22 Principal Component Analysis: Explained Through Ellipse

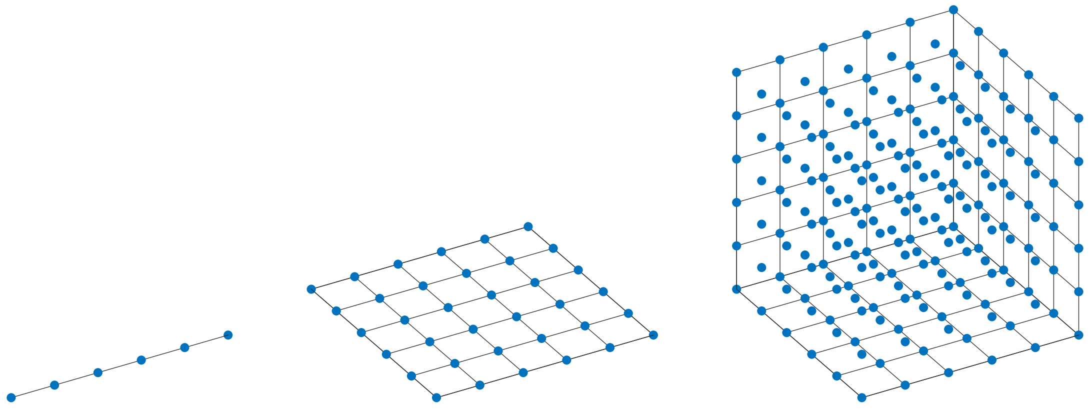
22.1 Seeing High-Dimensional Worlds More Clearly

22.1.1 The Curse of Dimensionality

Dimensionality reduction is a key idea in machine learning and data analysis. It refers to the process of mapping data from a high-dimensional space to a lower-dimensional one, while preserving as much useful information as possible. Many real-world datasets contain a large number of features, which can make them difficult to process, expensive to compute, and almost impossible to visualize.

As the number of features increases, we also face the curse of dimensionality: the volume of the feature space grows exponentially, causing data points to become extremely sparse. This sparsity makes pattern recognition and learning much harder. To build an intuition, imagine sampling just 6 points along each feature dimension.

In one dimension, you have 6 points (Figure 1 (a)). In two dimensions, those points form a grid of $6^2 = 36$ points (Figure 1 (b)). In three dimensions, you already have $6^3 = 216$ points (Figure 1 (c)).



(a) (b) (c)

Figure 1. Sampling in 1D, 2D, and 3D

22.1.2 Why We Reduce Dimensions

By the time you reach four dimensions (Figure 2), that number grows to $6^4 = 1296$, and with ten features it explodes to more than sixty million points. This rapid growth is the essence of the curse of dimensionality and explains why high-dimensional datasets can be so challenging to work with.

The goal of dimensionality reduction is to tame this complexity. By projecting high-dimensional data onto a lower-dimensional subspace, we can reduce computation, improve generalization, and make patterns easier to understand and visualize. At the same time, we hope to keep the key structure of the data intact.

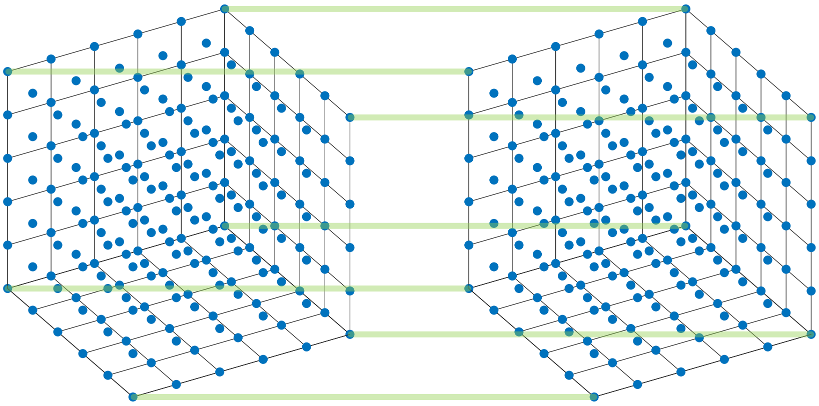


Figure 2. Sampling in 4D

22.1.3 PCA: The Linear Lens

One of the most widely used dimensionality reduction techniques is Principal Component Analysis (PCA). PCA finds a new set of orthogonal axes and projects the data onto them in a way that preserves the maximum possible variance. In essence, PCA removes redundancy and keeps only the most informative directions.

Dimensionality reduction is not limited to linear techniques. For datasets that lie on curved, low-dimensional structures embedded in high-dimensional space, manifold learning provides nonlinear approaches that attempt to uncover these hidden shapes. Although manifold learning methods will not be discussed in detail in this book, it is helpful to know that PCA is just one family within a broader set of dimensionality-reduction tools.

22.2 Understanding PCA Through Geometry

22.2.1 From Cups to Coordinates: The Intuition of Projection

At its core, Principal Component Analysis (PCA) is about finding a meaningful way to look at data. In mathematical terms, PCA projects data onto a new coordinate system—called the principal component space—where the axes represent the directions of maximum variation in the data. But to truly understand what this means, it helps to begin with something more tangible.

Imagine holding a cup in your hand. As shown in Figure 3, the cup is a three-dimensional object, yet when you take a photo of it, the image on your screen is two-dimensional. Depending on the angle from which you photograph it, the picture may show more or less detail about the cup's shape. To capture its structure accurately, you might take photos from multiple angles and then choose the view that reveals the most information (Figure 4).

This act of capturing a 3D object in 2D form is, in fact, a simple example of dimensionality reduction: representing 3D data in a 2D space by projection.

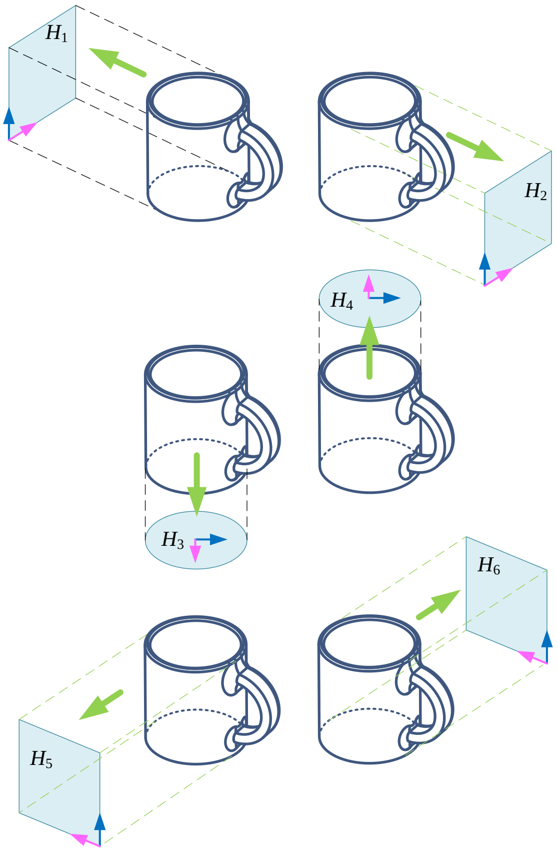


Figure 3. The cup and its six projection directions

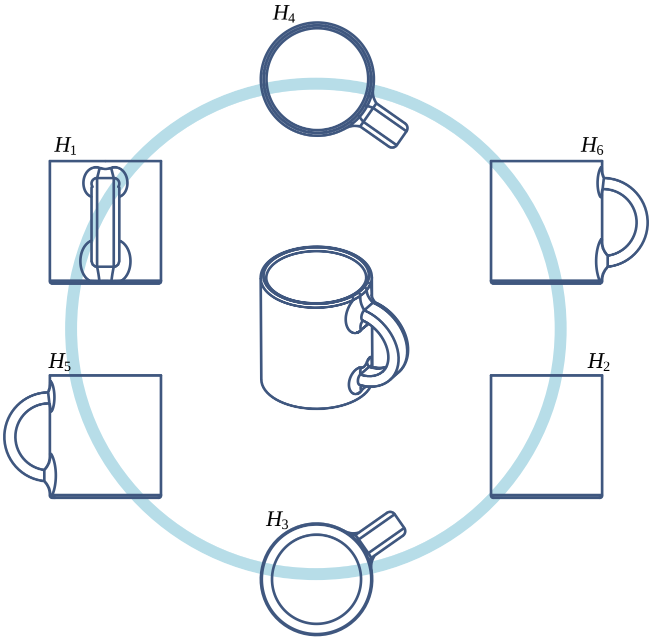


Figure 4. The cup projected onto six different planes

22.2.2 Preparing the Data: Centering and Scaling

This same idea applies to data. PCA searches for the best "view" or projection of high-dimensional data—one that captures as much variation as possible when projected onto a lower-dimensional space. To do that, we first need to prepare the data carefully.

Before performing PCA, we typically center the data by subtracting the mean of each feature. Geometrically, this means shifting the data so that its centroid lies at the origin. In Figure 5, this step is shown as moving the center of an ellipse (which represents the data distribution) to the origin. The shape of the ellipse is determined by the covariance matrix, which captures how different features vary together. You may recall from earlier chapters that the covariance matrix defines the orientation and spread of ellipses in a Gaussian distribution.

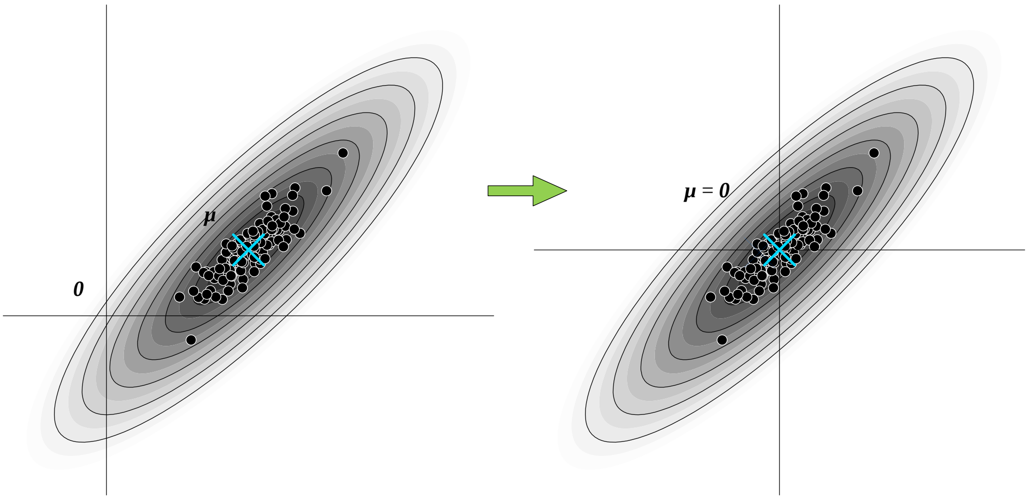


Figure 5. Centering the data by moving the centroid to the origin

Another essential preprocessing step is standardization. Real-world datasets often contain features with different units or scales—for example, height in centimeters and income in dollars. Without standardization, features with large numerical ranges could dominate the analysis. Standardization removes these scale effects by transforming each feature into a standardized score, known as a z-score, giving all features a mean of zero and a standard deviation of one. Mathematically, this involves two operations: translation (centering) and scaling (adjusting spread). After this step, PCA can fairly compare the importance of different features without bias toward those with larger variances.

22.2.3 What Projection Really Means

Now, to understand PCA conceptually, we must grasp what a projection means. Figure 6 illustrates a simple 2D case: projecting data onto the x_1-axis or x_2-axis. Each projection flattens the data onto one dimension, revealing the distribution of each feature separately. Since the data has been centered, the mean along each axis is zero.

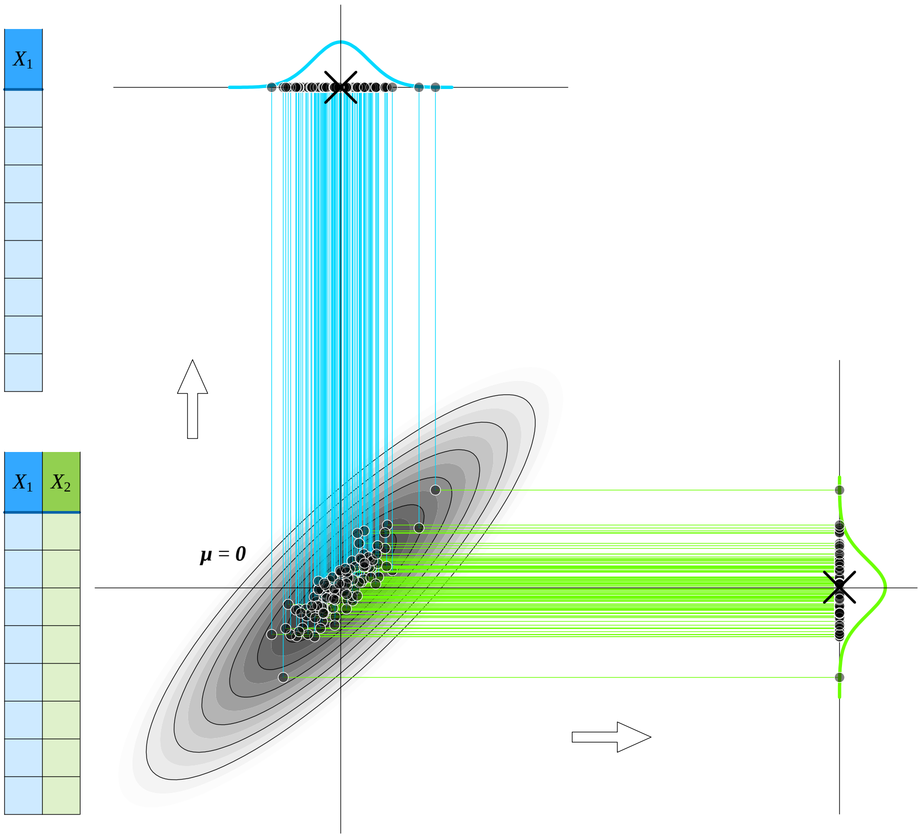


Figure 6. Projecting 2D data onto x_1- and x_2-axes and showing 1D distributions

22.2.4 Finding the Best View: Directions of Maximum Variance

PCA seeks the projection that best represents the variation in the data. The first principal component is the direction in which the projected data has the largest variance. In Figure 7, the data points are projected along sixteen different directions on the plane. By comparing the width (or spread) of the projected data in each direction—quantified by the standard deviation—we can identify the direction with the greatest spread. This direction, labeled as *C* or *K*, corresponds to the first principal component. The directions with the smallest spread (*G* and *O*) correspond to the least significant variation.

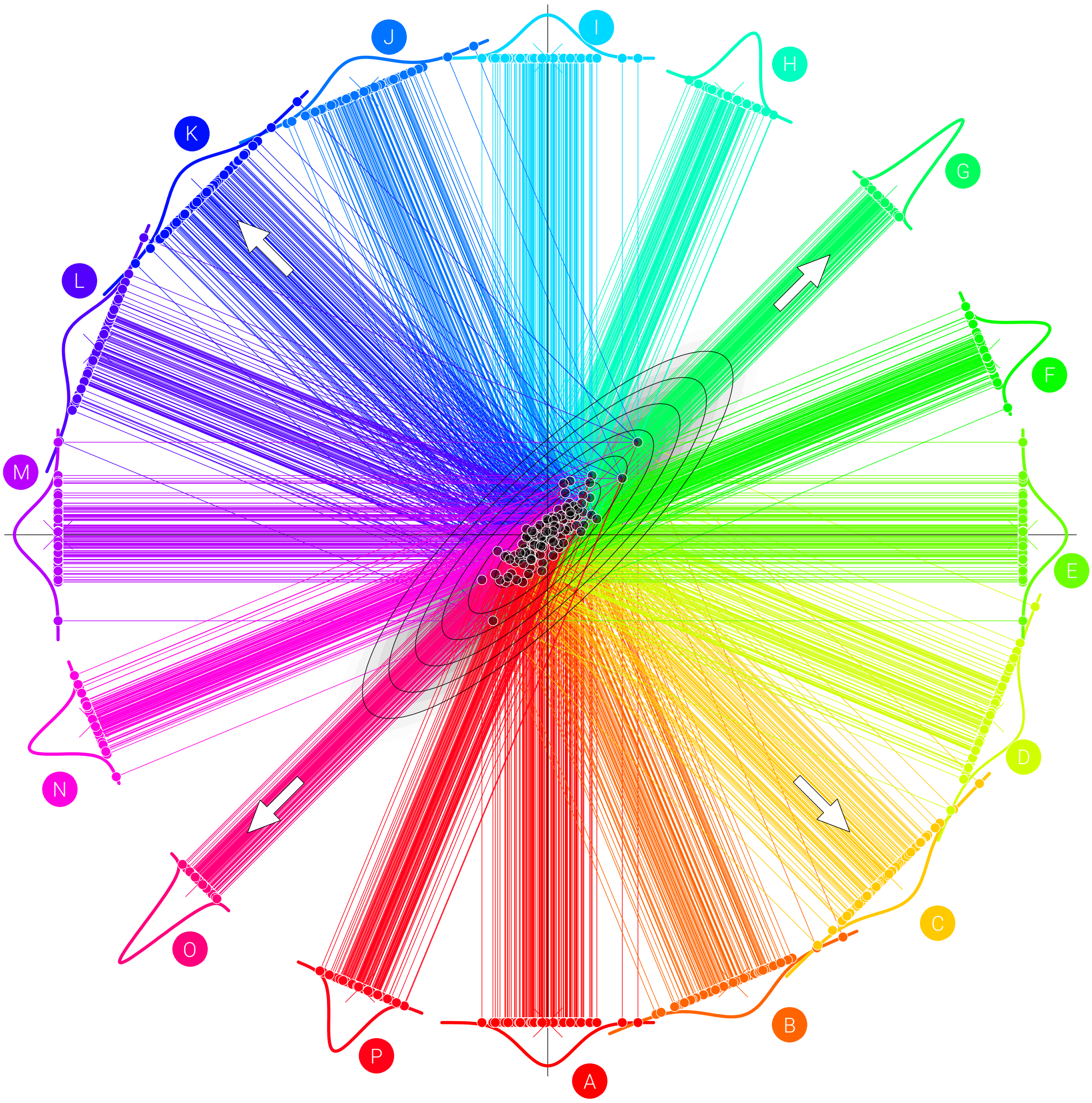


Figure 7. Projecting 2D data along sixteen different directions

22.2.5 Rotating the World: Aligning with the Data's True Axes

From another perspective, shown in Figure 8, PCA can be understood as rotating the coordinate system so that it aligns with the natural axes of the data's spread. In this rotated frame, the first principal component ($\boldsymbol{v}_1$, or PC1) aligns with the long axis of the ellipse, while the second principal component ($\boldsymbol{v}_2$, or PC2) aligns with the short axis. These two directions are always orthogonal.

We can think of this new coordinate system as $[\boldsymbol{v}_1, \boldsymbol{v}_2]$, which replaces the original system $[\boldsymbol{e}_1, \boldsymbol{e}_2]$. Rotating the coordinate system effectively "unlocks" the true shape of the data, transforming the tilted ellipse into one aligned with the axes, as shown in Figure 9.

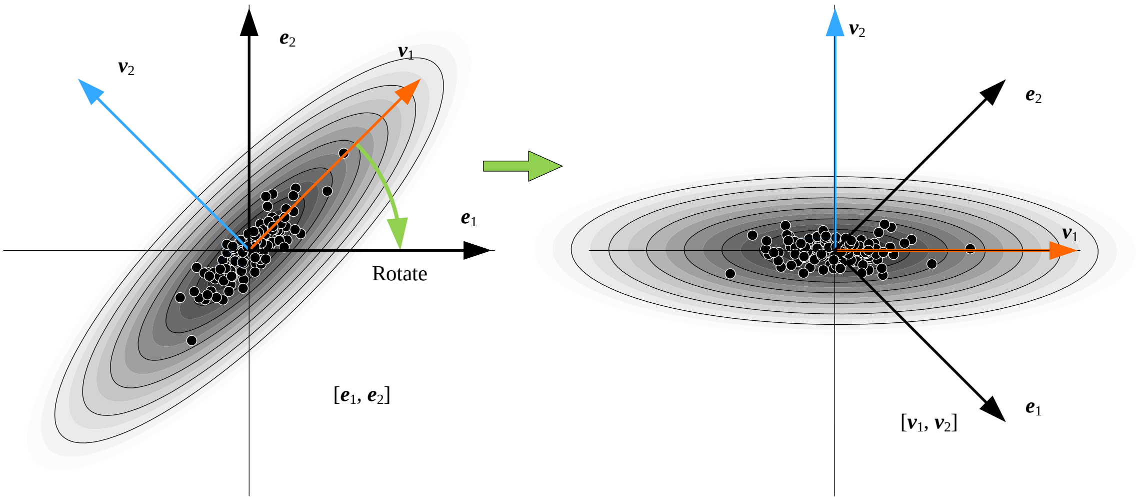


Figure 8. Rotating the coordinate system to align with principal components

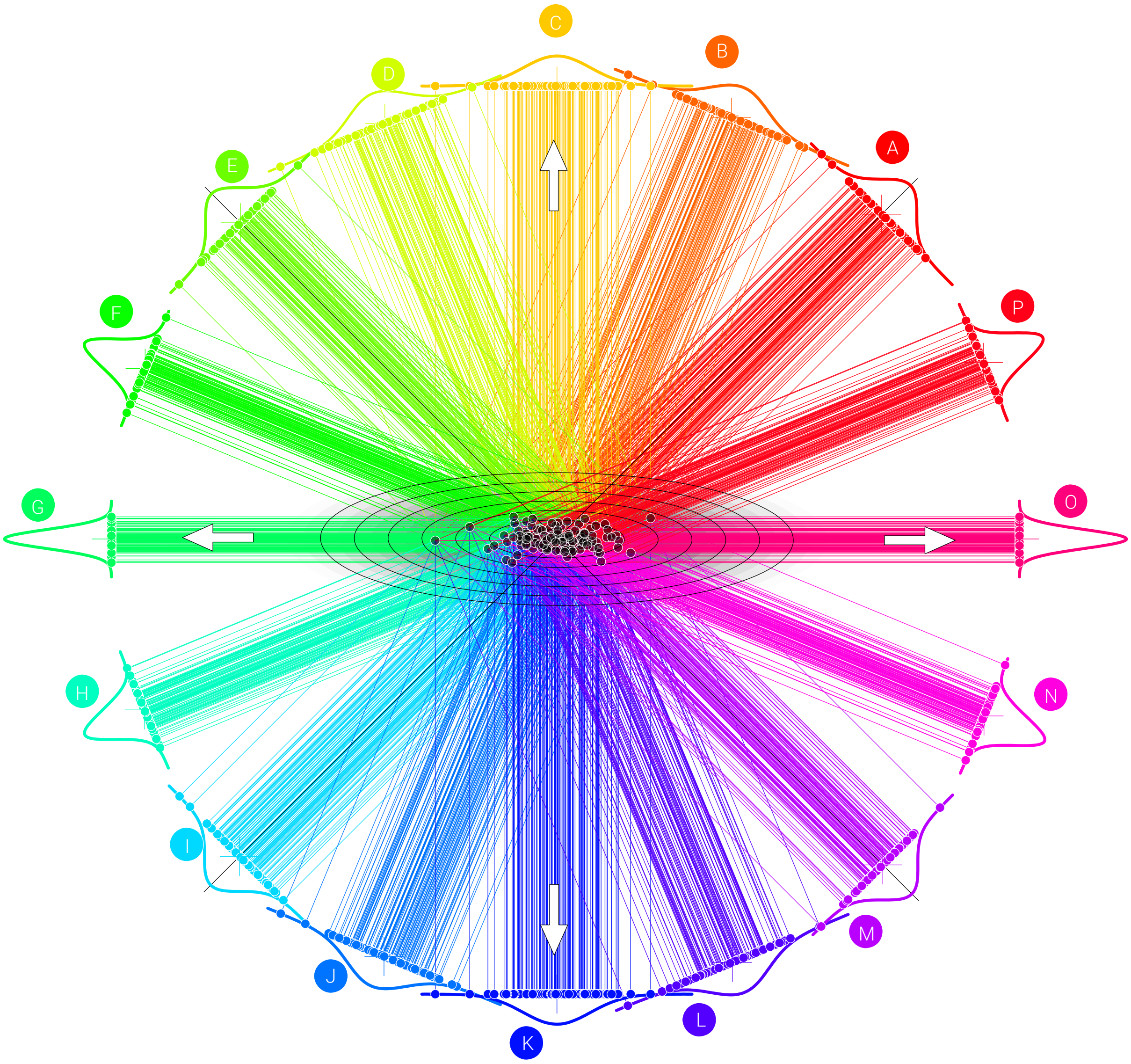


Figure 9. Viewing the data in the $[v_1, v_2]$ coordinate system

To compute these principal directions mathematically, we start with the covariance matrix $\boldsymbol{\Sigma}$ of the centered data. By performing eigenvalue decomposition on $\boldsymbol{\Sigma}$, we obtain its eigenvectors and eigenvalues. The eigenvectors define the principal directions ($\boldsymbol{v}_1$, $\boldsymbol{v}_2$, …), and the corresponding eigenvalues measure how much variance the data has along each direction. The first principal component corresponds to the largest eigenvalue. This workflow will be detailed in the next chapter.

In essence, PCA is a linear dimensionality reduction method: it represents each principal component as a linear combination of the original features. While this works well for data with linear relationships, it may fail to capture more complex, nonlinear patterns. To handle nonlinear structures, we can extend PCA using a technique known as Kernel Principal Component Analysis (Kernel PCA). Kernel PCA uses the kernel trick to map data into a higher-dimensional feature space, where linear relationships can represent nonlinear ones in the original space. By applying PCA in this transformed space, we can uncover patterns that traditional PCA would miss. Commonly used kernels include the radial basis function (RBF or Gaussian kernel), the polynomial kernel, and the sigmoid kernel. Later chapters will illustrate Kernel PCA visually and explain how it generalizes PCA to nonlinear data.

22.3 Visualizing PCA in Two Dimensions

22.3.1 The Shape of the Data: Covariance Ellipses

Figure 10 shows a scatter plot of the standardized dataset. The set of ellipses overlaid on the scatter points represents the covariance structure of the data. More precisely, these ellipses are contour lines of equal Mahalanobis distance. Unlike Euclidean distance, which simply measures straight-line distance, the Mahalanobis distance incorporates correlations between features. This means it can more accurately reflect the true shape and spread of the data distribution. Each ellipse in Figure 10 marks points that are equally distant from the center under this distance measure, and together they form a family of concentric contours.

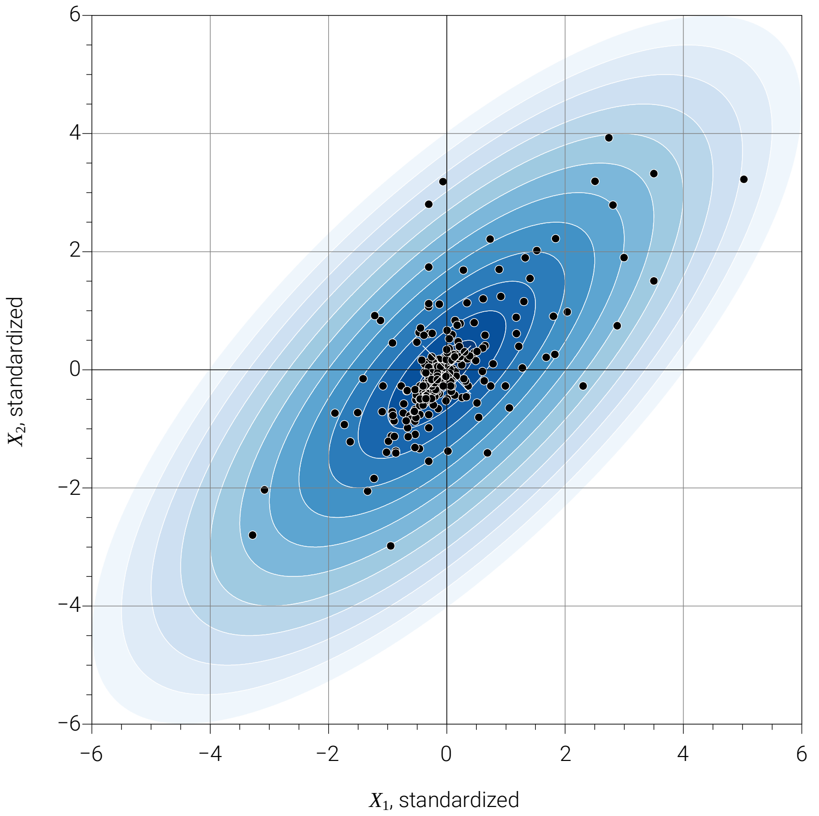


Figure 10. Standardized data with Mahalanobis distance contours. Figure generated by Ch22_01_PCA_explained_through_Ellipse.ipynb.

22.3.2 The Axes of Meaning: Principal Components

In Figure 11, we visualize the principal component directions. The first principal component, $\boldsymbol{v}_1$, aligns with the long axis of the ellipse, capturing the direction in which the data varies most. The second principal component, $\boldsymbol{v}_2$, lies along the short axis and captures the smallest variation. This matches the geometric intuition developed earlier: PCA rotates the coordinate system so that it aligns with the natural orientation of the data.

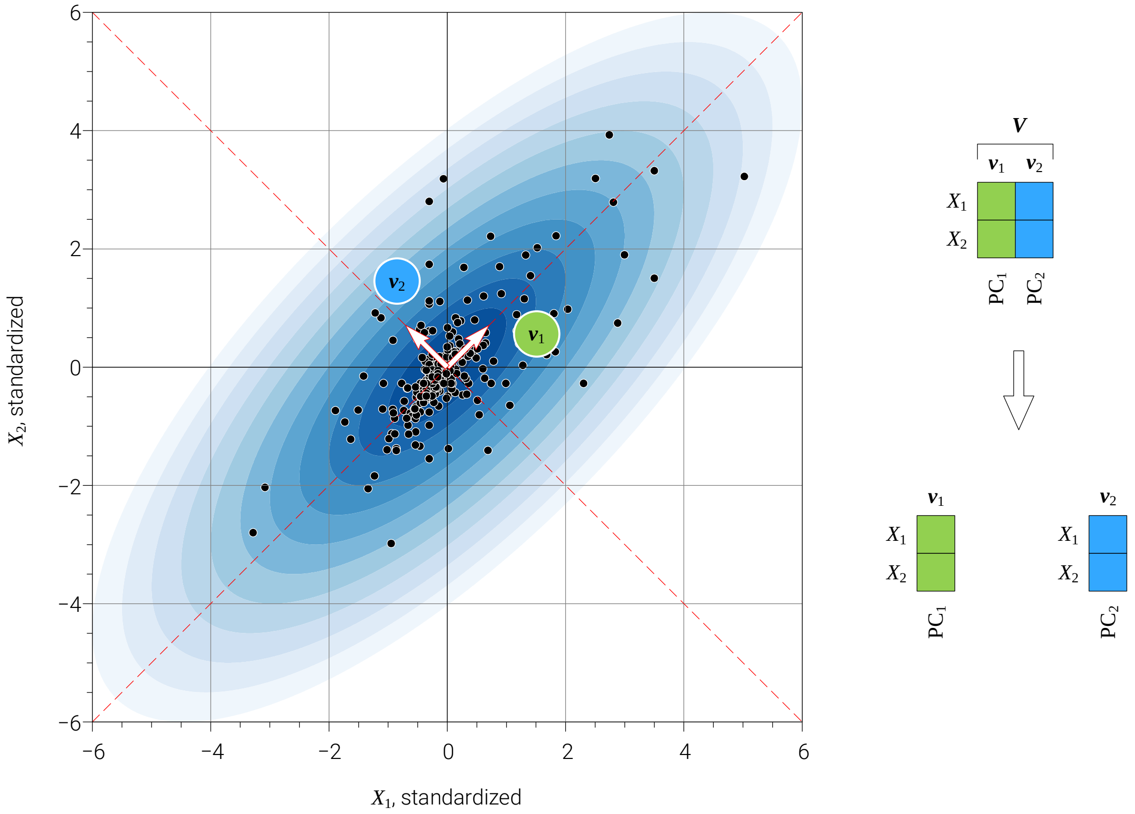


Figure 11. Principal component directions $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. Figure generated by Ch22_01_PCA_explained_through_Ellipse.ipynb.

22.3.3 Projecting onto the First Principal Component

Figure 12 shows the projection of the data onto the first principal component. Since $\boldsymbol{v}_1$ is the direction of maximum variance, this projection spreads the data out as much as possible along a single axis.

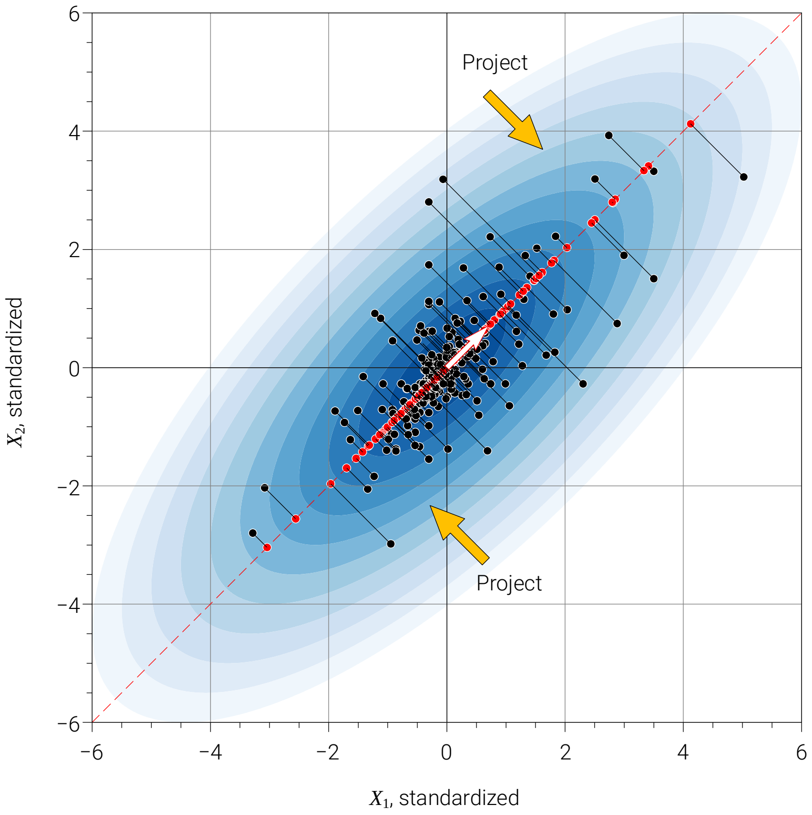


Figure 12. Projection onto the first principal component $\boldsymbol{v}_1$. Figure generated by Ch22_01_PCA_explained_through_Ellipse.ipynb.

22.3.4 Projecting onto the Second Principal Component

In contrast, Figure 13 shows the projection onto $\boldsymbol{v}_2$, where the variance is minimal and the data appears much more compressed. Because $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are perpendicular, they form a new orthogonal coordinate system.

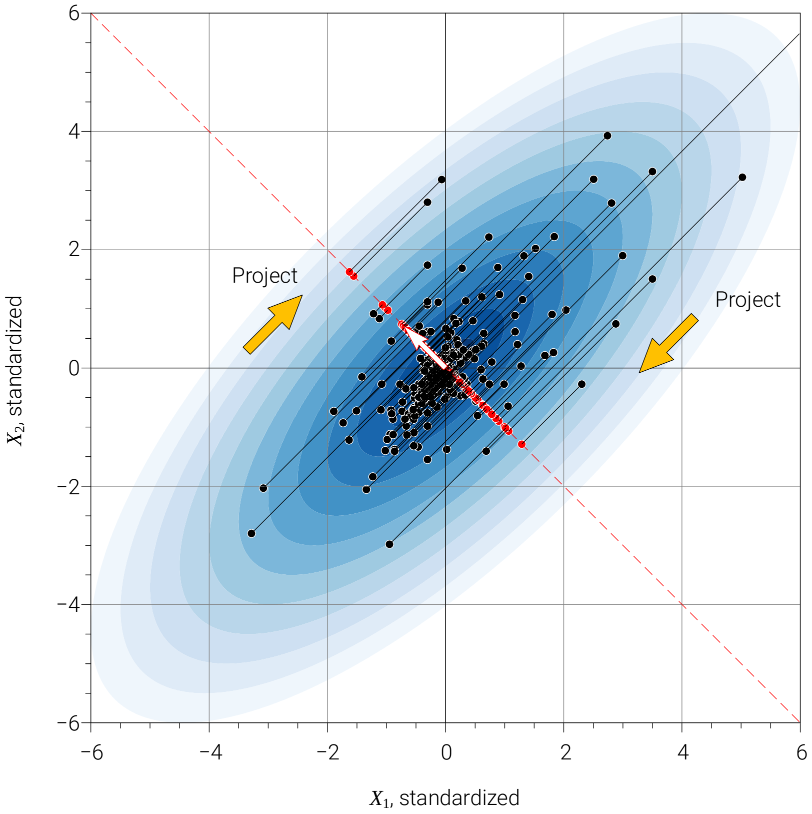


Figure 13. Projection onto the second principal component $\boldsymbol{v}_2$. Figure generated by Ch22_01_PCA_explained_through_Ellipse.ipynb.

22.3.5 Seeing Data Anew: The Rotated Coordinate System

When we view the data in this transformed system [$\boldsymbol{v}_1$, $\boldsymbol{v}_2$], as shown in Figure 14, the scatter plot becomes axis-aligned and easier to interpret.

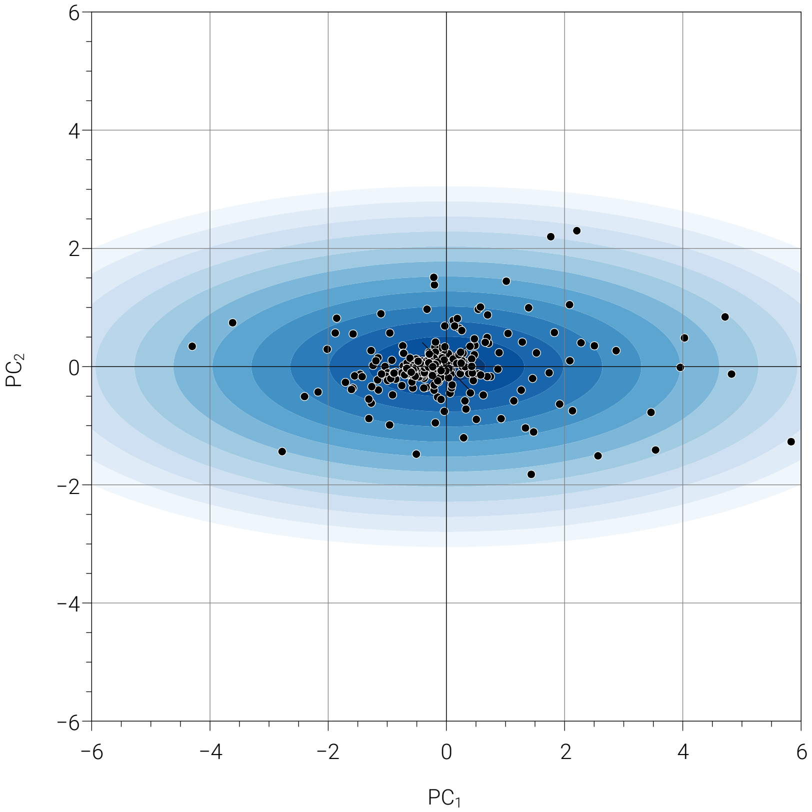


Figure 14. Scatter plot in the [$\boldsymbol{v}_1$, $\boldsymbol{v}_2$] coordinate system. Figure generated by Ch22_01_PCA_explained_through_Ellipse.ipynb.

22.4 Conclusion

This chapter introduces Principal Component Analysis as a geometric approach to understanding high dimensional data. It begins by explaining the curse of dimensionality and why reducing dimensions helps reveal structure hidden in complex datasets. Using the image of a three dimensional cup projected into two dimensions, it builds intuition for how PCA finds the most informative view of data.

The chapter explains centering and scaling, then shows how PCA identifies directions of maximum variance by rotating the coordinate system to align with the natural spread of data. Through ellipses, eigenvalues, and eigenvectors, readers see how PCA connects geometry and statistics.

Finally, two dimensional examples illustrate how projecting data onto principal components simplifies patterns and aids interpretation, while pointing to kernel PCA as an extension for nonlinear relationships.