Linear Algebra (XXVIII)

What we will cover today

Singular Value Decompositions and Approximation

Again we will see many tools, in particular orthogonality, are of great use.

Review

Actually SVD can be viewed as an extension of the spectral decompositions of symmetric matrices to arbitrary matrices, both for the statement and for the method of the proof.

Theorem (spectral decomposition)

Let S be a symmetric $n \times n$ matrix with eigenvalues $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$. Then there are n orthonormal vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in \mathbb{R}^n$ such that

$$S = \lambda_1 \boldsymbol{v}_1 \boldsymbol{v}_1^{\mathsf{T}} + \lambda_2 \boldsymbol{v}_2 \boldsymbol{v}_2^{\mathsf{T}} + \cdots + \lambda_n \boldsymbol{v}_n \boldsymbol{v}_n^{\mathsf{T}}.$$

More precisely, if $\mathrm{rank}(S) = r$, then $\lambda_{r+1} = \cdots = \lambda_n = 0$ and hence

$$S = \lambda_1 \boldsymbol{v}_1 \boldsymbol{v}_1^{\mathsf{T}} + \lambda_2 \boldsymbol{v}_2 \boldsymbol{v}_2^{\mathsf{T}} + \cdots + \lambda_r \boldsymbol{v}_r \boldsymbol{v}_r^{\mathsf{T}}.$$

Theorem (singular value decomposition)

Let A be an $m \times n$ matrix with singular values $\sigma_1, \ldots, \sigma_r \in \mathbb{R}$ with $r = \mathrm{rank}(A)$. Then there are orthonormal vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r \in \mathbb{R}^m$ and orthonormal $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r \in \mathbb{R}^n$ such that

$$A = \sigma_1 \boldsymbol{u}_1 \boldsymbol{v}_1^{\mathsf{T}} + \sigma_2 \boldsymbol{u}_2 \boldsymbol{v}_2^{\mathsf{T}} + \cdots + \sigma_r \boldsymbol{u}_r \boldsymbol{v}_r^{\mathsf{T}}.$$

Theorem (spectral decomposition)

Let S be a symmetric $n \times n$ matrix with eigenvalues $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$. Then there is an orthogonal $n \times n$ matrix Q such that

$$S = Q \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix} Q^{\mathsf{T}}.$$

Theorem (singular value decomposition)

Let A be a real $m \times n$ matrix with $r = \operatorname{rank}(A)$. Define an $m \times n$ matrix

$$\Sigma = \begin{bmatrix} D & \mathbf{0}_{r\times(n-r)} \\ \mathbf{0}_{(m-r)\times r} & \mathbf{0}_{(m-r)\times(n-r)} \end{bmatrix} \quad \textit{for } D = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_r \end{bmatrix},$$

where $\sigma_1, \ldots, \sigma_r$ are the singular values of A. Then for some orthogonal $m \times m$ matrix U and orthogonal $n \times n$ matrix V

$$A = U\Sigma V^{\mathsf{T}}.$$

Proof of SVD

Lemma

Let S be a real symmetric matrix. Then S has exactly $\mathrm{rank}(S)$ nonzero eigenvalues (counting repetition).

Lemma

Let A be an $m \times n$ matrix. Then

$$\mathrm{rank}(A^\intercal A) = \mathrm{rank}(AA^\intercal) = \mathrm{rank}(A).$$

Lemma

All eigenvalues of $A^\mathsf{T} A$ and AA^T are nonnegative.

In otherwords, both $A^\mathsf{T} A$ and AA^T are positive semidefinite.

Let

$$r = \operatorname{rank}(A),$$

and by the previous lemmas:

- $\operatorname{rank}(A^{\mathsf{T}}A) = \operatorname{rank}(AA^{\mathsf{T}}) = r$, thus each of them has r nonzero eigenvalues.
- Every eigenvalue of $A^{\mathsf{T}}A$ is nonnegative, so is every eigenvalue of AA^{T}.

Lemma

Both $A^{\mathsf{T}}A$ and AA^{T} have the same r positive eigenvalues (counting repetition). More precisely,

matrix	*positive eigenvalues*	*eigenvectors*
$A^{\mathsf{T}}A$	$\lambda_1, \ldots, \lambda_r$	$\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r$
AA^{T}	$\lambda_1, \ldots, \lambda_r$	$\frac{1}{\sqrt{\lambda_1}}A\boldsymbol{v}_1, \ldots, \frac{1}{\sqrt{\lambda_r}}A\boldsymbol{v}_r$

Definition

Let $\lambda_1, \ldots, \lambda_r$ be the nonzero (hence positive) eigenvalues of $A^\mathsf{T} A$ (hence the nonzero eigenvalues of AA^T). For every $i \in [r]$, we define

$$\sigma_i = \sqrt{\lambda_i}.$$

Then $\sigma_1, \ldots, \sigma_r$ are the singular values of A.

Theorem

Let A be a real $m \times n$ matrix with $r = \operatorname{rank}(A)$. Define an $m \times n$ matrix

$$\Sigma = \begin{bmatrix} D & \mathbf{0}_{r\times(n-r)} \\ \mathbf{0}_{(m-r)\times r} & \mathbf{0}_{(m-r)\times(n-r)} \end{bmatrix} \quad \textit{for } D = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_r \end{bmatrix},$$

where $\sigma_1, \dots, \sigma_r$ are the singular values of A. Then for some orthogonal $m \times m$ matrix U and orthogonal $n \times n$ matrix V

$$A = U\Sigma V^{\mathsf{T}}.$$

Let

$$\lambda_1, \ldots, \lambda_r, \lambda_{r+1}, \ldots, \lambda_n$$

be the eigenvalues of $A^\mathsf{T} A$ with

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$$

and $\lambda_{r+1} = \cdots = \lambda_n = 0$. Moreover, let

$$\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in \mathbb{R}^n$$

be the corresponding orthonormal eigenvectors. It follows that

$$V = \begin{bmatrix} \boldsymbol{v}_1 & \ldots \boldsymbol{v}_n \end{bmatrix}.$$

is an orthogonal $n \times n$ matrix.

Proof (2)

Next for every $i \in [r]$ let

$$\boldsymbol{u}_i = \frac{1}{\sqrt{\lambda_i}} A \boldsymbol{v}_i \in \mathbb{R}^m.$$

Thus

$$\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r$$

are orthonormal eigenvectors of AA^T with corresponding eigenvalues $\lambda_1, \ldots, \lambda_r$. We extend $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r$ to a basis

$$\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r, \boldsymbol{u}_{r+1}, \ldots, \boldsymbol{u}_m$$

for $\mathbb{R}^m$. By Gram-Schmidt, we can further assume that they are orthonormal. Therefore,

$$U = \begin{bmatrix} \boldsymbol{u}_1 & \ldots & \boldsymbol{u}_r & \boldsymbol{u}_{r+1} & \ldots & \boldsymbol{u}_m \end{bmatrix}$$

is orthogonal.

Let $i \in [n]$. If $i \le r$, then

$$A\boldsymbol{v}_i = \sqrt{\lambda_i}\boldsymbol{u}_i.$$

Otherwise, $i > r$, and then $\lambda_i = 0$. Since $\boldsymbol{v}_i$ is an eigenvector of $A^\mathsf{T} A$ for λ_i,

$$(A^\mathsf{T} A)\boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i = \mathbf{0}.$$

It follows that

$$0 = \boldsymbol{v}_i^\mathsf{T}(A^\mathsf{T} A)\boldsymbol{v}_i = A\boldsymbol{v}_i \cdot A\boldsymbol{v}_i = \|A\boldsymbol{v}_i\|,$$

thereby $A\boldsymbol{v}_i = \mathbf{0}$. We can thus deduce

$$AV = \begin{bmatrix} A\boldsymbol{v}_1 & \dots & A\boldsymbol{v}_r & A\boldsymbol{v}_{r+1} & \dots & A\boldsymbol{v}_n \end{bmatrix} = \begin{bmatrix} \sqrt{\lambda_1}\boldsymbol{u}_1 & \dots & \sqrt{\lambda_r}\boldsymbol{u}_r & \mathbf{0} & \dots & \mathbf{0} \end{bmatrix}$$

$$= \begin{bmatrix} \boldsymbol{u}_1 & \dots & \boldsymbol{u}_r & \boldsymbol{u}_{r+1} & \dots \boldsymbol{u}_m \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1} & & & & & \\ & \ddots & & & & \\ & & \sqrt{\lambda_r} & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix} = U\Sigma.$$

Since V is orthogonal, i.e., $V^{-1} = V^{\mathsf{T}}$, we conclude

$$A = U\Sigma V^{\mathsf{T}}.$$

$\square$

Bases and matrices in the SVD

Theorem

Let A be an $m \times n$ matrix with $r = \operatorname{rank}(A)$. Define an $m \times n$ matrix

$$\Sigma = \begin{bmatrix} D & \mathbf{0}_{r\times(n-r)} \\ \mathbf{0}_{(m-r)\times r} & \mathbf{0}_{(m-r)\times(n-r)} \end{bmatrix} \quad \text{for } D = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_r \end{bmatrix},$$

where $\sigma_1, \ldots, \sigma_r$ are the singular values of A. Then for some orthogonal $m \times m$ matrix U and orthogonal $n \times n$ matrix V

$$A = U\Sigma V^\mathsf{T}.$$

In particular, there are $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r \in \mathbb{R}^m$ and $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r \in \mathbb{R}^n$ such that

$$A = \sigma_1 \boldsymbol{u}_1 \boldsymbol{v}_1^\mathsf{T} + \sigma_2 \boldsymbol{u}_2 \boldsymbol{v}_2^\mathsf{T} + \cdots + \sigma_r \boldsymbol{u}_r \boldsymbol{v}_r^\mathsf{T}.$$

From

$$A = \begin{bmatrix} \boldsymbol{u}_1 & \cdots & \boldsymbol{u}_r & \boldsymbol{u}_{r+1} & \cdots & \boldsymbol{u}_m \end{bmatrix} \begin{bmatrix} \sigma_1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_1^\mathsf{T} \\ \vdots \\ \boldsymbol{v}_r^\mathsf{T} \\ \boldsymbol{v}_{r+1}^\mathsf{T} \\ \vdots \\ \boldsymbol{v}_n^\mathsf{T} \end{bmatrix}$$

$$= \sigma_1 \boldsymbol{u}_1 \boldsymbol{v}_1^\mathsf{T} + \sigma_2 \boldsymbol{u}_2 \boldsymbol{v}_2^\mathsf{T} + \cdots + \sigma_r \boldsymbol{u}_r \boldsymbol{v}_r^\mathsf{T}$$

we deduce:

Corollary

vectors	*orthonormal basis for the vector space*
$\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r$	$\boldsymbol{C}(A)$
$\boldsymbol{u}_{r+1}, \ldots, \boldsymbol{u}_m$	$\boldsymbol{N}(A^\mathsf{T})$
$\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r$	$\boldsymbol{C}(A^\mathsf{T})$
$\boldsymbol{v}_{r+1}, \ldots, \boldsymbol{v}_n$	$\boldsymbol{N}(A)$

Proof of $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r$ as a basis for $\boldsymbol{C}(A)$

As $\dim(\boldsymbol{C}(A)) = \operatorname{rank}(A)$, we only need to show

$$\boldsymbol{C}(A) \subseteq \operatorname{span}(\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r\}).$$

Recall

$$A = \sigma_1 \boldsymbol{u}_1 \boldsymbol{v}_1^\mathsf{T} + \sigma_2 \boldsymbol{u}_2 \boldsymbol{v}_2^\mathsf{T} + \cdots + \sigma_r \boldsymbol{u}_r \boldsymbol{v}_r^\mathsf{T}.$$

So (can you see?) every column of A is a linear combination of $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r$. $\square$

Proof of $\boldsymbol{v}_{r+1}, \ldots, \boldsymbol{v}_n$ as a basis for $\boldsymbol{N}(A)$

As $\dim(\boldsymbol{N}(A)) = n - \operatorname{rank}(A) = n - r$, we only need to show

$$A\boldsymbol{v}_i = \mathbf{0}$$

for every $r + 1 \leq i \leq n$. Again by

$$A = \sigma_1 \boldsymbol{u}_1 \boldsymbol{v}_1^\mathsf{T} + \sigma_2 \boldsymbol{u}_2 \boldsymbol{v}_2^\mathsf{T} + \cdots + \sigma_r \boldsymbol{u}_r \boldsymbol{v}_r^\mathsf{T},$$

together with the orthogonality of $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r, \ldots, \boldsymbol{v}_i, \ldots, \boldsymbol{v}_n$, we obtain

$$\begin{aligned} A\boldsymbol{v}_i &= \sigma_1 \boldsymbol{u}_1 \boldsymbol{v}_1^\mathsf{T} \boldsymbol{v}_i + \sigma_2 \boldsymbol{u}_2 \boldsymbol{v}_2^\mathsf{T} \boldsymbol{v}_i + \cdots + \sigma_r \boldsymbol{u}_r \boldsymbol{v}_r^\mathsf{T} \boldsymbol{v}_i \\ &= \sigma_1 \boldsymbol{u}_1 (\boldsymbol{v}_1 \cdot \boldsymbol{v}_i) + \sigma_2 \boldsymbol{u}_2 (\boldsymbol{v}_2 \cdot \boldsymbol{v}_i) + \cdots + \sigma_r \boldsymbol{u}_r (\boldsymbol{v}_r \cdot \boldsymbol{v}_i) \\ &= \mathbf{0}. \end{aligned}$$

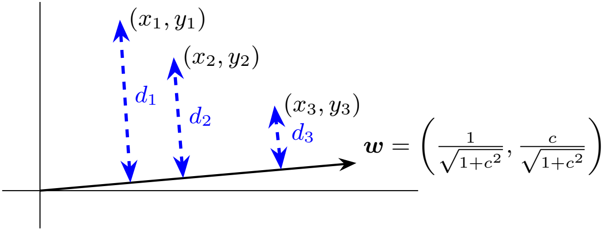
$\square$

Geometry and Applications of SVD

Assume that we have m points in $\mathbb{R}^2$

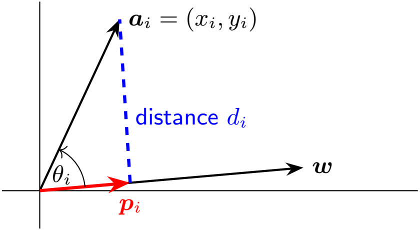
$$(x_1, y_1), \ldots, (x_m, y_m).$$

We want to find a line $y = cx$ for some $c \in \mathbb{R}$ such that (x_i, y_i)s are close to it as much as possible. Equivalently, this line can be determined by a unit vector $\boldsymbol{w} \in \mathbb{R}^2$



such that $d_1^2 + d_2^2 + \cdots + d_m^2$ is minimum.

For $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m \in \mathbb{R}^2$ we want to determine a unit vector $\boldsymbol{w} \in \mathbb{R}$



to minimize

$$\sum_{i \in [m]} d_i^2 = \sum_{i \in [m]} \left(\|\boldsymbol{a}_i\|^2 - \|\boldsymbol{p}_i\|^2 \right) = \sum_{i \in [m]} \|\boldsymbol{a}_i\|^2 - \sum_{i \in [m]} \|\boldsymbol{p}_i\|^2.$$

Equivalently, we maximize

$$\sum_{i \in [m]} \|\boldsymbol{p}_i\|^2 = \sum_{i \in [m]} (\|\boldsymbol{a}_i\| \cos\theta_i)^2 = \sum_{i \in [m]} \left(\|\boldsymbol{a}_i\| \frac{\boldsymbol{w} \cdot \boldsymbol{a}_i}{\|\boldsymbol{w}\| \|\boldsymbol{a}_i\|} \right)^2 = \sum_{i \in [m]} (\boldsymbol{w} \cdot \boldsymbol{a}_i)^2.$$

Each data now has n features. Given

$$\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m \in \mathbb{R}^n$$

we want to find a unit vector $\boldsymbol{w} \in \mathbb{R}^n$ to minimize

$$\sum_{i \in [m]} \big(\text{the distance between } \boldsymbol{a}_i \text{ and the line along } \boldsymbol{w}\big)^2.$$

It is equivalent to maximizing

$$\sum_{i \in [m]} (\boldsymbol{w} \cdot \boldsymbol{a}_i)^2,$$

whose precise proof is standard by now and will come later.

Define an $m \times n$ matrix by

$$A = \begin{bmatrix} \boldsymbol{a}_1^\mathsf{T} \\ \vdots \\ \boldsymbol{a}_m^\mathsf{T} \end{bmatrix}.$$

We look for a unit vector $\boldsymbol{w} \in \mathbb{R}^n$ such that

$$\sum_{i \in [m]} (\boldsymbol{w} \cdot \boldsymbol{a}_i)^2 = \sum_{i \in [m]} (\boldsymbol{a}_i \cdot \boldsymbol{w})^2 = \sum_{i \in [m]} \left(\boldsymbol{a}_i^\mathsf{T} \boldsymbol{w}\right)^2 = \|A\boldsymbol{w}\|^2$$

is maximized.

Theorem

Let A be an $m \times n$ matrix and σ_1 be its largest singular value. Then

$$\max_{\substack{\boldsymbol{w} \in \mathbb{R}^n \text{ with} \\ \|\boldsymbol{w}\| = 1}} \|A\boldsymbol{w}\|^2 = \sigma_1^2.$$

Moreover, the maximum is attained when $\boldsymbol{w} = \boldsymbol{v}_1$ where $\boldsymbol{v}_1$ appears in the SVD *of A*

$$A = U\Sigma V = U\Sigma \begin{bmatrix} \boldsymbol{v}_1^\mathsf{T} \\ \boldsymbol{v}_2^\mathsf{T} \\ \vdots \\ \boldsymbol{v}_n^\mathsf{T} \end{bmatrix}.$$

Orthogonality, revisited

Consider an orthonormal basis $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in \mathbb{R}^n$. Then every $\boldsymbol{w} \in \mathbb{R}^n$ can be expressed as a linear combination

$$\boldsymbol{w} = c_1\boldsymbol{v}_1 + \cdots + c_n\boldsymbol{v}_n,$$

i.e., the coordinate vector of $\boldsymbol{w}$ with respect to $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ is $(c_1, \ldots, c_n) \in \mathbb{R}^n$.

Using the orthonormality of $\boldsymbol{v}_i$s, each $c_i = \boldsymbol{v}_i \cdot \boldsymbol{w}$ and hence

$$c_i\boldsymbol{v}_i = (\boldsymbol{v}_i \cdot \boldsymbol{w})\boldsymbol{v}_i = \boldsymbol{v}_i^\mathsf{T}\boldsymbol{w}\boldsymbol{v}_i = \boldsymbol{v}_i\boldsymbol{v}_i^\mathsf{T}\boldsymbol{w},$$

i.e.,

$$\begin{aligned}\boldsymbol{w} &= (\boldsymbol{v}_1 \cdot \boldsymbol{w})\boldsymbol{v}_1 + \cdots + (\boldsymbol{v}_n \cdot \boldsymbol{w})\boldsymbol{v}_n = \boldsymbol{v}_1\boldsymbol{v}_1^T\boldsymbol{w} + \cdots + \boldsymbol{v}_n\boldsymbol{v}_n^\mathsf{T}\boldsymbol{w} \\ &= \begin{bmatrix}\boldsymbol{v}_1 & \ldots & \boldsymbol{v}_n\end{bmatrix}\begin{bmatrix}\boldsymbol{v}_1^\mathsf{T}\boldsymbol{v} \\ \vdots \\ \boldsymbol{v}_n^\mathsf{T}\boldsymbol{w}\end{bmatrix} = \underbrace{\begin{bmatrix}\boldsymbol{v}_1 & \ldots & \boldsymbol{v}_n\end{bmatrix}}_{Q}\begin{bmatrix}\boldsymbol{v}_1^\mathsf{T} \\ \vdots \\ \boldsymbol{v}_n^\mathsf{T}\end{bmatrix}\boldsymbol{w} = QQ^\mathsf{T}\boldsymbol{w}.\end{aligned}$$

This is precisely that

the least squares solution of $Q\boldsymbol{x} = \boldsymbol{b}$ is $\hat{\boldsymbol{x}} = Q^\mathsf{T}\boldsymbol{b}$.

Recall: projections using orthogonal Q

Let $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n \in \mathbb{R}^n$ be orthonormal and

$$Q = \begin{bmatrix} \boldsymbol{q}_1 & \ldots & \boldsymbol{q}_n \end{bmatrix}.$$

This is equivalent to that Q is an orthogonal matrix.

So the least squares solution of $Q\boldsymbol{x} = \boldsymbol{b}$ is $\hat{\boldsymbol{x}} = Q^{\mathsf{T}}\boldsymbol{b}$. The projection matrix is $QQ^{\mathsf{T}} = I$, hence

$$\boldsymbol{b} = QQ^{\mathsf{T}}\boldsymbol{b} = \begin{bmatrix} \boldsymbol{q}_1 & \ldots & \boldsymbol{q}_n \end{bmatrix} \begin{bmatrix} \boldsymbol{q}_1^{\mathsf{T}}\boldsymbol{b} \\ \vdots \\ \boldsymbol{q}_n^{\mathsf{T}}\boldsymbol{b} \end{bmatrix} = \boldsymbol{q}_1\boldsymbol{q}_1^{\mathsf{T}}\boldsymbol{b} + \ldots + \boldsymbol{q}_n\boldsymbol{q}_n^{\mathsf{T}}\boldsymbol{b}.$$

$\boldsymbol{b}$ is the sum of projections of $\boldsymbol{b}$ onto every line $\operatorname{span}(\{\boldsymbol{q}_i\})$.

We consider the vector space $\boldsymbol{V} = \mathbb{R}^n$ and two of its bases

$$\boldsymbol{e}_1, \dots, \boldsymbol{e}_n \quad \text{and} \quad \boldsymbol{v}_1, \dots, \boldsymbol{v}_n.$$

So the change of basis matrix M from $\boldsymbol{e}_1, \dots, \boldsymbol{e}_n$ to $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ satisfies

$$\begin{bmatrix} \boldsymbol{v}_1 & \dots & \boldsymbol{v}_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{e}_1 & \dots & \boldsymbol{e}_n \end{bmatrix} M.$$

It is easy to see (why?)

$$M = \begin{bmatrix} \boldsymbol{v}_1 & \dots & \boldsymbol{v}_n \end{bmatrix}.$$

Consider an arbitrary $\boldsymbol{w} = (w_1, \ldots, w_n) \in \boldsymbol{V}$. So the coordinate vector of $\boldsymbol{w}$ with respect to $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$ is $(w_1, \ldots, w_n) \in \mathbb{R}^n$, precisely $\boldsymbol{w}$ itself. Recall:

Theorem

Let $\overline{\boldsymbol{v}}$ and $\overline{\boldsymbol{v}}'$ be two bases for $\boldsymbol{V}$. Then for every $\boldsymbol{v} \in \mathbb{R}^n$

$$T_{\overline{\boldsymbol{v}}}(\boldsymbol{v}) = M T_{\overline{\boldsymbol{v}}'}(\boldsymbol{v}).$$

Hence the coordinate vector of $\boldsymbol{w}$ with respect to $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ is

$$M^{-1}\boldsymbol{w}.$$

If $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ are orthonormal, i.e.,

$$M = Q = \begin{bmatrix} \boldsymbol{v}_1 & \ldots & \boldsymbol{v}_n \end{bmatrix}$$

is orthogonal, then

$$M^{-1}\boldsymbol{w} = Q^{\mathsf{T}}\boldsymbol{w} \quad \text{and} \quad \boldsymbol{w} = QQ^{\mathsf{T}}\boldsymbol{w} = \boldsymbol{v}_1\boldsymbol{v}_1^T\boldsymbol{w} + \cdots + \boldsymbol{v}_n\boldsymbol{v}_n^{\mathsf{T}}\boldsymbol{w}.$$

Recall

Theorem

Let Q be an $n \times n$ orthogonal matrix and $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$.

(i) $\|Q\boldsymbol{x}\| = \|\boldsymbol{x}\|$, *hence* $\|Q^\mathsf{T}\boldsymbol{x}\| = \|\boldsymbol{x}\|$.

(ii) $Q\boldsymbol{x} \cdot Q\boldsymbol{y} = \boldsymbol{x} \cdot \boldsymbol{y}$, *hence* $Q^\mathsf{T}\boldsymbol{x} \cdot Q^\mathsf{T}\boldsymbol{y} = \boldsymbol{x} \cdot \boldsymbol{y}$.

Let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in \mathbb{R}^n$ be an orthonormal basis, or equivalently

$$Q = \begin{bmatrix} \boldsymbol{v}_1 & \ldots & \boldsymbol{v}_n \end{bmatrix}$$

is orthogonal. Consider an arbitrary vector $\boldsymbol{w} \in \mathbb{R}^n$.

(i) The coordinate vector of $\boldsymbol{w}$ with respect to $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in \mathbb{R}^n$ is $Q^\mathsf{T}\boldsymbol{w}$, i.e.,

$$\left(\boldsymbol{v}_1^\mathsf{T}\boldsymbol{w}, \ldots, \boldsymbol{v}_n^\mathsf{T}\boldsymbol{w}\right).$$

(ii) The length of the coordinate vector of $\boldsymbol{w}$ with respect to $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in \mathbb{R}^n$ equals to the length of $\boldsymbol{w}$, i.e.,

$$\|\boldsymbol{w}\| = \left\|Q^\mathsf{T}\boldsymbol{w}\right\| = \sqrt{(\boldsymbol{v}_1^\mathsf{T}\boldsymbol{w})^2 + \cdots + (\boldsymbol{v}_n^\mathsf{T}\boldsymbol{w})^2}.$$

Theorem

Let A be an $m \times n$ matrix and σ_1 be its largest singular value. Then

$$\max_{\substack{\boldsymbol{w} \in \mathbb{R}^n \text{ with} \\ \|\boldsymbol{w}\| = 1}} \|A\boldsymbol{w}\| = \sigma_1.$$

Moreover, the maximum is attained when $\boldsymbol{w} = \boldsymbol{v}_1$, where $\boldsymbol{v}_1$ appears in the SVD *of A*

$$A = U\Sigma V = U\Sigma \begin{bmatrix} \boldsymbol{v}_1^\top \\ \boldsymbol{v}_2^\top \\ \vdots \\ \boldsymbol{v}_n^\top \end{bmatrix}.$$

Let $\lambda_1 > \ldots > \lambda_r$ be the positive eigenvalues of $A^\mathsf{T} A$. For every $i \in [r]$, we define $\sigma_i = \sqrt{\lambda_i}$. Then $\sigma_1 > \ldots > \sigma_r$ are the singular values of A. Moreover, for some orthonormal $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m \in \mathbb{R}^m$ and $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in \mathbb{R}^n$

$$A = \begin{bmatrix} \boldsymbol{u}_1 & \cdots & \boldsymbol{u}_r & \boldsymbol{u}_{r+1} & \cdots & \boldsymbol{u}_m \end{bmatrix} \begin{bmatrix} \sigma_1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_1^\mathsf{T} \\ \vdots \\ \boldsymbol{v}_r^\mathsf{T} \\ \boldsymbol{v}_{r+1}^\mathsf{T} \\ \vdots \\ \boldsymbol{v}_n^\mathsf{T} \end{bmatrix}$$

$$= \sigma_1 \boldsymbol{u}_1 \boldsymbol{v}_1^\mathsf{T} + \sigma_2 \boldsymbol{u}_2 \boldsymbol{v}_2^\mathsf{T} + \cdots + \sigma_r \boldsymbol{u}_r \boldsymbol{v}_r^\mathsf{T}$$

Let $\boldsymbol{w} \in \mathbb{R}^n$ be a unit vector. So $\boldsymbol{w} = c_1\boldsymbol{v}_1 + \cdots + c_n\boldsymbol{v}_n$ for some $c_1, \ldots, c_n \in \mathbb{R}$ with $\sum_{i\in[n]} c_i^2 = 1$. Hence

$$\begin{aligned} A\boldsymbol{w} &= \left(\sigma_1\boldsymbol{u}_1\boldsymbol{v}_1^\mathsf{T} + \sigma_2\boldsymbol{u}_2\boldsymbol{v}_2^\mathsf{T} + \cdots + \sigma_r\boldsymbol{u}_r\boldsymbol{v}_r^\mathsf{T}\right)\left(c_1\boldsymbol{v}_1 + \cdots + c_n\boldsymbol{v}_n\right) \\ &= \sum_{i\in[r],j\in[n]} \sigma_i c_j \boldsymbol{u}_i\boldsymbol{v}_i^\mathsf{T}\boldsymbol{v}_j = \sum_{i\in[r]} \sigma_i c_i \boldsymbol{u}_i. \end{aligned}$$

Then the coordinate vector of $A\boldsymbol{w}$ with respect to the orthonormal basis $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r, \boldsymbol{u}_{r+1}, \ldots, \boldsymbol{u}_m$ is $(\sigma_1c_1, \ldots, \sigma_rc_r, 0, \ldots, 0)$. We deduce

$$\begin{aligned} \|A\boldsymbol{w}\|^2 &= \left\|\left(\sigma_1c_1, \ldots, \sigma_rc_r, 0, \ldots, 0\right)\right\|^2 = \sum_{i\in[r]} \sigma_i^2 c_i^2 \\ &\le \sigma_1^2 \sum_{i\in[r]} c_i^2 \le \sigma_1^2 \sum_{i\in[n]} c_i^2 = \sigma_1^2. \end{aligned}$$

And the equality can be attained by choosing $\boldsymbol{w} = \boldsymbol{v}_1$. $\square$

Fitting data by a k-dimension subspace

Fitting a set of data by a k-dimension subspace

Fix a constant $k \geq 1$. Given

$$\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m \in \mathbb{R}^n$$

we want to find a subspace $\boldsymbol{W} \subseteq \mathbb{R}^n$ with $\dim(\boldsymbol{W}) = k$ to minimize

$$\sum_{i \in [m]} \big(\text{the distance between } \boldsymbol{a}_i \text{ and } \boldsymbol{W}\big)^2.$$

Observe that $\boldsymbol{W}$ can be determined by any linearly independent

$$\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k \in \boldsymbol{W}.$$

And by Gram-Schmidt, $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$ can be chosen as orthonormal.

The distance between $\boldsymbol{a}_i$ and $\boldsymbol{W}$ is

the distance between $\boldsymbol{a}_i$ and the projection $\boldsymbol{p}_i$ of $\boldsymbol{a}_i$ onto $\boldsymbol{W}$.

Recall

1. We want to compute the projection $\boldsymbol{p}$ of $\boldsymbol{b}$ onto a subspace

$$\boldsymbol{V} = \operatorname{span}(\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n\})$$

   with linearly independent $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$. In particular $\boldsymbol{p} \in \boldsymbol{V}$.

2. We define $\boldsymbol{p}$ as a vector whose error vector $\boldsymbol{e} = \boldsymbol{b} - \boldsymbol{p}$ is perpendicular to $\boldsymbol{V}$.

3. We prove that the projection matrix is $P = A(A^\mathsf{T} A)^{-1} A^\mathsf{T}$, i.e.,

$$\boldsymbol{p} = P\boldsymbol{b} = A(A^\mathsf{T} A)^{-1} A^\mathsf{T} \boldsymbol{b}.$$

   Here, we have verified that $(A^\mathsf{T} A)^{-1}$ really exists provided $\operatorname{rank}(A) = n$.

   This also implies that $\boldsymbol{p}$ is the unique vector whose error vector $\boldsymbol{e} = \boldsymbol{b} - \boldsymbol{p}$ is perpendicular to $\boldsymbol{V}$.

4. As an exercise, we can show that for every $\boldsymbol{v} \in \boldsymbol{V}$

$$\|\boldsymbol{b} - \boldsymbol{v}\| = \min_{\boldsymbol{u} \in \boldsymbol{V}} \|\boldsymbol{b} - \boldsymbol{u}\| \quad \iff \quad \boldsymbol{v} = \boldsymbol{p}.$$

We want to determine $\boldsymbol{W}$ by choosing orthonormal

$$\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k \in \mathbb{R}^n.$$

To that end, let

$$Q = \begin{bmatrix} \boldsymbol{w}_1 & \ldots & \boldsymbol{w}_k \end{bmatrix}.$$

So

$$Q^\mathsf{T} Q = \begin{bmatrix} \boldsymbol{w}_1^\mathsf{T} \\ \vdots \\ \boldsymbol{w}_k^\mathsf{T} \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_1 & \ldots & \boldsymbol{w}_k \end{bmatrix} = \begin{bmatrix} \boldsymbol{w}_1^\mathsf{T} \boldsymbol{w}_1 & \cdots & \boldsymbol{w}_1^\mathsf{T} \boldsymbol{w}_k \\ \vdots & \ddots & \vdots \\ \boldsymbol{w}_k^\mathsf{T} \boldsymbol{w}_1 & \cdots & \boldsymbol{w}_k^\mathsf{T} \boldsymbol{w}_k \end{bmatrix} = I_{k \times k},$$

and

$$Q Q^\mathsf{T} = \begin{bmatrix} \boldsymbol{w}_1 & \ldots & \boldsymbol{w}_k \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_1^\mathsf{T} \\ \vdots \\ \boldsymbol{w}_k^\mathsf{T} \end{bmatrix} = \boldsymbol{w}_1 \boldsymbol{w}_1^\mathsf{T} + \cdots + \boldsymbol{w}_k \boldsymbol{w}_k^\mathsf{T}.$$

Then the projection of $\boldsymbol{a}_i$ onto $\boldsymbol{W} = \operatorname{span}(\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k\}) = \boldsymbol{C}(Q)$ is

$$\begin{aligned}\boldsymbol{p}_i &= Q(Q^\top Q)^{-1} Q^\top \boldsymbol{a}_i = QQ^\top \boldsymbol{a}_i = \boldsymbol{w}_1 \boldsymbol{w}_1^\top \boldsymbol{a}_i + \cdots + \boldsymbol{w}_k \boldsymbol{w}_k^\top \boldsymbol{a}_i \\ &= (\boldsymbol{w}_1 \cdot \boldsymbol{a}_i)\boldsymbol{w}_1 + \cdots + (\boldsymbol{w}_k \cdot \boldsymbol{a}_i)\boldsymbol{w}_k = \sum_{j \in [k]} (\boldsymbol{w}_j \cdot \boldsymbol{a}_i)\boldsymbol{w}_j.\end{aligned}$$

This is exactly

the sum of projections of $\boldsymbol{a}_i$ onto the lines $\operatorname{span}(\{\boldsymbol{w}_1\}), \ldots, \operatorname{span}(\{\boldsymbol{w}_k\})$.

So the distance between $\boldsymbol{a}_i$ and $\boldsymbol{W}$ is $d_i = \|\boldsymbol{a}_i - \boldsymbol{p}_i\|$, where $\boldsymbol{a}_i - \boldsymbol{p}_i$ is the previous error vector $\boldsymbol{e}$. Then (basically a previous exercise)

$$
\begin{aligned}
&\|\boldsymbol{a}_i - \boldsymbol{p}_i\|^2 \\
&= (\boldsymbol{a}_i - \boldsymbol{p}_i) \cdot (\boldsymbol{a}_i - \boldsymbol{p}_i) = (\boldsymbol{a}_i - \boldsymbol{p}_i)^\mathsf{T}(\boldsymbol{a}_i - \boldsymbol{p}_i) \\
&= \left(\boldsymbol{a}_i^\mathsf{T} - \boldsymbol{p}_i^\mathsf{T}\right)(\boldsymbol{a}_i - \boldsymbol{p}_i) = \boldsymbol{a}_i^\mathsf{T}\boldsymbol{a}_i - \boldsymbol{p}_i^\mathsf{T}\boldsymbol{a}_i - \boldsymbol{a}_i^\mathsf{T}\boldsymbol{p}_i + \boldsymbol{p}_i^\mathsf{T}\boldsymbol{p}_i \\
&= \boldsymbol{a}_i^\mathsf{T}\boldsymbol{a}_i - \boldsymbol{a}_i^\mathsf{T}\boldsymbol{p}_i - \boldsymbol{a}_i^\mathsf{T}\boldsymbol{p}_i + \boldsymbol{p}_i^\mathsf{T}\boldsymbol{p}_i = \boldsymbol{a}_i^\mathsf{T}\boldsymbol{a}_i - \boldsymbol{p}_i^\mathsf{T}\boldsymbol{p}_i - 2(\boldsymbol{a}_i^\mathsf{T}\boldsymbol{p}_i - \boldsymbol{p}_i^\mathsf{T}\boldsymbol{p}_i) \\
&= \boldsymbol{a}_i^\mathsf{T}\boldsymbol{a}_i - \boldsymbol{p}_i^\mathsf{T}\boldsymbol{p}_i - 2(\boldsymbol{a}_i - \boldsymbol{p}_i)^\mathsf{T}\boldsymbol{p}_i = \boldsymbol{a}_i^\mathsf{T}\boldsymbol{a}_i - \boldsymbol{p}_i^\mathsf{T}\boldsymbol{p}_i - 2(\boldsymbol{a}_i - \boldsymbol{p}_i) \cdot \boldsymbol{p}_i \\
&= \boldsymbol{a}_i^\mathsf{T}\boldsymbol{a}_i - \boldsymbol{p}_i^\mathsf{T}\boldsymbol{p}_i \qquad\qquad \text{(by } (\boldsymbol{a}_i - \boldsymbol{p}_i)\perp\boldsymbol{p}_i\text{)} \\
&= \|\boldsymbol{a}_i\|^2 - \|\boldsymbol{p}_i\|^2.
\end{aligned}
$$

By

$$d_i^2 = \|\boldsymbol{a}_i - \boldsymbol{p}_i\|^2 = \|\boldsymbol{a}_i\|^2 - \|\boldsymbol{p}_i\|^2 \quad \text{and} \quad \boldsymbol{p}_i = \sum_{j\in[k]} (\boldsymbol{w}_j \cdot \boldsymbol{a}_i)\boldsymbol{w}_j,$$

it is equivalent to maximizing

$$\sum_{i\in[m]} \|\boldsymbol{p}_i\|^2 = \sum_{i\in[m]} \left\| \sum_{j\in[k]} (\boldsymbol{w}_j \cdot \boldsymbol{a}_i)\boldsymbol{w}_j \right\|^2 = \sum_{i\in[m]} \sum_{j\in[k]} (\boldsymbol{w}_j \cdot \boldsymbol{a}_i)^2.$$

When $k = 1$ and $\boldsymbol{w}_1 = \boldsymbol{w}$, this is what we claimed: minimizing

$$\sum_{i\in[m]} \big(\text{the distance between } \boldsymbol{a}_i \text{ and the line along a unit vector } \boldsymbol{w}\big)^2.$$

is equivalent to maximizing

$$\sum_{i\in[m]} (\boldsymbol{w} \cdot \boldsymbol{a}_i)^2,$$

Recall we combine m points $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m \in \mathbb{R}^n$ into an $m \times n$ matrix

$$A = \begin{bmatrix} \boldsymbol{a}_1^\mathsf{T} \\ \vdots \\ \boldsymbol{a}_m^\mathsf{T} \end{bmatrix}.$$

We look for orthonormal vectors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k \in \mathbb{R}^n$ such that

$$\sum_{i \in [m]} \sum_{j \in [k]} (\boldsymbol{w}_j \cdot \boldsymbol{a}_i)^2 = \sum_{j \in [k]} \sum_{i \in [m]} (\boldsymbol{w}_j \cdot \boldsymbol{a}_i)^2 = \sum_{j \in [k]} \|A\boldsymbol{w}_j\|^2$$

is maximized.

Recall:

$$A = \begin{bmatrix} \boldsymbol{u}_1 & \cdots & \boldsymbol{u}_r & \boldsymbol{u}_{r+1} & \cdots & \boldsymbol{u}_m \end{bmatrix} \begin{bmatrix} \sigma_1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_1^\top \\ \vdots \\ \boldsymbol{v}_r^\top \\ \boldsymbol{v}_{r+1}^\top \\ \vdots \\ \boldsymbol{v}_n^\top \end{bmatrix}$$

$$= \sigma_1 \boldsymbol{u}_1 \boldsymbol{v}_1^\top + \sigma_2 \boldsymbol{u}_2 \boldsymbol{v}_2^\top + \cdots + \sigma_r \boldsymbol{u}_r \boldsymbol{v}_r^\top$$

Theorem

Assume $k \le r$. Then

$$\sum_{j \in [k]} \|A\boldsymbol{w}_j\|^2 \le \sum_{j \in [k]} \sigma_j^2.$$

And the equality holds for $\boldsymbol{w}_1 = \boldsymbol{v}_1$, ..., and $\boldsymbol{w}_k = \boldsymbol{v}_k$.