Geometric Foundations of Data Analysis I: Week 4

Joshua Maglione

29 September 2025

Contents

2.9 Coefficient of determination (R^2 values)

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 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 Theorem 2.4, we have $0 \leqslant SSR \leqslant SST$. □

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 $y = f(x_1, \dots, x_{p-1})$. 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.

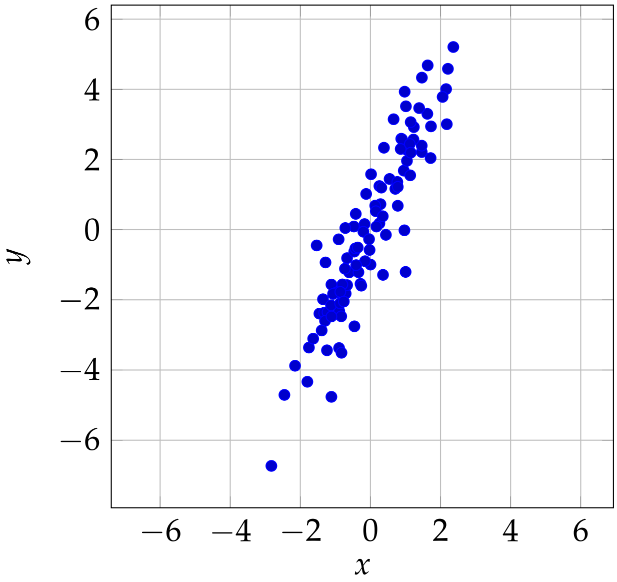

Figure 3.1: Some data in $\mathbb{R}^2$.

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 $\{(1, 0, \dots, 0), (0, 1, 0, \dots, 0), \dots, (0, \dots, 0, 1)\}$, and we want a basis $\{p_1, p_2, \dots, p_m\}$ 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.

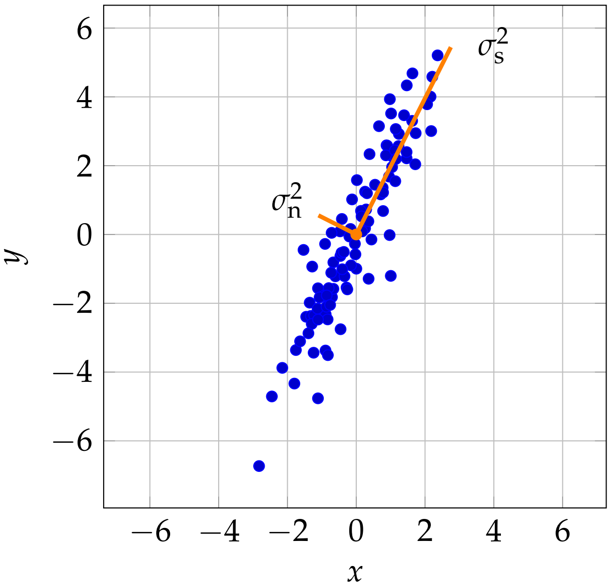

Figure 3.2: Signal and noise variances represented graphically.

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 of the same kind (like length)

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

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

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

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 trial. In other words, our data points are the columns of X, and each entry of a data point is a measurement of a particular type.

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.

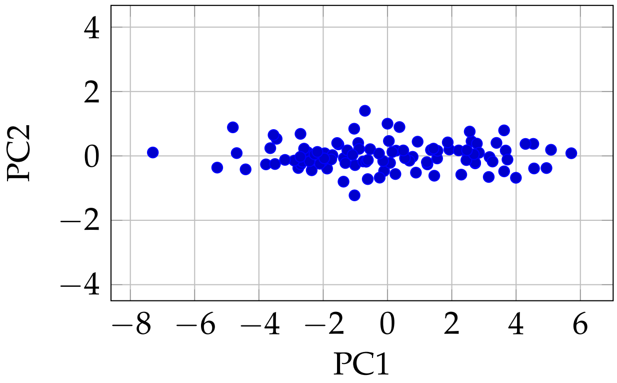

Figure 3.3: A new basis for our data.

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.