Geometry of quadratic forms, Mahalanobis distance, and PCA

Shaikh Ammar • 8 Dec 2025

Most treatments of PCA can be categorised into two: one that purely defines PCA in rigorous footing—explaining how the principal components are eigenvectors of the variance matrix and so on; and the other which purely talks of PCA geometrically mentioning nothing about how they actually correspond to the eigenvectors. In this post, I reconcile these two approaches: explaining PCA geometrically while still maintaining the rigour of it all.

Spectral decomposition

Among the class of diagonal, orthonormal and symmetric matrices, perhaps the easiest class of matrix to visualise is that of diagonal matrices. They just simply represent scaling the i-th axis by the i-th diagonal element. Orthonormal matrices are not hard to visualise either. They just represent the rotation which takes the i-th canonical basis vector (i.e., the vector with 1 in the i-th position and 0 elsewhere) to the i-th column vector of the said matrix (possibly with some reflection involved if the determinant is negative). Symmetric matrices don't seem to have such nice visualisation (after all symmetry seems wholly to be a property of the matrix itself and not the linear transformation it represents), but owing to the beauty of the *Spectral theorem*, one realises that symmetric (or more precisely, self-adjoint matrices) are just scaling transformations but in different axes. We briefly recall the Spectral theorem:

Theorem (Spectral theorem)**.** *Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be symmetric. Then, there exists an orthonormal matrix $\mathbf{P}$, and a diagonal matrix $\mathbf{\Lambda}$ such that*

$$\mathbf{A} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{\mathsf{T}}.$$

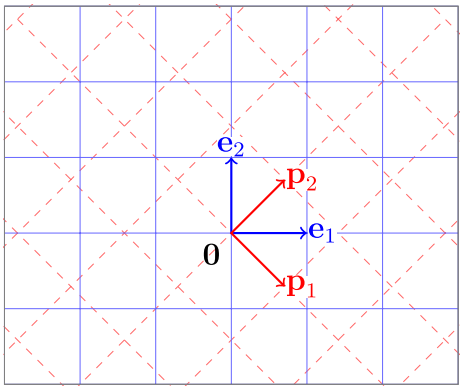
The danger of this theorem is understanding the algebra behind it, but not taking the moment to reflect what it means geometrically. First, notice that columns of $\mathbf{P}$ corresponds to eigenvectors of $\mathbf{A}$, and the scalars in $\mathbf{\Lambda}$, the corresponding eigenvalues (this can be seen by expressing an arbitrary eigenvector as $\mathbf{Px}$).

Now, note that $\mathbf{P}$ is the rotation that takes the canonical axes to the axes corresponding to the eigenvectors. Hence, $\mathbf{P}^\mathsf{T}$—the inverse of $\mathbf{P}$—is the rotation that takes the axes corresponding to eigenvectors to the canonical axes. So, the Spectral theorem basically is saying we can decompose the transformation described by $\mathbf{A}$ into: first, aligning its eigenvectors with the canonical axes so that they become the *new* axes, then, scaling it with the corresponding eigenvalues along these (new) axes, and finally rotating the eigenvectors back to how they were aligned.

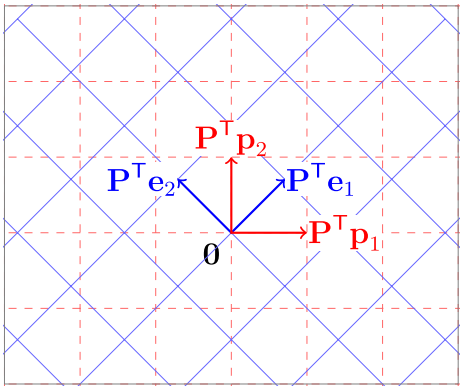
This sequence of transformation is illustrated for the matrix:

$$\begin{bmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ -\sqrt{2}/2 & \sqrt{2}/2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ -\sqrt{2}/2 & \sqrt{2}/2 \end{bmatrix}^\mathsf{T}.$$

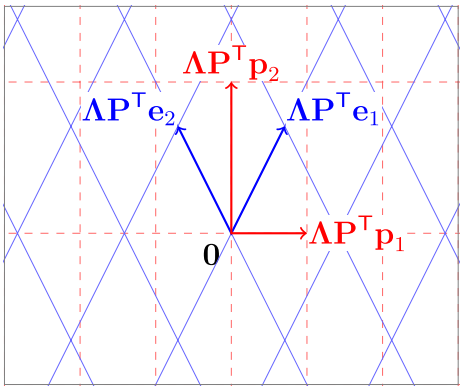
This is what we start with: $\mathbf{e}_i$ represents the i-th standard basis vector, and $\mathbf{p}_i$ the i-th eigenvector.



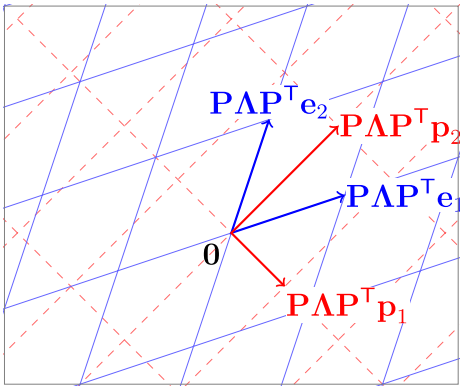
Then $\mathbf{P}^\mathsf{T}$ aligns the eigenvector to become the new axes.



Then, the scaling by $\mathbf{\Lambda}$.



Finally, $\mathbf{P}$ rotates the eigenvectors back to their original angle.



Quadratic forms

Let $\mathbf{A}$ be positive-definite. Then, expanding the equation $\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} = c^2$, it can easily be seen that it corresponds to an equation of an ellipse (or rather, ellipsoid), with axes given by the eigenvectors scaled by c times the reciprocal of the corresponding eigenvalue. A geometric argument in the same vain as the preceding section is more enlightening.

Let $\mathbf{A} = \mathbf{P\Lambda P}^\mathsf{T}$ be a spectral decomposition of $\mathbf{A}$. Since $\mathbf{A}$ is positive-definite, all its eigenvalues are positive and so, $\mathbf{\Lambda}$ has a square-root and is also, invertible. Now, the equation becomes $(\mathbf{\Lambda}^{1/2}\mathbf{P}^\mathsf{T}\mathbf{x})^\mathsf{T}(\mathbf{\Lambda}^{1/2}\mathbf{P}^\mathsf{T}\mathbf{x}) = c^2$. So, the solution of the equation involves all those vectors $\mathbf{x}$, which when rotated by $\mathbf{P}^\mathsf{T}$, and then scaled by $\mathbf{\Lambda}^{1/2}$, ends up in the sphere of radius c. So, if we scale the points in the sphere of radius c back down by $\mathbf{\Lambda}^{-1/2}$, and then rotate them back by $\mathbf{P}$, these points should end up in the locus of the equation i.e.,

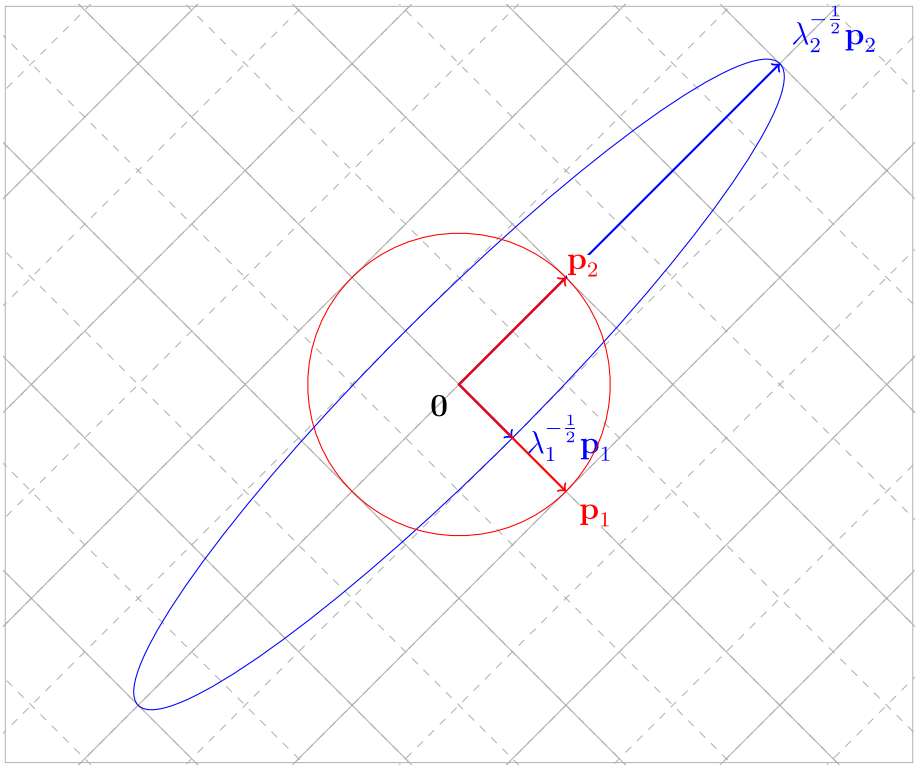
$$\{\mathbf{x} \mid \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} = c^2\} = \{\mathbf{P\Lambda}^{-1/2}\mathbf{x} \mid \mathbf{x}^\mathsf{T}\mathbf{x} = c^2\}.$$

When the (inverse) scaling is applied, it should be intuitive that the sphere now becomes an ellipse with the i-th axis given by i-th canonical basis vector scaled by c times the reciprocal of the square-root of $\mathbf{\Lambda}$. When the rotation is applied, the axes now correspond to the i-th eigenvector with the scaling just described.

The geometry of this is shown in for the matrix

$$\begin{bmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ -\sqrt{2}/2 & \sqrt{2}/2 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 1/3 \end{bmatrix}^2 \begin{bmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ -\sqrt{2}/2 & \sqrt{2}/2 \end{bmatrix}^\mathsf{T},$$

and $c^2 = 1$. The red circle denotes the unit circle. The blue ellipse is the locus of the quadratic form i.e., it what it is obtained after squishing the unit circle horizontally and stretching it threefold vertically, and then rotating it to the eigenvectors. It is important to note that the major (minor) axis is given by the eigenvector with smallest (largest) eigenvalue. Also, that the reciprocals of eigenvalues are the eigenvalues of the inverse.



Mahalanobis distance

The probability density function of a mulitvariate normal vector $\mathbf{x}$ of mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$ is given by

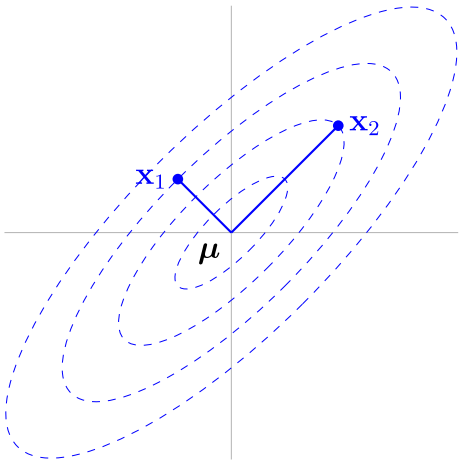
$$|2\pi\boldsymbol{\Sigma}|^{1/2} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}(\mathbf{x}-\boldsymbol{\mu})\right).$$

Notice that the locus constant density is given by the equation

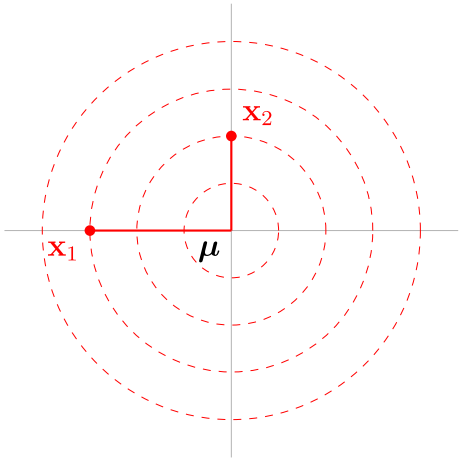
$$(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c^2.$$

As discussed, these are equations of ellipsoids with axes given by the eigenvectors of $\boldsymbol{\Sigma}$ and their lengths by c times the square-root of the corresponding eigenvalue, however, this time it is centred at $\boldsymbol{\mu}$. The quantity

$(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$ is the *Mahalanobis distance* between $\mathbf{x}$ and $\boldsymbol{\mu}$ which is a *statistical generalisation* of the Euclidean distance in a sense that will be made clear.



The figure above plots the loci of constant density for four different values. Now, consider the points $\mathbf{x}_1$ and $\mathbf{x}_2$. The latter is clearly farther away from the mean in the sense of Euclidean distance. However, statistically $\mathbf{x}_1$ is farther away in the sense that $\mathbf{x}_1$ is a less likely instance since it is situated in a locus of greater density. It'd be nice if we could transform the space, so that the ellipses become circles as in the figure below, and then measure the Euclidean distance of the points.



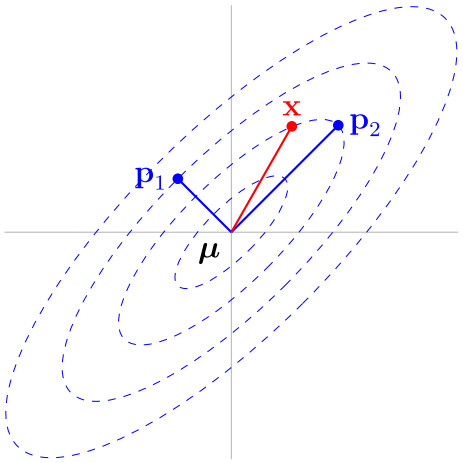
From discussion of quadratic forms, this is given by the transformation $\boldsymbol{\Lambda}^{-1/2}\mathbf{P}^{\mathsf{T}}$, where $\mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^{\mathsf{T}}$ is a spectral decomposition of $\boldsymbol{\Sigma}$ (not $\boldsymbol{\Sigma}^{-1}$). So, the transformed Euclidean distance now becomes

$$(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\mathbf{P}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Lambda}^{-1/2}\mathbf{P}^{\mathsf{T}}(\mathbf{x}-\boldsymbol{\mu}) = (\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}),$$

which is just the Mahalanobis distance.

Principal component analysis

Suppose we have a p-dimensional observation $\mathbf{x}$ from a mulitvariate normal distribution as below.



We want to lower the dimension of $\mathbf{x}$ to a one-dimensional one as some weighted linear combination $\mathbf{a}^\mathsf{T}\mathbf{x}$, $\|\mathbf{a}\| = 1$. This linear combination can be understood as the $\mathbf{a}$-th coordinate of $\mathbf{x}$ according to some orthonormal basis in which one of the basis vector is $\mathbf{a}$. In general, we may want to collapse into some dimension q, $1 \leq q < p$, in which case, we write the resultant combination as $\mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is an orthonormal linear transformation from $\mathbb{R}^p$ to $\mathbb{R}^q$. The question is what sort of $\mathbf{A}$ to choose.

In the case of $q = 1$, if we have multiple observation $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of unknown variance, then $\mathbf{a}$ should be chosen to maximise the sample variance of $\mathbf{a}^\mathsf{T}\mathbf{x}_1, \ldots, \mathbf{a}^\mathsf{T}\mathbf{x}_n$ since this would be separating the observations the most, and thus losing as little information as possible. Or, in the case of known variance, we ought to choose $\mathbf{a}$ as the one that maximises the variance of $\mathbf{a}^\mathsf{T}\mathbf{x}$ since this would explain the distribution the most. Looking at the figure, one immediately realises, the variance is maximised at $\mathbf{a} = \mathbf{p}_1$: the eigenvector corresponding to the largest eigenvalue of $\boldsymbol{\Sigma}$ since this is the axis where $\mathbf{x}$ varies the most—it's the widest part of the ellipse. And thus, $\mathbf{p}_1^\mathsf{T}\mathbf{x}$ is the *first* principal component of $\mathbf{x}$.

For $q = 2$, we should choose the *second* principal component $\mathbf{b}^\mathsf{T}\mathbf{x}$, $\|\mathbf{b}\| = 1$ as the one that maximises the variance same as before, but is wholly independent of $\mathbf{a}^\mathsf{T}\mathbf{x}$ since we want to minimise repetition of information as much as possible. Again, looking at the ellipse, one realises $\mathbf{b}$ should be chosen as $\mathbf{p}_2$—the eigenvector corresponding to the second-largest eigenvalue. And we proceed so on and so forth to define all the p principal components as $\mathbf{P}^\mathsf{T}\mathbf{x}$, where $\mathbf{P}$ is the orthonormal matrix of eigenvectors of $\boldsymbol{\Sigma}$, arranged in descending order of their eigenvalues.