Geometric Foundations of Data Analysis I

Joshua Maglione

September 25, 2023

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 **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. Hierarchical Clustering and Persistence,
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.)

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"?
- 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.

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

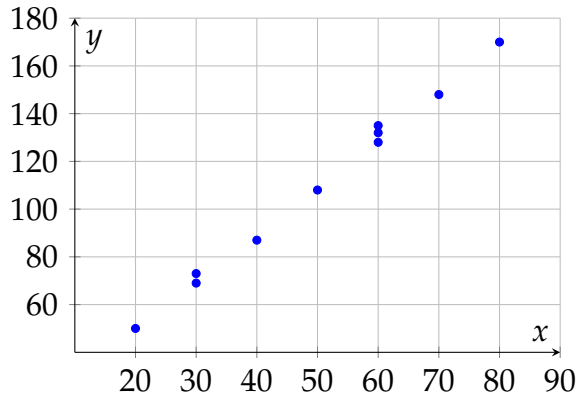


Figure 2.1: Data points exhibiting a linear relationship.

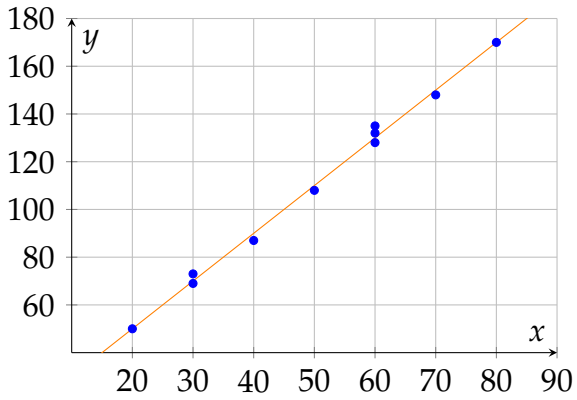


Figure 2.2: Line of best fit with data points.

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:

$$\begin{aligned} nb_0 + b_1 \sum x_i &= \sum y_i, \\ b_0 \sum x_i + b_1 \sum x_i^2 &= \sum x_i y_i. \end{aligned} \tag{2.1}$$

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:

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

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

2.3 In class exercises pt. I

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

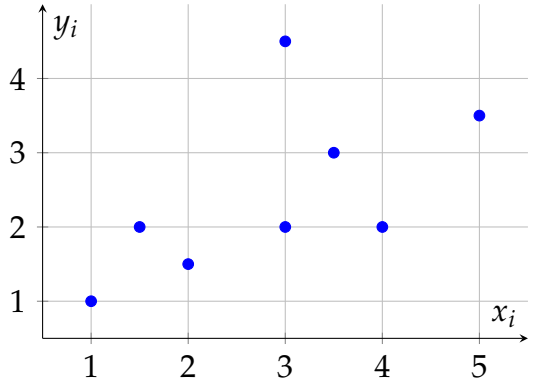
$$X^{\mathrm{t}} X = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}, \qquad X^{\mathrm{t}} Y = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}.$$

   (b) Show that $X^{\mathrm{t}}XB = X^{\mathrm{t}}Y$ is equivalent to Equation (2.1):

$$\begin{aligned} nb_0 + b_1 \sum x_i &= \sum y_i, \\ b_0 \sum x_i + b_1 \sum x_i^2 &= \sum x_i y_i. \end{aligned}$$

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.)

2.4 Plane of best fit

We consider two independent variables and one dependent variable now. Consider the following data points as given in Figure 2.3.

i	x_{i1}	x_{i2}	y_i
0	278	36	287
1	252	31	256
2	344	35	300
3	134	33	182
4	215	35	248
5	261	40	271
6	131	39	149
7	463	43	411
8	167	46	214
9	298	42	291
10	230	60	314
11	293	67	352
12	290	37	298
13	271	31	252
14	385	63	439
15	354	36	328

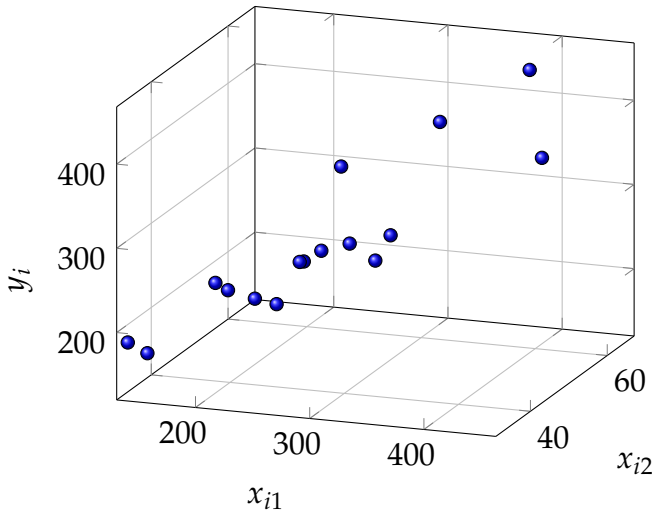


Figure 2.3: Data points in $\mathbb{R}^3$.

We can put some meaning to these data. For example, suppose a company is selling a product, and we have 16 populations of people labeled 0 through 15. The values x_{i1} are the population sizes in 100s of people; the values x_{i2} are the average yearly income in €1000 per capita; and the values y_i are the number of sales of the product. (There might be dependencies between population size and average income, but our model treats them as independent.)

It looks like though there is a plane of best fit for the data—thanks to the

suggestive viewing angle. Our goal is to find a plane, given by

$$y = b_0 + b_1 x_1 + b_2 x_2.$$

We can just do what we did last time. That is, for

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix}, \qquad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad B = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix},$$

we need to solve for B in the equation

$$X^t X B = X^t Y.$$

Thus, if $X^t X$ is invertible, there is a unique B, which is equal to $(X^t X)^{-1} X^t Y$. For our example, we have

$$X^t X = \begin{pmatrix} 16 & 4366 & 674 \\ 4366 & 1309480 & 187024 \\ 674 & 187024 & 30330 \end{pmatrix}, \qquad X^t Y = \begin{pmatrix} 4592 \\ 1343400 \\ 200571 \end{pmatrix}.$$

Therefore, the plane of best fit is approximately

$$y = -11.3 + 0.7 x_1 + 2.6 x_2.$$

Putting all the data together we have a plane of best fit as seen in Figure 2.4.

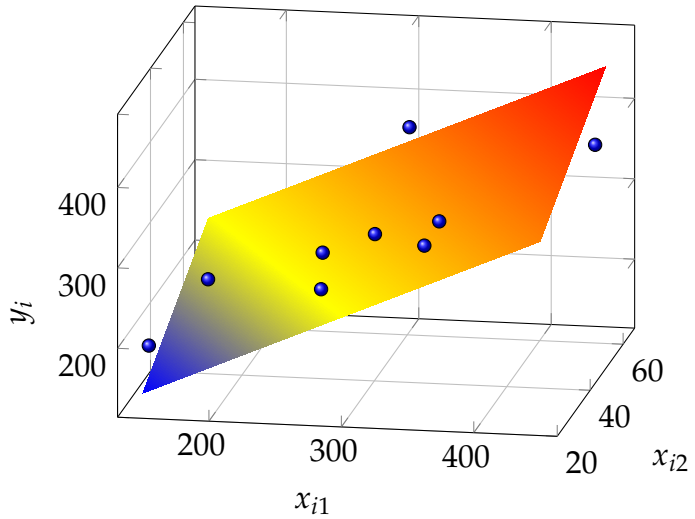


Figure 2.4: Data points together with the plane of best fit.

2.5 Hyperplane of best fit

Now we go to the general case. Suppose we have $p-1$ independent variables and 1 dependent variable, where $p \geqslant 2$. We assume we have n data points of the form

$$(x_{i1}, x_{i2}, \ldots, x_{i,p-1}, y_1) \in \mathbb{R}^p.$$

The least squares fitting for these data is a hyperplane of the form

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_{p-1} x_{p-1}.$$

To solve for the values b_i, we do as we did before. We define matrices

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,p-1} \end{pmatrix}, \qquad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad B = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{pmatrix}.$$

As before, the values we want are given by the equation

$$X^{\mathrm{t}} X B = X^{\mathrm{t}} Y. \tag{2.4}$$

2.6 Why Equation (2.4) works

The heart of least squares is (Euclidean) distance. The distance between two points $x = (x_1, \ldots, x_p)$ and $y = (y_1, \ldots, y_p)$ in $\mathbb{R}^p$ is

$$d(x, y) = \|x - y\| = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2}.$$

For a vector $v = (v_1, \ldots, v_p) \in \mathbb{R}^p$, the **length** of v is

$$\|v\| = d(0, v) = \sqrt{v_1^2 + v_2^2 + \cdots + v_p^2}.$$

Recall that the dot product of two (column) vectors u and v is

$$u \cdot v = u^{\mathrm{t}} v = u_1 v_1 + u_2 v_2 + \cdots + u_p v_p.$$

Thus, the length of v is $\|v\| = \sqrt{v \cdot v}$; in other words $\|v\|^2 = v \cdot v$. In addition, if $u \cdot v = 0$, we say that u and v are **orthogonal** (or perpendicular).

The goal of least squares is to *minimize distance*; more specifically to minimize $\|Y - XB\|$. Note that the column vector Y has entries that are the *actual* y_i *values*, and the column vector

$$XB = \begin{pmatrix} B \cdot (1, x_{11}, x_{12}, \ldots, x_{1,p-1}) \\ B \cdot (1, x_{21}, x_{22}, \ldots, x_{2,p-1}) \\ \vdots \\ B \cdot (1, x_{n1}, x_{n2}, \ldots, x_{n,p-1}) \end{pmatrix} = \begin{pmatrix} b_0 + b_1 x_{11} + b_2 x_{12} + \cdots + b_{p-1} x_{1,p-1} \\ b_0 + b_1 x_{21} + b_2 x_{22} + \cdots + b_{p-1} x_{2,p-1} \\ \vdots \\ b_0 + b_1 x_{n1} + b_2 x_{n2} + \cdots + b_{p-1} x_{n,p-1} \end{pmatrix}$$

Therefore, $\|Y - XB\|$ is the square root of a sum of squares of the form

$$y_i - b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_{p-1} x_{i,p-1}.$$

Hence minimizing $\|Y - XB\|$ is the same as minimizing $\|Y - XB\|^2$, which is a sum of *squares*.

Proposition 2.1. *The minimal distance $\|Y - XB\|$ is achieved by solving for B in*

$$X^{\mathrm{t}}XB = X^{\mathrm{t}}Y.$$

Proof. Consider the subspace $U = \{Xu \mid u \in \mathbb{R}^p\}$ of $\mathbb{R}^n$, and observe that our desired solution XB is contained in U. Since $\|Y - XB\|$ is minimal, we must have that the vector $Y - XB$ is orthogonal to all vectors contained in U.[1] That is, $(Xu) \cdot (Y - XB) = 0$ for all $u \in \mathbb{R}^p$. In other words, we have for all $u \in \mathbb{R}^p$,

$$\begin{aligned} 0 = (Xu)^{\mathrm{t}}(Y - XB) &= u^{\mathrm{t}}X^{\mathrm{t}}(Y - XB) \\ &= u^{\mathrm{t}}\left(X^{\mathrm{t}}Y - X^{\mathrm{t}}XB\right). \end{aligned}$$

Because $u^{\mathrm{t}}\left(X^{\mathrm{t}}Y - X^{\mathrm{t}}XB\right) = 0$ for all $u \in \mathbb{R}^p$, it follows that $X^{\mathrm{t}}Y - X^{\mathrm{t}}XB = 0$. □

2.7 In class exercises pt. II

1. Determine $X^{\mathrm{t}}X$ and $X^{\mathrm{t}}Y$ with

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,p-1} \end{pmatrix}, \qquad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

2. Using (1) and by taking partial derivatives of

$$S(b_0, \ldots, b_{p-1}) = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_{p-1} x_{i,p-1}))^2, \quad (2.5)$$

show that the hyperplane of best fit is obtained by solving $X^{\mathrm{t}}XB = X^{\mathrm{t}}Y$. (You could try this for $p = 3$ first.)

2.8 Nonlinear fittings

Although all of our examples so far have been linear fittings, we will demonstrate that least squares fittings works in the nonlinear case. What is important is that we have a candidate equation to fit. In the linear cases, we tried to fit

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_{p-1} x_{p-1}.$$

Suppose we have the following data as given in Figure 2.5. Instead of trying to fit the line $y = b_0 + b_1 x$, we could try to fit the parabola:

$$y = b_0 + b_1 x + b_2 x^2.$$

We can treat this the same way as before. Of course the quantities x and x^2 are *not* independent, but we can ignore this. Set

$$x_{i1} = x_i, \qquad x_{i2} = x_i^2.$$

x_i	y_i
2.27	2.50
5.06	-16.13
1.45	4.23
5.89	-22.46
0.48	1.37
-0.22	0.86
1.44	11.85
-1.77	-14.71
2.45	9.42
-1.54	-14.07
7.55	-55.62
1.76	4.45
5.16	-19.56
3.26	-2.79
3.23	5.20
0.85	8.09

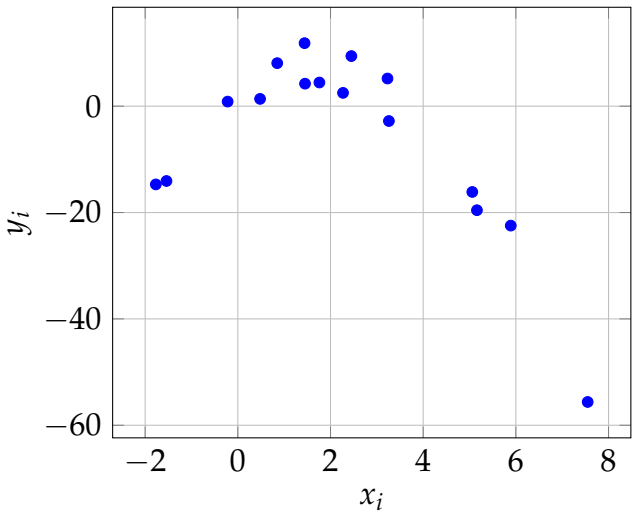


Figure 2.5: Data points demonstrating a nonlinear relationship.

Therefore, the hyperplane of best fit for the data (x_{i1}, x_{i2}, y_i) will give us the parabola of best fit. *Try this on your own!*

So one can fit any hypersurface $y = f(x_1, \ldots, x_{p-1})$ to the given data. The function f in this case is called the **regression function**. This general method of analysis is known as **regression analysis**. A few questions arise:

- Which surface is "best"?
- How can we quantify "best"?
- Even in the line case $(p = 2)$, how can we quantify how well data fits our line?

2.9 Coefficient of determination (r^2 values)

We are going to make precise how well our hyperplane fits our data. Recall that hyperplanes can be replaced by hypersurfaces; see Section 2.8. First we establish some notation. Suppose we have n data points $(x_{i1}, x_{i2}, \ldots, x_{i,p-1}, y_i) \in \mathbb{R}^p$. Then we define

$$\begin{aligned} \text{(Fitted value)} \qquad & \widehat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_{p-1} x_{i,p-1}, \\ \text{(Residual)} \qquad & e_i = y_i - \widehat{y}_i, \\ \text{(Sample mean)} \qquad & \overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i. \end{aligned}$$

[1] To see why this is true, see Section 6.3.1 of [1], which is all about orthogonal decompositions.

These yield vectors in $\mathbb{R}^n$ as follows

$$\widehat{Y} = \begin{pmatrix} \widehat{y}_1 \\ \vdots \\ \widehat{y}_n \end{pmatrix} = XB, \qquad E = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} = Y - \widehat{Y}, \qquad \overline{Y} = \begin{pmatrix} \overline{y} \\ \vdots \\ \overline{y} \end{pmatrix} = \overline{y} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

From our n data points, we have three points in $\mathbb{R}^n$ given by Y, $\widehat{Y}$, and $\overline{Y}$. Three points always lie on a plane, so the three point determine a triangle on such a plane. What does this triangle look like? If it is a triangle (and not a line or a single point), then the next lemma proves it must be a right triangle.

Lemma 2.2. *The vectors $E = Y - \widehat{Y}$ and $\widehat{Y} - \overline{Y}$ are orthogonal.*

Proof. Suppose $X^{\mathrm{t}}XB = X^{\mathrm{t}}Y$. We need to prove two equations. For the first,

$$0 = X^{\mathrm{t}}(Y - XB) = X^{\mathrm{t}}(Y - \widehat{Y}) = X^{\mathrm{t}}E.$$

Hence, $X^{\mathrm{t}}E = 0$. For the second,

$$\begin{aligned} \overline{Y}^{\mathrm{t}}E &= \overline{y}\sum_{i=1}^{n}(y_i - \widehat{y}_i) \\ &= \overline{y}\sum_{i=1}^{n}(y_i - (b_0 + b_1x_{i1} + b_2x_{i2} + \cdots + b_{p-1}x_{i,p-1})) \\ &= -\frac{\overline{y}}{2}\cdot\frac{\partial S}{\partial b_0} = 0, \end{aligned}$$

where S is defined in Equation (2.5), so $\overline{Y}^{\mathrm{t}}E = 0$. Thus, we have

$$(\widehat{Y} - \overline{Y}) \cdot E = (XB)^{\mathrm{t}}E - \overline{Y}^{\mathrm{t}}E = 0. \qquad \square$$

Remark 2.3. One can simplify the proof for Lemma 2.2 by applying an isometry to the data, so that $\overline{y} = 0$. That is, one only needs to prove that E and $\widehat{Y}$ are orthogonal.

The lengths of the differences of the vectors are important and have names:

$$\begin{aligned} \text{(Sums of Squares Total – SST)} &: \|Y - \overline{Y}\|^2, \\ \text{(Sums of Squares Error – SSE)} &: \|Y - \widehat{Y}\|^2, \\ \text{(Sums of Squares Regression – SSR)} &: \|\widehat{Y} - \overline{Y}\|^2. \end{aligned}$$

The value SST measures the *total variability* of the data set. For example, $\sqrt{SST} = \|Y - \overline{Y}\|$ is the distance from the actual data Y to the sample mean $\overline{Y}$. Using the same ideas, we can see that SSE measures the error of our regression and that SSR measures the distance from our regression to the sample mean.

Proposition 2.4.

$$SST = SSE + SSR.$$

Proof. Apply Lemma 2.2 and the Pythagorean Theorem:

$$\|Y - \overline{Y}\|^2 = \|Y - \widehat{Y}\|^2 + \|\widehat{Y} - \overline{Y}\|^2. \qquad \square$$

Now we can describe a quantity that measures how good our regression fits the given data.

Definition 2.5. The **coefficient of determination** (also known as the r^2**-value**) is

$$r^2 = \frac{SSR}{SST} = \frac{\|\widehat{Y} - \overline{Y}\|^2}{\|Y - \overline{Y}\|^2}.$$

Proposition 2.6. $0 \leqslant r^2 \leqslant 1$.

Proof. Since each SST and SSR are squares, they are nonnegative. By Proposition 2.4, we have $0 \leqslant SSR \leqslant SST$. $\square$

2.9.1 What do the extremes means?

The one case where r^2 is meaningless is when $SST = 0$. This implies both $SSR = SSE = 0$. Moreover, $Y = \overline{Y} = \overline{y}\mathbb{1}$, where $\mathbb{1}$ is the all ones column vector. Hence, every data point y_i is the same and, therefore, equal to the mean. Let's never return to this case.

We can have $SSR = 0$, which is equivalent to $r^2 = 0$. This implies that $\|\widehat{Y} - \overline{Y}\|^2 = 0$, so that $\widehat{Y} = \overline{Y}$. In other words, our prediction $\widehat{y}_i$ is just simply the mean. This means we have not found any relationship between the independent variables and the dependent variables.

In the other extreme we have $SSR = SST$, which is equivalent to $r^2 = 1$. This implies that $Y = \widehat{Y}$, so the given data lies (exactly) on the surface given by $y = f(x_1, \ldots, x_{p-1})$. That is, the regression function exactly predicts the data.

To summarize, when $r^2 = 0$, we cannot deduce any relationship between the independent and dependent variables, and when $r^2 = 1$, we understand completely the relationship between the independent and dependent variables. Very roughly speaking, the r^2 can be thought of as the ratio of how well the regression fits the data.

2.10 In class exercises pt. III

1. Prove the following.
   (a) $\|\overline{Y}\|^2 = n\overline{y}$.
   (b) $Y \cdot \overline{Y} = \widehat{Y} \cdot \overline{Y} = \|\overline{Y}\|^2$.
   (c) $Y \cdot \widehat{Y} = \|\widehat{Y}\|^2$.
2. Use (1) to show that
   (a) $SST = \|Y\|^2 - \|\overline{Y}\|^2$,

(b) $SSE = \|Y\|^2 - \|\widehat{Y}\|^2$,

(c) $SSR = \|\widehat{Y}\|^2 - \|\overline{Y}\|^2$.

3. What are the r^2 values for the examples above?

3 Principal component analysis

Principal component analysis (PCA) is a power method of analysis that comes standard in all data science tool kits. With little effort, one can reduce a complex data set to data that we can more easily see structure. More specifically, the goal of PCA is to find the "best" basis to express the data. In other words, our initial reference frame may not be the one that best expresses the structure of our data—PCA is a method to find the "best" reference frame.

3.1 Introducing PCA

We will start with a toy example, where the analysis is quite simple. Suppose we have many many data points in $\mathbb{R}^2$ as seen in Figure 3.1.

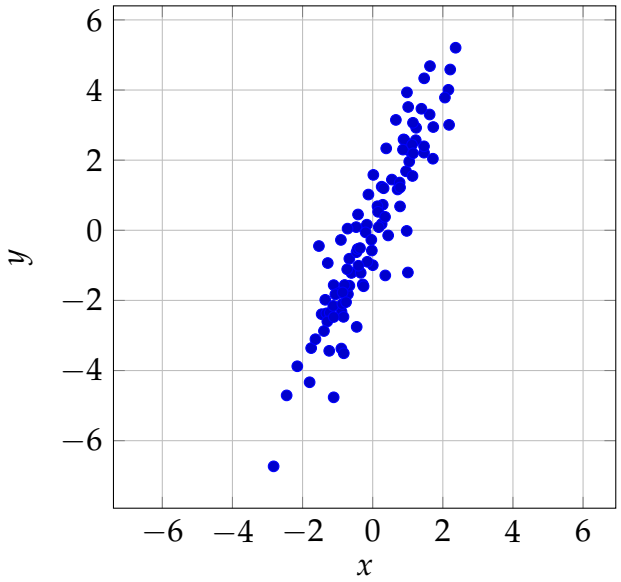


Figure 3.1: Some data in $\mathbb{R}^2$.

Let's assume that our data points live in $\mathbb{R}^m$, so that $m = 2$ in Figure 3.1. Suppose we write those data points in an $m \times n$ matrix X. Our current (and default) basis is $\{(1,0,\ldots,0),(0,1,0,\ldots,0),\ldots,(0,\ldots,0,1)\}$, and we want a basis $\{p_1, p_2, \ldots, p_m\}$ that better reflects the structure of our data. That is, we want an $m \times m$ matrix P, whose rows are the p_i, that provides us a better reference frame. Therefore, we want to transform our data X into a new data set Y such that

$$PX = Y.$$

The columns of X are the "old" data, and the columns of Y are the "new" data. If the p_i are row vectors and the x_i column vectors, then we want

$$PX = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_m \end{pmatrix} \begin{pmatrix} x_1 & x_2 & \cdots x_n \end{pmatrix} = \begin{pmatrix} p_1 \cdot x_1 & p_1 \cdot x_2 & \cdots & p_1 \cdot x_n \\ p_2 \cdot x_1 & p_2 \cdot x_2 & \cdots & p_2 \cdot x_n \\ \vdots & \vdots & \ddots & \vdots \\ p_m \cdot x_1 & p_m \cdot x_2 & \cdots & p_m \cdot x_n \end{pmatrix}.$$

Thus, if the columns of Y are written y_i, we have

$$y_i = \begin{pmatrix} p_1 \cdot x_i \\ p_2 \cdot x_i \\ \vdots \\ p_m \cdot x_i \end{pmatrix}. \tag{3.1}$$

Back to our example from Figure 3.1. We want data with a high *signal-to-noise* (SNR) ratio as to minimize noise. Assuming our data in Figure 3.1 was collected reasonably well, the direction of largest variance is the direction of most interesting dynamics. Therefore, the variance of signal, σ_s^2, would correspond to the length of the orange vector in fig. 3.2 pointing to the top right, and the variance of the noise, σ_n^2, would correspond to the length of the orange vector pointing to the top left.

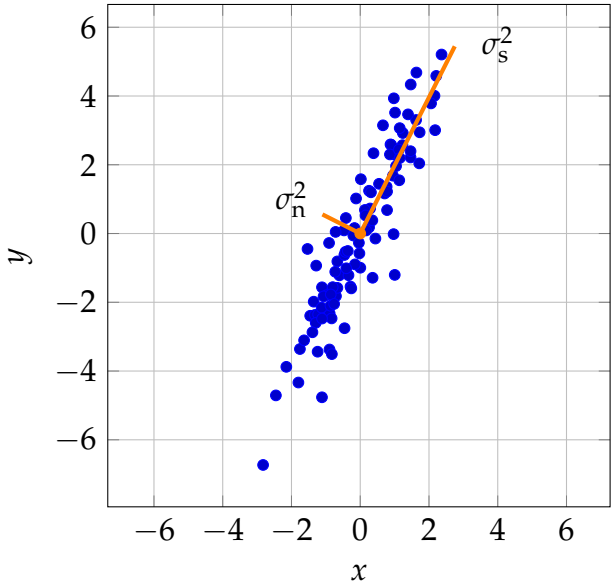


Figure 3.2: Signal and noise variances represented graphically.

Note also that in Figure 3.2 knowing the x value gives one a good approximation for the y value and vice versa. In this case, we might say that the data has a moderate amount of redundancy, whereas if the data points had a much higher r^2 value to its line of best fit, we would say the data have high redundancy. (And if the data had a much lower r^2 value, we would say the data have low redundancy.) One of the aims of PCA is to lower redundancy. For the 2-dimensional

case, this is simple—take the line of best fit, but for arbitrarily higher dimensions, this is not obvious.

Suppose we have n measurements

$$U = \{u_1, u_2, \ldots, u_n\} \qquad \text{and} \qquad V = \{v_1, v_2, \ldots, v_n\}$$

with mean equal to 0. The *variances* are equal to

$$\sigma_U^2 = \frac{1}{n}\sum_{i=1}^{n} u_i^2, \qquad\qquad \sigma_V^2 = \frac{1}{n}\sum_{i=1}^{n} v_i^2.$$

The *covariance* between the data sets U and V is

$$\sigma_{UV}^2 = \frac{1}{n}\sum_{i=1}^{n} u_i v_i.$$

The covariance measures the degree of the linear relationship between the two variables. Thus, a large positive value would imply that the data are positively correlated, and a large negative value would imply negatively correlated. And $\sigma_{UV}^2 = 0$ if and only if the data U and V are uncorrelated. Moreover the absolute magnitude of the covariance measures the degree of redundancy.

If instead we wrote $u = (u_1, u_2, \ldots, u_n)$ and $v = (v_1, v_2, \ldots, v_n)$ as row vectors, then

$$\sigma_{uv}^2 = \frac{1}{n} u v^{\mathrm{t}}. \tag{3.2}$$

We generalize from two vectors to m vectors. Write

$$X = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_m' \end{pmatrix}$$

where x_i' is a row vector. Note that the rows of X correspond to measurements of a particular type, and the columns of X correspond to all measurements of a particular trail. Using Equation (3.2), we define the **covariance matrix** of X to be

$$C_X = \frac{1}{n} X X^{\mathrm{t}}.$$

Lemma 3.1. *1. The matrix C_X is symmetric. That is, $C_X = C_X^{\mathrm{t}}$.*

2. The diagonal entries of C_X are variances.

3. The off-diagonal entries of C_X are covariances.

Recall our goal is to find a new (and better) basis $\{p_1, p_2, \ldots, p_m\}$. Namely, we want an invertible matrix P to turn our data X into a data set Y where we can better understand the structure. If we could do this on the level of covariance

matrices, then we could pick an ideal covariance matrix, one where the diagonal entries are large in absolute magnitude and where the off-diagonal entries are small in absolute magnitude. (Variances being high in magnitude suggest interesting dynamics, and covariances low in magnitude suggest low redundancy.) In other words, we would like to find a matrix P such that

$$C_Y = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_m \end{pmatrix}, \tag{3.3}$$

where $Y = PX$. Typically we have $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_m$.

The main work of PCA is finding such a matrix P. We will describe how to construct such a P later, but let us return to our running example.

The first principal component is in the direction of the largest variance, and the second principal component is in the orthogonal direction. So in $\mathbb{R}^2$, this is quite simple. Although we have not given all the data point explicitly, the covariance matrix is

$$C_X = \begin{pmatrix} 1.27 & 2.52 \\ 2.52 & 5.95 \end{pmatrix}.$$

The slope of the line in the direction of the highest variance is 1.98, and it passes through the point $(0, 0)$. By rotating and permuting, we get

$$P = \begin{pmatrix} 0.40 & 0.92 \\ 0.92 & -0.40 \end{pmatrix},$$

and the graph of the data $Y = PX$ is seen in Figure 3.3.

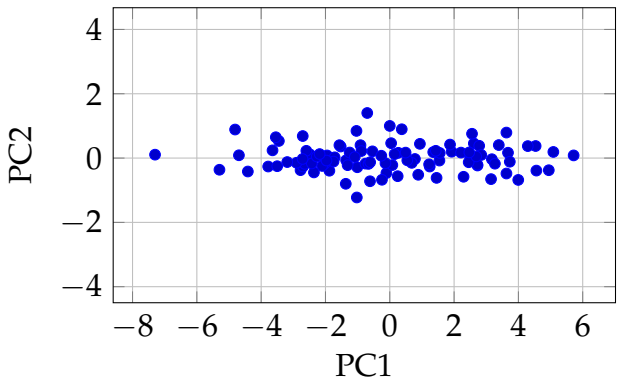


Figure 3.3: A new basis for our data.

Remark 3.2. There is a whole art of scaling data in a pre-processing stage that we will not explore in this course. Basically, if one variable ranges between ± 10 and another $\pm 10^3$, the second variable will bias the process simply by its scale. There are many different methods to rescale the data as not to lose (too much) information. *Throughout we will assume our data has roughly the same scale and not worry about rescaling*, but in practice this is an important issue.

3.1.1 Assumptions of PCA

Before we close this introduction to PCA, let us come back to some of the assumptions we have made along the way. There are three key assumptions we have made in introducing PCA. We will not focus much on these, but users of PCA should know that assumptions have been made. These statements may not necessarily hold for a particular data set.

1. *Linearity.*

   This allows us to reframe the problem as a change of basis problem.

2. *Large variance is important (and $SNR > 1$).*

   This can be a very strong assumption and really needs to take into account how the data was collected.

3. *Principal components are orthogonal.*

   This is not always the case, but orthogonality allows us to use linear algebra.

3.2 In class exercises pt. IV

1. (a) Show that C_X is symmetric.
   (b) Prove that the diagonal entries of C_X are variances and the off-diagonal entries are covariances.

2. Suppose X is an $m \times n$ matrix with sample mean $\overline{x} \in \mathbb{R}^m$. Let X' be the shifted data of X, so that its sample mean is 0. What is $C_{X'}$ in terms of X and $\overline{x}$?

3.3 Performing PCA

At the heart of performing PCA in practice is the following question. Recall that C_Y is a diagonal matrix; see the discussion around Equation (3.3).

Question 3.3. What is the relationship between C_X and C_Y if $Y = PX$?

Proof. From above, we have

$$C_Y = \frac{1}{n} YY^{\mathrm{t}} = \frac{1}{n}(PX)(PX)^{\mathrm{t}} = PC_X P^{\mathrm{t}}. \qquad \square$$

In order to perform a PCA, we need to find a matrix P such that $PC_X P^{\mathrm{t}}$ is diagonal. Importantly, we do not want to change the variances, so we want P to be *distance preserving*; that is, we want P to satisfy

$$\|Pv\| = \|v\| \tag{3.4}$$

for all vectors $v \in \mathbb{R}^m$. We say such a matrix P is an **isometry**. We can take Equation (3.4) and massage it, so that P must satisfy

$$v^{\mathrm{t}} P^{\mathrm{t}} Pv = (Pv) \cdot (Pv) = v \cdot v = v^{\mathrm{t}} v$$

for all $v \in \mathbb{R}^m$. Since this needs to hold for all vectors, it must hold for all pairs of basis vectors (e_i, e_j) for all $i, j \in \{1, \ldots, m\}$. Thus, distance preserving is equivalent to $P^t P = I_m$, but these matrices are called **orthogonal**. (Note: pairs of distinct columns of such a matrix P are pairwise orthogonal. Can you prove this?!)

Let us bring this back to the equation we established. We want P to be an orthogonal matrix such that

$$C_Y = P C_X P^t. \tag{3.5}$$

Moreover, we want C_Y to be a diagonal matrix. Since $P^t P = I_m$, it follows that $P^{-1} = P^t$, so using this identity we have

$$C_Y = P C_X P^{-1}.$$

Since C_Y is diagonal, this is accomplished through *eigendecomposition*. Therefore, the rows of P are eigenvectors, and the diagonal entries of C_Y are eigenvalues.

All the entries of C_Y are *real*, and we know that some matrices have complex eigenvalues. For example, the matrix

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

has eigenvalues i and $-i$, for $i = \sqrt{-1}$. **Is it always possible to find a real matrix P such that Equation (3.5) holds?** The answer is "Yes," and we will prove it later. For now, let us assume it is always possible.

We give a recipe for cooking up the principal components.

Given: n data points $x_i \in \mathbb{R}^m$ (of roughly the same scale),

Return: m principal components.

1. Compute the mean of each coordinate: $\mu_j = \sum_i x_{ij}$,
2. Organize the normalised data into a matrix $X = (x_{ij} - \mu_j)$,
3. Compute the covariance matrix C_X of X,
4. Compute the eigenvectors of C_X, and sort them based on their eigenvalues: largest is first and smallest is last.
5. Return the (ordered) orthonormal basis of eigenvectors.

PCA is a phrase used for this algorithm together with its analysis, and one of the main ways PCA is performed is by taking only the first k principal components (rather than all m). Here, k is usually determined by the *eigenvalues*.

Both C_X and C_Y have the same eigenvalues: $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_m$. Since the trace of matrix is the sum of its eigenvalues, both C_X and C_Y have the same trace. In other words,

$$\operatorname{tr}(C_X) = \sum_{i=1}^{m} \lambda_i,$$

and by Lemma 3.1, the sum of the variances is the sum of the eigenvalues. Because we view variability as an important measurement to keep track of, we can choose a k that both maximizes the amount of variability "seen" and while minimizing the value of k. This is a bit more of an art than a science, but one general rule could be to choose the smallest k such that

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{m} \lambda_i} \geqslant 0.95.$$

For such a k, one might say that the first k principal components capture 95% of the total variability.

So then do the rest of the principal components not matter?

3.4 Projections

The answer to the previous question is essentially "No", and it can be helpful to consider an idealised example. Suppose $k < m$ and

$$\lambda_{k+1} = \lambda_{k+2} = \cdots = \lambda_m = 0.$$

References

[1] Dan Margalit and Joseph Rabinoff, *Interactive Linear Algebra*, 2019, https://textbooks.math.gatech.edu/ila/.

[2] Jonathon Shlens, *A tutorial on principal component analysis*, 2014, arXiv:1404.1100.