Notes on Principal Component Analysis (PCA)

Gabriele Tolomei

March 28, 2023

In many use cases, such as computer vision or natural language processing, data is naturally represented by vectors in a very high-dimensional space (i.e., a space made of thousands or even millions of dimensions). High-dimensional data may suffer from a well-known problem, which is typically referred to as the *curse of dimensionality.* Very roughly, this refers to the inability to distinguish between data points that are close from those that are far away from each other since data points in high-dimensional space tend, in fact, to be all distant (and sparser) from each other.

Principal Component Analysis (PCA) is an effective technique to reduce data dimensionality, which identifies a low-dimensional (linear) sub-space so as to preserve as much as possible the "structure" (i.e., variance) of the data represented in the original, high-dimensional space.

Below, we review the theoretical foundations of this well-known feature extraction technique.

1 Preliminaries

We are given with a set of n data points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, each one laying on a d-dimensional space, such that $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,d})^T$, for all $i = \{1, \ldots, n\}$.

Initially, we associate a random variable X_j to each dimension (i.e., $j = \{1, \ldots, d\}$). Thus, we define the expected value of each X_j as the mean computed from the n observations of *that* specific j dimension (i.e., feature). In other words:

$$E[X_j] = \mu_j = \frac{1}{n}\sum_{i=1}^{n} x_{i,j}.$$

The mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_d)^T$ contains the sample mean of each dimension. Then, we can rewrite all the data points by "centering" them around the mean (i.e., we substract from each dimension its corresponding mean):

$$\boldsymbol{x}_i' = \boldsymbol{x}_i - \boldsymbol{\mu} = (x_{i,1} - \mu_1, \ldots, x_{i,d} - \mu_d)^T.$$

The values of each random variable X_j are changed accordingly, and therefore each original random variable X_j becomes X'_j, which will now have 0-mean:

$$E[X'_j] = \frac{1}{n}\sum_{i=1}^{n}(x_{i,j} - \mu_j) = \frac{1}{n}\left(\sum_{i=1}^{n} x_{i,j} - \sum_{i=1}^{n} \mu_j\right) =$$

$$\frac{1}{n}\left(\sum_{i=1}^{n} x_{i,j} - n\mu_j\right) = \frac{1}{n}\sum_{i=1}^{n} x_{i,j} - \frac{1}{n}n\mu_j = \mu_j - \mu_j = 0.$$

We can get to the same result by observing that $\sum_{i=1}^{n} x_{i,j} = n\mu_j$; thus:

$$\frac{1}{n}\left(\sum_{i=1}^{n} x_{i,j} - n\mu_j\right) = \frac{1}{n}(n\mu_j - n\mu_j) = 0.$$

In addition to that, we may also want to *standardize* each dimension so that they will eventually have 1-standard-deviation. In other words, for each dimension j, we compute the sample standard deviation s_j as follows:

$$s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{i,j} - \mu_j)^2}.$$

We can conveniently represent all the standard deviations into a single d-dimensional vector as we did for the sample mean, namely: $\boldsymbol{s} = (s_1, s_2, \ldots, s_d)^T$. Eventually, the i-th original data point $\boldsymbol{x}_i$ is standardized as follows:

$$\boldsymbol{z}_i = \frac{\boldsymbol{x}'_i}{\boldsymbol{s}} = \frac{\boldsymbol{x}_i - \boldsymbol{\mu}}{\boldsymbol{s}} = \Big(\underbrace{\frac{x_{i,1} - \mu_1}{s_1}}_{z_{i,1}}, \ldots, \underbrace{\frac{x_{i,d} - \mu_d}{s_d}}_{z_{i,d}}\Big)^T.$$

If we now associate a *new* random variable Z_j to each of those standardized dimensions, it will turn out that:

$$(i)\ E[Z_j] = 0; (ii)\ \sqrt{E[(Z_j - E[Z_j])^2]} = 1 \quad \forall j \in \{1, \ldots, d\}.$$

The former condition (i) easily derives from the fact that $E[X'_j] = 0$, but it can be explicitly verified as follows:

$$E[Z_j] = \frac{1}{n}\sum_{i=1}^{n}\frac{x_{i,j} - \mu_j}{s_j} = \frac{1}{ns_j}\left(\sum_{i=1}^{n} x_{i,j} - \sum_{i=1}^{n} \mu_j\right) = \frac{1}{ns_j}(n\mu_j - n\mu_j) = 0.$$

To verify the second condition (ii), notice that $E[Z_j] = 0$, therefore:

$$\sqrt{E[(Z_j - E[Z_j])^2]} = \sqrt{E[(Z_j - 0)^2]} = \sqrt{E[Z_j^2]} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} z_{i,j}^2}.$$

Let us rewrite the last term by expanding $z_{i,j}^2$:

$$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n} z_{i,j}^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\Big(\frac{x_{i,j}-\mu_j}{s_j}\Big)^2} = \sqrt{\frac{1}{n-1}\frac{1}{s_j^2}\sum_{i=1}^{n}(x_{i,j}-\mu_j)^2}.$$

Note that:

$$s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{i,j}-\mu_j)^2.$$

Therefore, we substitute it into the previous equation and obtain the following:

$$\sqrt{\frac{1}{\cancel{n-1}}\frac{\cancel{n-1}}{\cancel{\sum_{i=1}^{n}(x_{i,j}-\mu_j)^2}}\cancel{\sum_{i=1}^{n}(x_{i,j}-\mu_j)^2}} = \sqrt{1} = 1.$$

We can now conveniently arrange our n d-dimensional standardized data points into an $n \times d$ matrix Z, as follows:

$$Z = \begin{bmatrix} z_{1,1} & z_{1,2} & \cdots & z_{1,d} \\ z_{2,1} & z_{2,2} & \cdots & z_{2,d} \\ \vdots & \vdots & \vdots & \vdots \\ z_{n,1} & z_{n,2} & \cdots & z_{n,d} \end{bmatrix} = \begin{bmatrix} — & \boldsymbol{z}_1^T & — \\ — & \boldsymbol{z}_2^T & — \\ — & \cdots & — \\ — & \boldsymbol{z}_n^T & — \end{bmatrix} = \begin{bmatrix} | & | & | & | \\ \boldsymbol{f}_1 & \boldsymbol{f}_2 & \cdots & \boldsymbol{f}_d \\ | & | & | & | \end{bmatrix}$$

where $\boldsymbol{z}_i^T = (z_{i,1}, z_{i,2}, \ldots, z_{i,d})$ is the d-dimensional (row) vector corresponding to the i-th data point and $\boldsymbol{f}_j = (z_{1,j}, z_{2,j}, \ldots, z_{n,j})^T$ is the n-dimensional (column) vector corresponding to the observations of the j-th feature.

1.1 The Covariance Matrix

Generally speaking, the *covariance* between two random variables Z_j, Z_k captures how those two are related to each other. More formally:

$$\mathrm{Cov}(Z_j, Z_k) = E[(Z_j - E[Z_j])(Z_k - E[Z_k])].$$

Let us consider the set of d random variables associated with the d standardized dimensions above, i.e., $Z_1, Z_2, \ldots, Z_d$. Moreover, let us compute the covariance between each pair of variables, i.e., $\mathrm{Cov}(Z_j, Z_k)$ $\forall j, k \in \{1, \ldots, d\}$. In other words, in a d-dimensional space, we define the *covariance matrix* K as the $d \times d$ square matrix as follows:

$$K = \begin{bmatrix} \mathrm{Cov}(Z_1, Z_1) & \mathrm{Cov}(Z_1, Z_2) & \ldots & \mathrm{Cov}(Z_1, Z_d) \\ \mathrm{Cov}(Z_2, Z_1) & \mathrm{Cov}(Z_2, Z_2) & \ldots & \mathrm{Cov}(Z_2, Z_d) \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{Cov}(Z_d, Z_1) & \mathrm{Cov}(Z_d, Z_2) & \ldots & \mathrm{Cov}(Z_d, Z_d) \end{bmatrix}$$

Under the assumption that $E[Z_j] = 0$ for all j the covariance between any pair of dimensions turns into:

$$\text{Cov}(Z_j, Z_k) = E[Z_j Z_k] = \frac{1}{n-1} \sum_{i=1}^{n} z_{i,j} * z_{i,k}.$$

Note that $\text{Cov}(Z_j, Z_k) = \text{Cov}(Z_k, Z_j)$, i.e., the covariance matrix is *symmetric*. Moreover, observe that on the main diagonal of K, we have the following:

$$\text{Cov}(Z_j, Z_j) = E[Z_j Z_j] = \frac{1}{n-1} \sum_{i=1}^{n} z_{i,j} * z_{i,j} = \frac{1}{n-1} \sum_{i=1}^{n} z_{i,j}^2 = (E[Z_j - \underbrace{E[Z_j]}_{=0}])^2 = \text{Var}(Z_j).$$

Overall, we have:

$$K = \begin{bmatrix} \text{Var}(Z_1) & \text{Cov}(Z_1, Z_2) & \dots & \text{Cov}(Z_1, Z_d) \\ \text{Cov}(Z_2, Z_1) & \text{Var}(Z_2) & \dots & \text{Cov}(Z_2, Z_d) \\ \vdots & \vdots & \vdots & \vdots \\ \text{Cov}(Z_d, Z_1) & \text{Cov}(Z_d, Z_2) & \dots & \text{Var}(Z_d) \end{bmatrix}$$

It is also worth noticing that the covariance matrix K can be obtained as follows:

$$K = \frac{1}{n-1} * Z^T Z = \frac{1}{n-1} * \underbrace{\begin{bmatrix} — & \boldsymbol{f}_1^T & — \\ — & \boldsymbol{f}_2^T & — \\ — & \dots & — \\ — & \boldsymbol{f}_d^T & — \end{bmatrix}}_{Z^T} \underbrace{\begin{bmatrix} | & | & | & | \\ \boldsymbol{f}_1 & \boldsymbol{f}_2 & \dots & \boldsymbol{f}_d \\ | & | & | & | \end{bmatrix}}_{Z}$$

To convince yourself the above equivalence is true, let us see what happens when we compute $\boldsymbol{f}_j^T \cdot \boldsymbol{f}_j$, i.e., the dot product between two vectors representing the same generic j-th dimension:

$$\boldsymbol{f}_j^T \cdot \boldsymbol{f}_j = \sum_{i=1}^{n} z_{i,j} * z_{i,j} = \sum_{i=1}^{n} z_{i,j}^2.$$

The result above confirms that:

$$\frac{1}{n-1} \boldsymbol{f}_j^T \cdot \boldsymbol{f}_j = \frac{1}{n-1} \sum_{i=1}^{n} z_{i,j}^2 = \text{Var}(Z_j) = \text{Cov}(Z_j, Z_j).$$

Now, let us consider the dot product between the generic j-th and k-th dimension ($j \neq k$):

$$\boldsymbol{f}_j^T \cdot \boldsymbol{f}_k = \sum_{i=1}^{n} z_{i,j} * z_{i,k}.$$

Again, this results in the following:

$$\frac{1}{n-1} \boldsymbol{f}_j^T \cdot \boldsymbol{f}_k = \frac{1}{n-1} \sum_{i=1}^{n} z_{i,j} * z_{i,k} = \text{Cov}(Z_j, Z_k).$$

You may wonder why seeing $K = Z^T Z$ may be useful. In the following, we demonstrate that this is convenient to show that the covariance matrix K is *positive-semidefinite*. Furthermore, we will use this result to derive a nice property of the *eigenvalues* associated with that matrix (more on this later).

Definition 1. *A $d \times d$ symmetric real matrix A is said to be **positive-semidefinite** (or non-negative-definite) iff $\boldsymbol{v}^T A \boldsymbol{v} \geq 0 \ \forall \boldsymbol{v} \in \mathbb{R}^d$.*

Lemma 1. *The correlation matrix K is **positive-semidefinite**.*

Proof. Let us assume that K is *not* positive-semidefinite. According to Def. 1, that means it should exist *at least one* vector $\boldsymbol{w} \in \mathbb{R}^d$, such that $\boldsymbol{w}^T K \boldsymbol{w} < 0$. By using the fact that $K = Z^T Z$, we can rewrite this expression as follows:

$$\boldsymbol{w}^T K \boldsymbol{w} = \boldsymbol{w}^T Z^T Z \boldsymbol{w} = (Z\boldsymbol{w})^T (Z\boldsymbol{w}).$$

Notice that $Z\boldsymbol{w}$ is the result of a $n \times d$ matrix (Z) multiplied by a $d \times 1$ column vector ($\boldsymbol{w}$), i.e., an $n \times 1$ column vector, namely:

$$Z\boldsymbol{w} = (\boldsymbol{z}_1^T \cdot \boldsymbol{w}, \boldsymbol{z}_2^T \cdot \boldsymbol{w}, \dots, \boldsymbol{z}_n^T \cdot \boldsymbol{w})^T,$$

where each $\boldsymbol{z}_i^T \cdot \boldsymbol{w} = \sum_{j=1}^d z_{i,j} * w_j$. Let us define $\alpha_i = \boldsymbol{z}_i^T \cdot \boldsymbol{w}$ and, thus, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_n)^T$ then:

$$Z\boldsymbol{w} = (\alpha_1, \alpha_2, \dots, \alpha_n)^T = \boldsymbol{\alpha}.$$

Now, we can rewrite $(Z\boldsymbol{w})^T (Z\boldsymbol{w})$ as follows:

$$(Z\boldsymbol{w})^T (Z\boldsymbol{w}) = \boldsymbol{\alpha}^T \cdot \boldsymbol{\alpha} = \sum_{i=1}^n \alpha_i * \alpha_i = \sum_{i=1}^n \alpha_i^2.$$

In other words, we have shown that:

$$\boldsymbol{w}^T K \boldsymbol{w} = (Z\boldsymbol{w})^T (Z\boldsymbol{w}) = \sum_{i=1}^n \alpha_i^2.$$

Since $\sum_{i=1}^n \alpha_i^2$ is a sum of squares, it turns out it cannot be negative! In other words:

$$\sum_{i=1}^n \alpha_i^2 \geq 0 \Rightarrow \boldsymbol{w}^T K \boldsymbol{w} \geq 0,$$

which contradicts our initial assumption. Therefore, the correlation matrix K is indeed positive-semidefinite.

1.2 From a Visual Intuition to the Rigorous Definition of PCA

We have already "visually" convinced ourselves that in order to find the principal components of our d-dimensional space, we must find the *eigenvectors* (and their associated *eigenvalues*) of the covariance matrix K. In other words, we need to solve the following equation:

$$K\boldsymbol{e} = \lambda\boldsymbol{e}.$$

where $\boldsymbol{e}$ is a d-dimensional eigenvector and λ is the corresponding eigenvalue. The equation above can be rewritten as follows:

$$K\boldsymbol{e} - \lambda\boldsymbol{e} = \mathbf{0} \Rightarrow (K - \lambda I)\boldsymbol{e} = \mathbf{0}.$$

where I is a d-by-d *identity matrix.*
We, therefore, need to solve the above *homogeneous* system of equations; any homogeneous system always has a *trivial* solution (i.e., in the case above, the zero-vector $\boldsymbol{e} = \mathbf{0}$). The only way for a homogeneous system like the one above to have *also* non-trivial solutions is for its matrix $(K - \lambda I)$ to be *non-invertible.* Indeed, if $(K - \lambda I)$ is invertible then we can multiply by its inverse $(K - \lambda I)^{-1}$ both sides of the equation:

$$\cancel{(K - \lambda I)}\,\cancel{(K - \lambda I)^{-1}}\boldsymbol{e} = \mathbf{0}(K - \lambda I)^{-1}.$$

Eventually, the only solution we obtain is still $\boldsymbol{e} = \mathbf{0}$.

From linear algebra theory, we know that a square matrix like $(K - \lambda I)$ is invertible iff its determinant is *not* equal to 0; on the other hand, if the determinant of $(K - \lambda I)$ is equal to 0 then that matrix will not be invertible, and therefore the corresponding homogeneous system will also have a non-trivial solution.
As a result, we must solve for λ the following equation:

$$\det(K - \lambda I) = 0.$$

The equation above is called the *characteristic equation* or *characteristic polynomial* of K. It is a polynomial function in λ of degree d. So we know that this equation will not have more than d roots or solutions, therefore no more than d eigenvalues.
Suppose we find d eigenvalues $\lambda_1, \ldots, \lambda_d$ as the solutions of the characteristic equation above. Then, we can plug each of these eigenvalues to, in turn, figure out the corresponding eigenvector.
Finally, eigenvectors must be divided by their L^2-norm in order to normalize them as length-1 vectors.

In the next section, we formally prove the two statements below:

1. The eigenvectors maximize the variance among all possible data directions (i.e., data projections);

2. We pick the eigenvector with the largest eigenvalue $\lambda_{\max}$ as the first principal component because $\lambda_{\max}$ is exactly the variance of the data along that eigenvector (i.e., projecting data onto any other eigenvector will result in retaining a smaller variance).

2 Eigenvectors: Greatest Variance Direction

Let us consider again our set of n d-dimensional *standardized* input data points above $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$, i.e., $\boldsymbol{z}_i = (z_{i,1}, \ldots, z_{i,d})^T$.
Let us also assume $\boldsymbol{e} = (e_1, \ldots, e_d)$ is the d-dimensional vector pointing towards the direction of the greatest variance we aim to find. Suppose we want to project each vector $\boldsymbol{z}_i$ onto the vector $\boldsymbol{e}$. Generally speaking, the *vector projection* of a vector $\boldsymbol{a}$ on another (non-zero) vector $\boldsymbol{b}$ is a vector whose magnitude is the *scalar projection* of $\boldsymbol{a}$ on $\boldsymbol{b}$ with the same direction as $\boldsymbol{b}$. In other words, it is defined as:

$$\boldsymbol{a}_{\|} = (||\boldsymbol{a}|| \cos\theta)\hat{\boldsymbol{b}}.$$

where $(||\boldsymbol{a}|| \cos\theta)$ is the *scalar projection* of the vector projection (i.e., the scaling factor), assuming θ is the angle between $\boldsymbol{a}$ and $\boldsymbol{b}$, and $\hat{\boldsymbol{b}} = \frac{\boldsymbol{b}}{||\boldsymbol{b}||}$ is the unit vector having the same direction of $\boldsymbol{b}$. When θ is known, we can compute:

$$\cos\theta = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{||\boldsymbol{a}||||\boldsymbol{b}||}.$$

As such, the scalar projection can be computed in terms of the dot product by noticing that:

$$||\boldsymbol{a}|| \cos\theta = ||\boldsymbol{a}|| \frac{\boldsymbol{a}^T \cdot \boldsymbol{b}}{||\boldsymbol{a}||||\boldsymbol{b}||} = \frac{\boldsymbol{a}^T \cdot \boldsymbol{b}}{||\boldsymbol{b}||}.$$

Similarly, we can compute the vector projection as follows:

$$\boldsymbol{a}_{\|} = \underbrace{\frac{\boldsymbol{a}^T \cdot \boldsymbol{b}}{||\boldsymbol{b}||}}_{||\boldsymbol{a}|| \cos\theta} \underbrace{\frac{\boldsymbol{b}}{||\boldsymbol{b}||}}_{\hat{\boldsymbol{b}}} = \frac{\boldsymbol{a}^T \cdot \boldsymbol{b}}{||\boldsymbol{b}||^2}\boldsymbol{b}.$$

Going back to our setting, we consider the scalar projection of each $\boldsymbol{z}_i$ onto $\boldsymbol{e}$. By using the definition above, the scalar projection can be computed as follows:

$$\frac{\boldsymbol{z}_i^T \cdot \boldsymbol{e}}{||\boldsymbol{e}||}.$$

By enforcing $\boldsymbol{e}$ to be normalized (i.e., a unit-length vector, such that $||\boldsymbol{e}|| = 1$), the scalar projection simply turns into computing the dot product between $\boldsymbol{z}_i$ and $\boldsymbol{e}$:

$$\boldsymbol{z}_i^T \cdot \boldsymbol{e} = \sum_{j=1}^{d} z_{i,j} * e_j.$$

Let us now compute the variance of the scalar projections of *all* our n input data points, as measured along the direction of $\boldsymbol{e}$, which is:

$$V(\boldsymbol{e}) = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{z}_i^T \cdot \boldsymbol{e} - \mu)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\Big(\sum_{j=1}^{d} z_{i,j} * e_j - \mu\Big)^2,$$

where μ is the mean computed among all the scalar projections.
First of all, we will show that such $\mu = 0$. Indeed, by definition, we have:

$$\mu = \frac{1}{n}\sum_{i=1}^{n}\underbrace{\Big(\sum_{j=1}^{d} z_{i,j} * e_j\Big)}_{\boldsymbol{z}_i^T \cdot \boldsymbol{e}};$$

$$\mu = \sum_{j=1}^{d} e_j\left(\frac{1}{n}\sum_{i=1}^{n} z_{i,j}\right),$$

where e_j can be factored out from the internal summation, as it does not depend on i.
By our initial assumption (i.e., each input data point is standardized, therefore centered around the mean), we know that $\frac{1}{n}\sum_{i=1}^{n} z_{i,j} = 0$, for all $j \in \{1, \ldots, d\}$, and therefore $\mu = 0$.
Putting all together, the variance as measured when data points are projected onto $\boldsymbol{e}$ is:

$$V(\boldsymbol{e}) = \frac{1}{n-1}\sum_{i=1}^{n}\Big(\sum_{j=1}^{d} z_{i,j} * e_j\Big)^2.$$

Eventually, we want to find that vector $\boldsymbol{e}^*$, such that it *maximizes* the above quantity. In other words:

$$\boldsymbol{e}^* = \text{argmax}_{\boldsymbol{e}}\{V(\boldsymbol{e})\} = \text{argmax}_{\boldsymbol{e}}\left\{\frac{1}{n-1}\sum_{i=1}^{n}\Big(\sum_{j=1}^{d} z_{i,j} * e_j\Big)^2\right\}.$$

Now, if we had not put the constraint on the length of $\boldsymbol{e}$ – such that $||\boldsymbol{e}|| = 1$ – the optimization problem above would not be solvable, as we can always be able to find a vector whose magnitude increases the variance. Therefore, the actual problem we want to solve is a *constrained* optimization problem defined as follows:

$$\boldsymbol{e}^* = \text{argmax}_{\boldsymbol{e}}\left\{\frac{1}{n-1}\sum_{i=1}^{n}\Big(\sum_{j=1}^{d} z_{i,j} * e_j\Big)^2\right\}$$
$$\text{s.t. } ||\boldsymbol{e}|| = 1 \Rightarrow ||\boldsymbol{e}|| - 1 = 0.$$

Whenever we aim to maximize (or minimize) a function that is subject to an equality constraint like the one defined above, the method of Lagrange multipliers is used. As such, the optimization problem turns into solving the following objective:

$$\boldsymbol{e}^* = \text{argmax}_{\boldsymbol{e}}\left\{\frac{1}{n-1}\sum_{i=1}^{n}\Big(\sum_{j=1}^{d} z_{i,j} * e_j\Big)^2 - \lambda\Big[\underbrace{\Big(\sum_{j=1}^{d} e_j^2\Big) - 1}_{\text{constraint: } ||\boldsymbol{e}||-1=0}\Big]\right\}.$$

To solve the optimization problem above, we have to take the gradient of the function, set it to 0, and solve it for $\boldsymbol{e}$:

$$\nabla V(\boldsymbol{e}) = \nabla \left\{ \frac{1}{n-1} \sum_{i=1}^{n} \Big(\sum_{j=1}^{d} z_{i,j} * e_j \Big)^2 - \lambda \Big[\Big(\sum_{j=1}^{d} e_j^2 \Big) - 1 \Big] \right\}.$$

The gradient is just the d-dimensional vector of all the partial derivatives of $V(\boldsymbol{e})$ w.r.t. each dimension of $\boldsymbol{e}$, i.e., $e_1, \ldots, e_d$:

$$\nabla V(\boldsymbol{e}) = \left(\frac{\partial V(\boldsymbol{e})}{\partial e_1}, \ldots, \frac{\partial V(\boldsymbol{e})}{\partial e_d} \right).$$

The generic k-th component of the gradient is computed as follows:

$$\frac{\partial V(\boldsymbol{e})}{\partial e_k} = \frac{2}{n-1} \sum_{i=1}^{n} \Big(\sum_{j=1}^{d} z_{i,j} * e_j \Big) z_{i,k} - 2\lambda e_k.$$

In order to find $\boldsymbol{e}^*$, we have to set $\nabla V(\boldsymbol{e}) = \mathbf{0}$, which is equal to set *all* of its components $\frac{\partial V(\boldsymbol{e})}{\partial e_k} = 0$ simultaneously, and solve it for $\boldsymbol{e}$.
Let us work out how to solve the equation below for the k-th component and generalize it to all the components:

$$\frac{2}{n-1} \sum_{i=1}^{n} \Big(\sum_{j=1}^{d} z_{i,j} * e_j \Big) z_{i,k} - 2\lambda e_k = 0 \Rightarrow \frac{2}{n-1} \sum_{i=1}^{n} \Big(\sum_{j=1}^{d} z_{i,j} * e_j \Big) z_{i,k} = 2\lambda e_k$$

We can again factor out e_j from the summation, as it does not depend on i:

$$2 \sum_{j=1}^{d} e_j \underbrace{\frac{1}{n-1} \sum_{i=1}^{n} (z_{i,j} * z_{i,k})}_{\mathrm{Cov}(Z_j, Z_k) = \mathrm{Cov}(Z_k, Z_j)} = 2\lambda e_k.$$

$$=$$

$$\sum_{j=1}^{d} e_j \underbrace{\frac{1}{n-1} \sum_{i=1}^{n} (z_{i,j} * z_{i,j})}_{\mathrm{Cov}(Z_j, Z_k) = \mathrm{Cov}(Z_k, Z_j)} = \lambda e_k.$$

The equation above must hold contemporarily for all k, i.e., $k = 1 \ldots d$. Therefore, we have to solve the following system of equations:

$$\begin{cases} \sum_{j=1}^{d} \mathrm{Cov}(Z_1, Z_j) e_j = & \lambda e_1 \\ \sum_{j=1}^{d} \mathrm{Cov}(Z_2, Z_j) e_j = & \lambda e_2 \\ \qquad \cdots & \cdots \\ \sum_{j=1}^{d} \mathrm{Cov}(Z_d, Z_j) e_j = & \lambda e_d \end{cases}$$

Each equation above is the dot product of a row of the covariance matrix K with the vector $\boldsymbol{e}$.

$$\begin{cases} \boldsymbol{z}_1^T \cdot \boldsymbol{e} = & \lambda e_1 \\ \boldsymbol{z}_2^T \cdot \boldsymbol{e} = & \lambda e_2 \\ \quad \ldots & \ldots \\ \boldsymbol{z}_d^T \cdot \boldsymbol{e} = & \lambda e_d \end{cases}$$

In other words, we can rewrite the system of equations above in full-matrix form and find a solution to:

$$K\boldsymbol{e} = \lambda \boldsymbol{e}.$$

As we have already seen in the previous section, $\boldsymbol{e}$ is a non-trivial solution to the homogeneous system above if it is an eigenvector, and this proves what we wanted.
To summarize:

1. We started from computing the variance $V(\boldsymbol{e})$ of our input data points w.r.t. a generic projection vector $\boldsymbol{e}$;

2. We tried to maximize $V(\boldsymbol{e})$ by computing its gradient (constrained with the proper Lagrange multiplier) and setting it to 0;

3. Eventually, we found that the (non-trivial) solution to this problem corresponds exactly to $\boldsymbol{e}$ being an eigenvector of the covariance matrix.

3 Largest Eigenvalue: Principal Component

In the section above, we have formally proved that eigenvectors point toward the direction which maximizes the variance of data. In this section, we aim to prove why the principal component corresponds to the eigenvector with the largest eigenvalue[1].
Let us go back to our initial definition of variance along the (eigen)vector $\boldsymbol{e}$, which is defined as follows:

$$V(\boldsymbol{e}) = \frac{1}{n-1} \sum_{i=1}^{n} \Big(\sum_{j=1}^{d} z_{i,j} * e_j \Big)^2.$$

We rewrite the equation above by unrolling the sum of squares as follows:

$$V(\boldsymbol{e}) = \frac{1}{n-1} \sum_{i=1}^{n} \Big(\sum_{j=1}^{d} z_{i,j} * e_j \Big) \Big(\sum_{k=1}^{d} z_{i,k} * e_k \Big).$$

The order of summations can be (carefully) moved outside:

$$V(\boldsymbol{e}) = \sum_{k=1}^{d} \sum_{j=1}^{d} \left(\frac{1}{n-1} \sum_{i=1}^{n} z_{i,j} * z_{i,j} \right) e_j e_k.$$

[1] Remember that, in general, there are d distinct eigenvalues for a $d \times d$ covariance matrix.

Again, the most internal summation is just the covariance between random variables Z_j and Z_k (associated with the j-th and k-th dimension, respectively):

$$V(\boldsymbol{e}) = \sum_{k=1}^{d}\sum_{j=1}^{d}\left(\underbrace{\frac{1}{n-1}\sum_{i=1}^{n} z_{i,j} * z_{i,k}}_{\mathrm{Cov}(Z_j,Z_k)=\mathrm{Cov}(Z_k,Z_j)}\right)e_j e_k$$

Therefore:

$$V(\boldsymbol{e}) = \sum_{k=1}^{d}\left(\sum_{j=1}^{d}\mathrm{Cov}(Z_k, Z_j)e_j\right)e_k.$$

Again, the most internal summation corresponds to the dot product between the k-th row of the covariance matrix K and the vector $\boldsymbol{e}$. Since we now know that $\boldsymbol{e}$ is an eigenvector, by definition of it, we also know that for all $k = 1, \dots, d$, it must hold the following:

$$\sum_{j=1}^{d}\mathrm{Cov}(Z_k, Z_j)e_j = \lambda e_k.$$

Therefore:

$$V(\boldsymbol{e}) = \sum_{k=1}^{d}\Big(\lambda e_k\Big)e_k = \lambda\sum_{k=1}^{d} e_k^2 = \lambda||\boldsymbol{e}||.$$

Since $||\boldsymbol{e}|| = 1$, we obtain:

$$V(\boldsymbol{e}) = \lambda.$$

In other words, the variance along the unit-length eigenvector $\boldsymbol{e}$ is exactly equal to its corresponding eigenvalue λ.

As a consequence of this result, to find the $k \ll d$ principal components, we will just need to find up to d eigenvalues $\lambda_1, \dots, \lambda_d$ and their associated eigenvectors $\boldsymbol{e}_1, \dots, \boldsymbol{e}_d$. Then, we sort eigenvectors from the largest to the smallest eigenvalue, and we pick the first k of them. Indeed, the eigenvector associated with the largest eigenvalue ($\lambda_{\max}$) will correspond to the direction with the highest variance, the eigenvector associated with the second largest eigenvalue will correspond to the direction with the second highest variance, and so on and so forth.

Notice that the eigenvalues of the (standardized) covariance matrix *cannot* be negative. We can, of course, derive that from the fact that each eigenvalue λ_i carries a fraction of the total variance of the data when projected to the corresponding eigenvector $\boldsymbol{e}_i$. As the variance is non-negative, so does each λ_i. However, another way to prove this is by means of the result we have demonstrated with Lemma 1, namely that the covariance matrix is positive-semidefinite.

Theorem 1. *Any positive-semidefinite $d \times d$ real matrix A has non-negative eigenvalues.*

Proof. If A is a positive-semidefinite then the following must hold true:

$$\boldsymbol{v}^T A\boldsymbol{v} \geq 0 \; \forall \boldsymbol{v} \in \mathbb{R}^d.$$

Moreover, by definition of eigenvectors/eigenvalues:

$$A\boldsymbol{v} = \lambda\boldsymbol{v}.$$

If we multiply both sides of the equation above by $\boldsymbol{v}^T$, we obtain:

$$\boldsymbol{v}^T A\boldsymbol{v} = \boldsymbol{v}^T \lambda\boldsymbol{v} = \lambda\boldsymbol{v}^T\boldsymbol{v}.$$

Since $\boldsymbol{v}^T\boldsymbol{v} \geq 0$, then it must also be $\lambda \geq 0$ in order to satisfy $\boldsymbol{v}^T A\boldsymbol{v} \geq 0$.