения, в результате в системе останется больше уравнений, чем в первом методе, и решение будет точнее.

При этом, в результате данной операции возможно получить $Tol\big(argmax(Tol)\big) > 0$.

3. Реализация

Работа реализована на языке программирования Python 3.10 с использованием пакетов json, matplotlib, intvalpy.

Основные функции:

load_data – функция для считывания показаний датчиков

regression_type_first – функция для выполнения первого подхода решения задачи регрессии

regression_type_second - функция для выполнения второго подхода решения задачи регрессии

build_plots – функция для построения графиков

amount_of_neg – функция для поиска и удаления строк с отрицательным Tol

Для построения графиков, в том числе коридора совместности используется еще несколько методов:

unique – удаляет дубликаты из массива и округляет значения

clear_zero_rows - удаляет строки и элементы массивов, если все элементы близки к нулю (по сравнению с заданным порогом)

get_boundary_intervals – вычисляет границы интервалов

get_particular_points – находит особые точки на основе границ интервалов

get_intervals_path – находит последовательность точек на основе массива интервалов

lineqs – находит вершины множества решений

IntLinIncR2 – используется для отображения множества решений

4. Результаты

Для рассмотрения значений, каждому датчику в чипе были даны координаты в зависимости от номера канала и ячейки. Таким образом, датчик, получивший данные из канала j $(1 \leq j \leq 8)$ и находившийся в ячейке

$(1 \le j \le 1024)$ будет иметь координаты i, j. Рассматриваются данные для датчиков с координатами

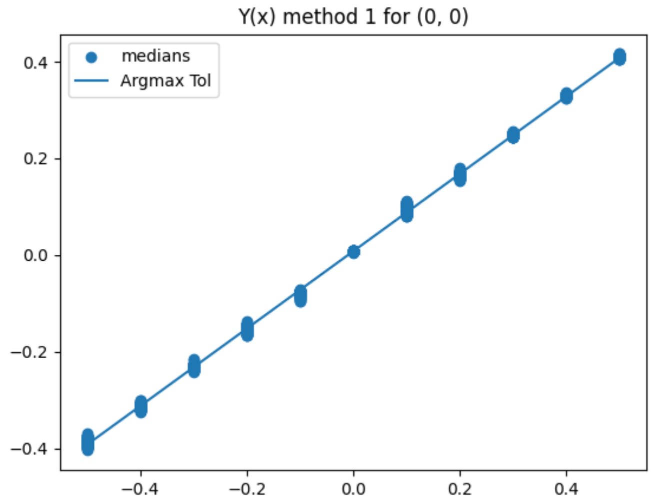

Рисунок 2. Калибровочная кривая для датчика (0,0), полученная первым методом

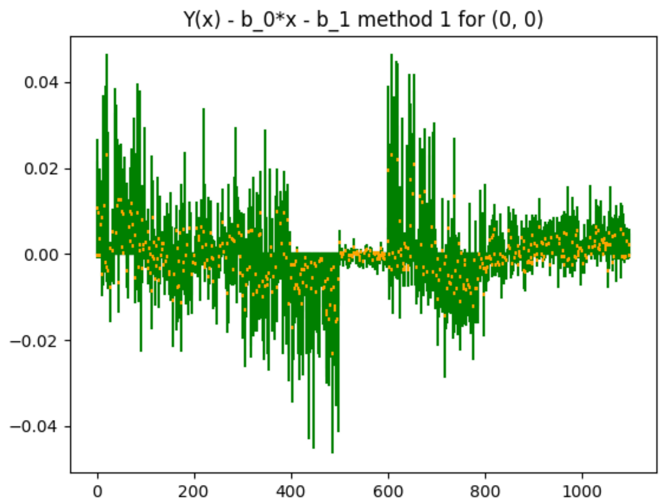

Рисунок 3. Разность между данными и калибровочной прямой для первого метода и датчика (0,0). Зеленым обозначен новый интервал, желтым - новый

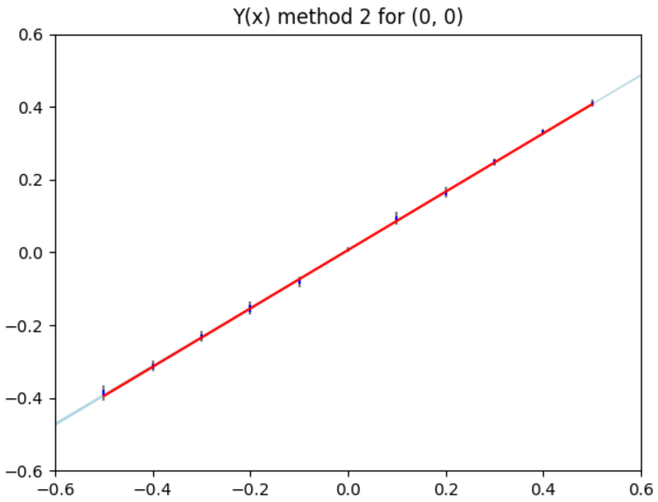

Рисунок 4. Калибровочная прямая полученная вторым методом для датчика (0,0) (обозначена красным цветов). Твины обозначены серым и синим цветом. Коридоры совместности обозначены голубым и светло-серым

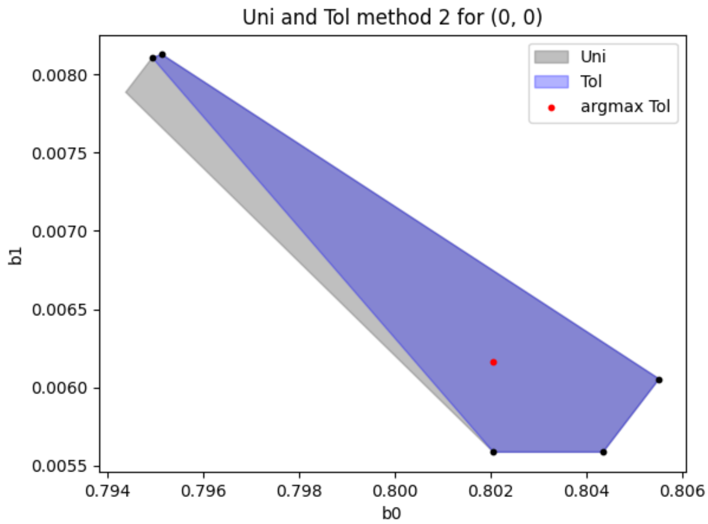

Рисунок 5. Uni, Tol и argmax(Tol) для датчика (0,0)

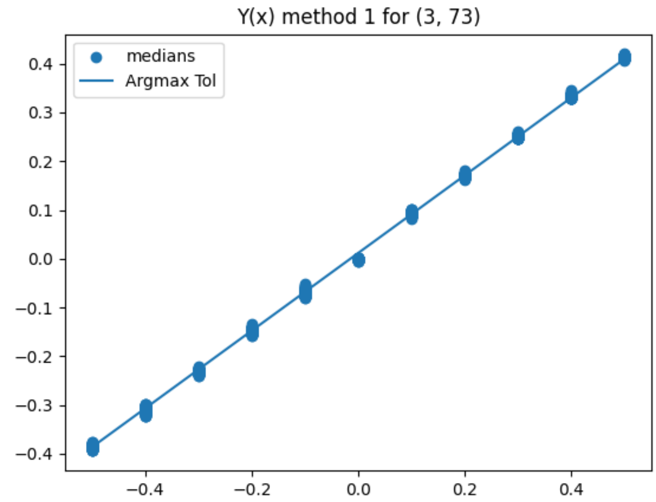

Рисунок 6. Калибровочная кривая для датчика (3,73), полученная первым методом

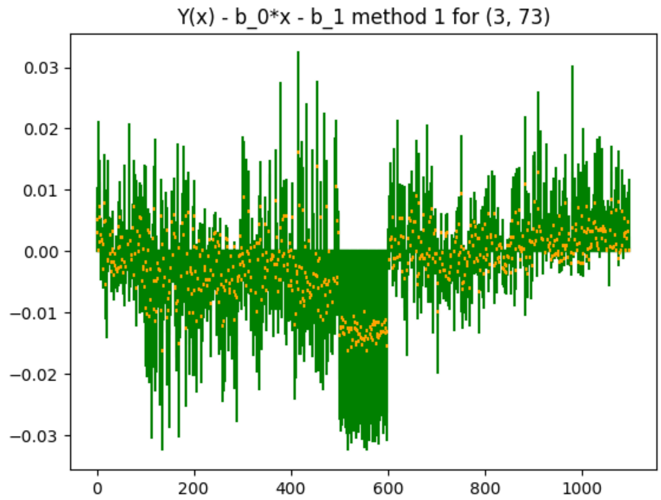

Рисунок 7. Разность между данными и калибровочной прямой для первого метода и датчика (3,73). Зеленым обозначен новый интервал, желтым - новый

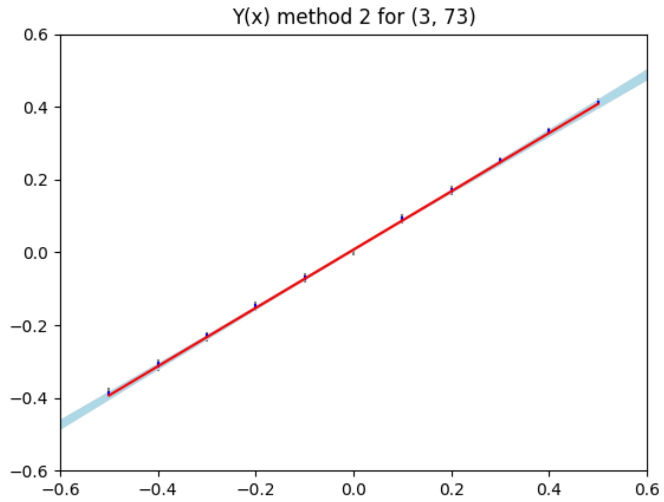

Рисунок 8. Калибровочная прямая полученная вторым методом для датчика (3,73) (обозначена красным цветов). Твины обозначены серым и синим цветом. Коридоры совместности обозначены голубым и светло-серым

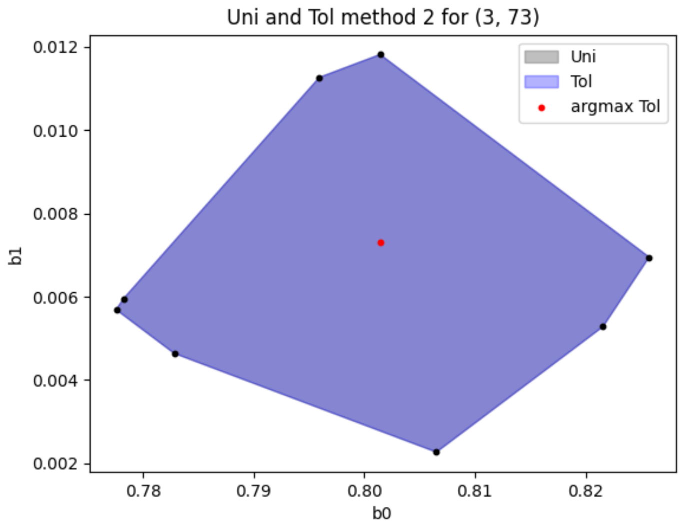

Рисунок 9. Uni, Tol и argmax(Tol) для датчика (3,73)

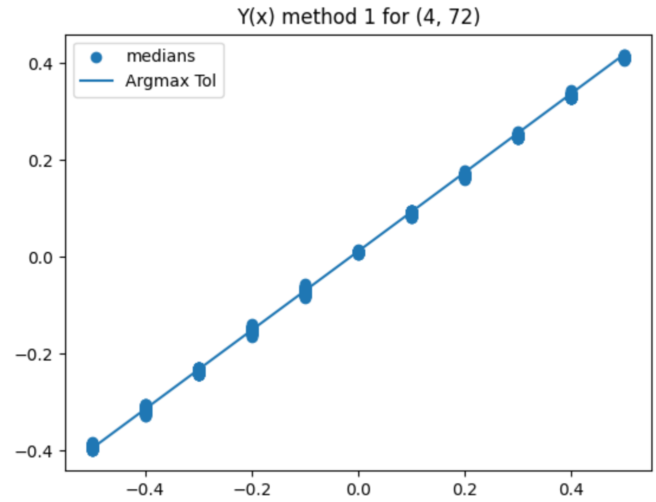

Рисунок 10. Калибровочная кривая для датчика (4, 72), полученная первым методом

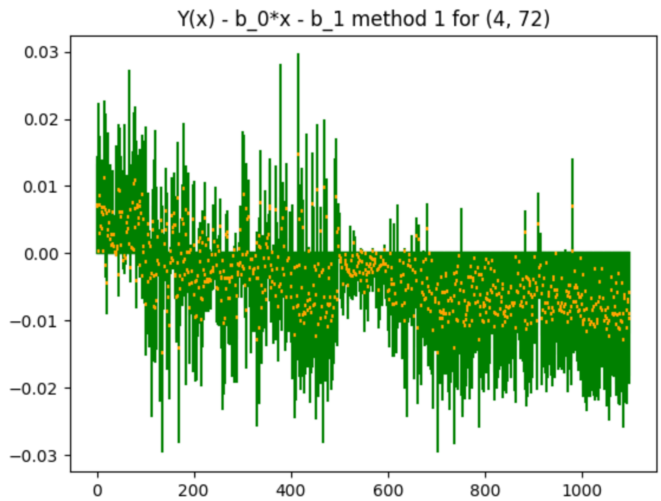

Рисунок 11. Разность между данными и калибровочной прямой для первого метода и датчика (4, 72). Зеленым обозначен новый интервал, желтым - новый

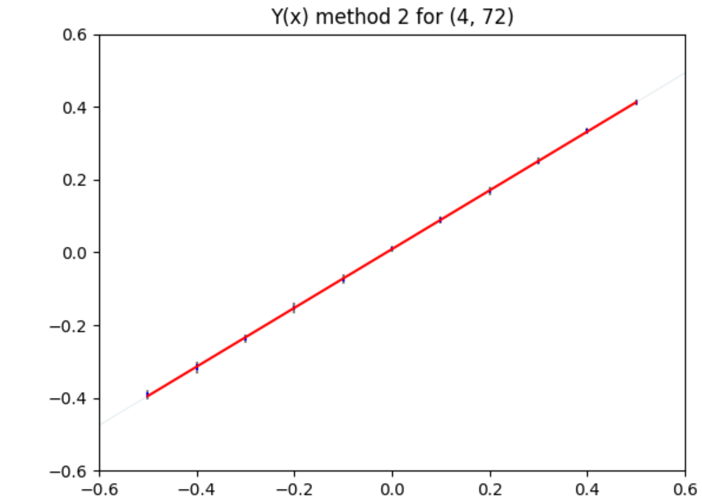

Рисунок 12. Калибровочная прямая полученная вторым методом для датчика (4, 72) (обозначена красным цветов). Твины обозначены серым и синим цветом. Коридоры совместности обозначены голубым и светло-серым

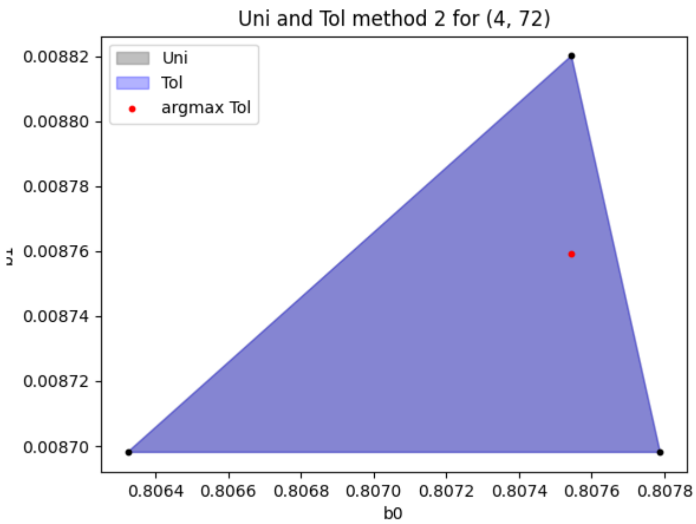

Рисунок 13. Uni, Tol и argmax(Tol) для датчика (4, 72)

Численные результаты представлены в таблице 1:

Таблица 1. Численные результаты

Номер датчика	Метод	β_0	β_1	Количество модифицированных интервалов
(0, 0)	1	0.816	0.011	1094
(0, 0)	2	0.808	0.009	0
(3, 73)	1	0.801	0.008	1085

(3, 73)	2	0.802	0.006	12
(4, 72)	1	0.797	0.012	1089
(4, 72)	2	0.801	0.007	32

5. Выводы

В ходе работы было реализовано решение задачи регрессии двумя методами: нахождением argmax(Tol) и нахождением оценки при помощи твинной арифметики. Можно заметить, что результаты являются близкими, но не совпадают.

Приложения

1. Репозиторий, содержащий программу реализации передачи сообщений и отчет
   https://github.com/AnastasyaSeveryukhina/interval-and-networks