Geometric Foundations of Data Analysis I: Week 5

Joshua Maglione

6 October 2025

Contents

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

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

In order to perform a PCA, we need to find a matrix P such that PC_XP^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

$$\|Pv\| = \|v\| \tag{3.1}$$

for all vectors $v \in \mathbb{R}^m$. We say such a matrix P is an **isometry**. We can take Equation (3.1) and massage it, so that P must satisfy

$$v^tP^tPv = (Pv) \cdot (Pv) = v \cdot v = v^tv$$

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 $P^tP = I_m$, 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

$$C_Y = PC_XP^t. \tag{3.2}$$

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

$$C_Y = PC_XP^{-1}.$$

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 $i = \sqrt{-1}$.

Question 3.4. Is it always possible to find a real matrix P such that Equation (3.2) holds?

The answer to Question 3.4 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

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{m} \lambda_i} \geqslant 0.95. \tag{3.3}$$

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