On the mathematical foundations and geometric interpretation of principal component analysis

D. Laszuk
Maria Curie-Sklodowska University, Faculty of Mathematics, Physics and Computer Science, Lublin, Poland
dominikalaszuk14@gmail.com

Principal Component Analysis (PCA) is one of the most widely used methods for understanding high-dimensional data. Given a dataset of n observations of p possibly correlated variables $X_1, \ldots, X_p$, the main goal of PCA is to replace them with a smaller number of new variables

$$Z_k = a_{k1}X_1 + a_{k2}X_2 + \cdots + a_{kp}X_p,$$

where the real coefficients a_{ki} satisfy $a_{k1}^2 + \cdots + a_{kp}^2 = 1$ and are chosen so that the Z_k are uncorrelated and successively capture as much variance as possible.

In this survey talk, we present PCA from three complementary perspectives: algebraic, geometric, and computational. We begin with its mathematical foundations, showing how principal components arise naturally from the eigenvalues and eigenvectors of the covariance matrix, and how the total variance of the data can be expressed in terms of these quantities.

Next, we focus on geometric intuition. We interpret PCA as a rotation of the coordinate system, where new axes are aligned with directions of maximal variance. This viewpoint allows us to visualize how the data is reorganized and why a small number of components can often provide a good approximation of the original dataset.

Finally, we demonstrate a simple implementation of PCA in Python on the MNIST dataset. This example illustrates how abstract mathematical ideas translate into practical algorithms used in data analysis and machine learning.

The aim of this work is to provide an accessible introduction to PCA that bridges abstract linear algebra, geometric intuition, and practical data analysis. The presentation is based primarily on [1] and [2].

References

[1] Mukhopadyay, P., Multivariate Statistical Analysis, World Scientific, Singapore, 2009, 549 pp.

[2] Deisenroth M. P., Faisal A. A., Ong C. S., Mathematics for Machine Learning, Cambridge University Press, Cambridge, 2020, 407 pp.