Principal Component Analysis and Application to k-Means

Aris Anagnostopoulos, Chris Schwiegelshohn

This is a set of notes that are based on some initial notes by Chris Schwiegelshohn and I have also used ideas from various books, especially by the book of Strang [2], and by Blum et al. [1]. I recommend you the former for gaining intuition on linear algebra, and the latter if you want to delve into various theoretical topics in data mining and data science.

The problem with the tecniques based on linear algebra is that because we live in a three-dimensional world, we generally do not have intuition about what happens in higher dimensions, which is where our data live as we typically represent them. Whereas some parts of our intuition in three dimensions carry to higher dimensions, many other parts fail. Thus we need to use math to understand what is gong on, and to gain new intuition about such spaces.

My advice when you study these notes, is to not just read the math, but to try to understand what each expression means, geometrically. For example[1] that, if the columns of $\mathbf{W}$ form an orthonormal basis of a k-dimensional subspace of $\mathbb{R}^d$, then

$$\left\|\mathbf{A}\mathbf{W_k}\mathbf{W_k}^T\right\|_F^2 = \sum_{i=1}^{n} \left\|\mathbf{A_{(i)}}\mathbf{W_k}\mathbf{W_k}^T\right\|_2^2$$

is the sum of the squares of the lengths of the projections of each row of $\mathbf{A}$ on this subspace, and that

$$\|\mathbf{A}\mathbf{W_k}\|_F^2 = \sum_{i=1}^{n} \left\|\mathbf{A_{(i)}}\mathbf{W_k}\right\|_2^2$$

is the sum over each row of $\mathbf{A}$ of the square sum of the squares of the lengths of the row's projection on the columns of $\mathbf{W}$. This, for example, means that by the Pythagorean theorem the two quantities are equal, something, that by just comparing the two formulae is not obvious.

1 Introduction

Assume that we have n d-dimensional data points. One way to represent them is using a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$, where each line $\mathbf{A_{(i)}} \in \mathbb{R}^d$ represents the ith data point (see Figure 1):

$$A = \begin{bmatrix} \text{———}\mathbf{A_{(1)}}\text{———} \\ \text{———}\mathbf{A_{(2)}}\text{———} \\ \vdots \\ \text{———}\mathbf{A_{(n)}}\text{———} \end{bmatrix}.$$

In some books you may see columns corresponding to points and rows to dimensions, then everything holds but considering $\mathbf{A}^T$ instead of $\mathbf{A}$.

[1] Of course this example does not make sense the first time that you read it but hopefully you will understand it after you study these notes.

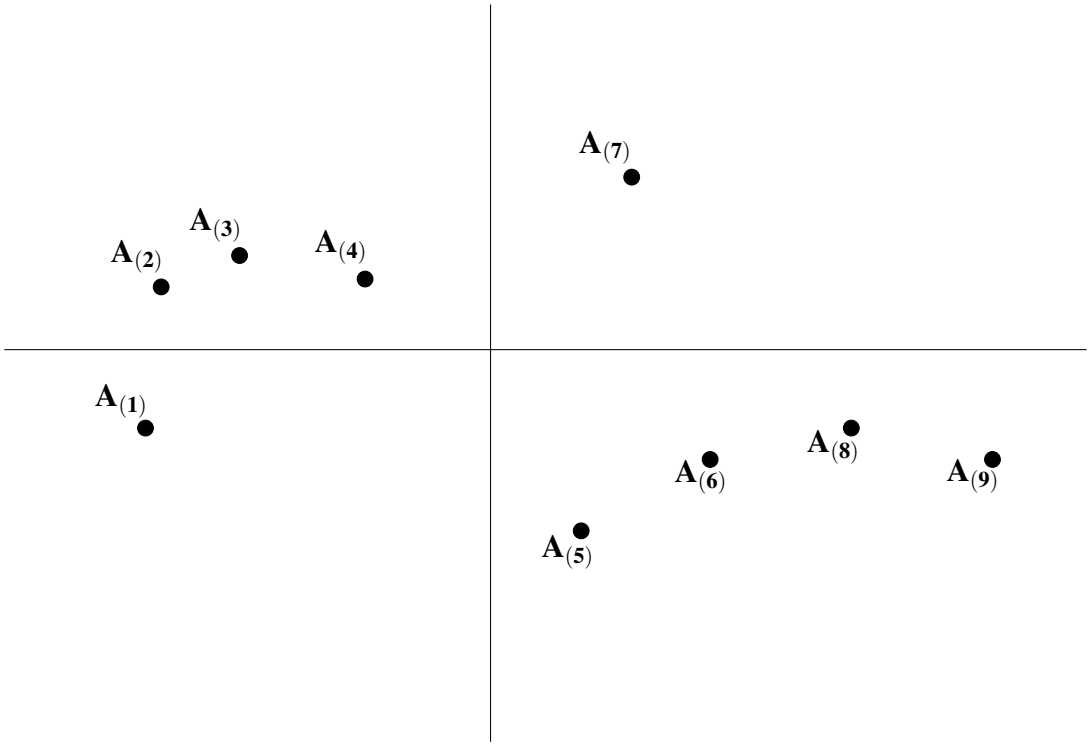


Figure 1: Representation of each point as a row of matrix $\mathbf{A}$.

This viewpoint of our dataset as a matrix, allows us to use tools from linear algebra to study our data.

In these notes we will see the prinicpal-component analysis (PCA), which is essentially the application of the singular-value decomposition to data analysis.

We start with some linear-algebra background.

2 Background in Linear Algebra

In this section we will present some basic notions from linear algebra, which will allow us to understand more easily the material. We start by presenting some notation.

2.1 Definitions and Notation

For a given matrix $\mathbf{A}$, we use $\mathbf{A}_{ij}$ do denote the individual element. We use $\mathbf{A}_{(\mathbf{i})}$ to denote the ith row of $\mathbf{A}$. We use $\mathbf{A}^{(\mathbf{j})}$ to denote the jth column of $\mathbf{A}$.

If $x, y \in \mathbb{R}^d$ we denote their dot product by $x^T y$.

Two vectors $\mathbf{v}_i$ and $\mathbf{v}_j$ are *orthogonal* to each other if $\mathbf{v}_i^T \mathbf{v}_j = 0$. A collection of k vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ is *orthonormal* if each vector has unit ℓ_2 norm ($\|\mathbf{v}_i\|_2 = 1$) and if each vector $\mathbf{v}_i$ is orthogonal to every other vector $\mathbf{v}_j$ in the collection.

A matrix $\mathbf{V} \in \mathbb{R}^{n \times d}$, with $n \geq d$, is *semi-orthogonal* if the set of its columns is orthonormal. Then we have that $\mathbf{V}^T \mathbf{V} = \mathbf{I} (= \mathbf{I}_d)$ (but $\mathbf{V}\mathbf{V}^T \neq \mathbf{I}_n$, unless $n = d$).

2.2 Matrix Multiplication

Consider the matrices

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 & -5 \\ 4 & 0 & 2 & 0 \\ 2 & -1 & 1 & 3 \end{bmatrix}$$

and

$$\mathbf{B} = \begin{bmatrix} 4 & 0 \\ 1 & 3 \\ 7 & -1 \\ 3 & 0 \end{bmatrix}$$

Then the usual way to compute the product $\mathbf{AB}$ is by setting the value of element $(\mathbf{AB})_{ij}$ to be the dot product of the ith row of $\mathbf{A}$ with the jth column of $\mathbf{B}$:

$$(AB)_{ij} = \mathbf{A_{(i)}} \cdot \mathbf{B^{(j)}}.$$

Thus we get

$$\mathbf{AB} = \begin{bmatrix} 1 & 2 & 3 & -5 \\ 4 & 0 & 2 & 0 \\ 2 & -1 & 1 & 3 \end{bmatrix} \cdot \begin{bmatrix} 4 & 0 \\ 1 & 3 \\ 7 & -1 \\ 3 & 0 \end{bmatrix} = \begin{bmatrix} 12 & 3 \\ 30 & -2 \\ 23 & -4 \end{bmatrix}$$

However, notice that we can express this multiplication as a summation of 4 rank-1 matrices, with the ith term being the product of the ith column of $\mathbf{A}$ with the ith row of $\mathbf{B}$:

$$\begin{aligned} \mathbf{AB} &= \begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix} \cdot \begin{bmatrix} 4 & 0 \end{bmatrix} + \begin{bmatrix} 2 \\ 0 \\ -1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 3 \end{bmatrix} + \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 7 & -1 \end{bmatrix} + \begin{bmatrix} -5 \\ 0 \\ 3 \end{bmatrix} \cdot \begin{bmatrix} 3 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 4 & 0 \\ 16 & 0 \\ 8 & 0 \end{bmatrix} + \begin{bmatrix} 2 & 6 \\ 0 & 0 \\ -1 & -3 \end{bmatrix} + \begin{bmatrix} 21 & -3 \\ 14 & -2 \\ 7 & -1 \end{bmatrix} + \begin{bmatrix} -15 & 0 \\ 0 & 0 \\ 9 & 0 \end{bmatrix} = \begin{bmatrix} 12 & 3 \\ 30 & -2 \\ 23 & -4 \end{bmatrix} \end{aligned}$$

More generally:

$$\begin{aligned} \mathbf{AB} &= \begin{bmatrix} | & | & & | \\ \mathbf{A^{(1)}} & \mathbf{A^{(2)}} & \cdots & \mathbf{A^{(d)}} \\ | & | & & | \end{bmatrix} \cdot \begin{bmatrix} \text{———}\mathbf{B_{(1)}}\text{———} \\ \text{———}\mathbf{B_{(2)}}\text{———} \\ \vdots \\ \text{———}\mathbf{B_{(d)}}\text{———} \end{bmatrix} \\ &= \begin{bmatrix} | \\ \mathbf{A^{(1)}} \\ | \end{bmatrix} \cdot \begin{bmatrix} \text{———}\mathbf{B_{(1)}}\text{———} \end{bmatrix} + \begin{bmatrix} | \\ \mathbf{A^{(2)}} \\ | \end{bmatrix} \cdot \begin{bmatrix} \text{———}\mathbf{B_{(2)}}\text{———} \end{bmatrix} + \cdots + \begin{bmatrix} | \\ \mathbf{A^{(d)}} \\ | \end{bmatrix} \cdot \begin{bmatrix} \text{———}\mathbf{B_{(d)}}\text{———} \end{bmatrix}. \end{aligned}$$

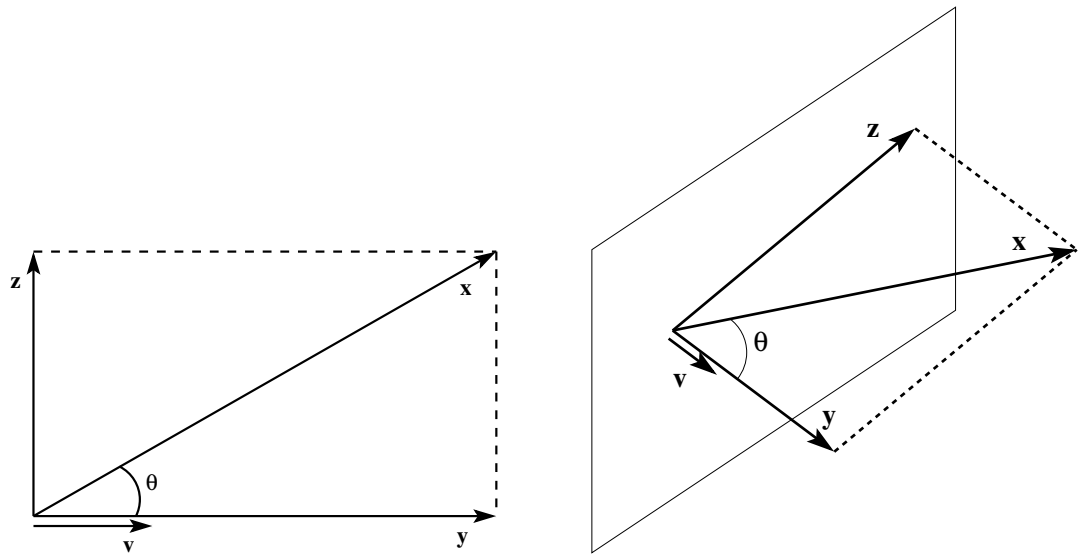
Mathematically, if $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $\mathbf{B} \in \mathbb{R}^{d \times \ell}$, then

$$\mathbf{AB} = \mathbf{A^{(1)}} \cdot \mathbf{B_{(1)}} + \mathbf{A^{(2)}} \cdot \mathbf{B_{(2)}} + \cdots + \mathbf{A^{(d)}} \cdot \mathbf{B_{(d)}} = \sum_{r=1}^{d} \mathbf{A^{(r)}} \cdot \mathbf{B_{(r)}}.$$

This is because in both ways of doing the multiplication, we obtain

$$(\mathbf{AB})_{ij} = \sum_{r=1}^{d} \mathbf{A}_{ir} \mathbf{B}_{rj}.$$

It turns out that this view of matrix multiplication is often very useful to understand what is going on when we work with matrices.



(a) $\mathbf{x} \in \mathbb{R}^2$, so the subspace orthogonal to $\mathbf{v}$ has dimension $2 - 1 = 1$.

(b) $\mathbf{x} \in \mathbb{R}^3$, so the subspace orthogonal to $\mathbf{v}$ has dimension $3 - 1 = 2$.

Figure 2: Projection of vector $\mathbf{x}$ on the direction of vector $\mathbf{v}$. Vector $\mathbf{x} = \mathbf{y}+\mathbf{z}$ can be decomposed to two parts: $\mathbf{y}$, the projection on $\mathbf{v}$, and $\mathbf{z}$, the projection on the subspace orthogonal to $\mathbf{v}$.

2.3 Projection onto a Vector

PCA is about projection of points and matrices on subspaces, so we start with the projection of a vector onto another vector. See Figure 2. Consider some vector $\mathbf{x} \in \mathbb{R}^d$ and some unit vector $\mathbf{v} \in \mathbb{R}^d$. Let $\mathbf{y}$ be the projection of $\mathbf{x}$ on $\mathbf{v}$. Then, if the angle between $\mathbf{x}$ and $\mathbf{v}$ is equal to θ, the length of the projection $\mathbf{y}$ equals just the dot product of $\mathbf{x}$ and $\mathbf{v}$:

$$\|\mathbf{y}\|_2 = \|\mathbf{x}\|_2 \cos\theta = \|\mathbf{x}\|_2 \frac{\mathbf{v}^T\mathbf{x}}{\|\mathbf{v}\|_2 \|\mathbf{x}\|_2} = \mathbf{v}^T\mathbf{x}.$$

Then the projection $\mathbf{y}$ equals the unit vector $\mathbf{v}$ times this length:

$$\mathbf{y} = \mathbf{v}\mathbf{v}^T\mathbf{x}.$$

Thus to project the point $\mathbf{x}$ on $\mathbf{v}$ it is enough to left-multiply $\mathbf{x}$ with $\mathbf{v}\mathbf{v}^T$; recall that for matrix multiplication the associative property is true: $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$.

Given that $\mathbf{A}$ contains our points as row vectors, let us see how we can project a point represented as a row vector. We consider the row vector $\mathbf{x}^T \in \mathbb{R}^{1\times d}$. Then, using the fact that $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$, we obtain the row vector:

$$\mathbf{y}^T = \mathbf{x}^T\mathbf{v}\mathbf{v}^T.$$

Therefore, the projection of the data point $\mathbf{A}_{(\mathbf{i})}$ on $\mathbf{v}$ is simply $\mathbf{A}_{(\mathbf{i})}\mathbf{v}\mathbf{v}^T$. So, if we want to project the entire dataset $\mathbf{A}$ on the direction $\mathbf{v}$, we can multiply to the right with $\mathbf{v}\mathbf{v}^T$:

$$\mathbf{A}\mathbf{v}\mathbf{v}^T.$$

The ith row of the resulting matrix is the projection of $\mathbf{A}_{(\mathbf{i})}$ on $\mathbf{v}$.

The projection, $\mathbf{z}$ (see Figure 2), of $\mathbf{x}$ on the $(d-1)$-dimensional subspace that is orthogonal to $\mathbf{v}$ is

$$\mathbf{z} = \mathbf{x} - \mathbf{v}\mathbf{v}^T\mathbf{x} = (\mathbf{I} - \mathbf{v}\mathbf{v}^T)\mathbf{x}.$$

Similarly, given our dataset $\mathbf{A}$, we can decompose it into two parts, the projection of each point on the direction of $\mathbf{v}$: $\mathbf{A}\mathbf{v}\mathbf{v}^T$, and to the subspace orthogonal to $\mathbf{v}$: $\mathbf{A}(\mathbf{I} - \mathbf{v}\mathbf{v}^T)$.

2.4 Projection on a Subspace

In the previous section we projected $\mathbf{x}$ on a single vector. We will now generalize by projecting to a higher-dimensional subspace. Consider the subspace defined (*spanned* as we say) by the orthonormal set of k vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$, where each $\mathbf{v}_i \in \mathbb{R}^d$. Then the projection $\mathbf{y}$ of $\mathbf{x}$ on the subspace spanded by the k vectors equals to the sum of the projections on each vector:

$$\mathbf{y} = \mathbf{v}_1\mathbf{v}_1^T\mathbf{x} + \cdots + \mathbf{v}_k\mathbf{v}_k^T\mathbf{x} = (\mathbf{v}_1\mathbf{v}_1^T + \cdots + \mathbf{v}_k\mathbf{v}_k^T)\mathbf{x}.$$

Define the $d \times k$ matrix $\mathbf{V_k}$:

$$\mathbf{V_k} = \begin{bmatrix} | & | & & | \\ \mathbf{v_1} & \mathbf{v_2} & \cdots & \mathbf{v_k} \\ | & | & & | \end{bmatrix}.$$

Then, from Section 2.2, we obtain that the projection of $\mathbf{x}$ on the subpsace of $\mathbb{R}^d$ defined by the k columns of the semi-orthogonal matrix $\mathbf{V_k}$ equals

$$\mathbf{y} = \mathbf{V_k}{\mathbf{V_k}}^T\mathbf{x}.$$

Of course this subspace has dimension k, equal to the rank of $\mathbf{V_k}$.

The projection $\mathbf{z}$ of $\mathbf{x}$ to the orthogonal subspace (which has dimension $d - k$) is

$$\mathbf{z} = \mathbf{x} - \mathbf{V_k}{\mathbf{V_k}}^T\mathbf{x} = (\mathbf{I} - \mathbf{V_k}{\mathbf{V_k}}^T)\mathbf{x}.$$

As with the case of Section 2.3, the projection of our $n \times d$ matrix $\mathbf{A}$ on the subspace defined by $\mathbf{V_k}$ equals to

$$\mathbf{A}\mathbf{V_k}{\mathbf{V_k}}^T,$$

and the projection to the orthogonal subspace equals to

$$\mathbf{A} - \mathbf{A}\mathbf{V_k}{\mathbf{V_k}}^T = \mathbf{A}(\mathbf{I} - \mathbf{V_k}{\mathbf{V_k}}^T).$$

Note that for any given subspace of dimension $k > 1$, there are infinite possible orthonormal bases: one can take an orthonormal basis $(\mathbf{v}_1, \ldots, \mathbf{v}_k)$ of the subspace and rotate it on the subspace to obtain another orthonormal basis $(\mathbf{v}'_1, \ldots, \mathbf{v}'_k)$ for the same subspace—think of the rotation of an orthonormal basis on the plane. Then, the projection of the dataset $\mathbf{A}$ on the subspace is the same no matter what basis we use, so for the corresponding matrix $\mathbf{V'_k}$ we have:

$$\mathbf{A}\mathbf{V_k}{\mathbf{V_k}}^T = \mathbf{A}\mathbf{V'_k}{\mathbf{V'_k}}^T$$

and

$$\mathbf{A}(\mathbf{I} - \mathbf{V}\mathbf{V}^T) = \mathbf{A}(\mathbf{I} - \mathbf{V}'\mathbf{V}'^T).$$

3 Variance

PCA is based on the variance of the dataset. First we revisit the familiar case where are data are points in $\mathbb{R}$, and then we generalize to higher dimensions.

3.1 One Dimension

Consider a set of n samples $\mathbf{x} = (x_1, \dots x_n)^T$, with $x_i \in \mathbb{R}$. The *empirical mean* of these samples is defined as

$$\frac{1}{n}\sum_{i=1} x_i.$$

We also define the expected second moment as

$$\frac{1}{n}\sum_{i=1} x_i^2,$$

and then the *empirical variance* is defined as

$$\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{j=1}^{n} x_j\right)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2,$$

recalling the formula for the variance: $\mathbf{E}\big[(X - \mathbf{E}[X])^2\big] = \mathbf{E}\big[X^2\big] - \mathbf{E}[X]^2$.

As a matter of fact, we often center the data such that the mean is 0 and then the variance reduces to $\frac{1}{n}\sum_{i=1}^{n} x_i^2$. In this case, we may interpret the variance as the (scaled) squared Euclidean norm of the vector containing the samples. Note that, by definition, the variance of the values x_i is the second moment of them after being centered. In other words, centering the data does not changes their variance. In general, for a vector $\mathbf{x} \in \mathbb{R}^n$, with $\mathbf{x} = (x_1, \dots, x_n)^T$, $\|\mathbf{x}\|_p = \sqrt[p]{\sum_{i=1}^{n} |x_i|^p}$, hence for normalized data sets, the variance equals the scaled 2-norm of $\mathbf{x}$: $\frac{1}{n}\|\mathbf{x}\|_2^2$. We also note that $\mathbf{x}^T\mathbf{x} = \|\mathbf{x}\|_2^2$ for any vector $\mathbf{x}$.

3.2 Multiple Dimensions

Let us now consider the notions of the previous section in higher dimensions, that is, the samples of $\mathbf{A_{(i)}}$ are no longer numbers, but vectors in $\mathbb{R}^d$. The empirical mean translates straightforwardly and is also commonly known as the centroid:

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{A_{(i)}}.$$

The notion of variance is not as easy to generalize. Ideally, we would like to retain the notion that the variance quantifies the spread of the data set with respect to the mean (or centroid). The difficulty of extending this notion is that the spread is different along different directions. This is properly captured by the *covariance matrix*. Here instead, our notion of generalization will be simpler, as we are looking for a single number, rather than the more complex spectral structure included in the covariance matrix. Instead, we define the *directional variance* along an arbitrary unit vector $\mathbf{v}$ as

$$\mathbf{Var_v}[\mathbf{A}] \triangleq \frac{1}{n}\sum_{i=1}^{n}\left(\left(\mathbf{A_{(i)}} - \frac{1}{n}\sum_{j=1}^{n}\mathbf{A_{(j)}}\right)\cdot\mathbf{v}\right)^2.$$

To understand the definition, note that, as we saw in Section 2.3, the length of the projection of each vector on the direction of $\mathbf{v}$, is given by the dot product $\mathbf{A_{(i)}}\cdot\mathbf{v}$—note that $\mathbf{A_{(i)}}$ is a row vector, so we don't take the transpose. Similarly, the length of the projection on $\mathbf{v}$ of the centroid of all the points $\mathbf{A_{(i)}}$ is $\left(\frac{1}{n}\sum_{j=1}^{n}\mathbf{A_{(j)}}\right)\cdot\mathbf{v}$. Therefore, the variance of the length of

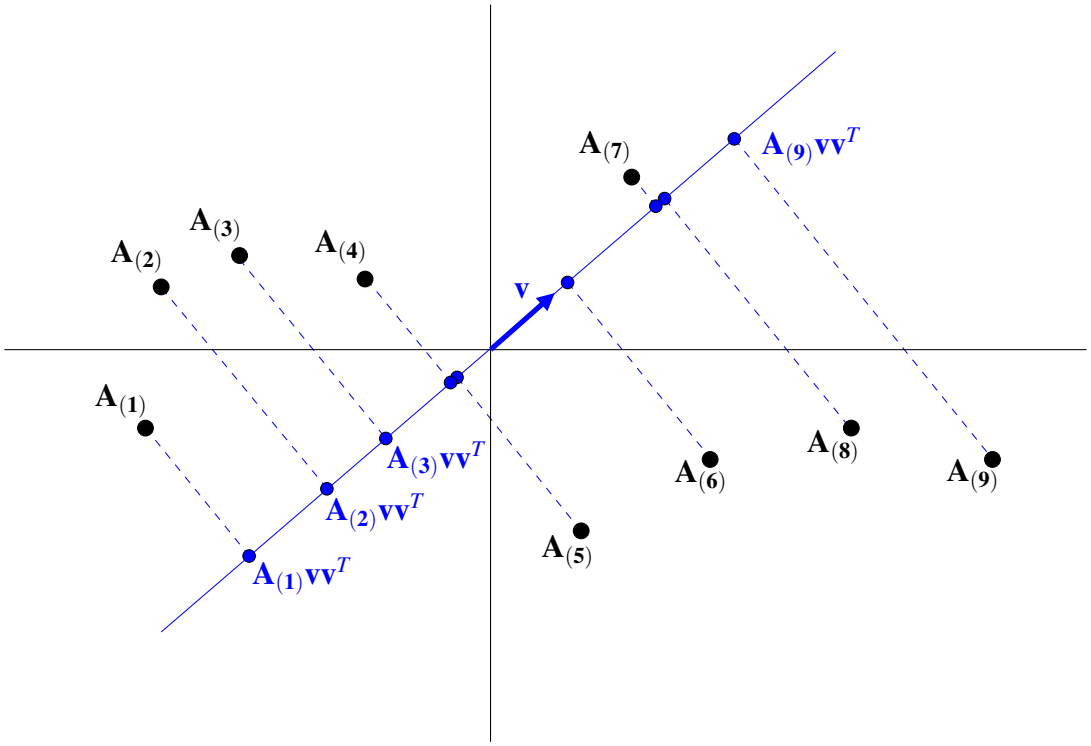


Figure 3: Projection of the dataset $\mathbf{A}$ onto the direction of vector $\mathbf{v}$. The projection of point $\mathbf{A_{(i)}}$ is $\mathbf{A_{(i)}vv}^T$ and the variance $\mathbf{Var_v}[\mathbf{A}]$ is the variance of the lengths of the projections, that is, the distances between the origin and the blue points.

the projections of the n points $\mathbf{A_{(i)}}$ on the direction of $\mathbf{v}$ is given by the above definition. See Figure 3.

Again, for centered inputs with $\sum_{j=1}^{n} \mathbf{A_{(j)}} = \mathbf{0}$, this reduces to

$$\mathbf{Var_v}[\mathbf{A}] = \frac{1}{n}\sum_{i=1}^{n} \left(\mathbf{A_{(i)}} \cdot \mathbf{v}\right)^2 = \frac{1}{n} \left\|\mathbf{Av}\right\|_2^2 .$$

Similarly to the one-dimensional case, centering the data does not changes their variance. Geometrically, this expression means that we project all points along the direction $\mathbf{v}$ and compute the variance of a (now) 1-dimensional set of samples. To capture the entire variance of the point set, we pick an arbitrary orthogonal basis $V = \{\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_d}\}$ of $\mathbb{R}^d$ and compute

$$\mathbf{Var}[\mathbf{A}] \triangleq \sum_{j=1}^{d} \mathbf{Var_{v_j}}[\mathbf{A}] = \frac{1}{n}\sum_{j=1}^{d} \left\|\mathbf{Av_j}\right\|_2^2 = \frac{1}{n}\sum_{j=1}^{d}\sum_{i=1}^{n} \left(\mathbf{A_{(i)}} \cdot \mathbf{v_j}\right)^2 .$$

The fact that in the above definition of $\mathbf{Var}[\mathbf{A}]$ there is no indication of the basis, implies that the variance does not depend on the basis chosen. We prove thin in the next lemma, which we prove for the general case where the points are not centered.

Lemma 1. *Consider two orthonormal bases of* $\mathbb{R}^d$*:* $V = \{\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_d}\}$ *and* $W = \{\mathbf{w_1}, \mathbf{w_2}, \ldots, \mathbf{w_d}\}$*. Then we have that*

$$\sum_{j=1}^{d} \mathbf{Var_{v_j}}\left[\mathbf{A_{(i)}}\right] = \sum_{j=1}^{d} \mathbf{Var_{w_j}}\left[\mathbf{A_{(i)}}\right] .$$

Proof. Because W is a basis for $\mathbb{R}^d$, each vector $\mathbf{v_k}$ can be expressed as a linear combination $\mathbf{v_k} = \sum_{j=1}^{d} \alpha_{k,j} \cdot \mathbf{w_j}$ and any vector $\mathbf{w_j}$ can be expressed as a linear combination $\mathbf{w_j} = \sum_{k=1}^{d} \beta_{j,k} \cdot \mathbf{v_k}$. Notice that, because $\mathbf{w_r}^T\mathbf{w_r} = 1$ and $\mathbf{w_r}^T\mathbf{w_j} = 0$ for $r \neq j$, we have that:

$$\mathbf{v_k}^T\mathbf{w_j} = \sum_{r=1}^{d} \alpha_{k,r} \cdot \mathbf{w_r}^T\mathbf{w_j} = \alpha_{k,j}$$

and, similarly, that

$$\mathbf{w_j}^T\mathbf{v_k} = \sum_{r=1}^{d} \beta_{j,r} \cdot \mathbf{v_r}^T\mathbf{v_k} = \beta_{j,k}.$$

Therefore, $\beta_{j,k} = \alpha_{k,j}$, so

$$\mathbf{w_j} = \sum_{k=1}^{d} \alpha_{k,j} \cdot \mathbf{v_k}.$$

Notice also that

$$\begin{aligned} 1 = \mathbf{w_j}^T\mathbf{w_j} &= \left(\sum_{k=1}^{d} \alpha_{k,j} \cdot \mathbf{v_k}\right)^T \cdot \left(\sum_{k=1}^{d} \alpha_{k,j} \cdot \mathbf{v_k}\right) \\ &= \sum_{k=1}^{d} \alpha_{k,j}^2 \mathbf{v_k}^T\mathbf{v_k} + \sum_{k=1}^{d} \sum_{\substack{r=1 \\ r\neq k}}^{d} \alpha_{k,r} \mathbf{v_k}^T\mathbf{v_r} = \sum_{k=1}^{d} \alpha_{k,j}^2, \end{aligned} \tag{1}$$

and for $j \neq r$

$$\begin{aligned} 0 = \mathbf{w_j}^T\mathbf{w_r} &= \left(\sum_{k=1}^{d} \alpha_{k,j} \cdot \mathbf{v_k}\right)^T \cdot \left(\sum_{k=1}^{d} \alpha_{k,r} \cdot \mathbf{v_k}\right) \\ &= \sum_{k=1}^{d} \alpha_{k,j}\alpha_{k,r} \mathbf{v_k}^T\mathbf{v_k} + \sum_{k=1}^{d} \sum_{\substack{\ell=1 \\ \ell\neq k}}^{d} \alpha_{k,j}\alpha_{\ell,r} \mathbf{v_k}^T\mathbf{v_\ell} = \sum_{k=1}^{d} \alpha_{k,j}\alpha_{k,r}. \end{aligned} \tag{2}$$

Then, let as define $\mathbf{x_i}^T = \mathbf{A_{(i)}} - \frac{1}{n}\sum_{j=1}^{n} \mathbf{A_{(j)}}$ and we have:

$$\begin{aligned} \mathbf{Var_{v_k}}\left[\mathbf{A_{(i)}}\right] &= \frac{1}{n}\sum_{i=1}^{n} \left(\left(\mathbf{A_{(i)}} - \frac{1}{n}\sum_{j=1}^{n} \mathbf{A_{(j)}}\right)\mathbf{v_k}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} \left(\mathbf{x_i}^T \cdot \mathbf{v_k}\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} \left(\mathbf{x_i}^T \sum_{j=1}^{d} \alpha_{k,j} \cdot \mathbf{w_j}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} \left(\sum_{j=1}^{d} \mathbf{x_i}^T \cdot \mathbf{w_j} \cdot \alpha_{k,j}\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d} \left(\mathbf{x_i}^T \cdot \mathbf{w_j}\right)^2 \alpha_{k,j}^2 + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{\substack{r=1 \\ r\neq j}}^{d} \mathbf{x_i}^T \cdot \mathbf{w_j} \cdot \mathbf{x_i}^T \cdot \mathbf{w_r} \cdot \alpha_{k,j}\alpha_{k,r} \\ &= \sum_{j=1}^{d} \left(\frac{1}{n}\sum_{i=1}^{n} \left(\mathbf{x_i}^T \cdot \mathbf{w_j}\right)^2\right) \alpha_{k,j}^2 + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{\substack{r=1 \\ r\neq j}}^{d} \mathbf{x_i}^T \cdot \mathbf{w_j} \cdot \mathbf{x_i}^T \cdot \mathbf{w_r} \cdot \alpha_{k,j}\alpha_{k,r} \\ &= \sum_{j=1}^{d} \mathbf{Var_{w_j}}\left[\mathbf{A_{(i)}}\right] \cdot \alpha_{k,j}^2 + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{\substack{r=1 \\ r\neq j}}^{d} \mathbf{x_i}^T \cdot \mathbf{w_j} \cdot \mathbf{x_i}^T \cdot \mathbf{w_r} \cdot \alpha_{k,j}\alpha_{k,r} \end{aligned}$$

Summing up over all $\mathbf{v_k}$, we we then obtain

$$\sum_{k=1}^{d} \mathbf{Var_{v_k}}\left[\mathbf{A_{(i)}}\right] = \sum_{j=1}^{d} \mathbf{Var_{w_j}}\left[\mathbf{A_{(i)}}\right] \cdot \sum_{k=1}^{d} \alpha_{k,j}^2 + \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} \sum_{\substack{r=1 \\ r \neq j}}^{d} \mathbf{x_i}^T \mathbf{w_j} \mathbf{x_i}^T \mathbf{w_r} \cdot \sum_{k=1}^{d} \alpha_{k,j} \alpha_{k,r}$$

$$= \sum_{j=1}^{d} \mathbf{Var_{w_j}}\left[\mathbf{A_{(i)}}\right],$$

using Equations (1) and (2). □

As in the one dimensional case, our notion of high dimensional variance has an algebraic interpretation. The *Frobenius norm* of a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ is defined as

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{d} A_{i,j}^2}.$$

If the centroid is equal to the origin, the squared Frobenius norm is, up to scale, equal to the multidimensional variance, as well as the 1-means cost. To see the former, consider the basis $\{\mathbf{e_k}\}_{k=1}^{n}$, where $\mathbf{e_k}$ is the vector that is equal to 1 at the kth coordinate and 0 everywhere else. We have

$$\mathbf{Var_{e_k}}\left[\mathbf{A_{(i)}}\right] = \mathbf{E}\left[\left(\mathbf{A_{(i)}} \mathbf{e_k}\right)^2\right] = \frac{1}{n} \sum_{i=1}^{n} \mathbf{A_{(i)}}_k^2 = \frac{1}{n} \sum_{i=1}^{n} \mathbf{A}_{ik}^2$$

and

$$\|\mathbf{A}\|_F^2 = \sum_{k=1}^{d} \sum_{i=1}^{n} \mathbf{A}_{ik}^2 = n \sum_{k=1}^{d} \mathbf{Var_{e_k}}\left[\mathbf{A_{(i)}}\right] = n\mathbf{Var}[\mathbf{A}].$$

Eigenvalues/Eigenvectors and Singular Values/Singular Vectors

Let us now make a pause to look at some important notions for matrices. We start by recalling the following definition:

Definition 2 (Eigenvectors and eigenvalues)**.** *Let $\mathbf{A} \in \mathbb{R}^{d \times d}$. A vector $v \in \mathbb{R}^d$ with unit Euclidean norm is a (right) eigenvector with eigenvalue e if*

- $\mathbf{Av} = e\mathbf{v}$ *and*
- $\mathbf{v}^T\mathbf{A} = e\mathbf{v}^T$.

The concepts of eigenvalues and eigenvectors are of the most central in linear algebra. However, they have the disadvantage that they are only applied on square matrices.

Therefore, for general matrices (such as our matrix $\mathbf{A}$) we have some more general concepts:

Definition 3 (Singular vectors and values)**.** *Let $\mathbf{A} \in \mathbb{R}^{n \times d}$. Two vectors $\mathbf{u} \in \mathbb{R}^n$ and $\mathbf{v} \in \mathbb{R}^d$ with unit Euclidean norm are respectively called left and right singular vectors of $\mathbf{A}$ if the following two equations hold*

- $\mathbf{Av} = \sigma\mathbf{u}$

- $\mathbf{u}^T\mathbf{A} = \sigma\mathbf{v}^T$.

σ is known as a singular value of $\mathbf{A}$.

Note that the concepts of singular values and vectors are related with those of eigenvalues and eigenvectors:

Proposition 4. *Let* $\mathbf{A}$ *be a matrix with right singular vector* $\mathbf{v}$ *and singular value* σ. *Then* $\mathbf{v}$ *is an eigenvector of* $\mathbf{A}^T\mathbf{A}$ *with eigenvalue* σ^2.

Proof. $\mathbf{A}^T\mathbf{A}\mathbf{v} = \mathbf{A}^T\mathbf{u}\sigma = (\mathbf{u}^T\mathbf{A})^T\sigma = (\sigma\mathbf{v}^T)^T\sigma = \sigma^2\mathbf{v}$ and $\mathbf{v}^T\mathbf{A}^T\mathbf{A} = \sigma\mathbf{u}^T\mathbf{A} = \sigma^2\mathbf{v}^T$. □

This shows that the largest eigenvalue of $\mathbf{A}^T\mathbf{A}$ and the squared largest singular value of $\mathbf{A}$ are equivalent. From linear algebra we know that:

Theorem 5. *Let* $\mathbf{A} \in \mathbb{R}^{n\times d}$ *be a matrix with rank* r. *Then there exist* r *triples* $\sigma_i \in \mathbb{R}$, $\mathbf{u_i} \in \mathbb{R}^n$, $\mathbf{v_i} \in \mathbb{R}^d$, *such that:*

- $A\mathbf{u_i} = \sigma_i\mathbf{v_i}$ *and* $A^T\mathbf{v_i} = \sigma_i\mathbf{u_i}^T$.
- $\sigma_i \neq 0$. *Furthermore we can choose* $\sigma_i > 0$, *and we typically define* $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$.
- *For* $i, j \in [r]$ *with* $i \neq j$ *we have that* $\mathbf{u_i}^T\mathbf{u_i} = 1$ *and* $\mathbf{u_i}^T\mathbf{u_j} = 0$. *In other words, the* $\mathbf{u_i}$*s form an orthonormal basis for an* r*-dimensional subspace of* $\mathbb{R}^n$.
- *For* $i, j \in [r]$ *with* $i \neq j$ *we have that* $\mathbf{v_i}^T\mathbf{v_i} = 1$ *and* $\mathbf{v_i}^T\mathbf{v_j} = 0$. *In other words, the* $\mathbf{v_i}$*s form an orthonormal basis for an* r*-dimensional subspace of* $\mathbb{R}^d$.
- $\mathbf{A}$ *can be written as*

$$\mathbf{A} = \sigma_1\mathbf{u_1}\mathbf{v_1}^T + \sigma_2\mathbf{u_2}\mathbf{v_2}^T + \cdots + \sigma_r\mathbf{u_r}\mathbf{v_r}^T = \sum_{i=1}^{r}\sigma_i\mathbf{u_i}\mathbf{v_i}^T. \tag{3}$$

Let us group together the $\mathbf{u_i}$s, the $\mathbf{v_i}$s, and the σ_is. We define the matrices $\mathbf{U_r} \in \mathbb{R}^{n\times r}$ and $\mathbf{V_r} \in \mathbb{R}^{d\times r}$ as

$$\mathbf{U_r} = \begin{bmatrix} | & | & & | \\ \mathbf{u_1} & \mathbf{u_2} & \cdots & \mathbf{u_r} \\ | & | & & | \end{bmatrix} \qquad \mathbf{V_r} = \begin{bmatrix} | & | & & | \\ \mathbf{v_1} & \mathbf{v_2} & \cdots & \mathbf{v_r} \\ | & | & & | \end{bmatrix},$$

and the matrix $\mathbf{\Sigma_r} \in \mathbb{R}^{r\times r}$ as

$$\mathbf{\Sigma_r} = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & \sigma_r \end{bmatrix}.$$

Note that we have $\mathbf{U_r}^T\mathbf{U_r} = \mathbf{I_r}$ and $\mathbf{V_r}^T\mathbf{V_r} = \mathbf{I_r}$ but $\mathbf{U_r}\mathbf{U_r}^T \neq \mathbf{I_n}$ for $r < n$ and $\mathbf{V_r}\mathbf{V_r}^T \neq \mathbf{I_d}$ for $d < n$.

Now we can write

$$\mathbf{A} = \sum_{i=1}^{r} \sigma_i \begin{bmatrix} | \\ \mathbf{u_1} \\ | \end{bmatrix} \cdot \begin{bmatrix} \text{———} \mathbf{v_i}^T \text{———} \end{bmatrix} = \sum_{i=1}^{r} \begin{bmatrix} | \\ \mathbf{u_1} \\ | \end{bmatrix} \cdot \begin{bmatrix} \text{———} \sigma_i \mathbf{v_i}^T \text{———} \end{bmatrix}$$

$$\overset{(a)}{=} \begin{bmatrix} | & | & & | \\ \mathbf{u_1} & \mathbf{u_2} & \cdots & \mathbf{u_r} \\ | & | & & | \end{bmatrix} \cdot \begin{bmatrix} \text{———} \sigma_1 \mathbf{v_1}^T \text{———} \\ \text{———} \sigma_2 \mathbf{v_2}^T \text{———} \\ \vdots \\ \text{———} \sigma_r \mathbf{v_r}^T \text{———} \end{bmatrix}$$

$$= \begin{bmatrix} | & | & & | \\ \mathbf{u_1} & \mathbf{u_2} & \cdots & \mathbf{u_r} \\ | & | & & | \end{bmatrix} \cdot \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & \sigma_r \end{bmatrix} \cdot \begin{bmatrix} \text{———} \mathbf{v_1}^T \text{———} \\ \text{———} \mathbf{v_2}^T \text{———} \\ \vdots \\ \text{———} \mathbf{v_r}^T \text{———} \end{bmatrix}$$

where (a) follows from Section 2.2. Therefore, we obtain

$$\mathbf{A} = \mathbf{U_r} \mathbf{\Sigma_r} \mathbf{V_r}^T. \tag{4}$$

Equations (3) and the equvallent (4) (and the one we will see later (5)) define the *singular value decomposition* (SVD) of matrix $\mathbf{A}$. Often we refer to the form of Equation (4) as the *reduced form* of the SVD. Later we will see the full form. The SVD is very important in data mining, because, as we will see, it shows how we can decompose our dataset into components that carry the information of the data in decreasing order.

Theorem 5 gives us r orthonormal left singular vectors $\mathbf{u_i}$s. We can choose $n - r$ more vectors $\mathbf{u_{r+1}}, \ldots, \mathbf{u_n}$ such that each of them has unit norm and is orthogonal to all the other $\mathbf{u_i}$s. (Note then these $n - r$ vectors form a basis for the nullspace of A.) This means, that the collection $\mathbf{u_1}, \ldots, \mathbf{u_n}$ is a basis for $\mathbb{R}^n$.

Similarly, we can choose $d - r$ more vectors $\mathbf{v_{r+1}}, \ldots, \mathbf{v_d}$, such that the entire collection $\mathbf{v_1}, \ldots, \mathbf{v_d}$ is a basis for $\mathbb{R}^d$.

We now define the matrices $\mathbf{U} \in \mathbb{R}^{n \times n}$ and $\mathbf{V} \in \mathbb{R}^{d \times d}$ as

$$\mathbf{U} = \begin{bmatrix} | & | & & | \\ \mathbf{u_1} & \mathbf{u_2} & \cdots & \mathbf{u_n} \\ | & | & & | \end{bmatrix} \qquad \mathbf{V} = \begin{bmatrix} | & | & & | \\ \mathbf{v_1} & \mathbf{v_2} & \cdots & \mathbf{v_d} \\ | & | & & | \end{bmatrix},$$

and the matrix $\mathbf{\Sigma} \in \mathbb{R}^{n \times d}$ as

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & \sigma_r & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix}.$$

Note that we have $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I_n}$ and $\mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I_d}$. We can then write:

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T. \tag{5}$$

This equation is equivalent to Equations (3) and (4), and we often refer to it as the (full) SVD.

4 Best-Fit Projections to Vectors and Subspaces

Principal Component Analysis is all about dimensionality reduction. As a tentative step, let us consider reducing the dimension down to 1. The main question is which direction is the most important one. Our notion of directional variance helps us in this regard. If a direction has extremely low directional variance, we can confidently say that the centroid (or origin if our data are normalized) will approximate the point set well enough. The most uncertainty is with respect to directions of high directional variance. Hence, if we are only allowed to choose a single direction, we should choose the one with maximum directional variance. Phrased as an optimization problem, we aim to solve the following (see Figure 4).

$$\max_{\mathbf{v}\in\mathbb{R}^d, \|\mathbf{v}\|=1} \mathbf{Var_v}\left[\mathbf{A_{(i)}}\right].$$

Again, this has an algebraic interpretation. Specifically, the maximum directional variance is (up to scaling) known as the squared spectral norm, where for any $n \times d$ matrix A the spectral norm is defined as

$$\|\mathbf{A}\|_2 = \max_{\mathbf{v}\in\mathbb{R}^d, \|\mathbf{v}\|_2=1} \|\mathbf{A}\mathbf{v}\|_2 = \max_{\mathbf{v}\in\mathbb{R}^d, \|\mathbf{v}\|_2=1} \sqrt{\sum_{i=1}^{n} \left(\mathbf{A_{(i)}}\mathbf{v}\right)^2}.$$

In the next theorem we show that the maximum directional variance equals the maximum eigenvector σ_1 and is achieved in the direction of the direction of $\mathbf{v_1}$.

Theorem 6. *Let* $\mathbf{A}$ *be a matrix. Then the spectral norm* $\|\mathbf{A}\|_2$ *is equal to the square of the largest singular value of* $\mathbf{A}$, σ_1^2 *and achieved in the direction of* $\mathbf{v_1}$.

Proof. Recall that σ_1 is the largest singula value with $\mathbf{u}_{i^*}$ and $\mathbf{v}_{i^*}$ being the corresponding right and left singular vectors of A. We have that $\|\mathbf{v}_1\|_2 = 1$ and that

$$\|\mathbf{A}\mathbf{v}\|_2^2 = \|\sigma_1\mathbf{u}_1\|_2^2 = \sigma_1^2\mathbf{u}_1^T\mathbf{u}_1 = \sigma_1^2.$$

Therefore we proved that for $\mathbf{v} = \mathbf{v}_1$ we have that $\|\mathbf{A}\mathbf{v}\|_2 = \sigma_1$, which means that $\|\mathbf{A}\|_2 \geq \sigma_1$. (Recall the definition $\|\mathbf{A}\|_2 = \max_{\mathbf{v}\in\mathbb{R}^d, \|\mathbf{v}\|_2=1} \|\mathbf{A}\mathbf{v}\|_2$.)

Now we show that $\|\mathbf{A}\|_2 \leq \sigma_1$. Consider any vector $\mathbf{v}$ with $\|\mathbf{v}\|_2 = 1$. We will prove that $\|\mathbf{A}\mathbf{v}\|_2 \leq \sigma_1$, and this will complete the proof. We have that $\|\mathbf{A}\mathbf{v}\|_2^2 = \mathbf{v}^T\mathbf{A}^T\mathbf{A}\mathbf{v}$. Let us consider $\mathbf{v}$ as a linear combination of the eigenvectors of $\mathbf{A}^T\mathbf{A}$, that is, $\mathbf{v} = \sum_{i=1}^d \alpha_i\mathbf{v_i}$ with $\sum_{i=1}^d \alpha_i^2 = 1$ (because $\|\mathbf{v}\|_2 = 1$) and $\{\mathbf{v}_1, \ldots, \mathbf{v_d}\}$ being an orthogonal basis of eigenvectors of

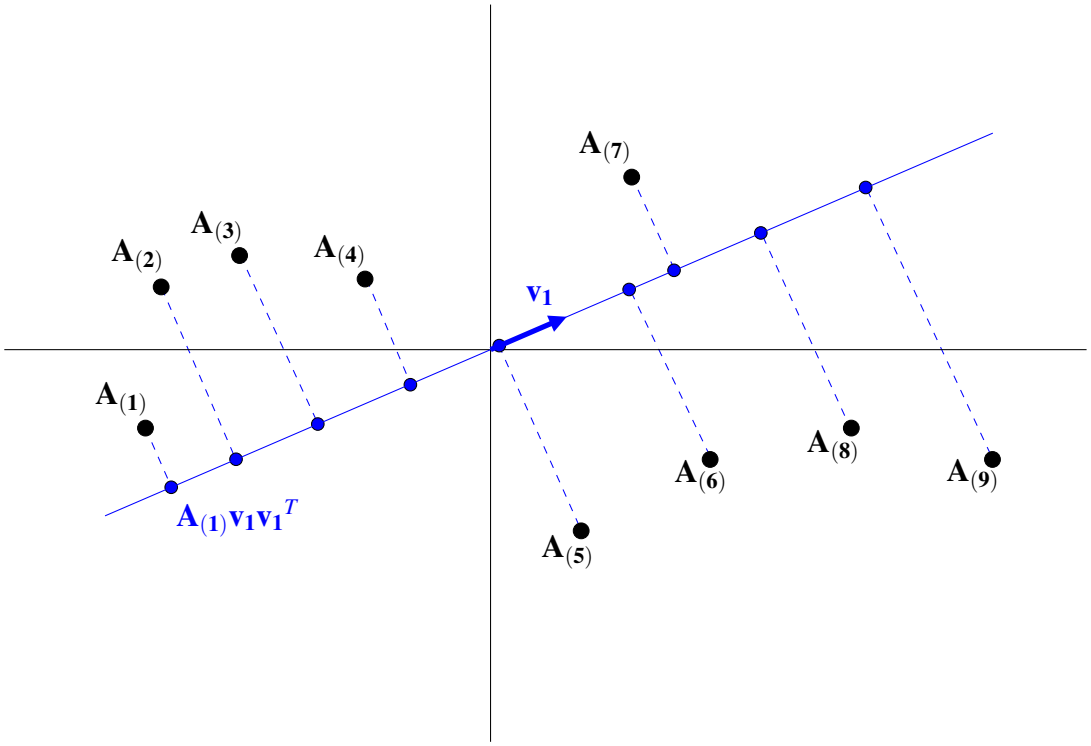


Figure 4: Projection of the dataset $\mathbf{A}$ onto the direction of vector $\mathbf{v_1}$, which is the vector $\mathbf{v}$ that maximizes the variance $\mathbf{Var_v}[\mathbf{A}]$. It also minimizes the sum of the squares of the distances of the points to the line defined by $\mathbf{v}$ (the dashed lines).

$\mathbf{A}^T\mathbf{A}$. Then,

$$\begin{aligned}
\mathbf{v}^T\mathbf{A}^T\mathbf{A}\mathbf{v} &= \left(\sum_{i=1}^{d}\alpha_i\mathbf{v}_i^T\right)\mathbf{A}^T\mathbf{A}\left(\sum_{j=1}^{d}\alpha_j\mathbf{v}_j\right) \\
&= \left(\sum_{i=1}^{d}\alpha_i\mathbf{v}_i^T\right)\left(\sum_{j=1}^{d}\alpha_j\mathbf{A}^T\mathbf{A}\mathbf{v}_j\right) \\
&= \left(\sum_{i=1}^{d}\alpha_i\mathbf{v}_i^T\right)\left(\sum_{j=1}^{d}\alpha_j\sigma_j^2\mathbf{v}_j\right) \\
&= \sum_{i=1}^{d}\alpha_i^2\sigma_i^2\mathbf{v}_i^T\mathbf{v}_i + \sum_{i=1}^{d}\sum_{\substack{j=1\\ j\neq i}}^{d}\alpha_i\alpha_j\sigma_i^2\mathbf{v_i}^T\mathbf{v_j},
\end{aligned}$$

where the second equality follows from Proposition 4. Because $\{\mathbf{v}_1, \ldots, \mathbf{v_d}\}$ is an orthonormal basis, we have that $\mathbf{v_i}^T\mathbf{v_i} = 1$ and $\mathbf{v_i}^T\mathbf{v_j} = 0$ for $i \neq j$. Hence,

$$\|\mathbf{A}\mathbf{v}\|_2^2 = \mathbf{v}^T\mathbf{A}^T\mathbf{A}\mathbf{v} = \sum_{i=1}^{d}\alpha_i^2\sigma_i^2 \leq \sum_{i=1}^{d}\alpha_i^2 \max_{1\leq j\leq d}\sigma_j^2 = \max_{1\leq j\leq d}\sigma_j^2 = \sigma_1^2.$$

□

Theorem 6 and Proposition 4 tell us that it is sufficient to find the largest eigenvalue of $\mathbf{A}^T\mathbf{A}$ and associated eigenvector to determine the maximum directional variance. There are various

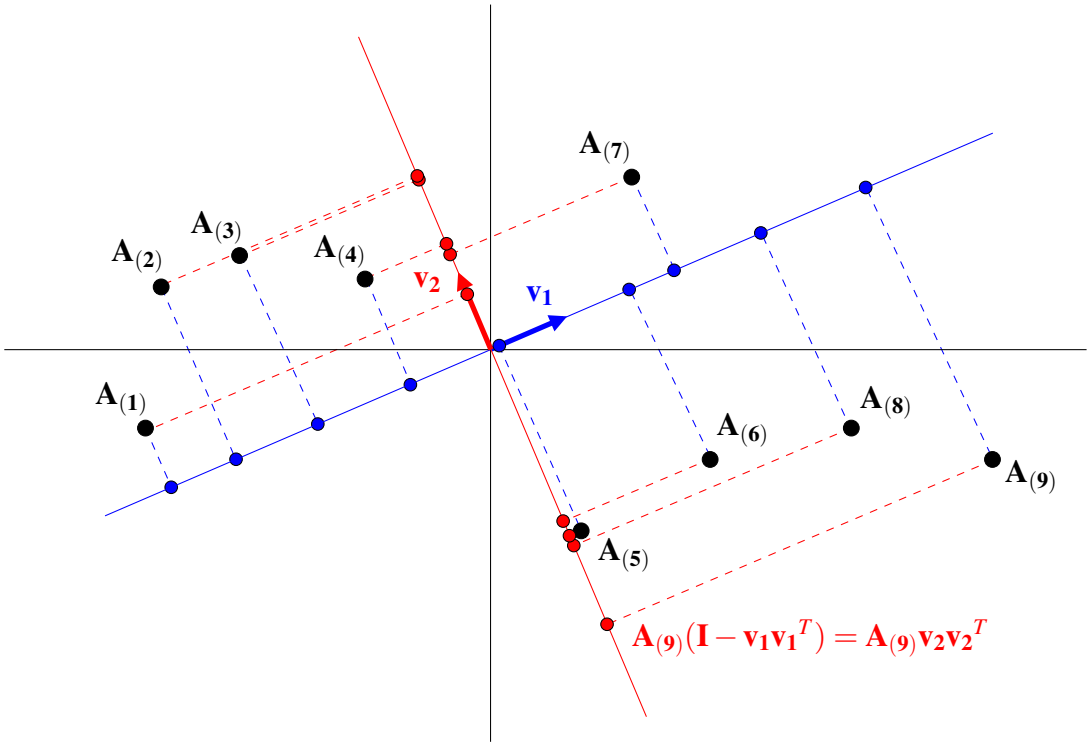


Figure 5: Projection of the dataset $\mathbf{A}$ onto the subspace that is orthogonal to $\mathbf{v_1}$. This is the "leftover" after the projection on $\mathbf{v_1}$ and is represented by the red points. The projection of point $\mathbf{A_{(i)}}$ on this subspace is the point $\mathbf{A_{(i)}}(\mathbf{I} - \mathbf{v_1}\mathbf{v_1}^T)$. Here, because the points lie on $\mathbb{R}^2$, the orthogonal subspace is simply the one spanned by the vector $\mathbf{v_2}$, so we have that the projected points are $\mathbf{A_{(i)}}(\mathbf{I} - \mathbf{v_1}\mathbf{v_1}^T) = \mathbf{A_{(i)}}\mathbf{v_2}\mathbf{v_2}^T$.

algorithms for computing eigenvalues: theoretically one can solve a system of linear equations, in practives there are numerical algorithms, such as the *power method*, which find them quite efficiently.

4.1 Best-Fit Subspaces

We now carry the discussion of the previous section further

2nd best direction. What if we want to take one more step? That is, consider the "leftover" after the projection (i.e., the projecion to the subspace orthogonal to $\mathbf{v_1}$); what is the direction of maximum variance? See Figure 5.

It is easy to see that the projection of the matrix to the orthogonal subspace to $\mathbf{v_1}$, which is the matrix $\mathbf{A}(\mathbf{I} - \mathbf{v_1}\mathbf{v_1}^T)$, has the same singular values and singular vectors of $\mathbf{A}$, with the

exception of σ_1 which has become 0. To see this, note that we can write:

$$\begin{aligned}\mathbf{A}(\mathbf{I}-\mathbf{v_1}\mathbf{v_1}^T) &= \sum_{i=1}^{r}\sigma_i\mathbf{u_i}\mathbf{v_i}^T(\mathbf{I}-\mathbf{v_1}\mathbf{v_1}^T)\\ &= \sum_{i=1}^{r}\left(\sigma_i\mathbf{u_i}\mathbf{v_i}^T-\sigma_i\mathbf{u_i}\mathbf{v_i}^T\mathbf{v_1}\mathbf{v_1}^T\right)\\ &= \sum_{i=1}^{r}\sigma_i\mathbf{u_i}\mathbf{v_i}^T-\sigma_1\mathbf{u_1}\mathbf{v_1}^T\\ &= \sum_{i=2}^{r}\sigma_i\mathbf{u_i}\mathbf{v_i}^T,\end{aligned}$$

begcause we have $\mathbf{v_i}^T\mathbf{v_1}=1$ for $i=1$, and 0 otherwise. Therefore, by applying Theorem 6 to $\mathbf{A}(\mathbf{I}-\mathbf{v_1}\mathbf{v_1}^T)$, whose highest singular value is σ_2, we obtain that the direction of maximum variance is $\mathbf{v_2}$ and the variance is σ_2^2/n. In other words, we have:

$$\left\|\mathbf{A}(\mathbf{I}-\mathbf{v_1}\mathbf{v_1}^T)\right\|_2=\max_{\mathbf{v}\in\mathbb{R}^d,\|\mathbf{v}\|_2=1}\left\|\mathbf{A}(\mathbf{I}-\mathbf{v_1}\mathbf{v_1}^T)\mathbf{v}\right\|_2=\sigma_2=\max_{\substack{\mathbf{v}\in\mathbb{R}^d,\|\mathbf{v}\|_2=1\\ \mathbf{v}^T\mathbf{v_1}=0}}\|\mathbf{A}\mathbf{v}\|_2. \tag{6}$$

kth best direction. We can extend this process for $k\le d$ steps. We can keep projecting the "leftover," and in a similar way we can show that the kth projection of maximum variance is along the direction of $\mathbf{v_k}$ and that we have (recall, from Section 2.4 that after we project to the vectors $\mathbf{v_1},\dots\mathbf{v_{k-1}}$, the "leftover" of $\mathbf{A}$ is $\mathbf{A}(\mathbf{I}-\mathbf{V_{k-1}}\mathbf{V_{k-1}}^T)$):

$$\left\|\mathbf{A}(\mathbf{I}-\mathbf{V_{k-1}}\mathbf{V_{k-1}}^T)\right\|_2=\max_{\mathbf{v}\in\mathbb{R}^d,\|\mathbf{v}\|_2=1}\left\|\mathbf{A}(\mathbf{I}-\mathbf{V_{k-1}}\mathbf{V_{k-1}}^T)\mathbf{v}\right\|_2=\sigma_k=\max_{\substack{\mathbf{v}\in\mathbb{R}^d,\|\mathbf{v}\|_2=1\\ \mathbf{v}^T\mathbf{v_1}=0,\dots,\mathbf{v}^T\mathbf{v_{k-1}}=1}}\|\mathbf{A}\mathbf{v}\|_2. \tag{7}$$

Variance along multiple dimensions. Let us ask now a slightly different question. What is the best 2-dimensional projection of $\mathbf{A}$? In other words, what is a projection that maximizes the directional variance of the projected points? Let's try to understand this. First, following the definition we used above, given some set of vectors $\{\mathbf{v_i}\}$, for any integer k we will use $\mathbf{V_k}$ to represent the matrix

$$\mathbf{V_k}=\begin{bmatrix}| & | & & |\\ \mathbf{v_1} & \mathbf{v_2} & \cdots & \mathbf{v_k}\\ | & | & & |\end{bmatrix}.$$

We define analogously the matrix $\mathbf{W_k}$ for a set of vectors $\{\mathbf{w_i}\}$.

Consider a data point $\mathbf{A_{(i)}}$, and a set of k orthonormal vectors $\mathbf{w_1},\dots,\mathbf{w_k}$. By the Pythagorean theorem, the length of the projection of the vector $\mathbf{A_{(i)}}$ on the subspace spanned by $\mathbf{w_1},\dots,\mathbf{w_k}$ is

$$\left\|\mathbf{A_{(i)}}\mathbf{W_k}\mathbf{W_k}^T\right\|_2^2=\sum_{j=1}^{k}\left\|\mathbf{A_{(i)}}\mathbf{w_k},\mathbf{w_k}^T\right\|_2^2.$$

But as we saw in Section 2.3, the length of each vector $\mathbf{A_{(i)}}\mathbf{w_k},\mathbf{w_k}^T$ is $\mathbf{A_{(i)}}\mathbf{w_k}$. Therefore we have that

$$\left\|\mathbf{A_{(i)}}\mathbf{W_k}\mathbf{W_k}^T\right\|_2^2=\sum_{j=1}^{k}\left(\mathbf{A_{(i)}}\mathbf{w_k}\right)^2=\left\|\left(\mathbf{A_{(i)}}\mathbf{w_1},\mathbf{A_{(i)}}\mathbf{w_2},\dots,\mathbf{A_{(i)}}\mathbf{w_k}\right)\right\|_2^2=\left\|\mathbf{A_{(i)}}\mathbf{W_k}\right\|_2^2.$$

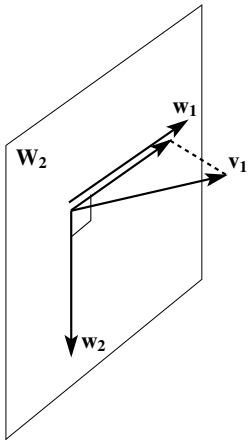


Figure 6: Choosing a basis for the subspace $\mathbf{W_2}$ such that $\mathbf{w_2}$ is orthogonal to $\mathbf{v_1}$.

Then, by summing over all the rows of $\mathbf{A}$, we have

$$\left\|\mathbf{A}\mathbf{W_k}\mathbf{W_k}^T\right\|_F^2 = \|\mathbf{A}\mathbf{W_k}\|_F^2 . \tag{8}$$

Best k-dimensional projection. Let's go back to the question of finding the 2-dimensional projection of $\mathbf{A}$ that maximizes the variance

$$\frac{1}{n}\|\mathbf{A}\mathbf{W_2}\|_F^2 = \frac{1}{n}\left(\|\mathbf{A}\mathbf{w_1}\|_2^2 + \|\mathbf{A}\mathbf{w_2}\|_2^2\right).$$

We will show next that the best subspace is the one spanned by the right singular vectors of $\mathbf{A}$, $\mathbf{v_1}$ and $\mathbf{v_2}$. Consider any other 2-dimensional subspace $\mathbf{W_2}$ and consider an orthonormal basis $(\mathbf{w_1}, \mathbf{w_2})$ for $\mathbf{W_2}$, such that $\mathbf{w_2}$ is perpendicural to $\mathbf{v_1}$ (i.e., $\mathbf{w_2}^T\mathbf{v_1} = 0$). Note that it is always possible to choose such a basis: If $\mathbf{W_2}$ is orthogonal to $\mathbf{v_1}$, then any orthonormal basis of $\mathbf{W_2}$ will do (each vector in $\mathbf{W_2}$ is orthogonal to $\mathbf{v_1}$. Otherwise, consider the projection of $\mathbf{v_1}$ onto $\mathbf{W_2}$ and let $\mathbf{w_1}$ be the vector along this projection and $\mathbf{w_2}$ be orthogonal to the projection; see Figure 6.

From Theorem 6 we have that

$$\|\mathbf{A}\mathbf{v_1}\|_2 \geq \|\mathbf{A}\mathbf{w_1}\|_2 ,$$

and from Equation (6) we have that

$$\|\mathbf{A}\mathbf{v_2}\|_2 \geq \|\mathbf{A}\mathbf{w_2}\|_2 .$$

Therefore,

$$\|\mathbf{A}\mathbf{V_2}\|_F^2 = \|\mathbf{A}\mathbf{v_1}\|_2^2 + \|\mathbf{A}\mathbf{v_2}\|_2^2 \geq \|\mathbf{A}\mathbf{w_1}\|_2^2 + \|\mathbf{A}\mathbf{w_2}\|_2^2 = \|\mathbf{A}\mathbf{W_2}\|_F^2 ,$$

proving that the subspace defined by the matrix $\mathbf{V_2}$ is the one that maximizes the variance.

Definition 7. *Given a matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $k \leq d$, the subspace that maximizes the directional variance of the points in $\mathbf{A}$ is called the* best-fit k*-dimensional subspace of $\mathbb{R}^d$ with respect to $\mathbf{A}$.*

We have seen that for $k = 1$ and $k = 2$, the best-fit subpace is the one spanned by the first k right singular vectors. We next show that this is more general.

Theorem 8. *For any $k \leq d$, the best fit subspace is the one spanned by the first k right singular vectors of $\mathbf{A}$. In particular, consider the vectors $\mathbf{v_1}, \ldots, \mathbf{v_k}$, and the corresponding matrix $\mathbf{V_k}$. Then, for any other orthonormal vectors $\mathbf{w_1}, \ldots, \mathbf{w_k}$ we have*

$$\|\mathbf{AV_k}\|_F \geq \|\mathbf{AW_k}\|_F .$$

Proof. The proof for general k is similar to the proof for $k = 2$ that we showd above. We will prove by induction. The case $k = 1$ is given by Theorem 6. Assume now that it hold for $k - 1$, that is, for any $(k-1)$-dimensional subspace $\mathbf{W_{k-1}}$ we have

$$\|\mathbf{AV_{k-1}}\|_F \geq \|\mathbf{AW_{k-1}}\|_F .$$

Consider now any k-dimensional subspace $\mathbf{W_k}$. Choose a basis for $\mathbf{W_k}$, such that $\mathbf{w_k}$ is orthogonal to all vectors $\mathbf{v_1}, \ldots, \mathbf{v_{k-1}}$; this can be done similarly to the case of $k = 2$. Then, from Equation (7) we have that

$$\|\mathbf{Av_k}\|_2 \geq \|\mathbf{Aw_k}\|_2 .$$

Therefore, we obtain

$$\begin{aligned}
\|\mathbf{AV_k}\|_F^2 &= \sum_{j=1}^{k-1} \|\mathbf{Av_j}\|_2^2 + \|\mathbf{Av_k}\|_2^2 \\
&= \|\mathbf{AV_{k-1}}\|_F^2 + \|\mathbf{Av_k}\|_2^2 \\
&\geq \|\mathbf{AW_{k-1}}\|_F^2 + \|\mathbf{Aw_k}\|_2^2 \\
&= \sum_{j=1}^{k-1} \|\mathbf{Aw_j}\|_2^2 + \|\mathbf{Aw_k}\|_2^2 \\
&= \|\mathbf{AW_k}\|_F^2 ,
\end{aligned}$$

proving that the subspace defined by the matrix $\mathbf{V_k}$ is the one that maximizes the variance. □

5 Principal Component Analysis

Having now gained intuition we can also see how we can apply it on data analysis. Given the data matrix $\mathbf{A}$, the SVD $\mathbf{A} = \sum_{j=1}^{r} \sigma_j \mathbf{u_j} \mathbf{v_j}^T$ decomposes $\mathbf{A}$ into its *principal components*, as we call them. Each component, which is a rank-1 matrix, captures some of the information of the data. For a given component $\sigma_j \mathbf{u_j} \mathbf{v_j}^T$, $\mathbf{v_j}^T$ is the direction of the component, $\mathbf{u_j}$ specifies how much of this direction is present in each of n datapoints, and σ_j indicates the contribution of the component to the data, Because the singular values are sorted in nonincreasing order, the first components contain most of the information, the "signal," and the last ones can be thought of as minor information or even "noise."

We define, for $k \leq d$ (and typically $k \leq r$),

$$\mathbf{A_k} = \sum_{j=1}^{k} \sigma_j \mathbf{u_j} \mathbf{v_j}^T .$$

We have seen various facts:

- $\mathbf{A_k}$ is the a matrix of rank at most k and has rank k if and only if $\sigma_k > 0$. To make the discussion simpler, we next assume that this is the case.

- $\mathbf{A_k}$ is the matrix of rank k that is closest to $\mathbf{A}$ in the sense that minimizes $\|\mathbf{A} - \mathbf{X}\|_F$ among all matrices $\mathbf{X}$ of rank at most k.
- It contains the projection of the points to the k-dimensional subspace of $\mathbb{R}^d$ that maximizes the variance.
- Equivalently, it contains the projection of the points to the k-dimensional subspace of $\mathbb{R}^d$ that minimizes the sum of the squared distances between each point and its projection.

We often work with $\mathbf{A_k}$ instead of the original matrix $\mathbf{A}$. There are multiple reasons to do this. First, by looking at $\mathbf{A_2}$ or even $\mathbf{A_3}$ we can visualize our data and gain some intuition. Another important reason is to remove the noise. The intuition is that noise is generally considered random so it does not have any particular direction in the feature space $\mathbb{R}^d$. This means that it is likely present in the lowest components. Therefore, by dropping the last components, we can cleanup the data. This is similar to a highpass filter in signal processing, which removes the high frequences of a signal, sometimes considered to be noise.

How many components should we keep? There is not a unique way to do this. At the end of the day, we need to try with different values and observe the data, see what results we obtain, and so on. In any case, some typical approaches, is to keep adding components until we have captured enough of the variance (e.g., 95%), or until the singular values drop significantly from the previous ones.

Before we finish the discussion, we note two important steps that we need to do before applying the SVD. First, as we mentioned, we need to center our data. If the data are not centered, then the Frobenius norm does not capture the variance of our data and we end up projecting our data set on different directions from the ones maximizing the variance.

A second step is normalization: We need to make the variance of each feature equal. Otherwise, some features with high variance, may "pull" towards them the directions of maximum variance. But what is wrong with this? The reason is that the variance of each feature depends on the unit used. Consider for instance two dataset matrices $\mathbf{A}$ and $\mathbf{B}$ where in $\mathbf{A}$ some particular column (say j) the values are represented in centimeters, and $\mathbf{B}$ contains exactly the same values, with the only difference that in this feature the values are represented in meters (i.e. $\mathbf{B^{(j)}} = \mathbf{A^{(j)}}/100$). The two matrices contain exactly the same information, yet in $\mathbf{A}$ the $\mathbf{v_1}$ will be much more aligned with feature j.

For these reasons, when we want to analyze our data matrix $\mathbf{A}$, we perform the following two preprocssing steps (in this order):

1. Center: For each row I, set $\mathbf{A_{(i)}} = \mathbf{A_{(i)}} - \sum_{j=1}^{n} \mathbf{A_{(j)}} \Big/ n$
2. Normalize: For each column j, set $\mathbf{A^{(j)}} = \mathbf{A^{(j)}} \Big/ \sqrt{\sum_{i=1}^{n} A_{i,j}^2}$

6 Applications to k-means

We first would like to consider the sister minimization problem. We first observe that by the Pythagorean theorem, for every i and for any subspace defined by the matrix $\mathbf{W_k} = [\mathbf{w_1}, \ldots, \mathbf{w_k}]$ (where, as usually, the vectors $\mathbf{w_j}$ form and orthonormal basis) we have

$$\left\|\mathbf{A_{(i)}}\right\|_2^2 = \left\|\mathbf{A_{(i)}}\mathbf{W}\mathbf{W}^T\right\|_2^2 + \left\|\mathbf{A_{(i)}} - \mathbf{A_{(i)}}\mathbf{W}\mathbf{W}^T\right\|_2^2,$$

and summing over all the rows i of $\mathbf{A}$ we have:

$$\begin{aligned}\|\mathbf{A}\|_F^2 &= \left\|\mathbf{A}\mathbf{W}\mathbf{W}^T\right\|_F^2 + \left\|\mathbf{A} - \mathbf{A}\mathbf{W}\mathbf{W}^T\right\|_F^2 \\ &= \|\mathbf{A}\mathbf{W}\|_F^2 + \left\|\mathbf{A} - \mathbf{A}\mathbf{W}\mathbf{W}^T\right\|_F^2,\end{aligned}$$

where the last equality follows from Eq. (8). Given that $\|\mathbf{A}\|_F^2$ is fixed, this means that maximizing $\|\mathbf{AW}\|_F^2$ (which is what we have been doing till now) is equivalent to minizing $\left\|\mathbf{A} - \mathbf{AWW}^T\right\|_F^2$. Minimizing

$$\left\|\mathbf{A} - \mathbf{AWW}^T\right\|_F^2$$

for a rank k subspace $\mathbf{W}$ is known in literature as finding the best rank-k approximation.

Note that the whole discussion that we have done until now, could have been done for the matrix $\mathbf{A}^T$, whose SVD is given by $\mathbf{A}^T = \mathbf{V}^T\mathbf{\Sigma}^T\mathbf{U}$. Applying the previous discussion to $\mathbf{A}^T$, we want to find a matrix $\mathbf{Z}$ of rank k that minimize

$$\left\|\mathbf{A}^T - \mathbf{AZZ}^T\right\|_F^2,$$

which is equivalent to minimizing

$$\left\|\mathbf{A} - \mathbf{ZZ}^T\mathbf{A}\right\|_F^2.$$

Thus, instead of working with the rows of $\mathbf{A}$ we can work with the columns of $\mathbf{A}$. We have by now seen that the best rank-k matrix $\mathbf{Z}$ is the matrix $\mathbf{U_k}$. Next we will see that also the k-means problem can be formulated as a problem of finding such a rank-k matrix $\mathbf{Z}$, but with some additional constraints on $\mathbf{Z}$.

First, consider the 1-means objective function, where we aim to find a point μ such that $\sum_{i=1}^n \left\|\mathbf{A_{(i)}} - \mathbf{c}\right\|^2$ is minimized. We know that $\mathbf{c} = \mu = \frac{1}{n}\sum_{i=1}^n \mathbf{A_{(i)}}$ is optimal. Can we express this problem algebraically? Indeed, we can. Let us rewrite the one means objective as

$$\sum_{i=1}^{n} \left\|\mathbf{A_{(i)}} - \mathbf{c}\right\|^2 = \|\mathbf{A} - \mathbf{C}\|_F^2$$

with the contraint that every row of $\mathbf{C}$ is identical. Consider the vector $\mathbf{X} = \frac{1}{\sqrt{n}} \cdot \mathbf{1}$. Then the optimal matrix $\mathbf{C}$, that is, the matrix where every row is μ can be expressed as $\mathbf{XX}^T\mathbf{A}$. Moreover, $\mathbf{X}$ is a unit vector. 1-means is therefore nothing but a constrained low-rank approximation problem.

For k-means we have a similar picture. Consider the n by k *clustering matrix* $\mathbf{X}$ defined as

$$X_i = \begin{cases} \frac{1}{\sqrt{|C_j|}} & \text{if point } A_i \text{ is in cluster } C_j \\ 0 & \text{otherwise} \end{cases}.$$

The columns of $\mathbf{X}$ are orthogonal, that is, the jth column $\mathbf{X^j}$ satisfies $\left\|\mathbf{X^j}\right\|_2 = 1$ and any column $\mathbf{X^i}$ has a zero entry whenever a column $\mathbf{X^j}$ has a nonzero entry. Notice how $\mathbf{X^i}(\mathbf{X^i})^T\mathbf{A}$ is mapped to the centroid of the cluster C_i. k-means can be therefore viewed as

$$\min_{\text{rank } k \text{ clustering matrix } \mathbf{X}} \left\|\mathbf{A} - \mathbf{XX}^T\mathbf{A}\right\|_F^2.$$

By lifting the constraint that $\mathbf{X}$ need be a clustering matrix, we are back to solving the low-rank approximation. Hence if we only cluster in the best k-dimensional subspace instead of the original d-dimensional space, we preserve most of the cost. This is made formal in the following theorem.

Theorem 9. *Let k be an integer and $A \in \mathbb{R}^{n\times d}$. Suppose we have an algorithm* Alg *that computes an α-approximation. Let $\mathbf{A_k} = \mathbf{U_k}\mathbf{A}{\mathbf{V_k}}^T$ be the best rank-k approximation. Then running* Alg *on $\mathbf{A_k}$ yields a $\alpha + 1$ approximation.*

Proof. Let $\mathbf{X}$ be the optimal clustering matrix. We observe that the optimal k-means cost $\left\|\mathbf{A}-\mathbf{X}\mathbf{X}^T\mathbf{A}\right\|_F^2$ is lower bounded by $\|\mathbf{A}-\mathbf{A_k}\|_F^2$. Let $\mathbf{Y}$ be the clustering matrix obtained by Alg. Then

$$\begin{aligned}\left\|\mathbf{A}-\mathbf{Y}\mathbf{Y}^T\mathbf{A}\right\|_F^2 &\leq \left\|\mathbf{A}-\mathbf{Y}\mathbf{Y}^T\mathbf{A_k}\right\|_F^2 \leq \left\|\mathbf{A_k}-\mathbf{Y}\mathbf{Y}^T\mathbf{A_k}\right\|_F^2 + \|\mathbf{A}-\mathbf{A_k}\|_F^2 \\ &\leq \alpha\cdot\left\|\mathbf{A_k}-\mathbf{X}\mathbf{X}^T\mathbf{A_k}\right\|_F^2 + \left\|\mathbf{A}-\mathbf{X}\mathbf{X}^T\mathbf{A}\right\|_F^2 \\ &\leq \alpha\cdot\left\|\mathbf{A}-\mathbf{X}\mathbf{X}^T\mathbf{A}\right\|_F^2 + \left\|\mathbf{A}-\mathbf{X}\mathbf{X}^T\mathbf{A}\right\|_F^2 \leq (\alpha+1)\left\|\mathbf{A}-\mathbf{X}\mathbf{X}^T\mathbf{A}\right\|_F^2.\end{aligned}$$

□

We remark that using $\mathbf{A_m}$ instead of $\mathbf{A_k}$ for $m > k/\varepsilon$, we obtain an $(\alpha+\varepsilon)$ approximation.

References

[1] A. Blum, J. Hopcroft, and R. Kannan. *Foundations of Data Science.* Cambridge University Press, 2020.

[2] G. Strang. *Linear Algegra and Learning from Data.* Wellesley – Cambridge Press, 2019.