San José State University

Math 250: Mathematical Data Visualization

# Matrix norm and low-rank approximation

Dr. Guangliang Chen

## Outline

- Review of vector norms
- Matrix norms and condition number
- Low-rank matrix approximation
- Applications

## Introduction

Recall that a **vector space** is a collection $\mathcal{V}$ of objects, called "vectors", which are endowed with two kinds of operations,

- **vector addition**: $\mathbf{u}+\mathbf{v}$, for any $\mathbf{u},\mathbf{v}\in\mathcal{V}$;
- **scalar multiplication**: $k\mathbf{v}$, for any $k\in\mathbb{R},\mathbf{v}\in\mathcal{V}$

subject to requirements such as

- *Associativity*: $(\mathbf{u}+\mathbf{v})+\mathbf{w}=\mathbf{u}+(\mathbf{v}+\mathbf{w})$
- *Commutativity*: $\mathbf{u}+\mathbf{v}=\mathbf{v}+\mathbf{u}$
- *Distributivity*: $k(\mathbf{u}+\mathbf{v})=k\mathbf{u}+k\mathbf{v}$, $(s+t)\mathbf{u}=s\mathbf{u}+t\mathbf{u}$

Below are some examples of vector spaces:

- Euclidean spaces ($\mathbb{R}^n$)
- The collection of all matrices of a fixed size ($\mathbb{R}^{m\times n}$)
- The collection of all functions from $\mathbb{R}$ to $\mathbb{R}$
- The collection of all polynomials
- The collection of all infinite sequences

## Vector norm

A **norm** on a vector space $\mathcal{V}$ is a function

$$\|\cdot\| : \mathcal{V} \to \mathbb{R}$$

that satisfies the following three conditions:

- $\|\mathbf{v}\| \geq 0$ for all $\mathbf{v} \in \mathcal{V}$, and $\|\mathbf{v}\| = 0$ if and only if $\mathbf{v} = \mathbf{0}$;
- $\|k\mathbf{v}\| = |k|\|\mathbf{v}\|$ for any scalar $k \in \mathbb{R}$ and vector $\mathbf{v} \in \mathcal{V}$;
- $\|\mathbf{v} + \mathbf{w}\| \leq \|\mathbf{v}\| + \|\mathbf{w}\|$ for any two vectors $\mathbf{v}, \mathbf{w} \in \mathcal{V}$.

Note that $\|\mathbf{v}\|$ can be thought of as the **length** or **magnitude** of $\mathbf{v}$.

## $\ell_p$ norms on Euclidean spaces $\mathbb{R}^d$

For any fixed $p \geq 1$, the $\ell_p$ norm on $\mathbb{R}^d$ is defined as

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^d |x_i|^p\right)^{1/p}, \quad \text{for all } \mathbf{x} \in \mathbb{R}^d.$$

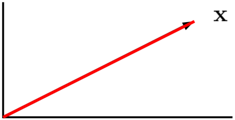
It is a rich family of vector norms.

*Remark*. For any $0 < p < 1$, the above function is no longer a vector norm, as it violates the third condition (convexity).

Three particular $\ell_p$ norms:

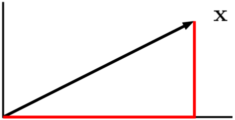
- **2-norm (Euclidean norm)**:

$$\|\mathbf{x}\|_2 = \sqrt{\sum x_i^2} = \sqrt{\mathbf{x}^T\mathbf{x}}$$



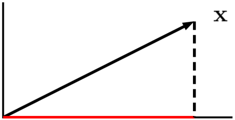

- **1-norm (Manhattan norm)**:

$$\|\mathbf{x}\|_1 = \sum |x_i|$$




- **$\infty$-norm (maximum norm)**:

$$\|\mathbf{x}\|_\infty = \max |x_i|$$

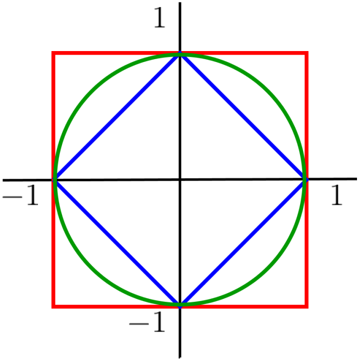


## Unit circles (under different $\ell_p$ norms)

Given any vector norm $\|\cdot\|$ on $\mathbb{R}^d$, the set of all vectors in $\mathbb{R}^d$ that have a unit norm is called a *unit circle* (under the given norm):

$$\{\mathbf{v} \in \mathbb{R}^d : \|\mathbf{v}\| = 1\}.$$

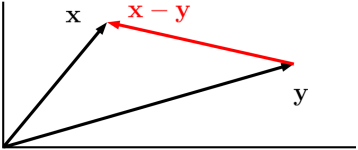
The figure on the right shows the unit circles in three different norms.




$\ell_\infty$, $\ell_2$ and $\ell_1$ unit circles

*Remark.* Any norm $\|\cdot\|$ on $\mathbb{R}^d$ can be used as a metric to measure the distance between two vectors:

$$\mathrm{dist}(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|, \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$$




For example, the Euclidean norm defines the Euclidean distance:

$$\mathrm{dist}_E(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2 = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}$$

## Matrix norm

A matrix norm is a norm on $\mathbb{R}^{m\times n}$ as a vector space (consisting of all matrices of the fixed size).

More specifically, a matrix norm is a function

$$\|\cdot\| : \mathbb{R}^{m\times n} \to \mathbb{R}$$

that satisfies the following three conditions:

- $\|\mathbf{A}\| \geq 0$ for all $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\|\mathbf{A}\| = 0$ if and only if $\mathbf{A} = \mathbf{O}$
- $\|k\mathbf{A}\| = |k| \cdot \|\mathbf{A}\|$ for any scalar $k \in \mathbb{R}$ and matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$
- $\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$ for any two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}$

Note that multiplication is also defined between matrices (with compatible sizes).

We say that a matrix norm $\|\cdot\|$ is **sub-multiplicative** if for any two matrices $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{B} \in \mathbb{R}^{n\times p}$,

$$\|\mathbf{AB}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{B}\|.$$

Note that some textbooks only regard sub-multiplicative matrix norms as matrix norms.

## The Frobenius norm

**Def 0.1.** The Frobenius norm of a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ is defined as

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2}$$

It is equivalent to the Euclidean 2-norm on vectorized matrices (i.e., $\mathbb{R}^{mn}$):

$$\|\mathbf{A}\|_F = \|\mathbf{A}(:)\|_2$$

Thus, it must satisfy all the three conditions of a norm.

**Example 0.1.** Let

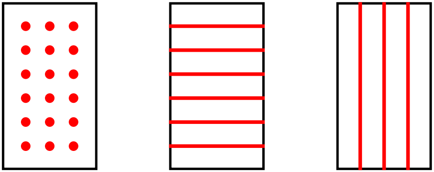
$$\mathbf{A} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

By direct calculation,

$$\|\mathbf{A}\|_F = \sqrt{1^2 + (-1)^2 + 0^2 + 1^2 + 1^2 + 0^2} = 2.$$

*Remark*. For any matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$,

$$\|\mathbf{A}\|_F^2 = \sum_{i=1}^m \|A_i\|_2^2 = \sum_{j=1}^n \|\mathbf{a}_j\|_2^2$$



*Theorem* 0.1. The matrix Frobenius norm is sub-multiplicative, that is, for any two matrices $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{B} \in \mathbb{R}^{n\times p}$,

$$\|\mathbf{AB}\|_F \leq \|\mathbf{A}\|_F \cdot \|\mathbf{B}\|_F.$$

*Proof.* Let $\mathbf{C} = \mathbf{AB} \in \mathbb{R}^{m\times p}$. By definition,

$$\|\mathbf{C}\|_F^2 = \sum_{i=1}^{m}\sum_{j=1}^{p} c_{ij}^2 = \sum_{i=1}^{m}\sum_{j=1}^{p} (A_i \mathbf{b}_j)^2.$$

Using the Cauchy-Schwarz inequality,

$$|\mathbf{x}\cdot\mathbf{y}| \leq \|\mathbf{x}\| \cdot \|\mathbf{y}\|, \quad \text{for all } \mathbf{x},\mathbf{y} \in \mathbb{R}^p, ,$$

we obtain that

$$\begin{aligned}\|\mathbf{C}\|_F^2 &\le \sum_{i=1}^m \sum_{j=1}^p \|A_i\|^2 \|\mathbf{b}_j\|^2 \\ &= \left(\sum_{i=1}^m \|A_i\|^2\right)\left(\sum_{j=1}^p \|\mathbf{b}_j\|^2\right) \\ &= \|\mathbf{A}\|_F^2 \, \|\mathbf{B}\|_F^2.\end{aligned}$$

Taking the square root of each side completes the proof. □

*Proposition* 0.2. For any matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$,

$$\|\mathbf{A}\|_F^2 = \operatorname{trace}(\mathbf{A}\mathbf{A}^T) = \operatorname{trace}(\mathbf{A}^T\mathbf{A})$$

*Proof.*

$$\operatorname{trace}(\mathbf{A}\mathbf{A}^T) = \sum_{i=1}^m A_i \cdot A_i^T = \sum_{i=1}^m \|A_i\|_2^2 = \|\mathbf{A}\|_F^2.$$

The other equality can be proved similarly, or instead using the matrix trace property:

$$\operatorname{trace}(\mathbf{A}\mathbf{B}) = \operatorname{trace}(\mathbf{B}\mathbf{A})$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{n\times m}$. □

*Theorem* 0.3. Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ be any matrix. Suppose its (nonzero) singular values are $\sigma_1 \geq \cdots \geq \sigma_r > 0$, where $r = \mathrm{rank}(\mathbf{A})$. Then

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{r} \sigma_i^2}$$

*Proof*. Consider the matrix $\mathbf{A}^T\mathbf{A}$. Its nonzero eigenvalues are $\lambda_i = \sigma_i^2$. According to the theorem on the preceding slide,

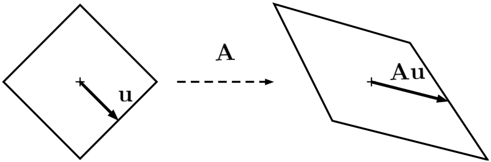
$$\|\mathbf{A}\|_F^2 = \mathrm{trace}(\mathbf{A}^T\mathbf{A}) = \sum_{i=1}^{r} \lambda_i = \sum_{i=1}^{r} \sigma_i^2.$$

## The matrix operator norm

A second matrix norm is the operator norm, which is induced by a vector norm on Euclidean spaces.

*Theorem* 0.4. For any vector norm $\|\cdot\|$ on Euclidean spaces, the following is a matrix norm on $\mathbb{R}^{m\times n}$:

$$\|\mathbf{A}\| \stackrel{\text{def}}{=} \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|} = \max_{\mathbf{u}\in\mathbb{R}^n:\|\mathbf{u}\|=1} \|\mathbf{A}\mathbf{u}\|$$



*Proof.* We need to verify the three conditions of a norm.

First, it is obvious that $\|\mathbf{A}\| \geq 0$ for any $\mathbf{A} \in \mathbb{R}^{m\times n}$. Suppose $\|\mathbf{A}\| = 0$. Then for any $\mathbf{x} \neq \mathbf{0}$, $\|\mathbf{A}\mathbf{x}\| = 0$, or equivalently, $\mathbf{A}\mathbf{x} = \mathbf{0}$. This implies that $\mathbf{A} = \mathbf{O}$. (The other direction is trivial)

Second, for any $k \in \mathbb{R}$,

$$\|k\mathbf{A}\| = \max_{\mathbf{u}\in\mathbb{R}^n:\|\mathbf{u}\|=1} \|(k\mathbf{A})\mathbf{u}\| = |k| \cdot \max_{\mathbf{u}\in\mathbb{R}^n:\|\mathbf{u}\|=1} \|\mathbf{A}\mathbf{u}\| = |k| \cdot \|\mathbf{A}\|.$$

Lastly, for any two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}$,

$$\begin{aligned}
\|\mathbf{A}+\mathbf{B}\| &= \max_{\mathbf{u}\in\mathbb{R}^n:\|\mathbf{u}\|=1} \|(\mathbf{A}+\mathbf{B})\mathbf{u}\| = \max_{\mathbf{u}\in\mathbb{R}^n:\|\mathbf{u}\|=1} \|\mathbf{A}\mathbf{u}+\mathbf{B}\mathbf{u}\| \\
&\le \max_{\mathbf{u}\in\mathbb{R}^n:\|\mathbf{u}\|=1} (\|\mathbf{A}\mathbf{u}\| + \|\mathbf{B}\mathbf{u}\|) \\
&\le \max_{\mathbf{u}\in\mathbb{R}^n:\|\mathbf{u}\|=1} \|\mathbf{A}\mathbf{u}\| + \max_{\mathbf{u}\in\mathbb{R}^n:\|\mathbf{u}\|=1} \|\mathbf{B}\mathbf{u}\| \\
&= \|\mathbf{A}\| + \|\mathbf{B}\|. \qquad \square
\end{aligned}$$

*Theorem* 0.5. For any norm on Euclidean spaces and its induced matrix operator norm, we have

$$\|\mathbf{A}\mathbf{x}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{x}\| \quad \text{for all } \mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{x} \in \mathbb{R}^n$$

*Proof.* For any particular vector $\mathbf{x} \neq \mathbf{0} \in \mathbb{R}^n$, by definition,

$$\frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|} \leq \|\mathbf{A}\| \longleftarrow \max_{\mathbf{y}\neq\mathbf{0}} \frac{\|\mathbf{A}\mathbf{y}\|}{\|\mathbf{y}\|}$$

This implies that

$$\|\mathbf{A}\mathbf{x}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{x}\|.$$

□

*Remark*. More generally, the matrix operator norm can be shown to be sub-multiplicative, i.e.,

$$\|\mathbf{AB}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{B}\|, \quad \text{for all } \mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{B} \in \mathbb{R}^{n\times p}$$

To see this, consider any nonzero $\mathbf{x} \in \mathbb{R}^p$. By using the preceding theorem, we have

$$\|\mathbf{ABx}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{Bx}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{B}\| \cdot \|\mathbf{x}\|$$

It follows that

$$\frac{\|\mathbf{ABx}\|}{\|\mathbf{x}\|} \leq \|\mathbf{A}\| \cdot \|\mathbf{B}\|.$$

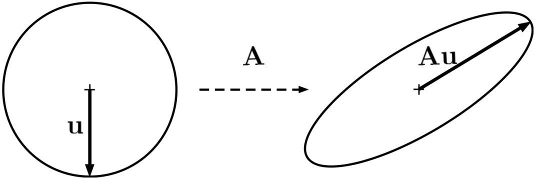
Taking the maximum of the left hand side over all nonzero $\mathbf{x}$ yields that

$$\|\mathbf{AB}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{B}\|$$

When the Euclidean norm (i.e., 2-norm) is used, the induced matrix operator norm is called the spectral norm.

**Def 0.2.** The **spectral norm** of a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ is defined as

$$\|\mathbf{A}\|_2 = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \max_{\mathbf{u}\in\mathbb{R}^n:\|\mathbf{u}\|_2=1} \|\mathbf{A}\mathbf{u}\|_2$$



*Remark*. The spectral norm of a row or column vector when regarded as a matrix always coincides with the Euclidean norm of the vector.

- Let $\mathbf{A} = [\mathbf{x}] \in \mathbb{R}^{n\times 1}$ be a single-column matrix. By definition, its spectral norm is

$$\|\mathbf{A}\|_2 = \|\mathbf{x}\cdot 1\|_2 = \|\mathbf{x}\|_2$$

- Let $\mathbf{A} = [\mathbf{x}^T] \in \mathbb{R}^{1\times n}$ be a single-row matrix. By definition, its spectral norm is

$$\|\mathbf{A}\|_2 = \max_{\mathbf{u}\in\mathbb{R}^n : \|\mathbf{u}\|_2=1} \|\mathbf{x}^T\mathbf{u}\|_2 = \|\mathbf{x}\|_2$$

*Theorem* 0.6. Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ be any matrix whose singular values (from large to small) are $\sigma_1 \geq \sigma_2 \geq \cdots$. Then

$$\|\mathbf{A}\|_2 = \sigma_1.$$

*Proof.* Consider the matrix $\mathbf{A}^T\mathbf{A}$ which is a positive semidefinite matrix with largest eigenvalue $\lambda_1 = \sigma_1^2$. We have

$$\|\mathbf{A}\|_2^2 = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|_2^2}{\|\mathbf{x}\|_2^2} = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} = \lambda_1 = \sigma_1^2,$$

where we used the Rayleigh quotient theorem. The maximizer is the largest right singular vector $\mathbf{v}_1$ of $\mathbf{A}$ (corresponding to $\sigma_1$). □

**Example 0.2.** For the matrix

$$\mathbf{A} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix},$$

we have

$$\|\mathbf{A}\|_2 = \sqrt{3}.$$

Note that we must have

$$\|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_F$$

for all matrices $\mathbf{A}$. Why?

We note that the Frobenius and spectral norms of a matrix correspond to the 2- and $\infty$-norms of the vector of singular values ($\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_r)$):

$$\|\mathbf{A}\|_F = \|\boldsymbol{\sigma}\|_2, \qquad \|\mathbf{A}\|_2 = \|\boldsymbol{\sigma}\|_\infty$$

The 1-norm of the singular value vector is called the nuclear norm of $\mathbf{A}$, which is very useful in convex programming.

**Def 0.3.** The **nuclear norm** of a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ is defined as

$$\|\mathbf{A}\|_* = \|\boldsymbol{\sigma}\|_1 = \sum \sigma_i.$$

**Example 0.3.** In the last example, $\|\mathbf{A}\|_* = \sqrt{3} + 1$.

# MATLAB function for matrix/vector norm

**norm – Matrix or vector norm.**

norm(X,2) returns the 2-norm of X.

norm(X) is the same as norm(X,2).

norm(X,'fro') returns the Frobenius norm of X.

In addition, for vectors...

norm(V,P) returns the p-norm of V defined as SUM(ABS(V).^P)^(1/P).

norm(V,Inf) returns the largest element of ABS(V).

## Condition number of a square matrix

Briefly speaking, the condition number of a square, invertible matrix is a measure of its near-singularity.

For example, both of the following matrices are invertible:

$$\mathbf{A} = \begin{pmatrix} 2 & 4 \\ 3 & 6.1 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 6 & -1 \\ -1 & 6.1 \end{pmatrix}$$

However, if we change the number 6.1 in $\mathbf{A}$ to 6, then $\mathbf{A}$ would become singular. In contrast, we need to change the number 6.1 in $\mathbf{B}$ to $\frac{1}{6}$ in order to make $\mathbf{B}$ singular. This shows that $\mathbf{A}$ is much closer to being singular than $\mathbf{B}$.

**Def 0.4.** Let $\|\cdot\|$ be any sub-multiplicative matrix norm. For any square, invertible matrix $\mathbf{A}$, the **condition number** of $\mathbf{A}$ (corresponding to this norm) is defined as

$$\kappa(\mathbf{A}) = \left\| \left( \frac{\mathbf{A}}{\|\mathbf{A}\|} \right)^{-1} \right\| = \|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\|$$

*Remark*. **Condition number has a lower bound of 1** (regardless of the matrix norm it corresponds to):

$$\kappa(\mathbf{A}) = \|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\| \geq \|\mathbf{A}\mathbf{A}^{-1}\| = \|\mathbf{I}\| = \max_{\mathbf{x} \neq \mathbf{0}} \frac{\|\mathbf{I}\mathbf{x}\|}{\|\mathbf{x}\|} = 1,$$

where in the inequality step we used the sub-multiplicative property.

*Theorem* 0.7. Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be any square, invertible matrix with singular values $\sigma_1 \geq \cdots \geq \sigma_n > 0$. Under the matrix spectral norm, the condition number of $\mathbf{A}$ is

$$\kappa(\mathbf{A}) = \frac{\sigma_1}{\sigma_n}.$$

*Proof.* Let the full SVD of the matrix $\mathbf{A}$ be $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. Since $\mathbf{A}$ is invertible, $\mathbf{\Sigma}$ is also invertible, and thus $\mathbf{A}^{-1} = \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^T$. This shows that the singular values of $\mathbf{A}^{-1}$ are $\frac{1}{\sigma_n} \geq \cdots \geq \frac{1}{\sigma_1} > 0$. It follows that

$$\kappa(\mathbf{A}) = \|\mathbf{A}\|_2 \cdot \|\mathbf{A}^{-1}\|_2 = \sigma_1 \cdot \frac{1}{\sigma_n} = \frac{\sigma_1}{\sigma_n}.$$

$\square$

*Remark*. The matrix condition number $\kappa(\mathbf{A})$ corresponding to the spectral norm has the following interpretations:

- We obtain again that $\kappa(\mathbf{A}) \geq 1$, which is because $\sigma_1 \geq \sigma_n$. In particular, $\kappa(\mathbf{A}) = 1$ if and only if all the singular values are equal: $\sigma_1 = \cdots = \sigma_n$.

- For a (nonzero) square, singular matrix $\mathbf{A}$, we must have $\sigma_n = 0$ (and $\sigma_1 > 0$). Therefore, $\kappa(\mathbf{A}) = \infty$.

- In general, a finite, large condition number means that the matrix is close to being singular. In this case, we say that the matrix $\mathbf{A}$ is ill-conditioned (for inversion). A rule of thumb is that

- $\mathbf{A}$ is severely ill-conditioned if $\kappa(\mathbf{A}) \geq 1000$ (inversion of the matrix would be numerically unstable);
- $\mathbf{A}$ is moderately ill-conditioned if $100 \leq \kappa(\mathbf{A}) < 1000$;
- $\mathbf{A}$ is not considered to be ill-conditioned if $\kappa(\mathbf{A}) < 100$ (in this case, inversion would be fine).

For example, for the two matrices on slide 29,

$$\kappa(\mathbf{A}) = 331.05, \quad \kappa(\mathbf{B}) = 1.40$$

Thus, $\mathbf{A}$ is much closer to being singular to $\mathbf{B}$ (the former is moderately ill-conditioned, while the latter is not).

Matlab implementation

**cond     Condition number with respect to inversion.**

cond(X) returns the 2-norm condition number (the ratio of the largest singular value of X to the smallest). Large condition numbers indicate a nearly singular matrix.

cond(X,P) returns the condition number of X in P-norm:

    NORM(X,P) * NORM(INV(X),P).

where P = 1, 2, inf, or 'fro'.

## Low-rank approximation of matrices

**Problem**. For any matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ and integer $k \geq 1$, find the rank-$k$ matrix $\mathbf{B}$ that is the closest to $\mathbf{A}$ (under a given norm such as Frobenius, or spectral):

$$\min_{\mathbf{B}\in\mathbb{R}^{m\times n}\,:\,\mathrm{rank}(\mathbf{B})=k} \|\mathbf{A}-\mathbf{B}\|$$

*Remark*. This problem arises in a number of tasks, e.g.,

- Data compression (and noise reduction)
- Matrix completion (and recommender systems)
- Orthogonal least squares fitting

*Theorem* 0.8 (Eckart–Young–Mirsky). Given $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $1 \le k \le r = \operatorname{rank}(\mathbf{A})$, let $\mathbf{A}_k$ be the truncated SVD of $\mathbf{A}$ with the largest $k$ terms: $\mathbf{A}_k = \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}_i^T$. Then $\mathbf{A}_k$ is the best rank-$k$ approximation to $\mathbf{A}$ in terms of both the Frobenius and spectral norms:

$$\min_{\mathbf{B}\,:\,\operatorname{rank}(\mathbf{B})=k} \|\mathbf{A}-\mathbf{B}\|_F = \|\mathbf{A}-\mathbf{A}_k\|_F = \sqrt{\sum_{i>k} \sigma_i^2}$$

$$\min_{\mathbf{B}\,:\,\operatorname{rank}(\mathbf{B})=k} \|\mathbf{A}-\mathbf{B}\|_2 = \|\mathbf{A}-\mathbf{A}_k\|_2 = \sigma_{k+1}.$$

*Remark*. The theorem still holds true if the equality constraint $\operatorname{rank}(\mathbf{B}) = k$ is relaxed to the inequality constraint $\operatorname{rank}(\mathbf{B}) \le k$ (which will also include all the lower-rank matrices).

**Example 0.4.** For the matrix

$$\mathbf{A} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix},$$

the best rank-1 approximation is

$$\mathbf{A}_1 = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T = \sqrt{3} \begin{pmatrix} \frac{2}{\sqrt{6}} \\ -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ -\frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{pmatrix}.$$

In this problem, the approximation error under either norm (spectral or Frobenius) is the same: $\|\mathbf{A} - \mathbf{A}_1\| = \sigma_2 = 1$.

## Application to image compression

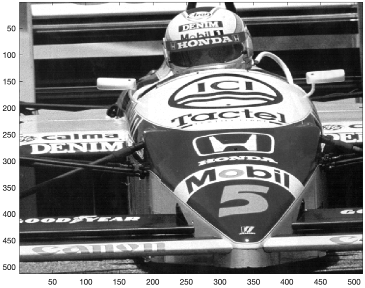
Digital images are stored as matrices, so we can apply SVD to obtain their low-rank approximations (and display them as images):

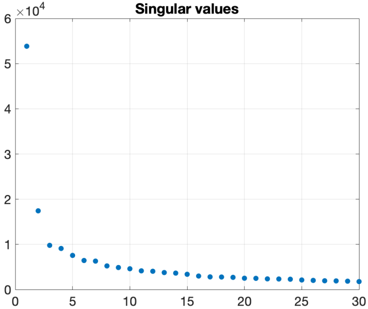
$$\mathbf{A}_{m\times n} \approx \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^T = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^T.$$

By storing $\mathbf{U}_k, \mathbf{\Sigma}_k, \mathbf{V}_k$ instead of $\mathbf{A}$, we can reduce the storage requirement from $mn$ to

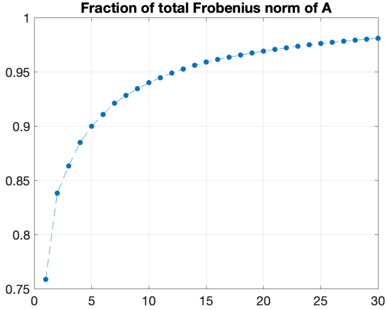
$$\underbrace{mk}_{\text{cost of } \mathbf{U}_k} + \underbrace{k}_{\text{cost of } \mathbf{\Sigma}_k} + \underbrace{nk}_{\text{cost of } \mathbf{V}_k} = (m+n+1)k.$$

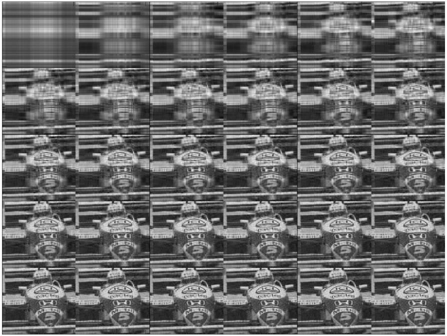
This is one magnitude smaller when $k \ll \min(m,n)$.


50
100
150
200
250
300
350
400
450
500
50
100
150
200
250
300
350
400
450
500


Singular values
$\times 10^4$
6
5
4
3
2
1
0
0
5
10
15
20
25
30

# Matrix norm and low-rank approximation

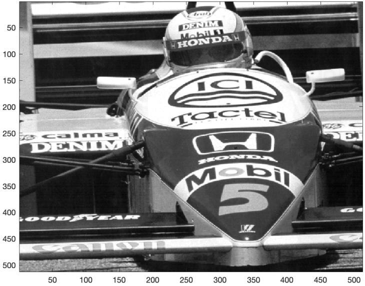


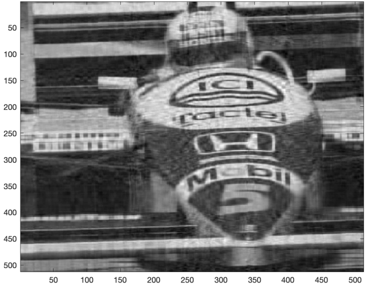



Fraction of total Frobenius norm is defined as

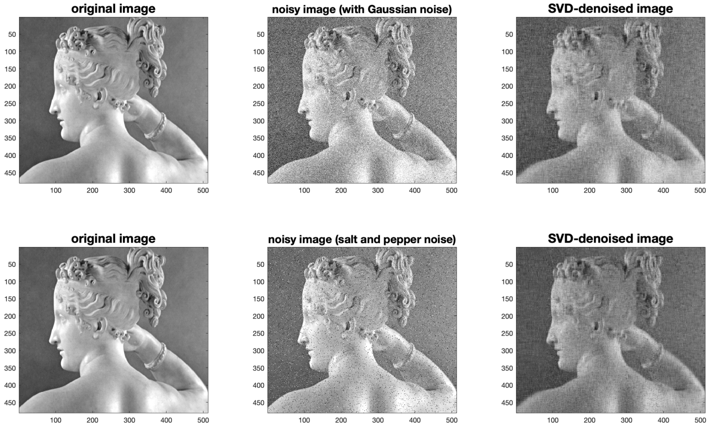
$$\frac{\|\mathbf{A}_k\|_F^2}{\|\mathbf{A}\|_F^2} = \frac{\sum_{i=1}^k \sigma_i^2}{\sum_{i=1}^r \sigma_i^2}, \quad \text{for all } k = 1, \ldots, r$$

Original image vs SVD-compressed image



# Application to image denoising



## The need of a redundant basis

The SVD basis is orthogonal (such that there is a unique representation), but it is too restrictive (not sufficiently representative).

There has been much research to use an overcomplete basis (called dictionary) for sparsely representing the data, e.g.,

- Tutorial on dictionary learning[1]
- A presentation on K-SVD[2]

[1] https://www.math.ucla.edu/~deanna/AMSnotes.pdf
[2] https://elad.cs.technion.ac.il/wp-content/uploads/2018/02/School-of-ICASSP-Sparse-Representations.pdf

## Application to recommender systems

| | Movie 1 | Movie 2 | Movie 3 | | $\cdots$ | | Movie $n$ |
|---|---|---|---|---|---|---|---|
| User 1 | | 4 | | | | | 3 |
| User 2 | 5 | | | 4 | | | |
| User 3 | | | 3 | 4 | | | 5 |
| $\vdots$ | | | | | | | |
| User $m$ | | 4 | 1 | | 3 | | |

See `https://web.stanford.edu/~hastie/TALKS/SVD_hastie.pdf`

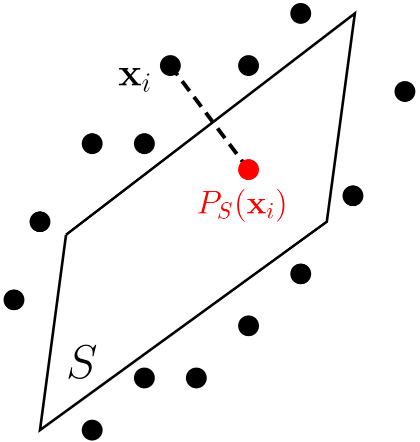
# Application to orthogonal least squares fitting

**Problem**: Given data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ and an integer $0 < k < d$, find the $k$-D orthogonal "best-fit" plane by solving

$$\min_S \sum_{i=1}^n \|\mathbf{x}_i - \mathcal{P}_S(\mathbf{x}_i)\|_2^2$$

*Remark*. This problem is different from ordinary linear regression:

- No predictor-response distinction
- Orthogonal (not vertical) fitting errors



*Theorem* 0.9. An orthogonal best-fit $k$-dimensional plane to the data $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T \in \mathbb{R}^{n\times d}$ is given by

$$\mathbf{x}(\boldsymbol{\alpha}) = \bar{\mathbf{x}} + \mathbf{V}_k \cdot \boldsymbol{\alpha}$$

where $\bar{\mathbf{x}}$ is the center of the data set

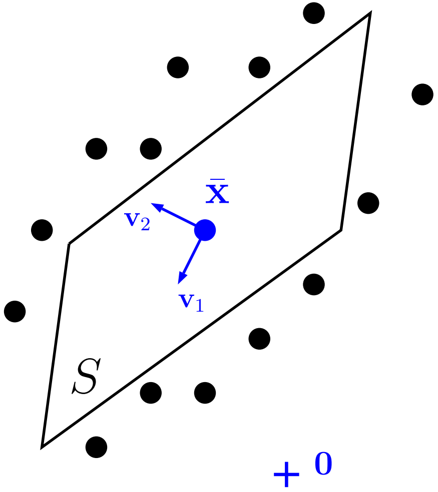
$$\bar{\mathbf{x}} = \frac{1}{n} \sum \mathbf{x}_i$$

and

$$\mathbf{V}_k = [\mathbf{v}_1 \dots \mathbf{v}_k] \in \mathbb{R}^{d\times k}$$

contains the top $k$ right singular vectors of the centered data matrix

$$\widetilde{\mathbf{X}} = \mathbf{X} - \mathbf{1}\bar{\mathbf{x}}^T = \mathbf{C}\mathbf{X}.$$



*Proof.* Suppose an arbitrary $k$-dimensional plane $\mathcal{S}$ is used to fit the data, with a fixed point $\mathbf{m} \in \mathbb{R}^d$, and an orthonormal basis

$$\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_k] \in \mathbb{R}^{d\times k}.$$

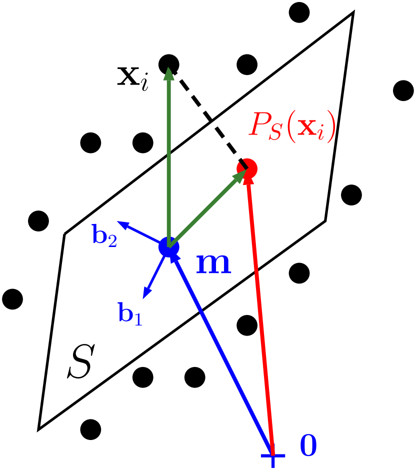
That is,

$\mathbf{B}^T\mathbf{B} = \mathbf{I}_k,$

$\mathbf{B}\mathbf{B}^T$ : orthogonal projection onto $S$

The projection of each data point $\mathbf{x}_i$ onto the candidate plane is

$$\mathcal{P}_S(\mathbf{x}_i) = \mathbf{m} + \mathbf{B}\mathbf{B}^T(\mathbf{x}_i - \mathbf{m}).$$



Accordingly, we may rewrite the original problem as

$$\min_{\substack{\mathbf{m}\in\mathbb{R}^d,\, \mathbf{B}\in\mathbb{R}^{d\times k}\\ \mathbf{B}^T\mathbf{B}=\mathbf{I}_k}} \sum_{i=1}^n \|\mathbf{x}_i - \mathbf{m} - \mathbf{B}\mathbf{B}^T(\mathbf{x}_i - \mathbf{m})\|^2$$

We show later that for any fixed choice $\mathbf{B}$, an optimal $\mathbf{m}$ is

$$\mathbf{m}^* = \frac{1}{n}\sum \mathbf{x}_i \stackrel{\text{def}}{=} \bar{\mathbf{x}}.$$

Plugging in $\bar{\mathbf{x}}$ for $\mathbf{m}$ and letting $\tilde{\mathbf{x}}_i = \mathbf{x}_i - \bar{\mathbf{x}}$ gives that

$$\min_{\mathbf{B}} \sum \|\tilde{\mathbf{x}}_i - \mathbf{B}\mathbf{B}^T\tilde{\mathbf{x}}_i\|^2.$$

In matrix notation, this becomes

$$\min_{\mathbf{B}} \|\widetilde{\mathbf{X}} - \widetilde{\mathbf{X}}\mathbf{B}\mathbf{B}^T\|_F^2, \qquad \text{where} \quad \widetilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n]^T \in \mathbb{R}^{n\times d}.$$

Since

$$\mathrm{rank}(\widetilde{\mathbf{X}}\mathbf{B}\mathbf{B}^T) \leq \mathrm{rank}(\mathbf{B}) = k,$$

any minimizer $\mathbf{B}$ should be such that

$$\widetilde{\mathbf{X}}\mathbf{B}\mathbf{B}^T = \widetilde{\mathbf{X}}_k,$$

where $\widetilde{\mathbf{X}}_k = \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^T$ is the best rank-$k$ approximation of $\widetilde{\mathbf{X}}$.

This equation also infinitely many solutions but a simple solution is

$$\mathbf{B} = \mathbf{V}_k.$$

Verify:

$$\widetilde{\mathbf{X}}\mathbf{V}_k\mathbf{V}_k^T = \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^T = \widetilde{\mathbf{X}}_k.$$

**Proof of $\mathbf{m}^* = \bar{\mathbf{x}}$:**

First, rewrite the above objective function as

$$g(\mathbf{m}) = \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{m} - \mathbf{B}\mathbf{B}^T(\mathbf{x}_i - \mathbf{m})\|^2 = \sum_{i=1}^{n} \|(\mathbf{I} - \mathbf{B}\mathbf{B}^T)(\mathbf{x}_i - \mathbf{m})\|^2$$

and apply the formula

$$\frac{\partial}{\partial \mathbf{x}} \|\mathbf{A}\mathbf{x}\|^2 = 2\mathbf{A}^T\mathbf{A}\mathbf{x}$$

to find its gradient:

$$\nabla g(\mathbf{m}) = -\sum 2(\mathbf{I} - \mathbf{B}\mathbf{B}^T)^T(\mathbf{I} - \mathbf{B}\mathbf{B}^T)(\mathbf{x}_i - \mathbf{m})$$

Note that $\mathbf{I} - \mathbf{B}\mathbf{B}^T$ is also an orthogonal projection matrix (onto the complement). Thus,

$$(\mathbf{I} - \mathbf{B}\mathbf{B}^T)^T(\mathbf{I} - \mathbf{B}\mathbf{B}^T) = (\mathbf{I} - \mathbf{B}\mathbf{B}^T)^2 = \mathbf{I} - \mathbf{B}\mathbf{B}^T.$$

It follows that

$$\nabla g(\mathbf{m}) = -\sum 2(\mathbf{I} - \mathbf{B}\mathbf{B}^T)(\mathbf{x}_i - \mathbf{m}) = -2(\mathbf{I} - \mathbf{B}\mathbf{B}^T)\left(\sum \mathbf{x}_i - n\mathbf{m}\right)$$

Any minimizer $\mathbf{m}$ must satisfy

$$2(\mathbf{I} - \mathbf{B}\mathbf{B}^T)\left(\sum \mathbf{x}_i - n\mathbf{m}\right) = 0$$

This equation has infinitely many solutions, but the simplest one is

$$\sum \mathbf{x}_i - n\mathbf{m} = \mathbf{0} \qquad \longrightarrow \qquad \mathbf{m} = \frac{1}{n}\sum \mathbf{x}_i.$$

**Example 0.5.** Find the orthogonal best-fit line for a data set of three points $(-3, 1), (-2, 3), (-1, 2)$.

*Solution*. First, the centroid of the data is $\bar{\mathbf{x}} = (-2, 2)$. Thus, the centered data matrix is

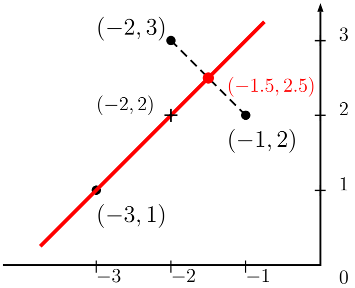
$$\widetilde{\mathbf{X}} = \begin{bmatrix} -1 & -1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \quad \xrightarrow{\text{svd}} \quad \mathbf{v}_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}.$$

Therefore, the orthogonal best-fit line is

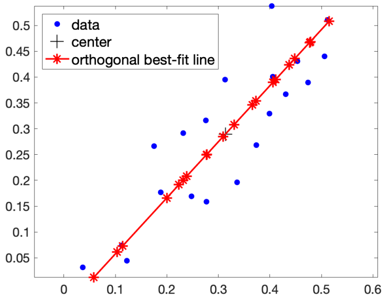
$$\mathbf{x}(t) = (-2, 2) + \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right) t,$$

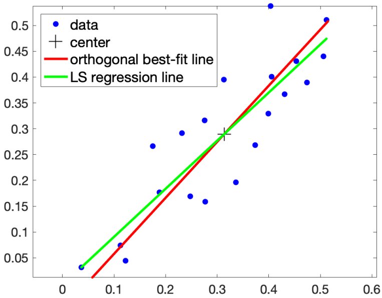
The projections of the original data onto the best-fit line are

$$\mathbf{1}\bar{\mathbf{x}}^T + \widetilde{\mathbf{X}}\mathbf{v}_1\mathbf{v}_1^T = \begin{bmatrix} -2 & 2 \\ -2 & 2 \\ -2 & 2 \end{bmatrix} + \begin{bmatrix} -1 & -1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} -3 & 1 \\ -\frac{3}{2} & \frac{5}{2} \\ -\frac{3}{2} & \frac{5}{2} \end{bmatrix}$$

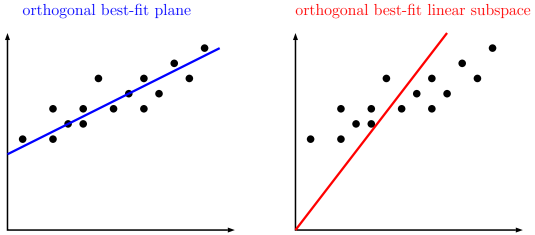


## Demonstration on another data set



## Orthogonal best-fit linear subspace



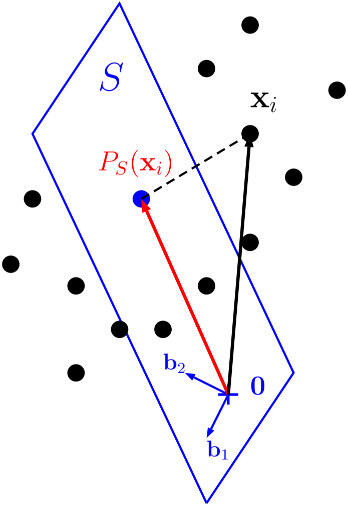

*Remark*. The orthogonal best-fit linear subspace in general differs from the orthogonal best-fit plane, with the latter fitting more closely the given data. Additionally, the orthogonal best-fit plane must go through the centroid of the data while the orthogonal best-fit linear subspace does not need to.

*Theorem* 0.10. Given a data set $\mathbf{X} = [\mathbf{x}_1 \ldots \mathbf{x}_n]^T \in \mathbb{R}^{n\times d}$ and an integer $0 < k < d$, a $k$-dimensional linear subspace that minimizes the orthogonal fitting error is given by

$$\mathbf{x}(\boldsymbol{\alpha}) = \mathbf{V}_k \cdot \boldsymbol{\alpha}, \quad \boldsymbol{\alpha} \in \mathbb{R}^k,$$

where $\mathbf{V}_k \in \mathbb{R}^{d\times k}$ contains the top right $k$ singular vectors of $\mathbf{X}$.

*Remark*. $\mathbf{X}$ is not centered when applying SVD to it.



*Proof.* Let $S$ be a linear subspace, with orthonormal basis $\mathbf{B} \in \mathbb{R}^{d\times k}$, used to fit the given data set. The orthogonal projection of an arbitrary data point $\mathbf{x}_i$ onto $S$ is

$$\mathcal{P}_S(\mathbf{x}_i) = \mathbf{B}\mathbf{B}^T\mathbf{x}_i, \quad i = 1, \ldots, n$$

The total orthogonal fitting error is thus

$$\sum_{i=1}^n \|\mathbf{x}_i - \mathcal{P}_S(\mathbf{x}_i)\|^2 = \sum_{i=1}^n \|\mathbf{x}_i - \mathbf{B}\mathbf{B}^T\mathbf{x}_i\|^2 = \|\mathbf{X} - \mathbf{X}\mathbf{B}\mathbf{B}^T\|_F^2.$$

To minimize the fitting error, we set

$$\mathbf{X}\mathbf{B}\mathbf{B}^T = \mathbf{X}_k = \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k$$

and find that $\mathbf{B} = \mathbf{V}_k$ solves the equation. $\square$

**Example 0.6.** Find the orthogonal best-fit linear line for the data set in the preceding example.

*Solution*. By direct calculation

$$\mathbf{X} = \begin{pmatrix} -3 & 1 \\ -2 & 3 \\ -1 & 2 \end{pmatrix} \quad \xrightarrow{\text{svd}} \quad \mathbf{v}_1 = \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}.$$

Therefore, the orthogonal best-fit linear line is

$$\mathbf{x}(t) = \left(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right) t,$$

and the projections of the original data onto the line are

$$\mathbf{X}\mathbf{v}_1\mathbf{v}_1^T = \begin{pmatrix} -3 & 1 \\ -2 & 3 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} = \begin{pmatrix} -2 & 2 \\ -\frac{5}{2} & \frac{5}{2} \\ -\frac{3}{2} & \frac{3}{2} \end{pmatrix}.$$