Notes on PCA and applications to k-means

Chris Schwiegelshohn

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

$$\mathbf{E}[x_i] \triangleq \frac{1}{n} \sum_{i=1} x_i$$

and the *empirical variance* is defined as

$$\mathbf{Var}[x_i] \triangleq \mathbf{E}\big[(x_i - \mathbf{E}[x_i])^2\big] = \mathbf{E}\big[x_i^2\big] - \mathbf{E}[x_i]^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 .$$

As a matter of fact, we often normalize 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. 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, $\mathbf{Var}[x_i] = \frac{1}{n}\|\mathbf{x}\|_2^2$, where $\mathbf{x}$ is the vector containing all the samples: $\mathbf{x} = (x_1, \dots x_n)$. We also note that $\mathbf{x}^T\mathbf{x} = \|\mathbf{x}\|_2^2$ for any vector $\mathbf{x}$.

Let us now consider these notions 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. 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*. 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

$$\mathbf{Var_v}\big[\mathbf{A_{(i)}}\big] = \mathbf{E}\big[(\mathbf{A_{(i)}} - \mathbf{E}\big[\mathbf{A_{(i)}}\big])^T\mathbf{v})^2\big] .$$

Again, for normalized inputs with $\mathbf{E}\big[\mathbf{A_{(i)}}\big] = \mathbf{0}$, this reduces to

$$\mathbf{Var_v}\big[\mathbf{A_{(i)}}\big] = \mathbf{E}\big[(\mathbf{A_{(i)}}^T\mathbf{v})^2\big] .$$

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 $V = \{\mathbf{v_1}, \mathbf{v_2}, \dots, \mathbf{v_d}\}$ of $\mathbb{R}^d$ and compute

$$\mathbf{Var}\big[\mathbf{A_{(i)}}\big] \triangleq \sum_{j=1}^{d} \mathbf{Var}\big[(\mathbf{A_{(i)}}^T\mathbf{v_j})^2\big] .$$

We note that this definition is invariant under any choice of orthogonal basis, that is, for two distinct candidates $V = \{\mathbf{v_1}, \mathbf{v_2}, \dots, \mathbf{v_d}\}$ and $W = \{\mathbf{w_1}, \mathbf{w_2}, \dots, \mathbf{w_d}\}$, the evaluation of

$\mathbf{Var}\big[\mathbf{A_{(i)}}\big]$ is still identical. Let us prove this. First, observe that 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

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

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,

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

if we define $\mathbf{x} = \mathbf{A_{(i)}} - \mathbf{E}\left[\mathbf{A_{(i)}}\right]$. Summing up over all $\mathbf{v_k}$, we we then obtain

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

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

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

and

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

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.

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

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 by d matrix A the spectral norm is defined as

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

The connection to algebra is deeper. We will see shortly that the vector $\mathbf{v}$ that induces the spectral norm of $\mathbf{A}$ is an eigenvector of $\mathbf{A}^T\mathbf{A}$ and a right singular vector of $\mathbf{A}$. The directional variance itself is (up to scaling) also the largest eigenvalue of $\mathbf{A}^T\mathbf{A}$. Let us first recall the following definitions.

Definition 1 (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 $\mathbf{A}$.

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 an eigenvector with eigenvalue e if*

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

Proposition 3. *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. $\mathbf{A}^T\mathbf{Av} = \mathbf{A}^T\mathbf{u}\sigma = \mathbf{v}^T\sigma^2$ 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. Let us further show that these are equivalent to the squared spectral norm.

Theorem 4. *Let $\mathbf{A}$ be a matrix. Then the spectral norm $\|\mathbf{A}\|_2$ is equal to the largest singular value of $\mathbf{A}$.*

Proof. Let $\mathbf{v}$ be the vector that induces the spectral norm of $\mathbf{A}$. Then $\|\mathbf{Av}\|_2^2 = \mathbf{v}^T\mathbf{A}^T\mathbf{Av}$. Let us consider $\mathbf{v}$ as a linear combination of eigenvectors of $\mathbf{A}^T\mathbf{A}$, that is, $\mathbf{v} = \sum_{i=1}^d \alpha_i\mathbf{v_i}$ with $\sum_{i=1}^d \alpha_i^2 = 1$ and $\{\mathbf{v}_1, \ldots, \mathbf{v_j}\}$ being an orthogonal basis of eigenvectors of $\mathbf{A}^T\mathbf{A}$. Then

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

Since V is an orthogonal basis $\mathbf{v_i}^T\mathbf{v_j} = 0$ and $\mathbf{v_i}^T\mathbf{v_i} = 1$. Hence

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

□

Theorem 4 and Proposition 3 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. Obviously, this is a very well studied problem and we will not discuss this in detail. We will only present a simple algorithm known as the power method.

1. Pick a random unit vector $\mathbf{x}$
2. Repeat $\mathbf{x} \leftarrow \frac{\mathbf{A}^T\mathbf{Ax}}{\|\mathbf{A}^T\mathbf{Ax}\|_2}$

To analyze this algorithm, we require two steps. First, we will show that a random unit vector $\mathbf{x}$ is not too far off from the target eigenvector $\mathbf{v}$. Second, we show that under certain assumptions, this method will quickly converge.

To pick a random unit vector in d dimensions, we choose d random and independent Gaussian variables $X_1, \ldots, X_d$ with mean 0 and variance 1. We then set $\mathbf{x}^T = (X_1, \ldots, X_d) \cdot \frac{1}{\sqrt{\sum_{i=1}^d X_i^2}}$. Then the following holds

Lemma 5. *Fix a unit vector* $\mathbf{v}$*. For any random unit vector* $\mathbf{x}$*, we have* $\mathbf{E}\left[(\mathbf{x}^T\mathbf{v})^2\right] \geq \frac{1}{d}$*.*

Proof. We first observe that because $\mathbf{x}$ is chosen uniformly at random from all unit vectors, we may assume $\mathbf{v}$ to be the unit vector along the first axis, that is, $\mathbf{v}^T = (1, 0 \ldots, 0)$. Then

$$\mathbf{E}\left[(\mathbf{x}^T\mathbf{v})^2\right] = \mathbf{E}\left[X_1^2 \cdot \frac{1}{\sum_{i=1}^d X_i^2}\right] = \mathbf{E}\left[\frac{1}{1 + \sum_{i=2}^d \frac{X_i^2}{X_1^2}}\right] \geq \frac{1}{\mathbf{E}\left[1 + \sum_{i=2}^d \frac{X_i^2}{X_1^2}\right]} = \frac{1}{1 + \sum_{i=2}^d \frac{\mathbf{E}\left[X_i^2\right]}{\mathbf{E}\left[X_1^2\right]}} = \frac{1}{d}$$

where the inequality follows from Jensen's inequality[1] and the last two equalities from the fact that X_1 and X_i (for $i \geq 2$) are independent and identically distributed. □

We cannot give good converge speeds of the power method in general. But if the gap between first and second eigenvalue is small, the method works pretty well. To see this, let $\mathbf{x} = \sum_{i=1}^d \alpha_i \mathbf{v_i}$ be the initial choice of $\mathbf{x}$. Let us now consider the vector $\mathbf{x^{(k)}}$, that is, the current candidate $\mathbf{x}$ after k iterations. We have (up to scaling)

$$\mathbf{x^{(k)}} = (\mathbf{A}^T\mathbf{A})^k\mathbf{x} = \sum_{i=1}^d \alpha_i \mathbf{v_i} \sigma_i^{2k} = \sigma_1^{2k} \cdot \left(\alpha_1 \mathbf{v_1} + \sum_{i=2}^d \alpha_i \mathbf{v_i} \left(\frac{\sigma_i}{\sigma_1}\right)^{2k}\right).$$

We note that since σ_1^2 is the largest eigenvalue, the term $\sum_{i=2}^d \alpha_i \mathbf{v_i} \left(\frac{\sigma_i}{\sigma_1}\right)^{2k}$ will eventually converge to zero. The time required for this to happen depends on the ratio between σ_1^2 and the next largest eigenvalue σ_2^2. In practice, we may assume that $\sigma_2 \cdot (1 + \varepsilon) < \sigma_1$ for a few reasons. First, data are often noisy, so the bad case of having an extremely small gap between the two eigenvalues is unlikely. Second, if the values are sufficiently close, we may simply be satisfied with any vector that is a linear combination of $\mathbf{v_1}$ and $\mathbf{v_2}$ (and possibly $\mathbf{v_3}, \mathbf{v_4}, \ldots$, if more eigenvalues are extremely close). We note that α_1 must not be equal to 0 for this algorithm to converge to the largest eigenvector. Hence, the random initialization.

Best-Fit Subspaces We now consider dimension reduction onto more than just a single dimension. Here, we reuse many of the notions we have seen before.

Again, formulated as an optimization problem, we are looking for k vectors $\mathbf{v_1}, \ldots, \mathbf{v_k}$ such that the directional variances $\sum_{j=1}^k \mathbf{Var_{v_j}}\left[\mathbf{A_{(i)}}\right]$ are maximized. Algebraically, we aim to find a $d \times k$ orthogonal matrix $\mathbf{V}$, such that $\|\mathbf{AV}\|_F^2$ is maximized. To see this, consider

$$n \cdot \sum_{j=1}^k \mathbf{Var_{v_j}}\left[\mathbf{A_{(i)}}\right] = \sum_{j=1}^k \|\mathbf{Av_j}\|_2^2 = \|\mathbf{AV}\|_F^2.$$

The solution to this problem essentially boils down to finding a the maximum directional variance, and then projecting the entire point set onto the orthogonal subspace.

[1] Jensen's inequality says that if a function $f(x)$ is convex, then for a random variable X we have that $\mathbf{E}[f(X)] \geq f(\mathbf{E}[X])$.

Theorem 6. *Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ be a matrix. Then the best fit subspace is characterized as follows.*

$$\begin{aligned}
\mathbf{v_1} &= \operatorname*{argmax}_{\|\mathbf{v}\|_2=1} \|\mathbf{Av}\|_2 \\
\mathbf{v_2} &= \operatorname*{argmax}_{\|\mathbf{v}\|_2=1, \mathbf{v}^T\mathbf{v_1}=0} \|\mathbf{Av}\|_2 \\
&\vdots \\
\mathbf{v_k} &= \operatorname*{argmax}_{\substack{\|\mathbf{v}\|_2=1, \\ \mathbf{v}^T\mathbf{v}_i=0, i\in\{1,\ldots,k-1\}}} \|\mathbf{Av}\|_2
\end{aligned}$$

Further, any matrix may be written as

$$\mathbf{UDV}^T$$

where $\mathbf{U}$ *and* $\mathbf{V}$ *are orthogonal matrices containing the singular vectors and* $\mathbf{D}$ *is a diagonal matrix whose entries contain the singular values. This decomposition is known as the singular value decomposition.*

Proof. We prove this by induction. The base case was proven in Theorem 4. Assume that W_k is an optimal subspace with basis vectors $\mathbf{w_1}, \ldots, \mathbf{w_k}$, and let $\mathbf{W_k}$ be the matrix in $\mathbb{R}^{k \times d}$ with rows $\mathbf{w_1}, \ldots, \mathbf{w_k}$.

We then have

$$\|\mathbf{AW_k}\|_F^2 = \sum_{i=1}^{k} \|\mathbf{Aw_i}\|_2^2 .$$

By assumption $\|\mathbf{Aw_i}\|_2^2 \leq \|\mathbf{Av_i}\|_2^2$ for all $i \in \{1, \ldots, k-1\}$. Hence

$$\|\mathbf{AW_k}\|_F^2 \leq \sum_{i=1}^{k-1} \|\mathbf{Av_i}\|_2^2 + \|\mathbf{Aw_k}\|_2^2 .$$

By optimality of $\mathbf{W_k}$, we can conclude that $\mathbf{W_k}$ is in the subspace spanned by $\{\mathbf{v_1}, \ldots, \mathbf{v_{k-1}}, \mathbf{w_k}\}$. Then there exists a vector $\mathbf{x}$ in that subspace orthogonal to $\mathbf{v_1}, \ldots, \mathbf{v_{k-1}}$ such that

$$\|\mathbf{AW_k}\|_F^2 - \sum_{i=1}^{k-1} \|\mathbf{Av_i}\|_2^2 = \|\mathbf{Ax}\|_2^2 .$$

Since $\mathbf{v_k}$ optimized over the entire space (and not just the space $\mathbf{W_k}$, we know that $\|\mathbf{Av_k}\|_2^2 \geq \|\mathbf{Ax}\|_2^2$, which concludes the proof of the first statement.

For the second statement, we simply let $k = d$. □

To compute the best subspace, we may run the power method k times, projecting onto the orthogonal subspace whenever we converge.

Applications to k-means We first would like to consider the sister minimization problem. We first observe that by the Pythagorean theorem, for any unit vector v we have

$$\|\mathbf{A}\|_F^2 = \left\|\mathbf{Avv}^T\right\|_2^2 + \left\|\mathbf{A}(I - \mathbf{vv}^T)\right\|_F^2 = \|\mathbf{Av}\|_2^2 + \left\|\mathbf{A} - \mathbf{Avv}^T\right\|_F^2 .$$

Hence maximizing $\|\mathbf{Av}\|_2^2$ is equivalent to minimizing $\left\|\mathbf{A} - \mathbf{Avv}^T\right\|_F^2$. These notions straightforwardly extend to subspaces, that is,

$$\|\mathbf{A}\|_F^2 = \|\mathbf{AW}\|_2^2 + \left\|\mathbf{A} - \mathbf{AWW}^T\right\|_F^2 .$$

Minimizing

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

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

Let us now study the relationship between $\mathbf{A}$ and $\mathbf{A}\mathbf{V_k}\mathbf{V_k}^T$. We first note that $\mathbf{A}\mathbf{V_k}\mathbf{V_k}^T = \mathbf{U}\mathbf{D}\mathbf{V}^T\mathbf{V_k}\mathbf{V_k}^T = \mathbf{U}\mathbf{D}\mathbf{V_k}^T = \mathbf{U_k}\mathbf{D_k}\mathbf{V_k}^T = \mathbf{U_k}\mathbf{U_k}^T\mathbf{A}$, where $\mathbf{U_k}$ is obtained from $\mathbf{U}$ be just taking the first k singular vectors and setting everything else to 0 and $\mathbf{D_k}$ is the diagonal matrix obtained from $\mathbf{D}$ such that all but the first k diagonal entries of $\mathbf{D}$ were set to 0. Note that we may also remove the 0 entries from $\mathbf{U_k}$ just like we do in $\mathbf{V_k}$, that is, consider $\mathbf{U_k}$ to be a $n \times k$ matrix. Maximizing $\|\mathbf{A}\mathbf{W}\|_F^2$ for some rank k column subspace $\mathbf{W}$ is therefore equivalent to maximizing $\|\mathbf{T}\mathbf{A}\|_F^2$ for some rank k row subspace $\mathbf{T}$.

Now 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 $\mathbf{X}\mathbf{X}^T\mathbf{A}$. 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 X^j has a non-zero 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

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

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 7. *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 $\mathbf{A_k} = \mathbf{U}\mathbf{A}\mathbf{V_k}^T$ 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.