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 $\boldsymbol{X}$ be a random vector with distribution $\mathbb{P}$. The mean of $\boldsymbol{X}$ is the "center of mass" of $\mathbb{P}$. The covariance of $\boldsymbol{X}$ 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 $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$, then it is common to arrange these vectors as rows of a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$. In this case, we can think of $\mathbb{P}$ as the uniform distribution over the n points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$. The mean of $\boldsymbol{X} \sim \mathbb{P}$ can be written as

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

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

$$\operatorname{cov}(\boldsymbol{X}) \;=\; \frac{1}{n} \boldsymbol{A}^\top \boldsymbol{A} - \left(\frac{1}{n} \boldsymbol{A}^\top \boldsymbol{1} \right) \left(\frac{1}{n} \boldsymbol{A}^\top \boldsymbol{1} \right)^\top \;=\; \frac{1}{n} \widetilde{\boldsymbol{A}}^\top \widetilde{\boldsymbol{A}}$$

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

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

$$\boldsymbol{z}^\top \operatorname{cov}(\boldsymbol{X}) \boldsymbol{z} \;=\; \boldsymbol{z}^\top \mathbb{E}\Big[(\boldsymbol{X} - \mathbb{E}(\boldsymbol{X}))(\boldsymbol{X} - \mathbb{E}(\boldsymbol{X}))^\top \Big] \boldsymbol{z} \;=\; \mathbb{E}\Big[\langle \boldsymbol{z}, \boldsymbol{X} - \mathbb{E}(\boldsymbol{X}) \rangle^2 \Big] \;\geq\; 0 \,.$$

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

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

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

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

$$\max_{\boldsymbol{u} \neq \boldsymbol{0}} \frac{\boldsymbol{u}^\top \boldsymbol{M} \boldsymbol{u}}{\boldsymbol{u}^\top \boldsymbol{u}} \;=\; \lambda_1 \,,$$

$$\min_{\boldsymbol{u} \neq \boldsymbol{0}} \frac{\boldsymbol{u}^\top \boldsymbol{M} \boldsymbol{u}}{\boldsymbol{u}^\top \boldsymbol{u}} \;=\; \lambda_d \,.$$

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

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

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

This final ratio represents a convex combination of the scalars $\lambda_1, \lambda_2, \ldots, \lambda_d$. Its largest value is λ_1, achieved by $\boldsymbol{w} = \boldsymbol{e}_1$ (and hence $\boldsymbol{u} = \boldsymbol{V}\boldsymbol{e}_1 = \boldsymbol{v}_1$), and its smallest value is λ_d, achieved by $\boldsymbol{w} = \boldsymbol{e}_d$ (and hence $\boldsymbol{u} = \boldsymbol{V}\boldsymbol{e}_d = \boldsymbol{v}_d$). ☐

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

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

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

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

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

$$\operatorname{cov}(\boldsymbol{U}^\top \boldsymbol{X}) = \boldsymbol{U}^\top \operatorname{cov}(\boldsymbol{X})\boldsymbol{U}.$$

The "total" variance in this subspace is often measured by the trace of the covariance: $\operatorname{tr}(\operatorname{cov}(\boldsymbol{U}^\top \boldsymbol{X}))$. 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 $\boldsymbol{U} \in \mathbb{R}^{d \times k}$, $\operatorname{tr}(\operatorname{cov}(\boldsymbol{U}^\top \boldsymbol{X})) = \mathbb{E}\,\|\boldsymbol{U}^\top(\boldsymbol{X} - \mathbb{E}(\boldsymbol{X}))\|_2^2$. Furthermore, if $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}$, then $\operatorname{tr}(\operatorname{cov}(\boldsymbol{U}^\top \boldsymbol{X})) = \mathbb{E}\,\|\boldsymbol{U}\boldsymbol{U}^\top(\boldsymbol{X} - \mathbb{E}(\boldsymbol{X}))\|_2^2$.*

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

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

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

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

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

where $c_j := \sum_{i=1}^k \langle \boldsymbol{v}_j, \boldsymbol{u}_i \rangle^2$ 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 $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_k$ to an orthonormal basis $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_d$ for $\mathbb{R}^d$, we have for each $j \in [d]$,

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

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

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

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 $\operatorname{span}(\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k) = \operatorname{span}(\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_k)$. 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 $\operatorname{span}(\boldsymbol{v}_{d-k+1}, \boldsymbol{v}_{d-k+2}, \ldots, \boldsymbol{v}_d) = \operatorname{span}(\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_k)$. □

We'll refer to the k largest eigenvalues of a symmetric matrix $\boldsymbol{M}$ as the *top-k eigenvalues of $\boldsymbol{M}$*, and the k smallest eigenvalues as the *bottom-k eigenvalues of $\boldsymbol{M}$*. 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 $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k$ be top-k eigenvectors of $\operatorname{cov}(\boldsymbol{X})$, and let $\boldsymbol{V}_k := [\boldsymbol{v}_1|\boldsymbol{v}_2|\cdots|\boldsymbol{v}_k]$. Then*

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

An orthogonal projection given by top-k eigenvectors of $\operatorname{cov}(\boldsymbol{X})$ 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 $\boldsymbol{X}$ 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 $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_k$—and a displacement vector $\boldsymbol{u}_0 \in \mathbb{R}^d$:

$$A \;=\; \{\boldsymbol{u}_0 + \alpha_1 \boldsymbol{u}_1 + \alpha_2 \boldsymbol{u}_2 + \cdots + \alpha_k \boldsymbol{u}_k : \alpha_1, \alpha_2, \ldots, \alpha_k \in \mathbb{R}\}\,.$$

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

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

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

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

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

The objective function is equivalent to

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

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

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

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

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

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

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

Theorem 5.4. *Let $\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_k$ be top-k eigenvectors of $\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^\top)$, and let $\boldsymbol{V}_k := [\boldsymbol{v}_1|\boldsymbol{v}_2|\cdots|\boldsymbol{v}_k]$. Then*

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

5.3 Noisy affine subspace recovery

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

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

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

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

1. Draw $J\in[n]$ uniformly at random.

2. Given J, draw $\boldsymbol{Z}\sim\mathrm{N}(\boldsymbol{0},\sigma_J^2\boldsymbol{I})$.

3. Set $\boldsymbol{X} := \boldsymbol{t}_J + \boldsymbol{Z}$.

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

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

$$\widehat{A} := \{\boldsymbol{v}_0 + \alpha_1 \boldsymbol{v}_1 + \alpha_2 \boldsymbol{v}_2 + \cdots + \alpha_k \boldsymbol{v}_k : \alpha_1, \alpha_2, \ldots, \alpha_k \in \mathbb{R}\}$$

contains $\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_n$.

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

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

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

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

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

5.4 Singular value decomposition

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Then

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

and $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_r$ are orthonormal.

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

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

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

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

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

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

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

Note that

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

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

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

Of course, one can extend $\boldsymbol{u}_1, \boldsymbol{u}_2, \dots, \boldsymbol{u}_r$ to an orthonormal basis for $\mathbb{R}^n$. Define the matrices $\boldsymbol{U} := [\boldsymbol{u}_1 | \boldsymbol{u}_2 | \cdots | \boldsymbol{u}_n] \in \mathbb{R}^{n \times n}$ and $\boldsymbol{V} := [\boldsymbol{v}_1 | \boldsymbol{v}_2 | \cdots | \boldsymbol{v}_d] \in \mathbb{R}^{d \times d}$. Also define $\boldsymbol{S} \in \mathbb{R}^{n \times d}$ to be the matrix whose only non-zero entries are $\sqrt{\lambda_i}$ in the (i,i)-th position, for $1 \leq i \leq r$. Then $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\top}$. This matrix factorization of $\boldsymbol{A}$ is typically called the *full SVD* of $\boldsymbol{A}$. (The vectors $\boldsymbol{u}_{r+1}, \boldsymbol{u}_{r+2}, \dots, \boldsymbol{u}_n$ and $\boldsymbol{v}_{r+1}, \boldsymbol{v}_{r+2}, \dots, \boldsymbol{v}_d$ are also regarded as singular vectors of $\boldsymbol{A}$; they correspond to the singular value equal to zero.)

Just as before, we'll refer to the k largest singular values of $\boldsymbol{A}$ as the *top-k singular values of* $\boldsymbol{A}$, and the k smallest singular values as the *bottom-k singular values of* $\boldsymbol{A}$. 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 $\boldsymbol{A}^{\top}\boldsymbol{A}$ are the right singular vectors $\boldsymbol{A}$, and the eigenvectors of $\boldsymbol{A}\boldsymbol{A}^{\top}$ are the left singular vectors of $\boldsymbol{A}$.

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

Variational characterization of singular values

Given the relationship between the singular values of $\boldsymbol{A}$ and the eigenvalues of $\boldsymbol{A}^{\top}\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^{\top}$, 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* $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ *be given by* $\boldsymbol{A} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$, *where* $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$. *For each* $i \in [r]$,

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

Relationship between eigendecomposition and SVD

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

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

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

Moore-Penrose pseudoinverse

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

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

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

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

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

$$\boldsymbol{A}\boldsymbol{A}^\dagger\boldsymbol{A} \;=\; \boldsymbol{A}\,.$$

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

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

$$\boldsymbol{A}^\dagger \;=\; (\boldsymbol{A}^\top\boldsymbol{A})^\dagger\boldsymbol{A}^\top$$

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

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

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

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

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

$$\operatorname{tr}(\boldsymbol{A}^\top \boldsymbol{B}) \;=\; \operatorname{tr}(\boldsymbol{B}\boldsymbol{A}^\top) \,,$$

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 $\boldsymbol{X}$, and denote it by $\|\boldsymbol{X}\|_{\mathrm{F}}$. 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 $\boldsymbol{X}$ has thin SVD $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$, where $\boldsymbol{S} = \operatorname{diag}(\sigma_1, \sigma_2, \dots, \sigma_r)$, and $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}$. Then its squared Frobenius norm is

$$\|\boldsymbol{X}\|_{\mathrm{F}}^2 \;=\; \operatorname{tr}(\boldsymbol{V}\boldsymbol{S}\boldsymbol{U}^\top \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top) \;=\; \operatorname{tr}(\boldsymbol{V}\boldsymbol{S}^2\boldsymbol{V}^\top) \;=\; \operatorname{tr}(\boldsymbol{S}^2\boldsymbol{V}^\top\boldsymbol{V}) \;=\; \operatorname{tr}(\boldsymbol{S}^2) \;=\; \sum_{i=1}^r \sigma_i^2 \,,$$

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

Best rank-k approximation in Frobenius norm

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

$$\boldsymbol{A}_k \;:=\; \sum_{i=1}^k \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top \,. \tag{5.1}$$

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

1. *The rows of $\boldsymbol{A}_k$ are the orthogonal projections of the corresponding rows of $\boldsymbol{A}$ to the k-dimensional subspace spanned by top-k right singular vectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k$ of $\boldsymbol{A}$.*
2. $\|\boldsymbol{A} - \boldsymbol{A}_k\|_{\mathrm{F}} \le \min\{\|\boldsymbol{A} - \boldsymbol{B}\|_{\mathrm{F}} : \boldsymbol{B} \in \mathbb{R}^{n\times d}, \operatorname{rank}(\boldsymbol{B}) \le k\}$.
3. *If $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n \in \mathbb{R}^d$ are the rows of $\boldsymbol{A}$, and $\hat{\boldsymbol{a}}_1, \hat{\boldsymbol{a}}_2, \ldots, \hat{\boldsymbol{a}}_n \in \mathbb{R}^d$ are the rows of $\boldsymbol{A}_k$, then*
$$\sum_{i=1}^n \|\boldsymbol{a}_i - \hat{\boldsymbol{a}}_i\|_2^2 \;\le\; \sum_{i=1}^n \|\boldsymbol{a}_i - \boldsymbol{b}_i\|_2^2$$
for any vectors $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_n \in \mathbb{R}^d$ that span a subspace of dimension at most k.

Proof. The orthogonal projection to the subspace W_k spanned by $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k$ is given by $\boldsymbol{x} \mapsto \boldsymbol{V}_k \boldsymbol{V}_k^\top \boldsymbol{x}$, where $\boldsymbol{V}_k := [\boldsymbol{v}_1|\boldsymbol{v}_2|\cdots|\boldsymbol{v}_k]$. Since $\boldsymbol{V}_k\boldsymbol{V}_k^\top \boldsymbol{v}_i = \boldsymbol{v}_i$ for $i \in [k]$ and $\boldsymbol{V}_k\boldsymbol{V}_k^\top \boldsymbol{v}_i = \boldsymbol{0}$ for $i > k$,
$$\boldsymbol{A}\boldsymbol{V}_k\boldsymbol{V}_k^\top \;=\; \sum_{i=1}^r \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top \boldsymbol{V}_k \boldsymbol{V}_k^\top \;=\; \sum_{i=1}^k \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top \;=\; \boldsymbol{A}_k\,.$$

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

Consider any matrix $\boldsymbol{B} \in \mathbb{R}^{n\times d}$ with $\operatorname{rank}(\boldsymbol{B}) \le k$, and let W be the subspace spanned by the rows of $\boldsymbol{B}$. Let $\boldsymbol{\Pi}_W$ denotes the orthogonal projector to W. Then clearly we have $\|\boldsymbol{A} - \boldsymbol{A}\boldsymbol{\Pi}_W\|_{\mathrm{F}} \le \|\boldsymbol{A} - \boldsymbol{B}\|_{\mathrm{F}}$. This means that

$$\min_{\substack{\boldsymbol{B}\in\mathbb{R}^{n\times d}:\\ \operatorname{rank}(\boldsymbol{B})\le k}} \|\boldsymbol{A} - \boldsymbol{B}\|_{\mathrm{F}}^2 = \min_{\substack{\text{subspace } W\subseteq\mathbb{R}^d:\\ \dim W \le k}} \|\boldsymbol{A} - \boldsymbol{A}\boldsymbol{\Pi}_W\|_{\mathrm{F}}^2 = \min_{\substack{\text{subspace } W\subseteq\mathbb{R}^d:\\ \dim W \le k}} \sum_{i=1}^n \|(\boldsymbol{I} - \boldsymbol{\Pi}_W)\boldsymbol{a}_i\|_2^2\,,$$

where $\boldsymbol{a}_i \in \mathbb{R}^d$ denotes the i-th row of $\boldsymbol{A}$. 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 $\boldsymbol{U}\boldsymbol{U}^\top$ for some $\boldsymbol{U} \in \mathbb{R}^{d\times k}$ satisfying $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}$, it follows that the expression above is the same as
$$\min_{\substack{\boldsymbol{U}\in\mathbb{R}^{d\times k}:\\ \boldsymbol{U}^\top\boldsymbol{U}=\boldsymbol{I}}} \sum_{i=1}^n \|(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{a}_i\|_2^2\,.$$
Observe that $\frac{1}{n}\sum_{i=1}^n \boldsymbol{a}_i\boldsymbol{a}_i^\top = \frac{1}{n}\boldsymbol{A}^\top\boldsymbol{A}$, so Theorem 5.6 implies that top-k eigenvectors of the $\frac{1}{n}\sum_{i=1}^n \boldsymbol{a}_i\boldsymbol{a}_i^\top$ are top-k right singular vectors of $\boldsymbol{A}$. By Theorem 5.4, the minimization problem above is achieved when $\boldsymbol{U} = \boldsymbol{V}_k$. This proves the second claim. The third claim is just a different interpretation of the second claim. □

Best rank-k approximation in spectral norm

Another important matrix norm is the *spectral norm*: for a matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$,
$$\|\boldsymbol{X}\|_2 \;:=\; \max_{\boldsymbol{u}\in S^{d-1}} \|\boldsymbol{X}\boldsymbol{u}\|_2\,.$$
By Theorem 5.6, the spectral norm of $\boldsymbol{X}$ is equal to its largest singular value.

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

- *For any $\boldsymbol{x} \in \mathbb{R}^d$,*
$$\|\boldsymbol{A}\boldsymbol{x}\|_2 \;\le\; \sigma_1 \|\boldsymbol{x}\|_2 \,.$$
- *For any $\boldsymbol{x}$ in the span of $\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_r$,*
$$\|\boldsymbol{A}\boldsymbol{x}\|_2 \;\ge\; \sigma_r \|\boldsymbol{x}\|_2 \,.$$

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

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

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

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

$$\|\boldsymbol{A}-\boldsymbol{A}_k\|_2 \;=\; \sigma_{k+1} \,.$$

Consider any matrix $\boldsymbol{B} \in \mathbb{R}^{n\times d}$ with $\operatorname{rank}(\boldsymbol{B}) \le k$. Its null space $\ker(\boldsymbol{B})$ has dimension at least $d - \operatorname{rank}(\boldsymbol{B}) \ge d - k$. On the other hand, the span W_{k+1} of $\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \dots, \boldsymbol{v}_{k+1}$ has dimension $k+1$. Therefore, there must be some non-zero vector $\boldsymbol{x} \in \ker(\boldsymbol{B}) \cap W_{k+1}$. For any such vector $\boldsymbol{x}$,

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

Therefore $\|\boldsymbol{A}-\boldsymbol{B}\|_2 \ge \|\boldsymbol{A}-\boldsymbol{A}_k\|_2$. $\square$