## 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 $X$ be a random vector with distribution $\mathbb{P}$. The mean of $X$ is the “center of mass” of $\mathbb{P}$. The covariance of $X$ 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 $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$, then it is common to arrange these vectors as rows of a matrix $A \in \mathbb{R}^{n \times d}$. In this case, we can think of $\mathbb{P}$ as the uniform distribution over the $n$ points $x_1, x_2, \ldots, x_n$. The mean of $X \sim \mathbb{P}$ can be written as

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

and the covariance of $X$ is

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

where $\tilde{A} = A - (1/n) 11^\top A$. We often call these the empirical mean and empirical covariance of the data $x_1, x_2, \ldots, x_n$.

Covariance matrices are always symmetric by definition. Moreover, they are always positive semidefinite, since for any non-zero $z \in \mathbb{R}^d$,

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

This also shows that for any unit vector $u$, the variance of $X$ in direction $u$ is

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

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

**Theorem 5.1.** Let $M \in \mathbb{R}^{d \times d}$ be a symmetric matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ and corresponding orthonormal eigenvectors $v_1, v_2, \ldots, v_d$. Then

$$
\max_{u \neq 0} \frac{u^\top Mu}{u^\top u} = \lambda_1,
\min_{u \neq 0} \frac{u^\top Mu}{u^\top u} = \lambda_d.
$$

These are achieved by $v_1$ and $v_d$, respectively. (The ratio $u^\top Mu / u^\top u$ is called the Rayleigh quotient associated with $M$ in direction $u$.)
Recall, the trace linear map $k$ of value of variance of $X$. Then

**Fact 5.1.** For any $\lambda$

**Corollary 5.1.** Let $v_1$ be a unit-length eigenvector of $\text{cov}(X)$ corresponding to the largest eigenvalue of $\text{cov}(X)$. Then

$$
\text{var}(\langle v_1, X \rangle) = \max_{\mathbf{u} \in S^{d-1}} \text{var}(\langle \mathbf{u}, X \rangle).
$$

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

$$
\mathbf{x} \mapsto \mathbf{U}^\top \mathbf{x}, \quad \text{where } \mathbf{U} = \begin{bmatrix}
\uparrow & \uparrow & \cdots & \uparrow \\
\mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_k
\end{bmatrix} \in \mathbb{R}^{d \times k}.
$$

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

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

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 $\text{tr}(\text{cov}(\mathbf{U}^\top \mathbf{X})) = \mathbb{E} \| \mathbf{U} \|_2^2$.

**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 $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d$. Then for any $k \in [d],$

$$
\max_{\mathbf{U} \in \mathbb{R}^{d \times k} : \mathbf{U}^\top \mathbf{U} = \mathbf{I}} \text{tr}(\mathbf{U}^\top \mathbf{M} \mathbf{U}) = \lambda_1 + \lambda_2 + \cdots + \lambda_k,
$$

$$
\min_{\mathbf{U} \in \mathbb{R}^{d \times k} : \mathbf{U}^\top \mathbf{U} = \mathbf{I}} \text{tr}(\mathbf{U}^\top \mathbf{M} \mathbf{U}) = \lambda_{d-k+1} + \lambda_{d-k+2} + \cdots + \lambda_d.
$$

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

**Proof.** Let $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k$ 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 \( c_j := \sum_{i=1}^{k} \langle v_j, u_i \rangle^2 \) 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 \( u_1, u_2, \ldots, u_k \) to an orthonormal basis \( u_1, u_2, \ldots, u_d \) for \( \mathbb{R}^d \), we have for each \( j \in [d] \),

\[
c_j = \sum_{i=1}^{k} \langle v_j, u_i \rangle^2 \leq \sum_{i=1}^{d} \langle v_j, u_i \rangle^2 = 1.
\]

Finally, since \( v_1, v_2, \ldots, v_d \) is an orthonormal basis for \( \mathbb{R}^d \),

\[
\sum_{j=1}^{d} c_j = \sum_{j=1}^{d} \sum_{i=1}^{k} \langle v_j, u_i \rangle^2 = \sum_{j=1}^{d} \sum_{i=1}^{k} \langle v_j, u_i \rangle^2 = \sum_{i=1}^{k} \| u_i \|_2^2 = k.
\]

The maximum value of \( \sum_{j=1}^{d} c_j \lambda_j \) over all choices of \( c_1, c_2, \ldots, c_d \in [0, 1] \) with \( \sum_{j=1}^{d} c_j = k \) is \( \lambda_1 + \lambda_2 + \cdots + \lambda_k \). This is achieved when \( c_1 = c_2 = \cdots = c_k = 1 \) and \( c_{k+1} = \cdots = c_d = 0 \), i.e., when \( \text{span}(v_1, v_2, \ldots, v_k) = \text{span}(u_1, u_2, \ldots, u_k) \). The minimum value of \( \sum_{j=1}^{d} c_j \lambda_j \) over all choices of \( c_1, c_2, \ldots, c_d \in [0, 1] \) with \( \sum_{j=1}^{d} c_j = k \) is \( \lambda_{d-k+1} + \lambda_{d-k+2} + \cdots + \lambda_d \). This is achieved when \( c_1 = \cdots = c_{d-k} = 0 \) and \( c_{d-k+1} = c_{d-k+2} = \cdots = c_d = 1 \), i.e., when \( \text{span}(v_{d-k+1}, v_{d-k+2}, \ldots, v_d) = \text{span}(u_1, u_2, \ldots, u_k) \).

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. Note that the choice of top-\( k \) (or bottom-\( k \)) eigenvectors is not necessarily unique.

**Corollary 5.2.** Let \( v_1, v_2, \ldots, v_k \) be top-\( k \) eigenvectors of \( \text{cov}(X) \), and let \( V_k := [v_1|v_2|\cdots|v_k] \). Then

\[
\text{tr}(\text{cov}(V_k^\top X)) = \max_{U \in \mathbb{R}^{d \times k} : U^\top U = I} \text{tr}(\text{cov}(U^\top X)).
\]

An orthogonal projection given by top-\( k \) eigenvectors of \( \text{cov}(X) \) 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 \( X \) 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 \( u_1, u_2, \ldots, u_k \)—and a displacement vector \( u_0 \in \mathbb{R}^d \):

\[
A = \{ u_0 + \alpha_1 u_1 + \alpha_2 u_2 + \cdots + \alpha_k u_k : \alpha_1, \alpha_2, \ldots, \alpha_k \in \mathbb{R} \}.
\]

Let \( U := [u_1|u_2|\cdots|u_k] \). Then, for any \( x \in \mathbb{R}^d \), the point in \( A \) closest to \( x \) is given by \( u_0 + UU^\top (x - u_0) \), and hence the squared distance from \( x \) to \( A \) is \( \|(I - UU^\top)(x - u_0)\|_2^2 \).

**Theorem 5.3.** Let \( v_1, v_2, \ldots, v_k \) be top-\( k \) eigenvectors of \( \text{cov}(X) \), let \( V_k := [v_1|v_2|\cdots|v_k] \), and \( v_0 := \mathbb{E}(X) \). Then

\[
\mathbb{E} \| (I - V_k V_k^\top) (X - v_0) \|_2^2 = \min_{U \in \mathbb{R}^{d \times k}, u_0 \in \mathbb{R}^d : U^\top U = I} \mathbb{E} \| (I - UU^\top) (X - u_0) \|_2^2.
\]
Proof. For any matrix $d \times d$ matrix $M$, the function $u_0 \mapsto \mathbb{E}\|M(X - u_0)\|_2^2$ is minimized when $Mu_0 = M\mathbb{E}(X)$ (Fact 5.2). Therefore, we can plug-in $\mathbb{E}(X)$ for $u_0$ in the minimization problem, whereupon it reduces to

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

The objective function is equivalent to

$$
\mathbb{E}\|(I - UU^T)(X - \mathbb{E}(X))\|_2^2 = \mathbb{E}\|X - \mathbb{E}(X)\|_2^2 - \mathbb{E}\|UU^T(X - \mathbb{E}(X))\|_2^2 = \mathbb{E}\|X - \mathbb{E}(X)\|_2^2 - \text{tr}(\text{cov}(U^TX)),
$$

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

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

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

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

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

A similar statement can be made about (linear) subspaces by using top-$k$ eigenvectors of $\mathbb{E}(XX^T)$ instead of $\text{cov}(X)$. This is sometimes called *uncentered PCA*.

**Theorem 5.4.** Let $v_1, v_2, \ldots, v_k$ be top-$k$ eigenvectors of $\mathbb{E}(XX^T)$, and let $V_k := [v_1 | v_2 | \cdots | v_k]$. Then

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

### 5.3 Noisy affine subspace recovery

Suppose there are $n$ points $t_1, t_2, \ldots, t_n \in \mathbb{R}^d$ that lie on an affine subspace $A_*$ of dimension $k$. In this scenario, you don’t directly observe the $t_i$; rather, you only observe noisy versions of these points: $Y_1, Y_2, \ldots, Y_n$, where for some $\sigma_1, \sigma_2, \ldots, \sigma_n > 0$,

$$
Y_j \sim N(t_j, \sigma_j^2 I) \quad \text{for all } j \in [n]
$$

and $Y_1, Y_2, \ldots, Y_n$ are independent. The observations $Y_1, Y_2, \ldots, Y_n$ no longer all lie in the affine subspace $A_*$, but by applying PCA to the empirical covariance of $Y_1, Y_2, \ldots, Y_n$, you can hope to approximately recover $A_*$. Regard $X$ as a random vector whose conditional distribution given the noisy points is uniform over $Y_1, Y_2, \ldots, Y_n$. In fact, the distribution of $X$ is given by the following generative process:

1. Draw $J \in [n]$ uniformly at random.
2. Given $J$, draw $Z \sim N(0, \sigma_J^2 I)$.
Note that the empirical covariance based on \( Y_1, Y_2, \ldots, Y_n \) is not exactly \( \text{cov}(X) \), but it can be a good approximation when \( n \) is large (with high probability). Similarly, the empirical average of \( Y_1, Y_2, \ldots, Y_n \) is a good approximation to \( \mathbb{E}(X) \) when \( n \) is large (with high probability). So here, we assume for simplicity that both \( \text{cov}(X) \) and \( \mathbb{E}(X) \) are known exactly. We show that PCA produces a \( k \)-dimensional affine subspace that contains all of the \( t_j \).

**Theorem 5.5.** Let \( X \) be the random vector as defined above, \( v_1, v_2, \ldots, v_k \) be top-\( k \) eigenvectors of \( \text{cov}(X) \), and \( v_0 := \mathbb{E}(X) \). Then the affine subspace

\[
\hat{A} := \{ v_0 + \alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_k v_k : \alpha_1, \alpha_2, \ldots, \alpha_k \in \mathbb{R} \}
\]

contains \( t_1, t_2, \ldots, t_n \).

**Proof.** Theorem 5.3 says that the matrix \( V_k := [v_1 | v_2 | \cdots | v_k] \) minimizes \( \mathbb{E} \| (I - UU^\top) (X - v_0) \|_2^2 \) (as a function of \( U \in \mathbb{R}^{d \times k} \), subject to \( U^\top U = I \)), or equivalently, maximizes \( \text{tr}(\text{cov}(U^\top X)) \). This maximization objective can be written as

\[
\text{tr}(\text{cov}(U^\top X)) = \mathbb{E} \| UU^\top (X - v_0) \|_2^2 \quad \text{(by Fact 5.1)}
\]

\[
= \frac{1}{n} \sum_{j=1}^n \mathbb{E} \left[\| UU^\top (t_j - v_0 + Z) \|_2^2 \bigg| J = j \right]
\]

\[
= \frac{1}{n} \sum_{j=1}^n \mathbb{E} \left[\| UU^\top (t_j - v_0) \|_2^2 + 2 \langle UU^\top (t_j - v_0), UU^\top Z \rangle + \| UU^\top Z \|_2^2 \bigg| J = j \right]
\]

\[
= \frac{1}{n} \sum_{j=1}^n \left\{ \| UU^\top (t_j - v_0) \|_2^2 + \mathbb{E} \left[\| UU^\top Z \|_2^2 \bigg| J = j \right] \right\}
\]

\[
= \frac{1}{n} \sum_{j=1}^n \left\{ \| UU^\top (t_j - v_0) \|_2^2 + k \sigma_j^2 \right\},
\]

where the penultimate step uses the fact that the conditional distribution of \( Z \) given \( J = j \) is \( N(0, \sigma_j^2 I) \), and the final step uses the fact that \( \| UU^\top Z \|_2^2 \) has the same conditional distribution (given \( J = j \)) as the squared length of a \( N(0, \sigma_j^2 I) \) random vector in \( \mathbb{R}^k \). Since \( UU^\top (t_j - v_0) \) is the orthogonal projection of \( t_j - v_0 \) onto the subspace spanned by the columns of \( U \) (call it \( W \)),

\[
\| UU^\top (t_j - v_0) \|_2^2 \leq \| t_j - v_0 \|_2^2 \quad \text{for all } j \in [n].
\]

The inequalities above are equalities precisely when \( t_j - v_0 \in W \) for all \( j \in [n] \). This is indeed the case for the subspace \( A_k = \{ v_0 \} \). Since \( V_k \) maximizes the objective, its columns must span a \( k \)-dimensional subspace \( \hat{W} \) that also contains all of the \( t_j - v_0 \); hence the affine subspace \( \hat{A} = \{ v_0 + x : x \in \hat{W} \} \) contains all of the \( t_j \).

### 5.4 Singular value decomposition

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

**Fact 5.3.** \( A^\top A \) and \( AA^\top \) are symmetric and positive semidefinite.
Let Lemma 5.1.

Lemma 5.1. Let \( \lambda \) be an eigenvalue of \( A^\top A \) with corresponding eigenvector \( v \).

- If \( \lambda > 0 \), then \( \lambda \) is a non-zero eigenvalue of \( AA^\top \) with corresponding eigenvector \( Av \).
- If \( \lambda = 0 \), then \( Av = 0 \).

Proof. First suppose \( \lambda > 0 \). Then

\[
AA^\top(Av) = A(A^\top Av) = A(\lambda v) = \lambda(Av),
\]

so \( \lambda \) is an eigenvalue of \( AA^\top \) with corresponding eigenvector \( Av \).

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

\[
\|Av\|_2^2 = v^\top A^\top Av = v^\top(\lambda v) = 0.
\]

Since only the zero vector has length 0, it must be that \( Av = 0 \).

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

Theorem 5.6 (Singular value decomposition). Let \( A \) be an \( n \times d \) matrix. Let \( v_1, v_2, \ldots, v_d \in \mathbb{R}^d \) be orthonormal eigenvectors of \( A^\top A \) corresponding to eigenvalues \( \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \). Let \( r \) be the number of positive \( \lambda_i \). Define

\[
u_i := \frac{Av_i}{\|Av_i\|_2} = \frac{Av_i}{\sqrt{v_i^\top A^\top Av_i}} = \frac{Av_i}{\sqrt{\lambda_i}} \quad \text{for each } i \in [r].
\]

Let \( u_{r+1}, u_{r+2}, \ldots, u_n \in \mathbb{R}^n \) be any orthonormal vectors that are orthogonal to \( \text{span}\{u_1, u_2, \ldots, u_r\} \).

Then

\[
A = \begin{bmatrix}
u_1 & \cdots & \nu_r & u_{r+1} & \cdots & u_n
\end{bmatrix} = \sqrt{\lambda_1} \begin{bmatrix}1 & \cdots & 1 & 0 & \cdots & 0
\end{bmatrix} = \begin{bmatrix}0 & \cdots & 0
\end{bmatrix}.
\]

Moreover, \( USV^\top = \sum_{i=1}^r \sqrt{\lambda_i} u_i v_i^\top \).

Proof. The proof of the second claim \( USV^\top = \sum_{i=1}^r \sqrt{\lambda_i} u_i v_i^\top \) is a straightforward computation.

To prove the first claim, that \( A = USV^\top \), it suffices to show that for some set of \( d \) linearly independent vectors \( q_1, q_2, \ldots, q_d \in \mathbb{R}^d \),

\[
Aq_j = \left( \sum_{i=1}^r \sqrt{\lambda_i} u_i v_i^\top \right) q_j \quad \text{for all } j \in [d].
\]

We’ll use \( v_1, v_2, \ldots, v_d \). Observe that

\[
Av_j = \begin{cases}
\sqrt{\lambda_j} u_j & \text{if } 1 \leq j \leq r, \\
0 & \text{if } r < j \leq d,
\end{cases}
\]
by definition of \( \mathbf{u}_i \) and by Lemma 5.1. Moreover,

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

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

We also note

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

where the last step follows since \( \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d \) are orthonormal. This, along with the choice of \( \mathbf{u}_{r+1}, \mathbf{u}_{r+2}, \ldots, \mathbf{u}_n \), implies that \( \mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n \) are orthonormal.

The decomposition of \( \mathbf{A} \) into the matrix product \( \mathbf{USV}^\top \) from Theorem 5.6 is called the singular value decomposition (SVD) of \( \mathbf{A} \). The columns of \( \mathbf{U} \) are the left singular vectors, and the columns of \( \mathbf{V} \) are the right singular vectors. The scalars \( \sqrt{\lambda_1} \geq \sqrt{\lambda_2} \geq \cdots \geq \sqrt{\lambda_r} \) are the (positive) singular values corresponding to the left/right singular vectors \( \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{u}_r \). The singular vectors \( \mathbf{u}_i \) and \( \mathbf{v}_i \) for \( i > r \) have 0 as a corresponding singular value.

The second representation, \( \mathbf{A} = \sum_{i=1}^{r} \sqrt{\lambda_i} \mathbf{u}_i \mathbf{v}_i^\top \) is called the thin SVD of \( \mathbf{A} \), as it can also be written as

\[
\mathbf{A} = \begin{bmatrix}
\mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_r
\end{bmatrix}
\begin{bmatrix}
\sqrt{\lambda_1} & \sqrt{\lambda_2} & \cdots & \sqrt{\lambda_r}
\end{bmatrix}
\begin{bmatrix}
\mathbf{v}_1^\top \\
\mathbf{v}_2^\top \\
\vdots \\
\mathbf{v}_r^\top
\end{bmatrix}
\]

The number \( r \) of positive \( \lambda_i \) is the rank of \( \mathbf{A} \), which is at most the smaller of \( n \) and \( d \).

Just as before, we'll refer to the \( k \) largest singular values of \( \mathbf{A} \) as the top-\( k \) singular values of \( \mathbf{A} \), and the \( k \) smallest singular values as the bottom-\( k \) singular values of \( \mathbf{A} \). We analogously use the term top-\( k \) (resp., bottom-\( k \)) singular vectors to refer to orthonormal singular vectors 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 \( \mathbf{AA}^\top \) are the left singular vectors of \( \mathbf{A} \).

Suppose there are \( n \) data points \( \mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n \in \mathbb{R}^d \), 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{XX}^\top) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{a}_i \mathbf{a}_i^\top = \frac{1}{n} \mathbf{A}^\top \mathbf{A}: \text{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 $A$ and the eigenvalues of $A^\top A$ and $A A^\top$, 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 $A \in \mathbb{R}^{n \times d}$ be given by $A = \sum_{i=1}^r \sigma_i u_i v_i^\top$, where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$. For each $i \in [r]$, 

$$
\sigma_i = \max_{x \in S^{d-1} : \langle v_i, x \rangle = 0 \forall j < i} y^\top A x = u_i^\top A v_i.
$$

Relationship between eigendecomposition and SVD

If $M \in \mathbb{R}^{d \times d}$ is symmetric and has eigendecomposition $M = \sum_{i=1}^d \lambda_i v_i v_i^\top$, then its singular values are the absolute values of the $\lambda_i$. We can take $v_1, v_2, \ldots, v_d$ as corresponding right singular vectors. For corresponding left singular vectors, we can take $u_i := v_i$ whenever $\lambda_i \geq 0$ (which is the case for all $i$ if $M$ is also psd), and $u_i := -v_i$ whenever $\lambda_i < 0$. Therefore, we have the following variational characterization of the singular values of $M$.

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

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

Moore-Penrose pseudoinverse

The SVD defines a kind of matrix inverse that is applicable to non-square matrices $A \in \mathbb{R}^{n \times d}$ (where possibly $n \neq d$). Let the SVD be given by $A = U S V^\top$, where $U \in \mathbb{R}^{n \times r}$ and $V \in \mathbb{R}^{d \times r}$ satisfy $U^\top U = V^\top V = I$, and $S \in \mathbb{R}^{r \times r}$ is diagonal with positive diagonal entries. Here, the rank of $A$ is $r$. The Moore-Penrose pseudoinverse of $A$ is given by 

$$
A^\dagger := VS^{-1}U^\top \in \mathbb{R}^{d \times n}.
$$

Note that $A^\dagger$ is well-defined: $S$ is invertible because its diagonal entries are all strictly positive.

What is the effect of multiplying $A$ by $A^\dagger$ on the left? Using the SVD of $A$,

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

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

$$
A A^\dagger A = A.
$$

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

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

$$
A^\dagger = (A^\top A)^{-1} A^\top
$$

and using the fact that $(A^\top A)^{\dagger} = (A^\top A)^{-1}$ when $A^\top A$ 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 $\langle X, Y \rangle := \text{tr}(X^T Y)$. 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:

\[
\langle cX + Y, Z \rangle = \text{tr}((cX + Y)^T Z) = \text{tr}(cX^T Z + Y^T Z) = c\text{tr}(X^T Z) + \text{tr}(Y^T Z) = c\langle X, Z \rangle + \langle Y, Z \rangle.
\]

The inner product naturally induces an associated norm $X \mapsto \sqrt{\langle X, X \rangle}$. Viewing $X \in \mathbb{R}^{n \times d}$ as a data matrix whose rows are the vectors $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$, we see that

\[
\langle X, X \rangle = \text{tr}(X^T X) = \text{tr}\left(\sum_{i=1}^n x_i x_i^T\right) = \sum_{i=1}^n \text{tr}(x_i x_i^T) = \sum_{i=1}^n \text{tr}(x_i^T x_i) = \sum_{i=1}^n \|x_i\|_2^2.
\]

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

\[
\text{tr}(A^T B) = \text{tr}(BA^T),
\]

which is called the cyclic property of the matrix trace. Therefore, the square of the induced norm is simply the sum-of-squares of the entries in the matrix. We call this norm the Frobenius norm of the matrix $X$, and denote it by $\|X\|_F$. It can be checked that this matrix inner product and norm are exactly the same as the Euclidean inner product and norm when you view the $n \times d$ matrices as $nd$-dimensional vectors obtained by stacking columns on top of each other (or rows side-by-side).

Suppose a matrix $X$ has thin SVD $X = USV^T$, where $S = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r)$, and $U^T U = V^T V = I$. Then its squared Frobenius norm is

\[
\|X\|_F^2 = \text{tr}(VSU^T USV^T) = \text{tr}(VS^2V^T) = \text{tr}(S^2V^T V) = \text{tr}(S^2) = \sum_{i=1}^r \sigma_i^2,
\]

the sum-of-squares of $X$’s singular values.

Best rank-$k$ approximation in Frobenius norm

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

\[
A_k := \sum_{i=1}^k \sigma_i u_i v_i^T. \tag{5.1}
\]

This matrix $A_k$ is the best rank-$k$ approximation to $A$ 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 $A \in \mathbb{R}^{n \times d}$ be any matrix, with SVD as given in Theorem 5.6, and $A_k$ as defined in (5.1). Then:

1. The rows of $A_k$ are the orthogonal projections of the corresponding rows of $A$ to the $k$-dimensional subspace spanned by top-$k$ right singular vectors $v_1, v_2, \ldots, v_k$ of $A$.

2. $\|A - A_k\|_F \leq \min\{\|A - B\|_F : B \in \mathbb{R}^{n \times d}, \text{rank}(B) \leq k\}$.

3. If $a_1, a_2, \ldots, a_n \in \mathbb{R}^d$ are the rows of $A$, and $\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_n \in \mathbb{R}^d$ are the rows of $A_k$, then

\[
\sum_{i=1}^n \|a_i - \hat{a}_i\|^2 \leq \sum_{i=1}^n \|a_i - b_i\|^2
\]

for any vectors $b_1, b_2, \ldots, b_n \in \mathbb{R}^d$ that span a subspace of dimension at most $k$.

**Proof.** The orthogonal projection to the subspace $W_k$ spanned by $v_1, v_2, \ldots, v_k$ is given by $x \mapsto V_k^T x$, where $V_k := [v_1 | v_2 | \cdots | v_k]$. Since $V_k V_k^T v_i = v_i$ for $i \in [k]$ and $V_k V_k^T v_i = 0$ for $i > k$,

\[
AV_k V_k^T = \sum_{i=1}^r \sigma_i u_i v_i^T V_k V_k^T = \sum_{i=1}^k \sigma_i u_i v_i^T = A_k.
\]

This equality says that the rows of $A_k$ are the orthogonal projections of the rows of $A$ onto $W_k$. This proves the first claim.

Consider any matrix $B \in \mathbb{R}^{n \times d}$ with $\text{rank}(B) \leq k$, and let $W$ be the subspace spanned by the rows of $B$. Let $\Pi_W$ denote the orthogonal projector to $W$. Then clearly we have $\|A - A \Pi_W\|_F \leq \|A - B\|_F$. This means that

\[
\min_{B \in \mathbb{R}^{n \times d} : \text{rank}(B) \leq k} \|A - B\|_F^2 = \min_{\text{subspace } W \subseteq \mathbb{R}^d : \dim W \leq k} \|A - A \Pi_W\|_F^2 = \min_{\text{subspace } W \subseteq \mathbb{R}^d : \dim W \leq k} \sum_{i=1}^n \|(I - \Pi_W) a_i\|^2.
\]

where $a_i \in \mathbb{R}^d$ denotes the $i$-th row of $A$. 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 $UU^T$ for some $U \in \mathbb{R}^{d \times k}$ satisfying $U^T U = I$, it follows that the expression above is the same as

\[
\min_{U \in \mathbb{R}^{d \times k} : U^T U = I} \sum_{i=1}^n \|(I - UU^T) a_i\|^2.
\]

Observe that $\frac{1}{n} \sum_{i=1}^n a_i a_i^T = \frac{1}{n} A^T A$, so Theorem 5.6 implies that top-$k$ eigenvectors of the $\frac{1}{n} \sum_{i=1}^n a_i a_i^T$ are top-$k$ right singular vectors of $A$. By Theorem 5.4 the minimization problem above is achieved when $U = V_k$. 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 $X \in \mathbb{R}^{n \times d}$,

\[
\|X\|_2 := \max_{u \in S^{d-1}} \|Xu\|_2.
\]

By Theorem 5.6 the spectral norm of $X$ is equal to its largest singular value.

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

- For any $x \in \mathbb{R}^d$,
  $$\|Ax\|_2 \leq \sigma_1 \|x\|_2.$$

- For any $x$ in the span of $v_1, v_2, \ldots, v_r$,
  $$\|Ax\|_2 \geq \sigma_r \|x\|_2.$$

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 $\|cx\|_2 = |c| \|x\|_2$ and $\|X + Y\|_2 \leq \|X\|_2 + \|Y\|_2$, and the only matrix with $\|X\|_2 = 0$ is $X = 0$. Because of this, the spectral norm also provides a metric between matrices, $\text{dist}(X, Y) = \|X - Y\|_2$, satisfying the properties given in Section 1.1.

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

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

Proof. Since the largest singular value of $A$ is $\sum_{i=k+1}^r \sigma_i u_i v_i^\top$ is $\sigma_{k+1}$, it follows that

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