PCA and change of basis

Thibaud Taillefumier

The most direct way to understand PCA is to consider the following linear algebra question: assuming the data is living in a d-dimensional vectorial space, what is the best choice of basis to represent the data? Intuitively, a "good" basis would be a basis in which we expect the data structure to be salient. However it is hard to imagine an automated procedure that produces such a basis without knowledge of the data characteristic "features "in the first place. Alternatively, we can try to find the basis in which the data covariance matrix is as simple as possible, that is under a diagonal form. Remember that the data covariance matrix is diagonal if the components of the centered data vector are uncorrelated. PCA achieves such a goal.

To see how it works, let us remember that a change of basis affects the coordinates of the data via a change of matrix P. Specifically, if $\boldsymbol{x}_i$ is the original data coordinate vector, the new coordinate vector $\boldsymbol{y}_i$ is obtained via matrix multiplication by P: $\boldsymbol{y}_i = P\boldsymbol{x}_i$. Incidentally, we can consider the data matrix in the new coordinates: $Y = PX$ where P is the same yet-to-be-defined change-of-basis matrix that simplifies the data covariance. To find P, we are going to use the fact that the covariance matrix of the new coordinates C_{YY} is related to the covariance matrix of the original coordinates C_{XX} by:

$$C_{YY} = \frac{1}{n-1} YY^T = \frac{1}{n-1}(PX)(PX)^T = P\left(\frac{1}{n-1}XX^T\right)P^T = PC_{XX}P^T\,.$$

Now from the previous lecture, we know that a good candidate basis should include the top eigenvector of C_{xx}. This suggests utilizing the spectral theorem to consider the full eigendecomposition of C_{XX}

$$C_{XX} = VDV^T\,, \quad \text{with} \quad D = \begin{bmatrix} s_1 & 0 & \dots & 0 \\ 0 & s_2 & \dots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \dots & s_d \end{bmatrix} \quad \text{and}$$

where the eigenvalues are such that $s_1 \geq s_2 \geq \ldots \geq s_d \geq 0$ and where the matrix V is orthogonal, i.e. $VV^T = I$. This allows us to rewrite the covariance C_{YY}

under the form

$$C_{YY} = PC_{XX}P^T = PVDV^TP^T = (PV)D(PV)^T\,,$$

which makes apparent what is the "good" choice for the change-of-basis matrix P. Choosing P as equal to the orthogonal matrix obtained via eigendecomposition, i.e. $P = V^{-1} = V^T$, yields

$$C_{YY} = (PV)D(PV)^T = (V^{-1}V)D(V^{-1}V)^T = IDI = D\,.$$

Thus, when considered in the basis defined by the eigenvectors of C_{XX}, the covariance of the data is equal to the diagonal matrix D, whose diagonal entries satisfies $s_1 \geq s_2 \geq \ldots \geq s_d \geq 0$. As intended, all the off-diagonal terms are zero, which means that the covariance between the data components in the eigenvector basis is zero: $\langle y_i y_j \rangle = 0$. Moreover, as V is an orthogonal matrix, we can interpret the components of $\boldsymbol{y}$ as the projection coefficients of $\boldsymbol{x}$ onto the eigenvector $\boldsymbol{v}_i$, $1 \leq i \leq d$, which constitute an orthonormal basis:

$$\boldsymbol{y} = V^T\boldsymbol{x} = \begin{bmatrix} \boldsymbol{v}_1^T \\ \boldsymbol{v}_2^T \\ \vdots \\ \boldsymbol{v}_d^T \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} \boldsymbol{v}_1^T\boldsymbol{x} \\ \boldsymbol{v}_2^T\boldsymbol{x} \\ \vdots \\ \boldsymbol{v}_d^T\boldsymbol{x} \end{bmatrix}.$$

In turn, we can interpret the eigenvalue s_i as the variance of the data when projected onto the eigenvector $\boldsymbol{v}_i$.

Depending on the field of studies, the eigenvectors $\boldsymbol{v}$ are also called principal components or singular vectors. These eigenvectors can be thought of as data "features" that can be retrieved from the data covariance matrix. Projecting the data on the first k eigenvectors produces a k-dimensional representation while preserving as much of the data variability as possible. Indeed, the data variability captured by the first k eigenvalues is the sum of the k first eigenvalues

$$\begin{aligned} \sum_{i=1}^{k} \mathbb{V}(y_i) &= \sum_{i=1}^{k} \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{v}_k^T\boldsymbol{x}_i)^2\,, \\ &= \sum_{i=1}^{k} \boldsymbol{v}_k^T \left(\frac{1}{n-1} XX^T\right) \boldsymbol{v}_k\,, \\ &= \sum_{i=1}^{k} \boldsymbol{v}_k^T C_{XX} \boldsymbol{v}_k\,, \\ &= \sum_{i=1}^{k} s_k\,, \end{aligned}$$

where we remember that the eigenvalues are ranked by decreasing order. The fraction of the data variability accounted by the first k components is given by

$$f_k = \frac{s_1 + \ldots + s_k}{s_1 + \ldots + s_k + \ldots + s_d},$$

where the denominator $s_1 + \ldots + s_k + \ldots + s_d$ is the total variance of the data. The closer f_k is to one the more faithful is the projection, i.e. the more accurate is the dimensional reduction.