Geometric Foundations of Data Analysis I: Week 1

Joshua Maglione

8 September 2025

Contents

1 Introduction

Data analysis and more broadly Data Science is a vast and important field within Computer Science and Mathematics. At the core, the goal is to make sense of data, which can be measurements, survey results, behavior patterns, etc. Often this data comes to us in a very "high dimension". That is, there are so many variables that it is impossible to visualize, and even in low dimensions, it may not be clear what the best conclusion is based on the analysis.

A few references seem to agree that the *total data* on all computers is something like $\mathbf{10^{23}}$ **bytes** or about **100 zettabytes**. While all of this data is not concentrated in one organization, we still require highly sophisticated tools to make sense of a huge amount of data. My goal with this course is that you will have a solid foundation with some standard tools. From this bedrock one could explore more sophisticated methods of data analysis more easily.

We will consider four key topics:

1. Least Squares Fitting,
2. Principal Component Analysis,
3. k-Means and Hierarchical Clustering,
4. Nearest Neighbors and the Johnson–Lindenstrauss Theorem

We will be working with the assumption that **the data we care about is preserved by orthogonal and linear transformations**. This is not true with all data—for example, one should not take (proper) linear combinations of people. However, for data like grams of different kinds of food, this is completely plausible. This assumption will not always be necessary, but we will just keep this in mind.

Half of our time will be spent bringing these ideas to life and getting our hands dirty. We will be working with Jupyter Notebooks to build familiarity with the concepts we will discuss. This will be done using Python and standard data analysis packages like Pandas.

2 Least squares fitting

This method of analysis is both simple and powerful. Like with most things in mathematics, without some good guiding examples, we can get lost in the formulas.

2.1 Build up

Suppose we have the following data as seen in Figure 2.1. (Maybe a company produces parts once per month in lots that vary in size depending on demand. We write i for the production run, x_i for the lot size, and y_i for the hours of labor.)

i	x_i	y_i
1	30	73
2	20	50
3	60	128
4	80	170
5	40	87
6	50	108
7	60	135
8	30	69
9	70	148
10	60	132

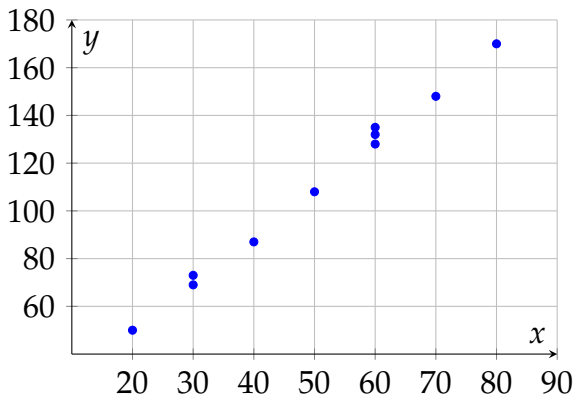


Figure 2.1: Data points exhibiting a linear relationship.

It seems clear from the plot that the data fits a geometric pattern—there is a linear phenomenon. If we try to find the line that is somehow closest to all the data points, we might draw something like in Figure 2.2.

Although the line is not a perfect fit, it seems to tell us something about the relationship between the lots size and the amount of hours.

Questions:

- What makes this line "better" than alternatives?
- How are we quantifying "better"?

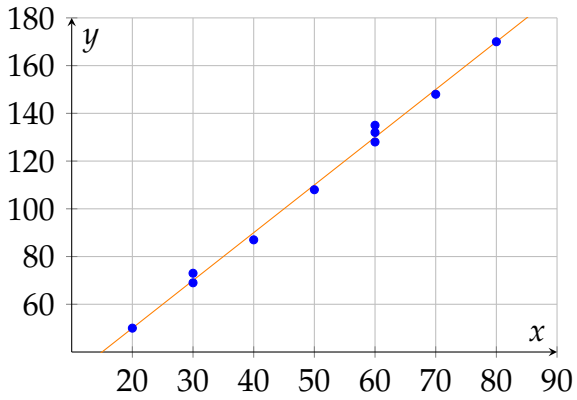


Figure 2.2: Line of best fit with data points.

- Why are we using a line?

We will answer the first question later, probably. Let us consider the question of quantifying "better". Least squares fitting is all about minimizing the squares of differences between the line and the actual data points. We will make this precise very soon.

2.2 Line of best fit

We know that the equation of a non-vertical line has the form

$$y = b_0 + b_1 x.$$

If we had n data points of the form (x_i, y_i), we could choose b_0 and b_1 to minimize the following sum

$$S(b_0, b_1) = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2 .$$

We can even solve for these values. Since S is a function in terms of b_0 and b_1, all possible minima occur when the partial derivatives of S are 0. In other words, the minima arise as values (b_0, b_1) such that

$$\frac{\partial S}{\partial b_0} = -2 \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i)) = 0,$$
$$\frac{\partial S}{\partial b_1} = -2 \sum_{i=1}^{n} (x_i y_i - x_i (b_0 + b_1 x_i)) = 0.$$

These equations are linear equations, so we can solve for these with techniques from linear algebra. This mean, we need to solve two equations in the unknown b_0 and b_1:

$$\begin{aligned} n b_0 + b_1 \sum x_i &= \sum y_i, \\ b_0 \sum x_i + b_1 \sum x_i^2 &= \sum x_i y_i. \end{aligned} \tag{2.1}$$

Using the data from our example in Section 2.1, we have

$$\sum x_i = 500, \qquad \sum y_i = 1100, \qquad \sum x_i^2 = 28400, \qquad \sum x_i y_i = 61800.$$

Thus, the equations we need to solve are

$$\begin{aligned} 10b_0 + 500b_1 &= 1100, \\ 500b_0 + 28400b_1 &= 61800, \end{aligned}$$

which yield $b_0 = 10$ and $b_1 = 2$. Going back to the context of the initial problem: this solution tells us that by increasing the lot size by one, we expect to increase the labor hours by two.

2.2.1 Written as matrices

Let us write the equations in (2.1) with matrices. This might seem like overkill at this stage, but it will set us up nicely to generalize. Let

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \qquad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad B = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}. \tag{2.2}$$

Therefore, the equations in (2.1) are equivalent to the single matrix equation:

$$X^t X B = X^t Y. \tag{2.3}$$

If $X^t X$ is invertible, then $B = (X^t X)^{-1} X^t Y$.

2.3 In class exercises pt. I

1. (a) With X, Y, and B as defined in Equation (2.2), show that

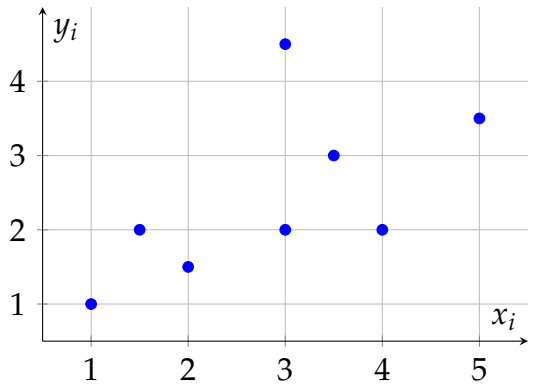
$$X^t X = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}, \qquad X^t Y = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}.$$

   (b) Show that $X^t X B = X^t Y$ is equivalent to Equation (2.1):

$$\begin{aligned} n b_0 + b_1 \sum x_i &= \sum y_i, \\ b_0 \sum x_i + b_1 \sum x_i^2 &= \sum x_i y_i. \end{aligned}$$

2. Find a least squares fitting line to the following data and draw in the line:

i	x_i	y_i
1	1.0	1.0
2	2.0	1.5
3	3.0	2.0
4	1.5	2.0
5	3.5	3.0
6	3.0	4.5
7	4.0	2.0
8	5.0	3.5



(Round b_0 and b_1 to the nearest half integer.)