

## Topic 5: Principal component analysis

### 5.1 Covariance matrices

Suppose we are interested in a population whose members are represented by vectors in  $\mathbb{R}^d$ . We model the population as a probability distribution  $\mathbb{P}$  over  $\mathbb{R}^d$ , and let  $\mathbf{X}$  be a random vector with distribution  $\mathbb{P}$ . The mean of  $\mathbf{X}$  is the “center of mass” of  $\mathbb{P}$ . The covariance of  $\mathbf{X}$  is also a kind of “center of mass”, but it turns out to reveal quite a lot of other information.

Note: if we have a finite collection of data points  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \in \mathbb{R}^d$ , then it is common to arrange these vectors as rows of a matrix  $\mathbf{A} \in \mathbb{R}^{n \times d}$ . In this case, we can think of  $\mathbb{P}$  as the uniform distribution over the  $n$  points  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ . The mean of  $\mathbf{X} \sim \mathbb{P}$  can be written as

$$\mathbb{E}(\mathbf{X}) = \frac{1}{n} \mathbf{A}^\top \mathbf{1},$$

and the covariance of  $\mathbf{X}$  is

$$\text{cov}(\mathbf{X}) = \frac{1}{n} \mathbf{A}^\top \mathbf{A} - \left( \frac{1}{n} \mathbf{A}^\top \mathbf{1} \right) \left( \frac{1}{n} \mathbf{A}^\top \mathbf{1} \right)^\top = \frac{1}{n} \tilde{\mathbf{A}}^\top \tilde{\mathbf{A}}$$

where  $\tilde{\mathbf{A}} = \mathbf{A} - (1/n) \mathbf{1} \mathbf{1}^\top \mathbf{A}$ . We often call these the *empirical mean* and *empirical covariance* of the data  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ .

Covariance matrices are symmetric and positive semidefinite, since for any

$$\mathbf{z}^\top \text{cov}(\mathbf{X}) \mathbf{z} = \mathbf{z}^\top \mathbb{E}[(\mathbf{X} - \mathbb{E}(\mathbf{X}))(\mathbf{X} - \mathbb{E}(\mathbf{X}))^\top] \mathbf{z} = \mathbb{E}[(\mathbf{z}^\top (\mathbf{X} - \mathbb{E}(\mathbf{X})))^2] \geq 0.$$

This also shows that for any unit vector  $\mathbf{u}$ , the variance of

$$\text{var}(\langle \mathbf{u}, \mathbf{X} \rangle) = \mathbb{E}[\langle \mathbf{u}, \mathbf{X} - \mathbb{E} \mathbf{X} \rangle^2] = \mathbf{u}^\top \text{cov}(\mathbf{X}) \mathbf{u}.$$

Consider the following question: in what direction does  $\mathbf{X}$  have the highest variance? It turns out this is given by an eigenvector corresponding to the largest eigenvalue of  $\text{cov}(\mathbf{X})$ . This follows the following *variational* characterization of eigenvalues of symmetric matrices.

**Theorem 5.1.** Let  $\mathbf{M} \in \mathbb{R}^{d \times d}$  be a symmetric matrix with eigenvalues  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$  and corresponding orthonormal eigenvectors  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_d$ . Then

$$\begin{aligned} \max_{\mathbf{u} \neq \mathbf{0}} \frac{\mathbf{u}^\top \mathbf{M} \mathbf{u}}{\mathbf{u}^\top \mathbf{u}} &= \lambda_1, \\ \min_{\mathbf{u} \neq \mathbf{0}} \frac{\mathbf{u}^\top \mathbf{M} \mathbf{u}}{\mathbf{u}^\top \mathbf{u}} &= \lambda_d. \end{aligned}$$

These are achieved by  $\mathbf{v}_1$  and  $\mathbf{v}_d$ , respectively. (The ratio  $\mathbf{u}^\top \mathbf{M} \mathbf{u} / \mathbf{u}^\top \mathbf{u}$  is called the Rayleigh quotient associated with  $\mathbf{M}$  in direction  $\mathbf{u}$ .)

*Proof.* Following Theorem 4.1, write the eigendecomposition of  $\mathbf{M}$  as  $\mathbf{M} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$  where  $\mathbf{V} = [\mathbf{v}_1 | \mathbf{v}_2 | \cdots | \mathbf{v}_d]$  is orthogonal and  $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_d)$  is diagonal. For any  $\mathbf{u} \neq \mathbf{0}$ ,

$$\begin{aligned} \frac{\mathbf{u}^\top \mathbf{M} \mathbf{u}}{\mathbf{u}^\top \mathbf{u}} &= \frac{\mathbf{u}^\top \mathbf{V} \mathbf{\Lambda} \mathbf{V}^\top \mathbf{u}}{\mathbf{u}^\top \mathbf{V} \mathbf{V}^\top \mathbf{u}} \quad (\text{since } \mathbf{V} \mathbf{V}^\top = \mathbf{I}) \\ &= \frac{\mathbf{w}^\top \mathbf{\Lambda} \mathbf{w}}{\mathbf{w}^\top \mathbf{w}} \quad (\text{using } \mathbf{w} := \mathbf{V}^\top \mathbf{u}) \\ &= \frac{w_1^2 \lambda_1 + w_2^2 \lambda_2 + \cdots + w_d^2 \lambda_d}{w_1^2 + w_2^2 + \cdots + w_d^2}. \end{aligned}$$

This final ratio represents a convex combination of the scalars  $\lambda_1, \lambda_2, \dots, \lambda_d$ . Its largest value is  $\lambda_1$ , achieved by  $\mathbf{w} = \mathbf{e}_1$  (and hence  $\mathbf{u} = \mathbf{V} \mathbf{e}_1 = \mathbf{v}_1$ ), and its smallest value is  $\lambda_d$ , achieved by  $\mathbf{w} = \mathbf{e}_d$  (and hence  $\mathbf{u} = \mathbf{V} \mathbf{e}_d = \mathbf{v}_d$ ).  $\square$

**Corollary 5.1.** Let  $\mathbf{v}_1$  be a unit-length eigenvector of  $\text{cov}(\mathbf{X})$  corresponding to the largest eigenvalue of  $\text{cov}(\mathbf{X})$ . Then

$$\text{var}(\langle \mathbf{v}_1, \mathbf{X} \rangle) = \max_{\mathbf{u} \in S^{d-1}} \text{var}(\langle \mathbf{u}, \mathbf{X} \rangle).$$

Now suppose we are interested in the  $k$ -dimensional subspace of  $\mathbb{R}^d$  that captures the “most” variance of  $\mathbf{X}$ . Recall that a  $k$ -dimensional subspace  $W \subseteq \mathbb{R}^d$  can always be specified by a collection of  $k$  orthonormal vectors  $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k \in W$ . By the orthogonal projection to  $W$ , we mean the linear map

$$\mathbf{x} \mapsto \mathbf{U} \mathbf{U}^\top \mathbf{x}$$

The covariance of  $\mathbf{U}^\top \mathbf{X}$ , a  $k \times k$  covariance matrix, is sim

$$\text{cov}(\mathbf{U}^\top \mathbf{X}) = \mathbf{U}^\top \text{cov}(\mathbf{X}) \mathbf{U}$$

The “total” variance in this subspace is often measured by the trace of the covariance:  $\text{tr}(\text{cov}(\mathbf{U}^\top \mathbf{X}))$ . Recall, the *trace* of a square matrix is the sum of its diagonal entries, and it is a linear function.

**Fact 5.1.** For any  $\mathbf{U} \in \mathbb{R}^{d \times k}$ ,  $\text{tr}(\text{cov}(\mathbf{U}^\top \mathbf{X})) = \mathbb{E} \|\mathbf{U}^\top (\mathbf{X} - \mathbb{E}(\mathbf{X}))\|_2^2$ . Furthermore, if  $\mathbf{U}^\top \mathbf{U} = \mathbf{I}$ , then  $\text{tr}(\text{cov}(\mathbf{U}^\top \mathbf{X})) = \mathbb{E} \|\mathbf{U} \mathbf{U}^\top (\mathbf{X} - \mathbb{E}(\mathbf{X}))\|_2^2$ .

**Theorem 5.2.** Let  $\mathbf{M} \in \mathbb{R}^{d \times d}$  be a symmetric matrix with eigenvalues  $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$  and corresponding orthonormal eigenvectors  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_d$ . Then for any  $k \in [d]$ ,

$$\begin{aligned} \max_{\mathbf{U} \in \mathbb{R}^{d \times k} : \mathbf{U}^\top \mathbf{U} = \mathbf{I}} \text{tr}(\mathbf{U}^\top \mathbf{M} \mathbf{U}) &= \lambda_1 + \lambda_2 + \cdots + \lambda_k, \\ \min_{\mathbf{U} \in \mathbb{R}^{d \times k} : \mathbf{U}^\top \mathbf{U} = \mathbf{I}} \text{tr}(\mathbf{U}^\top \mathbf{M} \mathbf{U}) &= \lambda_{d-k+1} + \lambda_{d-k+2} + \cdots + \lambda_d. \end{aligned}$$

The max is achieved by an orthogonal projection to the span of  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$ , and the min is achieved by an orthogonal projection to the span of  $\mathbf{v}_{d-k+1}, \mathbf{v}_{d-k+2}, \dots, \mathbf{v}_d$ .

*Proof.* Let  $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k$  denote the columns of  $\mathbf{U}$ . Then, writing  $\mathbf{M} = \sum_{j=1}^d \lambda_j \mathbf{v}_j \mathbf{v}_j^\top$  (Theorem 4.1),

$$\text{tr}(\mathbf{U}^\top \mathbf{M} \mathbf{U}) = \sum_{i=1}^k \mathbf{u}_i^\top \mathbf{M} \mathbf{u}_i = \sum_{i=1}^k \mathbf{u}_i^\top \left( \sum_{j=1}^d \lambda_j \mathbf{v}_j \mathbf{v}_j^\top \right) \mathbf{u}_i = \sum_{j=1}^d \lambda_j \sum_{i=1}^k \langle \mathbf{v}_j, \mathbf{u}_i \rangle^2 = \sum_{j=1}^d c_j \lambda_j$$

where  $c_j := \sum_{i=1}^k \langle \mathbf{v}_j, \mathbf{u}_i \rangle^2$  for each  $j \in [d]$ . We'll show that each  $c_j \in [0, 1]$ , and  $\sum_{j=1}^d c_j = k$ .

First, it is clear that  $c_j \geq 0$  for each  $j \in [d]$ . Next, extending  $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k$  to an orthonormal basis  $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_d$  for  $\mathbb{R}^d$ , we have for each  $j \in [d]$ ,

$$c_j = \sum_{i=1}^k \langle \mathbf{v}_j, \mathbf{u}_i \rangle^2 \leq \sum_{i=1}^d \langle \mathbf{v}_j, \mathbf{u}_i \rangle^2 = 1.$$

Finally, since  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_d$  is an orthonormal basis for  $\mathbb{R}^d$ ,

$$\sum_{j=1}^d c_j = \sum_{j=1}^d \sum_{i=1}^k \langle \mathbf{v}_j, \mathbf{u}_i \rangle^2 = \sum_{i=1}^k \sum_{j=1}^d \langle \mathbf{v}_j, \mathbf{u}_i \rangle^2 = \sum_{i=1}^k \|\mathbf{u}_i\|_2^2 = k.$$

The maximum value of  $\sum_{j=1}^d c_j \lambda_j$  over all choices of  $c_1, c_2, \dots, c_d \in [0, 1]$  with  $\sum_{j=1}^d c_j = k$  is  $\lambda_1 + \lambda_2 + \dots + \lambda_k$ . This is achieved when  $c_1 = c_2 = \dots = c_k = 1$  and  $c_{k+1} = \dots = c_d = 0$ , i.e., when  $\text{span}(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k) = \text{span}(\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k)$ . The minimum value of  $\sum_{j=1}^d c_j \lambda_j$  over all choices of  $c_1, c_2, \dots, c_d \in [0, 1]$  with  $\sum_{j=1}^d c_j = k$  is  $\lambda_{d-k+1} + \lambda_{d-k+2} + \dots + \lambda_d$ . This is achieved when  $c_1 = \dots = c_{d-k} = 0$  and  $c_{d-k+1} = c_{d-k+2} = \dots = c_d = 1$ , i.e., when  $\text{span}(\mathbf{v}_{d-k+1}, \mathbf{v}_{d-k+2}, \dots, \mathbf{v}_d) = \text{span}(\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k)$ .  $\square$

We'll refer to the  $k$  largest eigenvalues of a symmetric matrix  $M$  as the *top- $k$  eigenvalues* of  $M$ , and the  $k$  smallest eigenvalues as the *bottom- $k$  eigenvalues* of  $M$ . We analogously use the term *top- $k$  (resp., bottom- $k$ ) eigenvectors* to refer to orthonormal eigenvectors corresponding to the top- $k$  (resp., bottom- $k$ ) eigenvalues. These eigenvectors are not necessarily unique.

**Corollary 5.2.** Let  $\mathbf{V}_k := [\mathbf{v}_1 | \mathbf{v}_2 | \dots | \mathbf{v}_k]$ . Then

$$\text{tr}(\text{cov}(\mathbf{V}_k^\top \mathbf{X})) = \max_{\mathbf{U} \in \mathbb{R}^{d \times k}, \mathbf{U}^\top \mathbf{U} = \mathbf{I}} \text{tr}(\text{cov}(\mathbf{U}^\top \mathbf{X})).$$

An orthogonal projection given by top- $k$  eigenvectors of  $\text{cov}(\mathbf{X})$  is called a (*rank- $k$  principal component analysis (PCA) projection*). Corollary 5.2 reveals an important property of a PCA projection: it maximizes the variance captured by the subspace.

## 5.2 Best affine and linear subspaces

PCA has another important property: it gives an affine subspace  $A \subseteq \mathbb{R}^d$  that minimizes the expected squared distance between  $\mathbf{X}$  and  $A$ .

Recall that a  $k$ -dimensional *affine subspace*  $A$  is specified by a  $k$ -dimensional (linear) subspace  $W \subseteq \mathbb{R}^d$ —say, with orthonormal basis  $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k$ —and a displacement vector  $\mathbf{u}_0 \in \mathbb{R}^d$ :

$$A = \{\mathbf{u}_0 + \alpha_1 \mathbf{u}_1 + \alpha_2 \mathbf{u}_2 + \dots + \alpha_k \mathbf{u}_k : \alpha_1, \alpha_2, \dots, \alpha_k \in \mathbb{R}\}.$$

Let  $\mathbf{U} := [\mathbf{u}_1 | \mathbf{u}_2 | \dots | \mathbf{u}_k]$ . Then, for any  $\mathbf{x} \in \mathbb{R}^d$ , the point in  $A$  closest to  $\mathbf{x}$  is given by  $\mathbf{u}_0 + \mathbf{U}\mathbf{U}^\top(\mathbf{x} - \mathbf{u}_0)$ , and hence the squared distance from  $\mathbf{x}$  to  $A$  is  $\|(I - \mathbf{U}\mathbf{U}^\top)(\mathbf{x} - \mathbf{u}_0)\|_2^2$ .

**Theorem 5.3.** Let  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$  be top- $k$  eigenvectors of  $\text{cov}(\mathbf{X})$ , let  $\mathbf{V}_k := [\mathbf{v}_1 | \mathbf{v}_2 | \dots | \mathbf{v}_k]$ , and  $\mathbf{v}_0 := \mathbb{E}(\mathbf{X})$ . Then

$$\mathbb{E} \|(I - \mathbf{V}_k \mathbf{V}_k^\top)(\mathbf{X} - \mathbf{v}_0)\|_2^2 = \min_{\substack{\mathbf{U} \in \mathbb{R}^{d \times k}, \mathbf{u}_0 \in \mathbb{R}^d, \\ \mathbf{U}^\top \mathbf{U} = \mathbf{I}}} \mathbb{E} \|(I - \mathbf{U}\mathbf{U}^\top)(\mathbf{X} - \mathbf{u}_0)\|_2^2.$$

*Proof.* For any matrix  $d \times d$  matrix  $\mathbf{M}$ , the function  $\mathbf{u}_0 \mapsto \mathbb{E} \|\mathbf{M}(\mathbf{X} - \mathbf{u}_0)\|_2^2$  is minimized when  $\mathbf{M}\mathbf{u}_0 = \mathbf{M}\mathbb{E}(\mathbf{X})$  (Fact 5.2). Therefore, we can plug-in  $\mathbb{E}(\mathbf{X})$  for  $\mathbf{u}_0$  in the minimization problem, whereupon it reduces to

$$\min_{\mathbf{U} \in \mathbb{R}^{d \times k} : \mathbf{U}^\top \mathbf{U} = \mathbf{I}} \mathbb{E} \|(\mathbf{I} - \mathbf{U}\mathbf{U}^\top)(\mathbf{X} - \mathbb{E}(\mathbf{X}))\|_2^2.$$

The objective function is equivalent to

$$\begin{aligned} \mathbb{E} \|(\mathbf{I} - \mathbf{U}\mathbf{U}^\top)(\mathbf{X} - \mathbb{E}(\mathbf{X}))\|_2^2 &= \mathbb{E} \|\mathbf{X} - \mathbb{E}(\mathbf{X})\|_2^2 - \mathbb{E} \|\mathbf{U}\mathbf{U}^\top(\mathbf{X} - \mathbb{E}(\mathbf{X}))\|_2^2 \\ &= \mathbb{E} \|\mathbf{X} - \mathbb{E}(\mathbf{X})\|_2^2 - \text{tr}(\text{cov}(\mathbf{U}^\top \mathbf{X})), \end{aligned}$$

where the second equality comes from Fact 5.1. Therefore, minimizing the objective is equivalent to maximizing  $\text{tr}(\text{cov}(\mathbf{U}^\top \mathbf{X}))$ , which is achieved by PCA (Corollary 5.2).  $\square$

The proof of Theorem 5.3 depends on the following simple but useful fact.

**Fact 5.2** (Bias-variance decomposition). *Let  $\mathbf{Y}$  be a random vector in  $\mathbb{R}^d$ , and  $\mathbf{b} \in \mathbb{R}^d$  be any fixed vector. Then*

$$\mathbb{E} \|\mathbf{Y} - \mathbf{b}\|_2^2 = \mathbb{E} \|\mathbf{Y} - \mathbb{E}(\mathbf{Y})\|_2^2 + \|\mathbb{E}(\mathbf{Y}) - \mathbf{b}\|_2^2$$

(which, as a function of  $\mathbf{b}$ , is minimized when  $\mathbf{b} = \mathbb{E}(\mathbf{Y})$ ).

A similar statement can be made about (linear) subspaces by using top- $k$  eigenvectors of  $\mathbb{E}(\mathbf{X}\mathbf{X}^\top)$  instead of  $\text{cov}$ .

**Theorem 5.4.** *Let  $\mathbf{v}_k := [\mathbf{v}_1 | \mathbf{v}_2 | \dots | \mathbf{v}_k]$ . Then*

$$\mathbb{E} \|(\mathbf{I} - \mathbf{V}_k \mathbf{V}_k^\top) \mathbf{X}\|_2^2 = \min_{\mathbf{U} \in \mathbb{R}^{d \times k} : \mathbf{U}^\top \mathbf{U} = \mathbf{I}} \mathbb{E} \|\mathbf{X} - \mathbf{U} \mathbf{U}^\top \mathbf{X}\|_2^2.$$

### 5.3 Noisy affine subspace recovery

Suppose there are  $n$  points  $\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_n \in \mathbb{R}^d$  that lie on an affine subspace  $A_\star$  of dimension  $k$ . In this scenario, you don't directly observe the  $\mathbf{t}_i$ ; rather, you only observe noisy versions of these points:  $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n$ , where for some  $\sigma_1, \sigma_2, \dots, \sigma_n > 0$ ,

$$\mathbf{Y}_j \sim \mathcal{N}(\mathbf{t}_j, \sigma_j^2 \mathbf{I}) \quad \text{for all } j \in [n]$$

and  $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n$  are independent. The observations  $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n$  no longer all lie in the affine subspace  $A_\star$ , but by applying PCA to the empirical covariance of  $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n$ , you can hope to approximately recover  $A_\star$ .

Regard  $\mathbf{X}$  as a random vector whose conditional distribution given the noisy points is uniform over  $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n$ . In fact, the distribution of  $\mathbf{X}$  is given by the following generative process:

1. Draw  $J \in [n]$  uniformly at random.
2. Given  $J$ , draw  $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \sigma_J^2 \mathbf{I})$ .
3. Set  $\mathbf{X} := \mathbf{t}_J + \mathbf{Z}$ .

Note that the empirical covariance based on  $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n$  is not exactly  $\text{cov}(\mathbf{X})$ , but it can be a good approximation when  $n$  is large (with high probability). Similarly, the empirical average of  $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n$  is a good approximation to  $\mathbb{E}(\mathbf{X})$  when  $n$  is large (with high probability). So here, we assume for simplicity that both  $\text{cov}(\mathbf{X})$  and  $\mathbb{E}(\mathbf{X})$  are known exactly. We show that PCA produces a  $k$ -dimensional affine subspace that contains all of the  $\mathbf{t}_j$ .

**Theorem 5.5.** *Let  $\mathbf{X}$  be the random vector as defined above,  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$  be top- $k$  eigenvectors of  $\text{cov}(\mathbf{X})$ , and  $\mathbf{v}_0 := \mathbb{E}(\mathbf{X})$ . Then the affine subspace*

$$\hat{A} := \{\mathbf{v}_0 + \alpha_1 \mathbf{v}_1 + \alpha_2 \mathbf{v}_2 + \dots + \alpha_k \mathbf{v}_k : \alpha_1, \alpha_2, \dots, \alpha_k \in \mathbb{R}\}$$

*contains  $\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_n$ .*

*Proof.* Theorem 5.3 says that the matrix  $\mathbf{V}_k := [\mathbf{v}_1 | \mathbf{v}_2 | \dots | \mathbf{v}_k]$  minimizes  $\mathbb{E} \|(\mathbf{I} - \mathbf{U}\mathbf{U}^\top)(\mathbf{X} - \mathbf{v}_0)\|_2^2$  (as a function of  $\mathbf{U} \in \mathbb{R}^{d \times k}$ , subject to  $\mathbf{U}^\top \mathbf{U} = \mathbf{I}$ ), or equivalently, maximizes  $\text{tr}(\text{cov}(\mathbf{U}^\top \mathbf{X}))$ . This maximization objective can be written as

$$\begin{aligned} \text{tr}(\text{cov}(\mathbf{U}^\top \mathbf{X})) &= \mathbb{E} \|\mathbf{U}\mathbf{U}^\top(\mathbf{X} - \mathbf{v}_0)\|_2^2 \quad (\text{by Fact 5.1}) \\ &= \frac{1}{n} \sum_{j=1}^n \mathbb{E} \left[ \|\mathbf{U}\mathbf{U}^\top(\mathbf{t}_j - \mathbf{v}_0 + \mathbf{Z})\|_2^2 \mid J = j \right] \\ &= \frac{1}{n} \sum_{j=1}^n \mathbb{E} \left[ \|\mathbf{U}\mathbf{U}^\top(\mathbf{t}_j - \mathbf{v}_0)\|_2^2 + 2 \langle \mathbf{U}\mathbf{U}^\top(\mathbf{t}_j - \mathbf{v}_0), \mathbf{U}\mathbf{U}^\top \mathbf{Z} \rangle + \|\mathbf{U}\mathbf{U}^\top \mathbf{Z}\|_2^2 \mid J = j \right] \\ &= \frac{1}{n} \sum_{j=1}^n \left\{ \|\mathbf{U}\mathbf{U}^\top(\mathbf{t}_j - \mathbf{v}_0)\|_2^2 + k\sigma_j^2 \right\}, \end{aligned}$$

where the penultimate step uses the fact that the conditional distribution of  $\mathbf{Z}$  given  $J = j$  is  $N(\mathbf{0}, \sigma_j^2 \mathbf{I})$ , and the final step uses the fact that  $\|\mathbf{U}\mathbf{U}^\top \mathbf{Z}\|_2^2$  has the same conditional distribution (given  $J = j$ ) as the squared length of a  $N(\mathbf{0}, \sigma_j^2 \mathbf{I})$  random vector in  $\mathbb{R}^k$ . Since  $\mathbf{U}\mathbf{U}^\top(\mathbf{t}_j - \mathbf{v}_0)$  is the orthogonal projection of  $\mathbf{t}_j - \mathbf{v}_0$  onto the subspace spanned by the columns of  $\mathbf{U}$  (call it  $W$ ),

$$\|\mathbf{U}\mathbf{U}^\top(\mathbf{t}_j - \mathbf{v}_0)\|_2^2 \leq \|\mathbf{t}_j - \mathbf{v}_0\|_2^2 \quad \text{for all } j \in [n].$$

The inequalities above are equalities precisely when  $\mathbf{t}_j - \mathbf{v}_0 \in W$  for all  $j \in [n]$ . This is indeed the case for the subspace  $A_\star - \{\mathbf{v}_0\}$ . Since  $\mathbf{V}_k$  maximizes the objective, its columns must span a  $k$ -dimensional subspace  $\widehat{W}$  that also contains all of the  $\mathbf{t}_j - \mathbf{v}_0$ ; hence the affine subspace  $\hat{A} = \{\mathbf{v}_0 + \mathbf{x} : \mathbf{x} \in \widehat{W}\}$  contains all of the  $\mathbf{t}_j$ .  $\square$

## 5.4 Singular value decomposition

Let  $\mathbf{A}$  be any  $n \times d$  matrix. Our aim is to define an extremely useful decomposition of  $\mathbf{A}$  called the *singular value decomposition (SVD)*. Our derivation starts by considering two related matrices,  $\mathbf{A}^\top \mathbf{A}$  and  $\mathbf{A}\mathbf{A}^\top$ ; their eigendecompositions will lead to the SVD of  $\mathbf{A}$ .

**Fact 5.3.**  $\mathbf{A}^\top \mathbf{A}$  and  $\mathbf{A}\mathbf{A}^\top$  are symmetric and positive semidefinite.

It is clear that the eigenvalues of  $\mathbf{A}^\top \mathbf{A}$  and  $\mathbf{A}\mathbf{A}^\top$  are non-negative. In fact, the non-zero eigenvalues of  $\mathbf{A}^\top \mathbf{A}$  and  $\mathbf{A}\mathbf{A}^\top$  are exactly the same.

**Lemma 5.1.** *Let  $\lambda$  be an eigenvalue of  $\mathbf{A}^\top \mathbf{A}$  with corresponding eigenvector  $\mathbf{v}$ .*

- *If  $\lambda > 0$ , then  $\lambda$  is a non-zero eigenvalue of  $\mathbf{A}\mathbf{A}^\top$  with corresponding eigenvector  $\mathbf{A}\mathbf{v}$ .*
- *If  $\lambda = 0$ , then  $\mathbf{A}\mathbf{v} = \mathbf{0}$ .*

*Proof.* First suppose  $\lambda > 0$ . Then

$$\mathbf{A}\mathbf{A}^\top(\mathbf{A}\mathbf{v}) = \mathbf{A}(\mathbf{A}^\top \mathbf{A}\mathbf{v}) = \mathbf{A}(\lambda \mathbf{v}) = \lambda(\mathbf{A}\mathbf{v}),$$

so  $\lambda$  is an eigenvalue of  $\mathbf{A}\mathbf{A}^\top$  with corresponding eigenvector  $\mathbf{A}\mathbf{v}$ .

Now suppose  $\lambda = 0$  (which is the only remaining case, as per Fact 5.3). Then

$$\|\mathbf{A}\mathbf{v}\|_2^2 = \mathbf{v}^\top \mathbf{A}^\top \mathbf{A}\mathbf{v} = \mathbf{v}^\top (\lambda \mathbf{v}) = 0.$$

Since only the zero vector has length 0, it must be that  $\mathbf{A}\mathbf{v} = \mathbf{0}$ . □

(We can apply Lemma 5.1 to both  $\mathbf{A}$  and  $\mathbf{A}^\top$  to conclude that  $\mathbf{A}^\top \mathbf{A}$  and  $\mathbf{A}\mathbf{A}^\top$  have the same non-zero eigenvalues.)

**Theorem 5.6** (Singular value decomposition). *Let  $\mathbf{A}$  be an  $n \times d$  matrix. Let  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_d \in \mathbb{R}^d$  be orthonormal eigenvectors of  $\mathbf{A}^\top \mathbf{A}$  with eigenvalues  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d \geq 0$ . Let  $r$  be the number of positive eigenvalues.*

$$\mathbf{u}_i := \frac{\mathbf{A}\mathbf{v}_i}{\|\mathbf{A}\mathbf{v}_i\|_2} = \frac{\mathbf{A}\mathbf{v}_i}{\sqrt{\mathbf{v}_i^\top \mathbf{A}^\top \mathbf{A}\mathbf{v}_i}} = \frac{\mathbf{A}\mathbf{v}_i}{\sqrt{\lambda_i}} \quad \text{for each } i \in [r].$$

Then

$$\mathbf{A} = \underbrace{\begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_r \\ \downarrow & \downarrow & & \downarrow \end{bmatrix}}_{n \times r} \underbrace{\begin{bmatrix} \sqrt{\lambda_1} & & & \\ & \sqrt{\lambda_2} & & \\ & & \ddots & \\ & & & \sqrt{\lambda_r} \end{bmatrix}}_{r \times r} \underbrace{\begin{bmatrix} \leftarrow & \mathbf{v}_1^\top & \rightarrow \\ \leftarrow & \mathbf{v}_2^\top & \rightarrow \\ & \vdots & \\ \leftarrow & \mathbf{v}_r^\top & \rightarrow \end{bmatrix}}_{r \times d} = \sum_{i=1}^r \sqrt{\lambda_i} \mathbf{u}_i \mathbf{v}_i^\top,$$

and  $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_r$  are orthonormal.

*Proof.* It suffices to show that for some set of  $d$  linearly independent vectors  $\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_d \in \mathbb{R}^d$ ,

$$\mathbf{A}\mathbf{q}_j = \left( \sum_{i=1}^r \sqrt{\lambda_i} \mathbf{u}_i \mathbf{v}_i^\top \right) \mathbf{q}_j \quad \text{for all } j \in [d].$$

We'll use  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_d$ . Observe that

$$\mathbf{A}\mathbf{v}_j = \begin{cases} \sqrt{\lambda_j} \mathbf{u}_j & \text{if } 1 \leq j \leq r, \\ \mathbf{0} & \text{if } r < j \leq d, \end{cases}$$

by definition of  $\mathbf{u}_i$  and by Lemma 5.1. Moreover,

$$\left( \sum_{i=1}^r \sqrt{\lambda_i} \mathbf{u}_i \mathbf{v}_i^\top \right) \mathbf{v}_j = \sum_{i=1}^r \sqrt{\lambda_i} \langle \mathbf{v}_j, \mathbf{v}_i \rangle \mathbf{u}_i = \begin{cases} \sqrt{\lambda_j} \mathbf{u}_j & \text{if } 1 \leq j \leq r, \\ \mathbf{0} & \text{if } r < j \leq d, \end{cases}$$

since  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_d$  are orthonormal. We conclude that  $\mathbf{A}\mathbf{v}_j = (\sum_{i=1}^r \sqrt{\lambda_i} \mathbf{u}_i \mathbf{v}_i^\top) \mathbf{v}_j$  for all  $j \in [d]$ , and hence  $\mathbf{A} = \sum_{i=1}^r \sqrt{\lambda_i} \mathbf{u}_i \mathbf{v}_i^\top$ .

Note that

$$\mathbf{u}_i^\top \mathbf{u}_j = \frac{\mathbf{v}_i^\top \mathbf{A}^\top \mathbf{A} \mathbf{v}_j}{\sqrt{\lambda_i \lambda_j}} = \frac{\lambda_j \mathbf{v}_i^\top \mathbf{v}_j}{\sqrt{\lambda_i \lambda_j}} = 0 \quad \text{for all } 1 \leq i < j \leq r,$$

where the last step follows since  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_d$  are orthonormal. This implies that  $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_r$  are orthonormal.  $\square$

The decomposition of  $\mathbf{A}$  into the sum  $\mathbf{A} = \sum_{i=1}^r \sqrt{\lambda_i} \mathbf{u}_i \mathbf{v}_i^\top$  from Theorem 5.6 is called the *singular value decomposition (SVD)* of  $\mathbf{A}$ . The  $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_r$  are the *left singular vectors*, and the  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r$  are the *right singular vectors*. The scalars  $\sqrt{\lambda_1} \geq \sqrt{\lambda_2} \geq \dots \geq \sqrt{\lambda_r}$  are the (positive) *singular values* corresponding to the left/right singular vectors  $(\mathbf{u}_1, \mathbf{v}_1), (\mathbf{u}_2, \mathbf{v}_2), \dots, (\mathbf{u}_r, \mathbf{v}_r)$ . The representation  $\mathbf{A} = \sum_{i=1}^r \sqrt{\lambda_i} \mathbf{u}_i \mathbf{v}_i^\top$  is actually typically called the *thin SVD* of  $\mathbf{A}$ . The number  $r$  of positive  $\lambda_i$  is the *rank* of  $\mathbf{A}$ , which is at most the smaller of  $n$  and  $d$ .

Of course, one can extend  $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_r$  to an orthonormal basis for  $\mathbb{R}^n$ . Define the matrices  $\mathbf{U} := [\mathbf{u}_1 | \mathbf{u}_2 | \dots | \mathbf{u}_n] \in \mathbb{R}^{n \times n}$  and  $\mathbf{V} := [\mathbf{v}_1 \ \mathbf{v}_2 \ \dots \ \mathbf{v}_d] \in \mathbb{R}^{d \times d}$ . Also define  $\mathbf{S} \in \mathbb{R}^{n \times d}$  to be the matrix whose only non-zero en-  
tries are  $\sigma_1, \sigma_2, \dots, \sigma_r$ . Then  $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$ .

This matrix factorization and  $\mathbf{v}_{r+1}, \mathbf{v}_{r+2}, \dots, \mathbf{v}_n$  (value equal to zero.) <https://eduassistpro.github.io/> correspond to the singular vectors  $\mathbf{u}_{r+1}, \mathbf{u}_{r+2}, \dots, \mathbf{u}_n$

Just as before, we'll refer to the  $k$  largest singular values of  $\mathbf{A}$ , and the  $k$  smallest singular values as the *bottom- $k$*  singular values of  $\mathbf{A}$ , and the  $k$  largest singular values as the *top- $k$*  singular values of  $\mathbf{A}$ . We also use the term *top- $k$*  (resp., *bottom- $k$* ) *singular vectors* to refer to or corresponding to the top- $k$  (resp., bottom- $k$ ) singular values. Again, the choice of top- $k$  (or bottom- $k$ ) singular vectors is not necessarily unique.

## Relationship between PCA and SVD

As seen above, the eigenvectors of  $\mathbf{A}^\top \mathbf{A}$  are the right singular vectors  $\mathbf{A}$ , and the eigenvectors of  $\mathbf{A}\mathbf{A}^\top$  are the left singular vectors of  $\mathbf{A}$ .

Suppose there are  $n$  data points  $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n \in \mathbb{R}^d$ , arranged as the rows of the matrix  $\mathbf{A} \in \mathbb{R}^{n \times d}$ . Now regard  $\mathbf{X}$  as a random vector with the uniform distribution on the  $n$  data points. Then  $\mathbb{E}(\mathbf{X}\mathbf{X}^\top) = \frac{1}{n} \sum_{i=1}^n \mathbf{a}_i \mathbf{a}_i^\top = \frac{1}{n} \mathbf{A}^\top \mathbf{A}$ : top- $k$  eigenvectors of  $\frac{1}{n} \mathbf{A}^\top \mathbf{A}$  are top- $k$  right singular vectors of  $\mathbf{A}$ . Hence, rank- $k$  uncentered PCA (as in Theorem 5.4) corresponds to the subspace spanned by the top- $k$  right singular vectors of  $\mathbf{A}$ .

## Variational characterization of singular values

Given the relationship between the singular values of  $\mathbf{A}$  and the eigenvalues of  $\mathbf{A}^\top \mathbf{A}$  and  $\mathbf{A} \mathbf{A}^\top$ , it is easy to obtain variational characterizations of the singular values. We can also obtain the characterization directly.

**Fact 5.4.** Let the SVD of a matrix  $\mathbf{A} \in \mathbb{R}^{n \times d}$  be given by  $\mathbf{A} = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$ , where  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$ . For each  $i \in [r]$ ,

$$\sigma_i = \max_{\substack{\mathbf{x} \in S^{d-1}: \langle \mathbf{v}_j, \mathbf{x} \rangle = 0 \forall j < i \\ \mathbf{y} \in S^{n-1}: \langle \mathbf{u}_j, \mathbf{y} \rangle = 0 \forall j < i}} \mathbf{y}^\top \mathbf{A} \mathbf{x} = \mathbf{u}_i^\top \mathbf{A} \mathbf{v}_i.$$

### Relationship between eigendecomposition and SVD

If  $\mathbf{M} \in \mathbb{R}^{d \times d}$  is symmetric and has eigendecomposition  $\mathbf{M} = \sum_{i=1}^d \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$ , then its singular values are the absolute values of the  $\lambda_i$ . We can take  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_d$  as corresponding right singular vectors. For corresponding left singular vectors, we can take  $\mathbf{u}_i := \mathbf{v}_i$  whenever  $\lambda_i \geq 0$  (which is the case for all  $i$  if  $\mathbf{M}$  is also psd), and  $\mathbf{u}_i := -\mathbf{v}_i$  whenever  $\lambda_i < 0$ . Therefore, we have the following variational characterization of the singular values of  $\mathbf{M}$ .

**Fact 5.5.** Let the eigendecomposition of a symmetric matrix  $\mathbf{M} \in \mathbb{R}^{d \times d}$  be given by  $\mathbf{M} = \sum_{i=1}^d \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$ , where  $|\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_d|$ . For each  $i \in [d]$ ,

$$|\lambda_i| = \max_{\substack{\mathbf{x} \in S^{d-1}: \langle \mathbf{v}_j, \mathbf{x} \rangle = 0 \forall j < i \\ \mathbf{y} \in S^{d-1}: \langle \mathbf{v}_j, \mathbf{y} \rangle = 0 \forall j < i}} \mathbf{y}^\top \mathbf{M} \mathbf{x} = \max_{\mathbf{x} \in S^{d-1}: \langle \mathbf{v}_j, \mathbf{x} \rangle = 0 \forall j < i} |\mathbf{x}^\top \mathbf{M} \mathbf{x}| = |\mathbf{v}_i^\top \mathbf{M} \mathbf{v}_i|.$$

### Moore-Penrose pseudoinverse

The SVD defines a kind of matrix

(where possibly  $n \neq d$ ) satisfy  $\mathbf{U}^\top \mathbf{U} = \mathbf{V}^\top \mathbf{V} = \mathbf{I}$

of  $\mathbf{A}$  is  $r$ . The Moore-Penrose pseudoinverse of  $\mathbf{A}$  is given by

$\mathbf{A} \in \mathbb{R}^{n \times d}$   
 $\mathbf{U} \in \mathbb{R}^{n \times r}$  and  $\mathbf{V} \in \mathbb{R}^{d \times r}$   
 es. Here, the rank

Add WeChat edu\_assist\_pro

Note that  $\mathbf{A}^\dagger$  is well-defined:  $\mathbf{S}$  is invertible because its diagonal entries are all strictly positive. What is the effect of multiplying  $\mathbf{A}$  by  $\mathbf{A}^\dagger$  on the left? Using the SVD of  $\mathbf{A}$ ,

$$\mathbf{A}^\dagger \mathbf{A} = \mathbf{V} \mathbf{S}^{-1} \mathbf{U}^\top \mathbf{U} \mathbf{S} \mathbf{V}^\top = \mathbf{V} \mathbf{V}^\top \in \mathbb{R}^{d \times d},$$

which is the orthogonal projection to the row space of  $\mathbf{A}$ . In particular, this means that

$$\mathbf{A} \mathbf{A}^\dagger \mathbf{A} = \mathbf{A}.$$

Similarly,  $\mathbf{A} \mathbf{A}^\dagger = \mathbf{U} \mathbf{U}^\top \in \mathbb{R}^{n \times n}$ , the orthogonal projection to the column space of  $\mathbf{A}$ . Note that if  $r = d$ , then  $\mathbf{A}^\dagger \mathbf{A} = \mathbf{I}$ , as the row space of  $\mathbf{A}$  is simply  $\mathbb{R}^d$ ; similarly, if  $r = n$ , then  $\mathbf{A} \mathbf{A}^\dagger = \mathbf{I}$ .

The Moore-Penrose pseudoinverse is also related to least squares. For any  $\mathbf{y} \in \mathbb{R}^n$ , the vector  $\mathbf{A} \mathbf{A}^\dagger \mathbf{y}$  is the orthogonal projection of  $\mathbf{y}$  onto the column space of  $\mathbf{A}$ . This means that  $\min_{\mathbf{x} \in \mathbb{R}^d} \|\mathbf{A} \mathbf{x} - \mathbf{y}\|_2^2$  is minimized by  $\mathbf{x} = \mathbf{A}^\dagger \mathbf{y}$ . The more familiar expression for the least squares solution  $\mathbf{x} = (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{y}$  only applies in the special case where  $\mathbf{A}^\top \mathbf{A}$  is invertible. The connection to the general form of a solution can be seen by using the easily verified identity

$$\mathbf{A}^\dagger = (\mathbf{A}^\top \mathbf{A})^\dagger \mathbf{A}^\top$$

and using the fact that  $(\mathbf{A}^\top \mathbf{A})^\dagger = (\mathbf{A}^\top \mathbf{A})^{-1}$  when  $\mathbf{A}^\top \mathbf{A}$  is invertible.



## 5.5 Matrix norms and low rank SVD

### Matrix inner product and the Frobenius norm

The space of  $n \times d$  real matrices is a real vector space in its own right, and it can, in fact, be viewed as a Euclidean space with inner product given by  $\langle \mathbf{X}, \mathbf{Y} \rangle := \text{tr}(\mathbf{X}^\top \mathbf{Y})$ . It can be checked that this indeed is a valid inner product. For instance, the fact that the trace function is linear can be used to establish linearity in the first argument:

$$\begin{aligned} \langle c\mathbf{X} + \mathbf{Y}, \mathbf{Z} \rangle &= \text{tr}((c\mathbf{X} + \mathbf{Y})^\top \mathbf{Z}) \\ &= \text{tr}(c\mathbf{X}^\top \mathbf{Z} + \mathbf{Y}^\top \mathbf{Z}) \\ &= c \text{tr}(\mathbf{X}^\top \mathbf{Z}) + \text{tr}(\mathbf{Y}^\top \mathbf{Z}) = c\langle \mathbf{X}, \mathbf{Z} \rangle + \langle \mathbf{Y}, \mathbf{Z} \rangle. \end{aligned}$$

The inner product naturally induces an associated norm  $\mathbf{X} \mapsto \sqrt{\langle \mathbf{X}, \mathbf{X} \rangle}$ . Viewing  $\mathbf{X} \in \mathbb{R}^{n \times d}$  as a data matrix whose rows are the vectors  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \in \mathbb{R}^d$ , we see that

$$\langle \mathbf{X}, \mathbf{X} \rangle = \text{tr}(\mathbf{X}^\top \mathbf{X}) = \text{tr}\left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top\right) = \sum_{i=1}^n \text{tr}(\mathbf{x}_i \mathbf{x}_i^\top) = \sum_{i=1}^n \text{tr}(\mathbf{x}_i^\top \mathbf{x}_i) = \sum_{i=1}^n \|\mathbf{x}_i\|_2^2.$$

Above, we make use of the fact that for any matrices  $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times d}$

which is called the *cy* <https://eduassistpro.github.io/> used norm is simply the sum-of-squares *Frobenius norm* of the matrix  $\mathbf{X}$ , and denote it by  $\|\mathbf{X}\|_F$ . It can be checked that  $\|\mathbf{X}\|_F$  is the same as the Euclidean inner product and norm when  $\mathbf{X}$  is an  $n \times d$  matrix as  $n \times d$  matrices as  $nd$ -dimensional vectors obtained by stacking columns on top of each other.

Suppose a matrix  $\mathbf{X}$  has thin SVD  $\mathbf{X} = \mathbf{U} \mathbf{S} \mathbf{V}^\top$ , with  $\sigma_r$ , and  $\mathbf{U}^\top \mathbf{U} = \mathbf{V}^\top \mathbf{V} = \mathbf{I}$ . Then its squared Frobenius norm is

$$\|\mathbf{X}\|_F^2 = \text{tr}(\mathbf{V} \mathbf{S} \mathbf{U}^\top \mathbf{U} \mathbf{S} \mathbf{V}^\top) = \text{tr}(\mathbf{V} \mathbf{S}^2 \mathbf{V}^\top) = \text{tr}(\mathbf{S}^2 \mathbf{V}^\top \mathbf{V}) = \text{tr}(\mathbf{S}^2) = \sum_{i=1}^r \sigma_i^2,$$

the sum-of-squares of  $\mathbf{X}$ 's singular values.

### Best rank- $k$ approximation in Frobenius norm

Let the SVD of a matrix  $\mathbf{A} \in \mathbb{R}^{n \times d}$  be given by  $\mathbf{A} = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$ . Here, we assume  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$ . For any  $k \leq r$ , a *rank- $k$  SVD* of  $\mathbf{A}$  is obtained by just keeping the first  $k$  components (corresponding to the  $k$  largest singular values), and this yields a matrix  $\mathbf{A}_k \in \mathbb{R}^{n \times d}$  with rank  $k$ :

$$\mathbf{A}_k := \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}_i^\top. \quad (5.1)$$

This matrix  $\mathbf{A}_k$  is the best rank- $k$  approximation to  $\mathbf{A}$  in the sense that it minimizes the Frobenius norm error over all matrices of rank (at most)  $k$ . This is remarkable because the set of matrices of rank at most  $k$  is not a set over which it is typically easy to optimize. (For instance, it is not a convex set.)

**Theorem 5.7.** Let  $\mathbf{A} \in \mathbb{R}^{n \times d}$  be any matrix, with SVD as given in Theorem 5.6, and  $\mathbf{A}_k$  as defined in (5.1). Then:

1. The rows of  $\mathbf{A}_k$  are the orthogonal projections of the corresponding rows of  $\mathbf{A}$  to the  $k$ -dimensional subspace spanned by top- $k$  right singular vectors  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$  of  $\mathbf{A}$ .
2.  $\|\mathbf{A} - \mathbf{A}_k\|_F \leq \min\{\|\mathbf{A} - \mathbf{B}\|_F : \mathbf{B} \in \mathbb{R}^{n \times d}, \text{rank}(\mathbf{B}) \leq k\}$ .
3. If  $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n \in \mathbb{R}^d$  are the rows of  $\mathbf{A}$ , and  $\hat{\mathbf{a}}_1, \hat{\mathbf{a}}_2, \dots, \hat{\mathbf{a}}_n \in \mathbb{R}^d$  are the rows of  $\mathbf{A}_k$ , then

$$\sum_{i=1}^n \|\mathbf{a}_i - \hat{\mathbf{a}}_i\|_2^2 \leq \sum_{i=1}^n \|\mathbf{a}_i - \mathbf{b}_i\|_2^2$$

for any vectors  $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n \in \mathbb{R}^d$  that span a subspace of dimension at most  $k$ .

*Proof.* The orthogonal projection to the subspace  $W_k$  spanned by  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$  is given by  $\mathbf{x} \mapsto \mathbf{V}_k \mathbf{V}_k^\top \mathbf{x}$ , where  $\mathbf{V}_k := [\mathbf{v}_1 | \mathbf{v}_2 | \dots | \mathbf{v}_k]$ . Since  $\mathbf{V}_k \mathbf{V}_k^\top \mathbf{v}_i = \mathbf{v}_i$  for  $i \in [k]$  and  $\mathbf{V}_k \mathbf{V}_k^\top \mathbf{v}_i = \mathbf{0}$  for  $i > k$ ,

$$\mathbf{A} \mathbf{V}_k \mathbf{V}_k^\top = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^\top \mathbf{V}_k \mathbf{V}_k^\top = \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}_i^\top = \mathbf{A}_k.$$

This equality says that the rows of  $\mathbf{A}_k$  are the orthogonal projections of the rows of  $\mathbf{A}$  onto  $W_k$ . This proves the first claim.

Consider any matrix  $\mathbf{B}$ . Let  $\Pi_W$  denote the orthogonal projector onto the space spanned by the rows of  $\mathbf{B}$ . Then  $\|\mathbf{A} - \mathbf{B}\|_F^2 = \|\mathbf{A} - \mathbf{A} \Pi_W\|_F^2 + \|\mathbf{A} \Pi_W - \mathbf{B}\|_F^2$ . This means that

$$\min_{\substack{\mathbf{B} \in \mathbb{R}^{n \times d}, \\ \text{rank}(\mathbf{B}) \leq k}} \|\mathbf{A} - \mathbf{B}\|_F^2 = \min_{\substack{\text{subspace } W \subseteq \mathbb{R}^d, \\ \dim W \leq k}} \|\mathbf{A} - \mathbf{A} \Pi_W\|_F^2 = \min_{\substack{\mathbf{U} \in \mathbb{R}^{d \times k}, \\ \mathbf{U}^\top \mathbf{U} = \mathbf{I}}} \sum_{i=1}^n \|(I - \mathbf{U} \mathbf{U}^\top) \mathbf{a}_i\|_2^2,$$

where  $\mathbf{a}_i \in \mathbb{R}^d$  denotes the  $i$ -th row of  $\mathbf{A}$ . In fact, it is clear that we can take the minimization over subspaces  $W$  with  $\dim W = k$ . Since the orthogonal projector to a subspace of dimension  $k$  is of the form  $\mathbf{U} \mathbf{U}^\top$  for some  $\mathbf{U} \in \mathbb{R}^{d \times k}$  satisfying  $\mathbf{U}^\top \mathbf{U} = \mathbf{I}$ , it follows that the expression above is the same as

$$\min_{\substack{\mathbf{U} \in \mathbb{R}^{d \times k}, \\ \mathbf{U}^\top \mathbf{U} = \mathbf{I}}} \sum_{i=1}^n \|(I - \mathbf{U} \mathbf{U}^\top) \mathbf{a}_i\|_2^2.$$

Observe that  $\frac{1}{n} \sum_{i=1}^n \mathbf{a}_i \mathbf{a}_i^\top = \frac{1}{n} \mathbf{A}^\top \mathbf{A}$ , so Theorem 5.6 implies that top- $k$  eigenvectors of the  $\frac{1}{n} \sum_{i=1}^n \mathbf{a}_i \mathbf{a}_i^\top$  are top- $k$  right singular vectors of  $\mathbf{A}$ . By Theorem 5.4, the minimization problem above is achieved when  $\mathbf{U} = \mathbf{V}_k$ . This proves the second claim. The third claim is just a different interpretation of the second claim.  $\square$

### Best rank- $k$ approximation in spectral norm

Another important matrix norm is the *spectral norm*: for a matrix  $\mathbf{X} \in \mathbb{R}^{n \times d}$ ,

$$\|\mathbf{X}\|_2 := \max_{\mathbf{u} \in S^{d-1}} \|\mathbf{X} \mathbf{u}\|_2.$$

By Theorem 5.6, the spectral norm of  $\mathbf{X}$  is equal to its largest singular value.

**Fact 5.6.** Let the SVD of a matrix  $\mathbf{A} \in \mathbb{R}^{n \times d}$  be as given in Theorem 5.6, with  $r = \text{rank}(\mathbf{A})$ .

- For any  $\mathbf{x} \in \mathbb{R}^d$ ,

$$\|\mathbf{Ax}\|_2 \leq \sigma_1 \|\mathbf{x}\|_2.$$

- For any  $\mathbf{x}$  in the span of  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r$ ,

$$\|\mathbf{Ax}\|_2 \geq \sigma_r \|\mathbf{x}\|_2.$$

Unlike the Frobenius norm, the spectral norm does not arise from a matrix inner product. Nevertheless, it can be checked that it has the required properties of a norm: it satisfies  $\|c\mathbf{X}\|_2 = |c|\|\mathbf{X}\|_2$  and  $\|\mathbf{X} + \mathbf{Y}\|_2 \leq \|\mathbf{X}\|_2 + \|\mathbf{Y}\|_2$ , and the only matrix with  $\|\mathbf{X}\|_2 = 0$  is  $\mathbf{X} = \mathbf{0}$ . Because of this, the spectral norm also provides a metric between matrices,  $\text{dist}(\mathbf{X}, \mathbf{Y}) = \|\mathbf{X} - \mathbf{Y}\|_2$ , satisfying the properties given in Section 1.1.

The rank- $k$  SVD of a matrix  $\mathbf{A}$  also provides the best rank- $k$  approximation in terms of spectral norm error.

**Theorem 5.8.** Let  $\mathbf{A} \in \mathbb{R}^{n \times d}$  be any matrix, with SVD as given in Theorem 5.6, and  $\mathbf{A}_k$  as defined in (5.1). Then  $\|\mathbf{A} - \mathbf{A}_k\|_2 \leq \min\{\|\mathbf{A} - \mathbf{B}\|_2 : \mathbf{B} \in \mathbb{R}^{n \times d}, \text{rank}(\mathbf{B}) \leq k\}$ .

*Proof.* Since the largest singular value of  $\mathbf{A} - \mathbf{A}_k = \sum_{i=k+1}^r \sigma_i \mathbf{u}_i \mathbf{u}_i^\top$  is  $\sigma_{k+1}$ , it follows that

$$\|\mathbf{A} - \mathbf{A}_k\|_2 = \sigma_{k+1}.$$

Consider any matrix  $\mathbf{B}$  with  $\text{rank}(\mathbf{B}) \leq k$ . The matrix  $\mathbf{A} - \mathbf{B}$  has dimension at least  $d - \text{rank}(\mathbf{B}) \geq d - k$ . The matrix  $\mathbf{A} - \mathbf{A}_k$  has dimension at least  $d - \text{rank}(\mathbf{A}_k) \geq d - k$ . Therefore, there must be some non-zero vector  $\mathbf{x} \in \ker(\mathbf{B}) \cap \ker(\mathbf{A} - \mathbf{A}_k)$ . For any such vector  $\mathbf{x}$ ,

$$\begin{aligned} \|\mathbf{A} - \mathbf{B}\|_2 &\geq \frac{\|(\mathbf{A} - \mathbf{B})\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \\ &\geq \frac{\|\mathbf{Ax}\|_2}{\|\mathbf{x}\|_2} \quad (\text{since } \mathbf{x} \text{ is in the null space of } \mathbf{B}) \\ &= \frac{\sqrt{\|\mathbf{A}_{k+1}\mathbf{x}\|_2^2 + \|(\mathbf{A} - \mathbf{A}_{k+1})\mathbf{x}\|_2^2}}{\|\mathbf{x}\|_2} \\ &\geq \frac{\|\mathbf{A}_{k+1}\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \\ &\geq \sigma_{k+1} \quad (\text{by Fact 5.6}). \end{aligned}$$

Therefore  $\|\mathbf{A} - \mathbf{B}\|_2 \geq \|\mathbf{A} - \mathbf{A}_k\|_2$ . □