Principal Component Analysis

Thibaud Taillefumier

1 Problem setting

Suppose you are presented with data under the form of a sequence of d-dimensional vectors $\boldsymbol{x}_i$, $1 \leq i \leq n$. This is a very general setting. For instance, the data sequence can be a collection of membrane potential waveforms (as in spike sorting) or a set of images represented as a vector of pixels (as in machine learning). Taking the convention to represent data as a column vector

$$\boldsymbol{x}_i = \begin{bmatrix} x_{1,i} \\ x_{2,i} \\ \vdots \\ x_{d,i} \end{bmatrix},$$

we can pull all the data together and form the data matrix

$$X = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \dots & \boldsymbol{x}_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,n} \\ x_{2,1} & x_{2,2} & \dots & x_{2,n} \\ \vdots & & & \vdots \\ x_{d,1} & x_{d,2} & \dots & x_{d,n} \end{bmatrix},$$

where each column is a data sample. Analyzing—and hopefully understanding—the result of an experiment often consists in uncovering regularity or structure in the data matrix. Unfortunately, measured data is often "messy" in the sense that it is too high-dimensional for us to detect structure by direct inspection and in the sense that noise and/or redundancy often impairs data visualization.

Principal Component Analysis (PCA) is a handy tool to reveal structure via dimensionality reduction and denoising of the data. In a nutshell and loosely speaking, PCA consists in detecting characteristics "features" of the data that can be ranked by degree of relevance: the more relevant a feature, the more it explains the variability of the data. PCA is successful when considering only a few of the most relevant "features" is enough to describe the data satisfactorily. Thus, successful

PCA offers the possibility to perform dimensionality reduction as the data can be represented in a space whose dimension is specified by the number of kept "features". At the same time, successful PCA can be seen since denoising of the data as the ignored "features" are most likely due to noise or redundancy in the data collection process.

Before making the above statements more precise, we need to first make the assumption that our data vector has zero mean. Such an assumption incurs no loss of generality as we can always subtract the sample mean from the original data samples to form a zero-mean vector sequence:

$$\boldsymbol{x}_i \longleftarrow \boldsymbol{x}_i - \langle \boldsymbol{x}_i \rangle \, .$$

The approach taken by PCA is to look for data "features" in the data covariance matrix defined as

$$C_{XX} = \frac{1}{n-1} X X^T = \begin{bmatrix} \langle x_{1,i}^2 \rangle & \langle x_{1,i} x_{2,i} \rangle & \dots & \langle x_{1,i} x_{d,i} \rangle \\ \langle x_{2,i} x_{1,i} \rangle & \langle x_{2,i}^2 \rangle & \dots & \langle x_{2,i} x_{d,i} \rangle \\ \vdots & & & \vdots \\ \langle x_{d,i} x_{1,i} \rangle & \langle x_{d,i} x_{2,i} \rangle & \dots & \langle x_{d,i}^2 \rangle \end{bmatrix}$$

for zero mean data. The reason for looking for features of X in the covariance matrix C_{XX} is that if a few "features" are enough to characterize the data, we expect the data to lie on a low-dimensional manifold, as opposed to fill the whole d-dimensional space it lives in. If that low-dimensional manifold is not too convoluted, it will lie into a low-dimensional vectorial space and the covariance matrix will capture the few directions along which the data is primarily varying. Notice that an intrinsic limitation to PCA is that it can only detect linear features and as such, it is not well-suited to discover data features that result from highly non-linear transformation of the raw data.

2 The idea behind PCA

Given a data matrix X, let us look for the direction, i.e. the unit vector $\boldsymbol{v}$, such that the orthogonal projection of the data onto $\boldsymbol{v}$ best captures the overall variability of the data. The projection coefficients of each data sample $\boldsymbol{x}_i$ onto $\boldsymbol{v}$ defines a collection of numbers c_i, $1 \leq i \leq n$, which can be written in vectorial form as

$$\begin{bmatrix} c_1 & c_2 & \dots & c_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}^T \boldsymbol{x}_1 & \boldsymbol{v}^T \boldsymbol{x}_2 & \dots & \boldsymbol{v}^T \boldsymbol{x}_n \end{bmatrix} = \boldsymbol{v}^T X \, .$$

The variance of the data accounted by the vector $\boldsymbol{v}$ is defined as the variance the projection coefficients c, which can be expressed as:

$$\mathbb{V}(c) = \frac{1}{n-1}\sum_{i=1}^{n} c_i^2 = \frac{1}{n-1}\begin{bmatrix} c_1 & c_2 & \dots & c_n \end{bmatrix}\begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} = \frac{1}{n-1}\boldsymbol{v}^T X(\boldsymbol{v}^T X)^T .$$

Now, our problem is to find the unit vector $\boldsymbol{v}^\star$ for which $\mathbb{V}(c)$ is maximum, which can be stated formally as

$$\boldsymbol{v}^\star = \arg\max_{\|v\|^2=1} \mathbb{V}(c) ,$$

where the suffix expression under $\max$ indicates that we restrain our search to unit vectors. In order to tackle the above optimization problem, it is convenient to first reformulate the variance $\mathbb{V}(c)$ in terms of the covariance matrix C_{XX}, which contains all the information required for PCA:

$$\mathbb{V}(c) = \frac{1}{n-1}\boldsymbol{v}^T X(\boldsymbol{v}^T X)^T = \boldsymbol{v}^T\left(\frac{1}{n-1}XX^T\right)\boldsymbol{v} = \boldsymbol{v}^T C_{XX}\boldsymbol{v} ,$$

Our optimization problem consists then in finding

$$\boldsymbol{v}^\star = \arg\max_{\|v\|^2=1} \boldsymbol{v}^T C_{XX}\boldsymbol{v} .$$

The only difficulty involved in this problem is the fact that we restrain ourselves to unit vector. Without this constraint, we would just look for $\boldsymbol{v}^\star$ by setting the derivative of $\boldsymbol{v}^T C_{XX}\boldsymbol{v}$ with respect to the components of $\boldsymbol{v}$ to zero and solve the resulting system of equations. Also not formally exact, it turns out that this approach gives the right answer anyway. To see why, we need to realize that we can take into account the constraint of unit length by optimizing the function

$$\begin{aligned} F(\boldsymbol{v},\lambda) &= \boldsymbol{v}^T C_{XX}\boldsymbol{v} - \lambda\left(\|\boldsymbol{v}\|^2 - 1\right) , \\ &= \sum_i\sum_j (C_{XX})_{ij} v_i v_j - \lambda\left(\sum_i v_i^2 - 1\right) , \end{aligned}$$

which depends on $\boldsymbol{v}$ and a new parameter λ called the Lagrangian multiplier. Notice that if $\|\boldsymbol{v}\| = 1$, we have $F(\boldsymbol{v},\lambda) = \boldsymbol{v}^T C_{XX}\boldsymbol{v}$. If $\|\boldsymbol{v}\| \neq 1$, it is always possible to make F as positive or negative as possible by varying λ and there is no optimum. These loose observations are the reason for introducing the function F

and a complete justification of this fact is beyond the scope of this class. Let us directly proceed with the optimization of F by first computing the derivatives with respect to v_k:

$$\begin{aligned}\frac{\partial F(\boldsymbol{v},\lambda)}{\partial v_k} &= \sum_i (C_{XX})_{ik} v_i + \sum_j (C_{XX})_{kj} v_j - 2\lambda v_k \,, \\ &= 2\sum_i (C_{XX})_{ki} v_i - 2\lambda v_k \,,\end{aligned}$$

where we have used the fact that C_{XX} is a symmetric matrix. Setting these derivatives to zero yields a system of d equations:

$$\frac{\partial F(\boldsymbol{v},\lambda)}{\partial v_k} = 0 \quad \Leftrightarrow \quad \sum_i (C_{XX})_{ik} v_i = \lambda v_k \,.$$

This system can be conveniently written under matrix form as

$$C_{XX}\boldsymbol{v} = \lambda \boldsymbol{v} \,,$$

making apparent the fact that the vector $\boldsymbol{v}$ that maximizes the projected variance is an eigenvector of C_{XX}. Moreover, setting the derivative of F with respect to λ to zero yields

$$\frac{\partial F(\boldsymbol{v},\lambda)}{\partial \lambda} = \sum_i v_i^2 - 1 = \|\boldsymbol{v}\|^2 - 1 \,,$$

which is not a problem since eigenvectors are defined modulo their length: we can always chose a unit eigenvector. Now which eigenvector to choose? By the spectral theorem, we know that, in general, symmetric matrices have d distinct real eigenvalues s_i, $1 \leq i \leq d$. Moreover for covariance matrices, we know that these eigenvalues are all positives. Suppose, we pick an eigenvector $\boldsymbol{v}$ of unit length associated to eigenvalue λ. Then we have

$$F(\boldsymbol{v},\lambda) = \boldsymbol{v}^T C_{XX} \boldsymbol{v} = \lambda \boldsymbol{v}^T \boldsymbol{v} = \lambda \,,$$

where the first and last equalities are due to the fact $\|\boldsymbol{v}\| = 1$ and the second equality is due to the fact $\boldsymbol{v}$ is a λ-eigenvector. This shows that, to maximize $\boldsymbol{v}^T C_{XX} \boldsymbol{v}$, one has to choose the eigenvector associated with the top eigenvalue of the spectral decomposition of C_{XX}.

The above analysis shows that we can extract from the covariance matrix a particular direction that best accounts for the data variability. The key idea is to use linear algebra to find that direction as the top eigenvector of the data covariance matrix. The following section generalizes this idea to considering all the eigenvectors of the covariance matrix and introduces PCA as the "best" change of basis to account for data variability.

3 PCA and change of basis

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 section, 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 eigen decomposition 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 U is orthogonal, i.e. $UU^T = I$. 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 eigen decomposition,

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 component of $\boldsymbol{y}$ as the projection coefficient of $\boldsymbol{x}$ onto the eigenvector $\boldsymbol{v}_i$, $1 \leq i \leq d$ which constitutes 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 singular value 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

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

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.