Matrices

DS-GA 1013 / MATH-GA 2824 Optimization-based Data Analysis

http://www.cims.nyu.edu/~cfgranda/pages/OBDA_fall17/index.html

Carlos Fernandez-Granda

Basic properties

Singular value decomposition

Denoising

Collaborative filtering

Principal component analysis

Probabilistic interpretation

Dimensionality reduction

Eigendecomposition

Column and row space

The column space $\operatorname{col}(A)$ of a matrix A is the span of its columns

The row space $\operatorname{row}(A)$ is the span of its rows.

Rank

For any matrix A

$$\dim(\text{col}(A)) = \dim(\text{row}(A))$$

This is the rank of A

Orthogonal column spaces

If the column spaces of $A, B \in \mathbb{R}^{m \times n}$ are orthogonal then

$$\langle A, B \rangle = 0$$

Proof:

$$\langle A, B \rangle := \operatorname{tr}\left(A^T B\right) = \sum_{i=1}^{n} \langle A_{:,i}, B_{:,i} \rangle = 0$$

Consequence:

$$||A + B||_{\mathrm{F}}^2 = ||A||_{\mathrm{F}}^2 + ||B||_{\mathrm{F}}^2$$

Linear maps

Given two vector spaces $\mathcal{V}$ and $\mathcal{R}$ associated to the same scalar field, a linear map $f : \mathcal{V} \rightarrow \mathcal{R}$ is a map from vectors in $\mathcal{V}$ to vectors in $\mathcal{R}$ such that for any scalar α and any vectors $\vec{x}_1, \vec{x}_2 \in \mathcal{V}$

$$f\left(\vec{x}_1 + \vec{x}_2\right) = f\left(\vec{x}_1\right) + f\left(\vec{x}_2\right)$$

$$f\left(\alpha\, \vec{x}_1\right) = \alpha\, f\left(\vec{x}_1\right)$$

Matrix-vector product

The product of a matrix $A \in \mathbb{C}^{m \times n}$ and a vector $\vec{x} \in \mathbb{C}^n$ is a vector $A\vec{x} \in \mathbb{C}^m$, such that

$$(A\vec{x})[i] = \sum_{j=1}^{n} A_{ij}\vec{x}[j]$$

$A\vec{x}$ is a linear combination of the columns of A

$$A\vec{x} = \sum_{j=1}^{n} \vec{x}[j] A_{:j}$$

Equivalence between matrices and linear maps

For finite m, n every linear map $f : \mathbb{C}^m \to \mathbb{C}^n$ can be uniquely represented by a matrix $F \in \mathbb{C}^{m \times n}$

Proof

The matrix is

$$F := \begin{bmatrix} f\left(\vec{e}_1\right) & f\left(\vec{e}_2\right) & \cdots & f\left(\vec{e}_n\right) \end{bmatrix},$$

the columns are the result of applying f to the standard basis

For any vector $\vec{x} \in \mathbb{C}^n$

$$\begin{aligned} f(x) &= f\left(\sum_{i=1}^{n} \vec{x}[i]\vec{e}_i\right) \\ &= \sum_{i=1}^{n} \vec{x}[i] f\left(\vec{e}_i\right) \\ &= F\vec{x} \end{aligned}$$

Projecting and lifting

When a matrix $\mathbb{C}^{m \times n}$ is *fat*, i.e., $n > m$, we say that it projects vectors onto a lower dimensional space (this is not an orthogonal projection!)

When a matrix is *tall*, i.e., $m > n$, we say that it lifts vectors to a higher-dimensional space

Adjoint

Given two vector spaces $\mathcal{V}$ and $\mathcal{R}$ with inner products $\langle \cdot, \cdot \rangle_{\mathcal{V}}$ and $\langle \cdot, \cdot \rangle_{\mathcal{R}}$, the adjoint $f^* : \mathcal{R} \rightarrow \mathcal{V}$ of a linear map $f : \mathcal{V} \rightarrow \mathcal{R}$ satisfies

$$\langle f(\vec{x}), \vec{y} \rangle_{\mathcal{R}} = \langle \vec{x}, f^*(\vec{y}) \rangle_{\mathcal{V}}$$

for all $\vec{x} \in \mathcal{V}$ and $\vec{y} \in \mathcal{R}$

Conjugate transpose

The entries of the conjugate transpose $A^* \in \mathbb{C}^{n\times m}$ of $A \in \mathbb{C}^{m\times n}$ are

$$(A^*)_{ij} = \overline{A_{ji}}, \quad 1 \le i \le n,\ 1 \le j \le m.$$

If the entries are all real, this is just the transpose

The adjoint $f^* : \mathbb{C}^n \to \mathbb{C}^m$ of a $f : \mathbb{C}^m \to \mathbb{C}^n$ represented by a matrix F corresponds to F^*

Symmetric and Hermitian

A symmetric matrix is equal to its transpose

A Hermitian or self-adjoint matrix is equal to its conjugate transpose

Range

Let $\mathcal{V}$ and $\mathcal{R}$ be vector spaces associated to the same scalar field, the range of a map $f : \mathcal{V} \to \mathcal{R}$ is the set of vectors in $\mathcal{R}$ reached by f

$$\text{range}(f) := \{\vec{y} \mid \vec{y} = f(\vec{x}) \quad \text{for some } \vec{x} \in \mathcal{V}\}$$

The range of a matrix is the range of its associated linear map

The range is the column space

For any matrix $A \in \mathbb{C}^{m \times n}$

$$\text{range}\,(A) = \text{col}\,(A)$$

Proof:

$\text{col}\,(A) \subseteq \text{range}\,(A)$ because $A_{:i} = A\vec{e}_i$ for $1 \leq i \leq n$

$\text{range}\,(A) \subseteq \text{col}\,(A)$ because $A\vec{x}$ is a linear combination of the columns of A for any $\vec{x} \in \mathbb{C}^n$

Null space

Let $\mathcal{V}$ and $\mathcal{R}$ be vector spaces, the null space of a map $f : \mathcal{V} \to \mathcal{R}$ is the set of vectors that are mapped to zero:

$$\operatorname{null}(f) := \left\{ \vec{x} \mid f(\vec{x}) = \vec{0} \right\}$$

The null space of a matrix is the null space of its associated linear map

The null space of a linear map is a subspace

Null space and row space

For any matrix $A \in \mathbb{R}^{m \times n}$

$$\text{null}(A) = \text{row}(A)^{\perp}$$

Proof:

Any vector $\vec{x} \in \text{row}(A)$ can be written as $\vec{x} = A^T \vec{z}$, for some $\vec{z} \in \mathbb{R}^m$

If $y \in \text{null}(A)$ then

$$\langle \vec{y}, \vec{x} \rangle =$$

Null space and row space

For any matrix $A \in \mathbb{R}^{m \times n}$

$$\text{null}\,(A) = \text{row}\,(A)^{\perp}$$

Proof:

Any vector $\vec{x} \in \text{row}\,(A)$ can be written as $\vec{x} = A^T \vec{z}$, for some $\vec{z} \in \mathbb{R}^m$

If $y \in \text{null}\,(A)$ then

$$\langle \vec{y}, \vec{x} \rangle = \left\langle \vec{y}, A^T \vec{z} \right\rangle$$

Null space and row space

For any matrix $A \in \mathbb{R}^{m \times n}$

$$\operatorname{null}(A) = \operatorname{row}(A)^{\perp}$$

Proof:

Any vector $\vec{x} \in \operatorname{row}(A)$ can be written as $\vec{x} = A^T \vec{z}$, for some $\vec{z} \in \mathbb{R}^m$

If $y \in \operatorname{null}(A)$ then

$$\begin{aligned} \langle \vec{y}, \vec{x} \rangle &= \left\langle \vec{y}, A^T \vec{z} \right\rangle \\ &= \langle A\vec{y}, \vec{z} \rangle \end{aligned}$$

Null space and row space

For any matrix $A \in \mathbb{R}^{m \times n}$

$$\text{null}(A) = \text{row}(A)^{\perp}$$

Proof:

Any vector $\vec{x} \in \text{row}(A)$ can be written as $\vec{x} = A^T \vec{z}$, for some $\vec{z} \in \mathbb{R}^m$

If $y \in \text{null}(A)$ then

$$\begin{aligned} \langle \vec{y}, \vec{x} \rangle &= \left\langle \vec{y}, A^T \vec{z} \right\rangle \\ &= \langle A\vec{y}, \vec{z} \rangle \\ &= 0 \end{aligned}$$

Null space and range

Let $A \in \mathbb{R}^{m \times n}$

$$\dim(\mathrm{range}(A)) + \dim(\mathrm{null}(A)) = n$$

Identity matrix

The identity matrix of dimensions $n \times n$ is

$$I = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

For any $\vec{x} \in \mathbb{C}^n$, $I\vec{x} = \vec{x}$

Matrix inverse

The inverse of $A \in \mathbb{C}^{n \times n}$ is a matrix $A^{-1} \in \mathbb{C}^{n \times n}$ such that

$$AA^{-1} = A^{-1}A = I$$

Orthogonal matrices

An orthogonal matrix is a square matrix such that

$$U^T U = U U^T = I$$

The columns $U_{:1}$, $U_{:2}$, ..., $U_{:n}$ form an orthonormal basis

For any $\vec{x} \in \mathbb{R}^n$

$$\vec{x} = U U^T \vec{x} = \sum_{i=1}^{n} \langle U_{:i}, \vec{x} \rangle \, U_{:i}$$

Product of orthogonal matrices

If $U, V \in \mathbb{R}^{n \times n}$ are orthogonal matrices, then UV is also an orthogonal matrix

$$(UV)^T(UV) = V^T U^T UV = I$$

Orthogonal matrices

Orthogonal matrices change the direction of vectors, not their magnitude

$$||U\vec{x}||_2^2$$

Orthogonal matrices

Orthogonal matrices change the direction of vectors, not their magnitude

$$||U\vec{x}||_2^2 = \vec{x}^T U^T U \vec{x}$$

Orthogonal matrices

Orthogonal matrices change the direction of vectors, not their magnitude

$$
\begin{aligned}
||U\vec{x}||_2^2 &= \vec{x}^T U^T U \vec{x} \\
&= \vec{x}^T \vec{x}
\end{aligned}
$$

Orthogonal matrices

Orthogonal matrices change the direction of vectors, not their magnitude

$$\begin{aligned} ||U\vec{x}||_2^2 &= \vec{x}^T U^T U \vec{x} \\ &= \vec{x}^T \vec{x} \\ &= ||\vec{x}||_2^2 \end{aligned}$$

Orthogonal-projection matrix

Given a subspace $\mathcal{S} \subseteq \mathbb{R}^n$ of dimension d, the matrix

$$P := UU^T$$

where the columns of $U_{:1}$, $U_{:2}$, ..., $U_{:d}$ are an orthonormal basis of $\mathcal{S}$, maps any vector $\vec{x}$ to $\mathcal{P}_{\mathcal{S}}\,\vec{x}$

$$\begin{aligned} P\vec{x} &= UU^T\vec{x} \\ &= \sum_{i=1}^{d} \langle U_{:i}, \vec{x} \rangle \, U_{:i} \\ &= \mathcal{P}_{\mathcal{S}}\,\vec{x} \end{aligned}$$

Singular value decomposition

Every rank r real matrix $A \in R^{m \times n}$, has a singular-value decomposition (SVD) of the form

$$A = \begin{bmatrix} \vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_r \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & \sigma_r \end{bmatrix} \begin{bmatrix} \vec{v}_1^T \\ \vec{v}_2^T \\ \vdots \\ \vec{v}_r^T \end{bmatrix}$$

$$= USV^T$$

Singular value decomposition

- The singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$ are positive real numbers
- The left singular vectors $\vec{u}_1, \vec{u}_2, \ldots \vec{u}_r$ form an orthonormal set
- The right singular vectors $\vec{v}_1, \vec{v}_2, \ldots \vec{v}_r$ also form an orthonormal set
- The SVD is unique if all the singular values are different
- If $\sigma_i = \sigma_{i+1} = \ldots = \sigma_{i+k}$, then $\vec{u}_i, \ldots, \vec{u}_{i+k}$ can be replaced by any orthonormal basis of their span (the same holds for $\vec{v}_i, \ldots, \vec{v}_{i+k}$)
- The SVD of an $m \times n$ matrix with $m \geq n$ can be computed in $\mathcal{O}\left(mn^2\right)$

Column and row space

- The left singular vectors $\vec{u}_1$, $\vec{u}_2$, $\ldots \vec{u}_r$ are a basis for the column space
- The right singular vectors $\vec{v}_1$, $\vec{v}_2$, $\ldots \vec{v}_r$ are a basis for the row space

Proof:

$\text{span}\left(\vec{u}_1, \ldots, \vec{u}_r\right) \subseteq \text{col}\left(A\right)$ since $\vec{u}_i = A\left(\sigma_i^{-1}\vec{v}_i\right)$

$\text{col}\left(A\right) \subseteq \text{span}\left(\vec{u}_1, \ldots, \vec{u}_r\right)$ because $A_{:i} = U\left(SV^T\vec{e}_i\right)$

Singular value decomposition

$$A = [\underbrace{\vec{u}_1 \cdots \vec{u}_r}_{\text{Basis of range}(A)} \quad \vec{u}_{r+1} \cdots \vec{u}_n] \begin{bmatrix} \sigma_1 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & \sigma_r & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ & & \cdots & \cdots & & \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix} [\underbrace{\vec{v}_1\ \vec{v}_2 \cdots \vec{v}_r}_{\text{Basis of row}(A)} \ \underbrace{\vec{v}_{r+1} \cdots \vec{v}_n}_{\text{Basis of null}(A)}]^T$$

Rank and numerical rank

The rank of a matrix is equal to the number of nonzero singular values

Given a tolerance $\epsilon > 0$, the numerical rank is the number of singular values greater than ϵ

Linear maps

The SVD decomposes the action of a matrix $A \in \mathbb{R}^{m \times n}$ on a vector $\vec{x} \in \mathbb{R}^n$ into:

1. Rotation

$$V^T \vec{x} = \sum_{i=1}^{n} \langle \vec{v}_i, \vec{x} \rangle \vec{e}_i$$

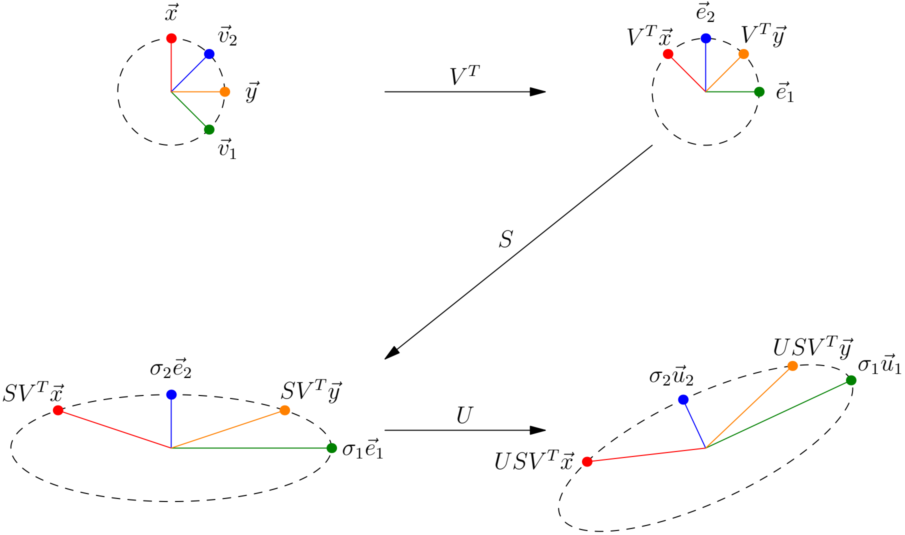
2. Scaling

$$SV^T \vec{x} = \sum_{i=1}^{n} \sigma_i \langle \vec{v}_i, \vec{x} \rangle \vec{e}_i$$

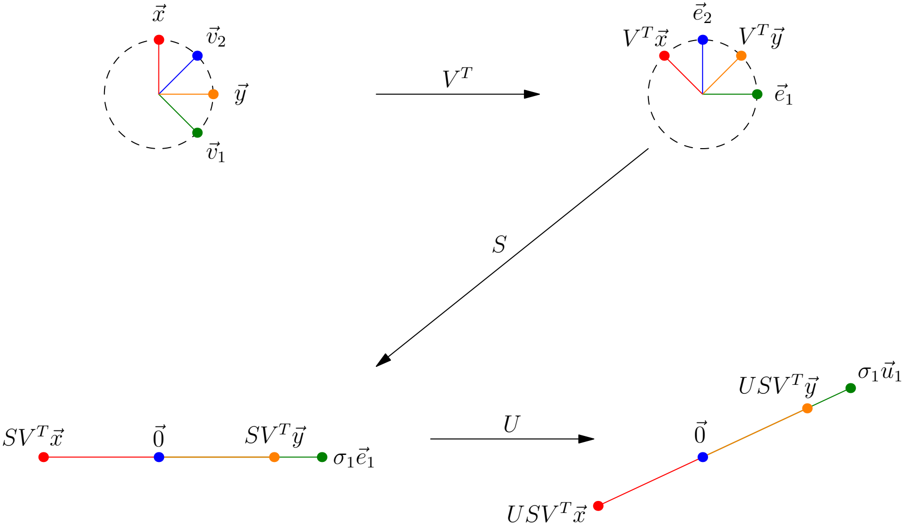
3. Rotation

$$USV^T \vec{x} = \sum_{i=1}^{n} \sigma_i \langle \vec{v}_i, \vec{x} \rangle \vec{u}_i$$

Linear maps


$\vec{x}$
$\vec{v}_2$
$\vec{y}$
$\vec{v}_1$
V^T
$\vec{e}_2$
$V^T\vec{x}$
$V^T\vec{y}$
$\vec{e}_1$
S
$\sigma_2\vec{e}_2$
$SV^T\vec{x}$
$SV^T\vec{y}$
$\sigma_1\vec{e}_1$
U
$USV^T\vec{y}$
$\sigma_2\vec{u}_2$
$\sigma_1\vec{u}_1$
$USV^T\vec{x}$

Linear maps



$\vec{x}$
$\vec{v}_2$
$\vec{y}$
$\vec{v}_1$
V^T
$\vec{e}_2$
$V^T\vec{x}$
$V^T\vec{y}$
$\vec{e}_1$
S
$SV^T\vec{x}$
$\vec{0}$
$SV^T\vec{y}$
$\sigma_1\vec{e}_1$
U
$USV^T\vec{y}$
$\sigma_1\vec{u}_1$
$\vec{0}$
$USV^T\vec{x}$

Singular values

For any orthogonal matrices $\widetilde{U} \in \mathbb{R}^{m \times m}$ and $\widetilde{V} \in \mathbb{R}^{n \times n}$ the singular values of $\widetilde{U}A$ and $A\widetilde{V}$ are the same as those of A

Proof:

$$\overline{U} := \widetilde{U}U$$
$$\overline{V}^T := V^T\widetilde{V}$$

are orthogonal, so $\overline{U}SV^T$ and $US\overline{V}^T$ are valid SVDs for $\widetilde{U}A$ and $A\widetilde{V}$

Singular values

The singular values satisfy

$$\sigma_1 = \max_{\{||\vec{x}||_2=1 \mid \vec{x}\in\mathbb{R}^n\}} ||A\vec{x}||_2$$

$$= \max_{\{||\vec{y}||_2=1 \mid \vec{y}\in\mathbb{R}^m\}} \left|\left|A^T\vec{y}\right|\right|_2$$

$$\sigma_i = \max_{\{||\vec{x}||_2=1 \mid \vec{x}\in\mathbb{R}^n, \vec{x}\perp\vec{u}_1,\ldots,\vec{u}_{i-1}\}} ||A\vec{x}||_2$$

$$= \max_{\{||\vec{y}||_2=1 \mid \vec{y}\in\mathbb{R}^m, \vec{y}\perp\vec{v}_1,\ldots,\vec{v}_{i-1}\}} \left|\left|A^T\vec{y}\right|\right|_2, \qquad 2 \leq i \leq \min\{m, n\}$$

Proof

Consider $\vec{x} \in \mathbb{R}^n$ such that $||\vec{x}||_2 = 1$ and for a fixed $1 \leq i \leq n$

$$\vec{x} \perp \vec{v}_1, \ldots, \vec{v}_{i-1}$$

We decompose $\vec{x}$ into

$$\vec{x} = \sum_{j=i}^{n} \alpha_j \vec{v}_j + \mathcal{P}_{\text{row}(A)^\perp} \vec{x}$$

where $1 = ||\vec{x}||_2^2 \geq \sum_{j=i}^{n} \alpha_j^2$

Proof

$$||A\vec{x}||_2^2$$

Proof

$$||A\vec{x}||_2^2 = \left\langle \sum_{k=1}^{n} \sigma_k \langle \vec{v}_k, \vec{x} \rangle \vec{u}_k, \sum_{k=1}^{n} \sigma_k \langle \vec{v}_k, \vec{x} \rangle \vec{u}_k \right\rangle$$

Proof

$$||A\vec{x}||_2^2 = \left\langle \sum_{k=1}^{n} \sigma_k \langle \vec{v}_k, \vec{x} \rangle \vec{u}_k, \sum_{k=1}^{n} \sigma_k \langle \vec{v}_k, \vec{x} \rangle \vec{u}_k \right\rangle$$

$$= \sum_{k=1}^{n} \sigma_k^2 \langle \vec{v}_k, \vec{x} \rangle^2 \quad \text{because } \vec{u}_1, \ldots, \vec{u}_n \text{ are orthonormal}$$

Proof

$$
\begin{aligned}
||A\vec{x}||_2^2 &= \left\langle \sum_{k=1}^{n} \sigma_k \langle \vec{v}_k, \vec{x} \rangle \, \vec{u}_k, \sum_{k=1}^{n} \sigma_k \langle \vec{v}_k, \vec{x} \rangle \, \vec{u}_k \right\rangle \\
&= \sum_{k=1}^{n} \sigma_k^2 \langle \vec{v}_k, \vec{x} \rangle^2 \quad \text{because } \vec{u}_1, \ldots, \vec{u}_n \text{ are orthonormal} \\
&= \sum_{k=1}^{n} \sigma_k^2 \left\langle \vec{v}_k, \sum_{j=i}^{n} \alpha_j \vec{v}_j + \mathcal{P}_{\text{row}(A)^\perp} \, \vec{x} \right\rangle^2
\end{aligned}
$$

Proof

$$
\begin{aligned}
||A\vec{x}||_2^2 &= \left\langle \sum_{k=1}^{n} \sigma_k \left\langle \vec{v}_k, \vec{x} \right\rangle \vec{u}_k, \sum_{k=1}^{n} \sigma_k \left\langle \vec{v}_k, \vec{x} \right\rangle \vec{u}_k \right\rangle \\
&= \sum_{k=1}^{n} \sigma_k^2 \left\langle \vec{v}_k, \vec{x} \right\rangle^2 \quad \text{because } \vec{u}_1, \ldots, \vec{u}_n \text{ are orthonormal} \\
&= \sum_{k=1}^{n} \sigma_k^2 \left\langle \vec{v}_k, \sum_{j=i}^{n} \alpha_j \vec{v}_j + \mathcal{P}_{\text{row}(A)^\perp} \vec{x} \right\rangle^2 \\
&= \sum_{j=i}^{n} \sigma_j^2 \alpha_j^2 \quad \text{because } \vec{v}_1, \ldots, \vec{v}_n \text{ are orthonormal}
\end{aligned}
$$

Proof

$$
\begin{aligned}
||A\vec{x}||_2^2 &= \left\langle \sum_{k=1}^{n} \sigma_k \langle \vec{v}_k, \vec{x} \rangle \vec{u}_k, \sum_{k=1}^{n} \sigma_k \langle \vec{v}_k, \vec{x} \rangle \vec{u}_k \right\rangle \\
&= \sum_{k=1}^{n} \sigma_k^2 \langle \vec{v}_k, \vec{x} \rangle^2 \quad \text{because } \vec{u}_1, \ldots, \vec{u}_n \text{ are orthonormal} \\
&= \sum_{k=1}^{n} \sigma_k^2 \left\langle \vec{v}_k, \sum_{j=i}^{n} \alpha_j \vec{v}_j + \mathcal{P}_{\text{row}(A)^\perp} \vec{x} \right\rangle^2 \\
&= \sum_{j=i}^{n} \sigma_j^2 \alpha_j^2 \quad \text{because } \vec{v}_1, \ldots, \vec{v}_n \text{ are orthonormal} \\
&\leq \sigma_i^2 \sum_{j=i}^{n} \alpha_j^2 \quad \text{because } \sigma_i \geq \sigma_{i+1} \geq \ldots \geq \sigma_n
\end{aligned}
$$

Proof

$$
\begin{aligned}
||A\vec{x}||_2^2 &= \left\langle \sum_{k=1}^{n} \sigma_k \left\langle \vec{v}_k, \vec{x} \right\rangle \vec{u}_k, \sum_{k=1}^{n} \sigma_k \left\langle \vec{v}_k, \vec{x} \right\rangle \vec{u}_k \right\rangle \\
&= \sum_{k=1}^{n} \sigma_k^2 \left\langle \vec{v}_k, \vec{x} \right\rangle^2 \quad \text{because } \vec{u}_1, \ldots, \vec{u}_n \text{ are orthonormal} \\
&= \sum_{k=1}^{n} \sigma_k^2 \left\langle \vec{v}_k, \sum_{j=i}^{n} \alpha_j \vec{v}_j + \mathcal{P}_{\text{row}(A)^\perp} \vec{x} \right\rangle^2 \\
&= \sum_{j=i}^{n} \sigma_j^2 \alpha_j^2 \quad \text{because } \vec{v}_1, \ldots, \vec{v}_n \text{ are orthonormal} \\
&\leq \sigma_i^2 \sum_{j=i}^{n} \alpha_j^2 \quad \text{because } \sigma_i \geq \sigma_{i+1} \geq \ldots \geq \sigma_n \\
&\leq \sigma_i^2
\end{aligned}
$$

Singular vectors

The right singular vectors satisfy

$$\vec{v}_1 = \arg\max_{\{||\vec{x}||_2=1 \mid \vec{x}\in\mathbb{R}^n\}} ||A\vec{x}||_2$$

$$\vec{v}_i = \arg\max_{\{||\vec{x}||_2=1 \mid \vec{x}\in\mathbb{R}^n, \vec{x}\perp\vec{v}_1,\dots,\vec{v}_{i-1}\}} ||A\vec{x}||_2 , \qquad 2 \leq i \leq m$$

and the left singular vectors satisfy

$$\vec{u}_1 = \arg\max_{\{||\vec{y}||_2=1 \mid \vec{y}\in\mathbb{R}^m\}} \left|\left|A^T\vec{y}\right|\right|_2$$

$$\vec{u}_i = \arg\max_{\{||\vec{y}||_2=1 \mid \vec{y}\in\mathbb{R}^m, \vec{y}\perp\vec{u}_1,\dots,\vec{u}_{i-1}\}} \left|\left|A^T\vec{y}\right|\right|_2 , \qquad 2 \leq i \leq n$$

Proof

$\vec{v}_i$ achieves the maximum

$$||A\vec{v}_i||_2^2$$

Same proof for $\vec{u}_i$ replacing A by A^T

Proof

$\vec{v}_i$ achieves the maximum

$$||A\vec{v}_i||_2^2 = \left\langle \sum_{k=1}^{n} \sigma_k \langle \vec{v}_k, \vec{v}_i \rangle \vec{u}_k, \sum_{k=1}^{n} \sigma_k \langle \vec{v}_k, \vec{v}_i \rangle \vec{u}_k \right\rangle$$

Same proof for $\vec{u}_i$ replacing A by A^T

Proof

$\vec{v}_i$ achieves the maximum

$$
\begin{aligned}
||A\vec{v}_i||_2^2 &= \left\langle \sum_{k=1}^{n} \sigma_k \left\langle \vec{v}_k, \vec{v}_i \right\rangle \vec{u}_k, \sum_{k=1}^{n} \sigma_k \left\langle \vec{v}_k, \vec{v}_i \right\rangle \vec{u}_k \right\rangle \\
&= \sum_{k=1}^{n} \sigma_k^2 \left\langle \vec{v}_k, \vec{v}_i \right\rangle^2
\end{aligned}
$$

Same proof for $\vec{u}_i$ replacing A by A^T

Proof

$\vec{v}_i$ achieves the maximum

$$\begin{aligned} ||A\vec{v}_i||_2^2 &= \left\langle \sum_{k=1}^{n} \sigma_k \langle \vec{v}_k, \vec{v}_i \rangle \vec{u}_k, \sum_{k=1}^{n} \sigma_k \langle \vec{v}_k, \vec{v}_i \rangle \vec{u}_k \right\rangle \\ &= \sum_{k=1}^{n} \sigma_k^2 \langle \vec{v}_k, \vec{v}_i \rangle^2 \\ &= \sigma_i^2 \end{aligned}$$

Same proof for $\vec{u}_i$ replacing A by A^T

Optimal subspace for orthogonal projection

Given a set of vectors $\vec{a}_1, \vec{a}_2, \ldots \vec{a}_n$ and a fixed dimension $k \leq n$, the SVD of

$$A := \begin{bmatrix} \vec{a}_1 & \vec{a}_2 & \cdots & \vec{a}_n \end{bmatrix} \in \mathbb{R}^{m \times n}$$

provides the k-dimensional subspace that captures the most energy

$$\sum_{i=1}^{n} \left|\left|\mathcal{P}_{\text{span}(\vec{u}_1, \vec{u}_2, \ldots, \vec{u}_k)}\, \vec{a}_i\right|\right|_2^2 \geq \sum_{i=1}^{n} \left|\left|\mathcal{P}_{\mathcal{S}}\, \vec{a}_i\right|\right|_2^2$$

for any subspace $\mathcal{S}$ of dimension k

Proof

Because $\vec{u}_1, \vec{u}_2, \ldots, \vec{u}_k$ are orthonormal

$$\sum_{i=1}^{n} \left|\left|\mathcal{P}_{\text{span}(\vec{u}_1,\vec{u}_2,\ldots,\vec{u}_k)}\, \vec{a}_i\right|\right|_2^2$$

Proof

Because $\vec{u}_1, \vec{u}_2, \ldots, \vec{u}_k$ are orthonormal

$$\sum_{i=1}^{n} \left|\left|\mathcal{P}_{\text{span}(\vec{u}_1,\vec{u}_2,\ldots,\vec{u}_k)}\, \vec{a}_i\right|\right|_2^2 = \sum_{i=1}^{n} \sum_{j=1}^{k} \langle \vec{u}_j, \vec{a}_i \rangle^2$$

Proof

Because $\vec{u}_1, \vec{u}_2, \ldots, \vec{u}_k$ are orthonormal

$$\begin{aligned}\sum_{i=1}^{n} \left|\left|\mathcal{P}_{\text{span}(\vec{u}_1,\vec{u}_2,\ldots,\vec{u}_k)}\, \vec{a}_i\right|\right|_2^2 &= \sum_{i=1}^{n} \sum_{j=1}^{k} \langle \vec{u}_j, \vec{a}_i \rangle^2 \\ &= \sum_{j=1}^{k} \left|\left|A^T \vec{u}_j\right|\right|_2^2\end{aligned}$$

Proof

Because $\vec{u}_1, \vec{u}_2, \ldots, \vec{u}_k$ are orthonormal

$$\sum_{i=1}^{n} \left|\left|\mathcal{P}_{\text{span}(\vec{u}_1,\vec{u}_2,\ldots,\vec{u}_k)}\, \vec{a}_i\right|\right|_2^2 = \sum_{i=1}^{n} \sum_{j=1}^{k} \langle \vec{u}_j, \vec{a}_i \rangle^2$$

$$= \sum_{j=1}^{k} \left|\left|A^T \vec{u}_j\right|\right|_2^2$$

Induction on k

Proof

Because $\vec{u}_1, \vec{u}_2, \ldots, \vec{u}_k$ are orthonormal

$$\sum_{i=1}^{n} \left|\left|\mathcal{P}_{\text{span}(\vec{u}_1,\vec{u}_2,\ldots,\vec{u}_k)}\, \vec{a}_i\right|\right|_2^2 = \sum_{i=1}^{n}\sum_{j=1}^{k} \langle \vec{u}_j, \vec{a}_i \rangle^2$$

$$= \sum_{j=1}^{k} \left|\left|A^T \vec{u}_j\right|\right|_2^2$$

Induction on k

The base case $k = 1$ follows from

$$\vec{u}_1 = \arg\max_{\{||\vec{y}||_2=1 \,\mid\, \vec{y}\in\mathbb{R}^m\}} \left|\left|A^T \vec{y}\right|\right|_2$$

Proof

Let $\mathcal{S}$ be a subspace of dimension k

$\mathcal{S} \cap \operatorname{span}(\vec{u}_1, \ldots, \vec{u}_{k-1})^{\perp}$ contains a nonzero vector $\vec{b}$

If $\dim(\mathcal{V})$ *has dimension* n*,* $\mathcal{S}_1, \mathcal{S}_2 \subseteq \mathcal{V}$ *and* $\dim(\mathcal{S}_1) + \dim(\mathcal{S}_2) > n$*, then* $\dim(\mathcal{S}_1 \cap \mathcal{S}_2) \geq 1$

Proof

Let $\mathcal{S}$ be a subspace of dimension k

$\mathcal{S} \cap \operatorname{span}(\vec{u}_1, \ldots, \vec{u}_{k-1})^{\perp}$ contains a nonzero vector $\vec{b}$

If $\dim(\mathcal{V})$ *has dimension* n*,* $\mathcal{S}_1, \mathcal{S}_2 \subseteq \mathcal{V}$ *and* $\dim(\mathcal{S}_1) + \dim(\mathcal{S}_2) > n$*, then* $\dim(\mathcal{S}_1 \cap \mathcal{S}_2) \geq 1$

There exists an orthonormal basis $\vec{b}_1, \vec{b}_2, \ldots, \vec{b}_k$ for $\mathcal{S}$ such that $\vec{b}_k := \vec{b}$ is orthogonal to $\vec{u}_1, \vec{u}_2, \ldots, \vec{u}_{k-1}$

Induction hypothesis

$$
\begin{aligned}
\sum_{i=1}^{k-1} \left|\left|A^T \vec{u}_i\right|\right|_2^2 &= \sum_{i=1}^{n} \left|\left|\mathcal{P}_{\text{span}(\vec{u}_1,\vec{u}_2,\ldots,\vec{u}_{k-1})}\, \vec{a}_i\right|\right|_2^2 \\
&\geq \sum_{i=1}^{n} \left|\left|\mathcal{P}_{\text{span}\left(\vec{b}_1,\vec{b}_2,\ldots,\vec{b}_{k-1}\right)}\, \vec{a}_i\right|\right|_2^2 \\
&= \sum_{i=1}^{k-1} \left|\left|A^T \vec{b}_i\right|\right|_2^2
\end{aligned}
$$

Proof

Recall that

$$\vec{u}_k = \underset{\left\{ ||\vec{y}||_2 = 1 \;|\; \vec{y} \in \mathbb{R}^m,\, \vec{y} \perp \vec{u}_1, \ldots, \vec{u}_{k-1} \right\}}{\arg\max} \left|\left| A^T \vec{y} \right|\right|_2$$

which implies

$$\left|\left| A^T \vec{u}_k \right|\right|_2^2 \geq \left|\left| A^T \vec{b}_k \right|\right|_2^2$$

Proof

Recall that

$$\vec{u}_k = \arg\max_{\left\{||\vec{y}||_2=1 \;|\; \vec{y}\in\mathbb{R}^m,\, \vec{y}\perp\vec{u}_1,\ldots,\vec{u}_{k-1}\right\}} \left|\left|A^T\vec{y}\right|\right|_2$$

which implies

$$\left|\left|A^T\vec{u}_k\right|\right|_2^2 \geq \left|\left|A^T\vec{b}_k\right|\right|_2^2$$

We conclude

$$\begin{aligned}
\sum_{i=1}^{n} \left|\left|\mathcal{P}_{\text{span}(\vec{u}_1,\vec{u}_2,\ldots,\vec{u}_k)}\,\vec{a}_i\right|\right|_2^2 &= \sum_{i=1}^{k} \left|\left|A^T\vec{u}_i\right|\right|_2^2 \\
&\geq \sum_{i=1}^{k} \left|\left|A^T\vec{b}_i\right|\right|_2^2 \\
&= \sum_{i=1}^{n} \left|\left|\mathcal{P}_{\mathcal{S}}\,\vec{a}_i\right|\right|_2^2
\end{aligned}$$

Best rank-k approximation

Let USV^T be the SVD of a matrix $A \in \mathbb{R}^{m \times n}$

The truncated SVD $U_{:,1:k} S_{1:k,1:k} V_{:,1:k}^T$ is the best rank-k approximation

$$U_{:,1:k} S_{1:k,1:k} V_{:,1:k}^T = \underset{\{\widetilde{A} \,|\, \text{rank}(\tilde{A})=k\}}{\arg\min} \left|\left| A - \widetilde{A} \right|\right|_{\text{F}}$$

Proof

Let $\widetilde{A}$ be an arbitrary matrix in $\mathbb{R}^{m \times n}$ with $\text{rank}(\widetilde{A}) = k$

Let $\widetilde{U} \in \mathbb{R}^{m \times k}$ be a matrix with orthonormal columns such that $\text{col}(\widetilde{U}) = \text{col}(\widetilde{A})$

$$\begin{aligned} \left|\left| U_{:,1:k} U_{:,1:k}^T A \right|\right|_F^2 &= \sum_{i=1}^{n} \left|\left| \mathcal{P}_{\text{col}\left(U_{:,1:k}\right)} \vec{a}_i \right|\right|_2^2 \\ &\geq \sum_{i=1}^{n} \left|\left| \mathcal{P}_{\text{col}\left(\widetilde{U}\right)} \vec{a}_i \right|\right|_2^2 \\ &= \left|\left| \widetilde{U} \widetilde{U}^T A \right|\right|_F^2 \end{aligned}$$

Orthogonal column spaces

If the column spaces of $A, B \in \mathbb{R}^{m \times n}$ are orthogonal then

$$||A + B||_F^2 = ||A||_F^2 + ||B||_F^2$$

Proof

$$\text{col}\left(A - \widetilde{U}\widetilde{U}^T A\right) \text{ is orthogonal to } \text{col}\left(\widetilde{A}\right) = \text{col}\left(\widetilde{U}\right)$$

$$\left|\left|A - \widetilde{A}\right|\right|_{\text{F}}^2 = \left|\left|A - \widetilde{U}\widetilde{U}^T A\right|\right|_{\text{F}}^2 + \left|\left|\widetilde{A} - \widetilde{U}\widetilde{U}^T A\right|\right|_{\text{F}}^2$$

Proof

$\text{col}\left(A - \widetilde{U}\widetilde{U}^T A\right)$ is orthogonal to $\text{col}\left(\widetilde{A}\right) = \text{col}\left(\widetilde{U}\right)$

$$
\begin{aligned}
\left|\left|A - \widetilde{A}\right|\right|_{\text{F}}^2 &= \left|\left|A - \widetilde{U}\widetilde{U}^T A\right|\right|_{\text{F}}^2 + \left|\left|\widetilde{A} - \widetilde{U}\widetilde{U}^T A\right|\right|_{\text{F}}^2 \\
&\geq \left|\left|A - \widetilde{U}\widetilde{U}^T A\right|\right|_{\text{F}}^2
\end{aligned}
$$

Proof

$\text{col}\left(A - \widetilde{U}\widetilde{U}^T A\right)$ is orthogonal to $\text{col}\left(\widetilde{A}\right) = \text{col}\left(\widetilde{U}\right)$

$$\begin{aligned} \left|\left|A - \widetilde{A}\right|\right|_F^2 &= \left|\left|A - \widetilde{U}\widetilde{U}^T A\right|\right|_F^2 + \left|\left|\widetilde{A} - \widetilde{U}\widetilde{U}^T A\right|\right|_F^2 \\ &\geq \left|\left|A - \widetilde{U}\widetilde{U}^T A\right|\right|_F^2 \\ &= \left|\left|A\right|\right|_F^2 - \left|\left|\widetilde{U}\widetilde{U}^T A\right|\right|_F^2 \end{aligned}$$

Proof

$\text{col}\left(A - \widetilde{U}\widetilde{U}^T A\right)$ is orthogonal to $\text{col}\left(\widetilde{A}\right) = \text{col}\left(\widetilde{U}\right)$

$$\begin{aligned}
\left|\left|A - \widetilde{A}\right|\right|_F^2 &= \left|\left|A - \widetilde{U}\widetilde{U}^T A\right|\right|_F^2 + \left|\left|\widetilde{A} - \widetilde{U}\widetilde{U}^T A\right|\right|_F^2 \\
&\geq \left|\left|A - \widetilde{U}\widetilde{U}^T A\right|\right|_F^2 \\
&= ||A||_F^2 - \left|\left|\widetilde{U}\widetilde{U}^T A\right|\right|_F^2 \\
&\geq ||A||_F^2 - \left|\left|U_{:,1:k} U_{:,1:k}^T A\right|\right|_F^2
\end{aligned}$$

Proof

$\text{col}\left(A - \widetilde{U}\widetilde{U}^T A\right)$ is orthogonal to $\text{col}\left(\widetilde{A}\right) = \text{col}\left(\widetilde{U}\right)$

$$
\begin{aligned}
\left|\left|A - \widetilde{A}\right|\right|_{\mathsf{F}}^2 &= \left|\left|A - \widetilde{U}\widetilde{U}^T A\right|\right|_{\mathsf{F}}^2 + \left|\left|\widetilde{A} - \widetilde{U}\widetilde{U}^T A\right|\right|_{\mathsf{F}}^2 \\
&\geq \left|\left|A - \widetilde{U}\widetilde{U}^T A\right|\right|_{\mathsf{F}}^2 \\
&= ||A||_{\mathsf{F}}^2 - \left|\left|\widetilde{U}\widetilde{U}^T A\right|\right|_{\mathsf{F}}^2 \\
&\geq ||A||_{\mathsf{F}}^2 - \left|\left|U_{:,1:k} U_{:,1:k}^T A\right|\right|_{\mathsf{F}}^2 \\
&= \left|\left|A - U_{:,1:k} U_{:,1:k}^T A\right|\right|_{\mathsf{F}}^2
\end{aligned}
$$

Reminder

For any pair of $m \times n$ matrices A and B

$$\operatorname{tr}\left(B^T A\right) := \operatorname{tr}\left(AB^T\right)$$

Frobenius norm

For any matrix $A \in \mathbb{R}^{m \times n}$, with singular values $\sigma_1, \ldots, \sigma_{\min\{m,n\}}$

$$||A||_{\mathsf{F}} = \sqrt{\sum_{i=1}^{\min\{m,n\}} \sigma_i^2}.$$

Frobenius norm

For any matrix $A \in \mathbb{R}^{m \times n}$, with singular values $\sigma_1, \ldots, \sigma_{\min\{m,n\}}$

$$||A||_\mathsf{F} = \sqrt{\sum_{i=1}^{\min\{m,n\}} \sigma_i^2}.$$

Proof:

$$||A||_\mathsf{F}^2 = \operatorname{tr}\left(A^T A\right)$$

Frobenius norm

For any matrix $A \in \mathbb{R}^{m \times n}$, with singular values $\sigma_1, \ldots, \sigma_{\min\{m,n\}}$

$$||A||_{\text{F}} = \sqrt{\sum_{i=1}^{\min\{m,n\}} \sigma_i^2}.$$

Proof:

$$\begin{aligned} ||A||_{\text{F}}^2 &= \text{tr}\left(A^T A\right) \\ &= \text{tr}\left(VSU^T USV^T\right) \end{aligned}$$

Frobenius norm

For any matrix $A \in \mathbb{R}^{m \times n}$, with singular values $\sigma_1, \ldots, \sigma_{\min\{m,n\}}$

$$||A||_{\text{F}} = \sqrt{\sum_{i=1}^{\min\{m,n\}} \sigma_i^2}.$$

Proof:

$$\begin{aligned} ||A||_{\text{F}}^2 &= \text{tr}\left(A^T A\right) \\ &= \text{tr}\left(VSU^T USV^T\right) \\ &= \text{tr}\left(VSSV^T\right) \qquad \text{because } U^T U = I \end{aligned}$$

Frobenius norm

For any matrix $A \in \mathbb{R}^{m \times n}$, with singular values $\sigma_1, \ldots, \sigma_{\min\{m,n\}}$

$$||A||_{\mathrm{F}} = \sqrt{\sum_{i=1}^{\min\{m,n\}} \sigma_i^2}.$$

Proof:

$$\begin{aligned}
||A||_{\mathrm{F}}^2 &= \operatorname{tr}\left(A^T A\right) \\
&= \operatorname{tr}\left(VSU^T USV^T\right) \\
&= \operatorname{tr}\left(VSSV^T\right) \qquad \text{because } U^T U = I \\
&= \operatorname{tr}\left(V^T VSS\right)
\end{aligned}$$

Frobenius norm

For any matrix $A \in \mathbb{R}^{m \times n}$, with singular values $\sigma_1, \ldots, \sigma_{\min\{m,n\}}$

$$||A||_{\mathsf{F}} = \sqrt{\sum_{i=1}^{\min\{m,n\}} \sigma_i^2}.$$

Proof:

$$\begin{aligned}
||A||_{\mathsf{F}}^2 &= \operatorname{tr}\left(A^T A\right) \\
&= \operatorname{tr}\left(VSU^T USV^T\right) \\
&= \operatorname{tr}\left(VSSV^T\right) \qquad \text{because } U^T U = I \\
&= \operatorname{tr}\left(V^T VSS\right) \\
&= \operatorname{tr}\left(SS\right) \qquad \text{because } V^T V = I
\end{aligned}$$

Operator norm

The operator norm of a linear map and the corresponding matrix $A \in \mathbb{R}^{m \times n}$ is

$$
\begin{aligned}
||A|| &:= \max_{\left\{ ||\vec{x}||_2 = 1 \ | \ \vec{x} \in \mathbb{R}^n \right\}} ||A\,\vec{x}||_2 \\
&= \sigma_1
\end{aligned}
$$

Nuclear norm

The nuclear norm of a matrix $A \in \mathbb{R}^{m \times n}$ with singular values $\sigma_1, \ldots, \sigma_{\min\{m,n\}}$ is

$$||A||_* := \sum_{i=1}^{\min\{m,n\}} \sigma_i$$

Multiplication by orthogonal matrices

For any orthogonal matrices $\widetilde{U} \in \mathbb{R}^{m \times m}$ and $\widetilde{V} \in \mathbb{R}^{n \times n}$ the singular values of $\widetilde{U}A$ and $A\widetilde{V}$ are the same as those of A

Consequence:

The operator, Frobenius and nuclear norm of $\widetilde{U}A$ and $A\widetilde{V}$ are the same as those of A

Hölder's inequality for matrices

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

$$||A||_* = \sup_{\{||B|| \leq 1 \mid B \in \mathbb{R}^{m \times n}\}} \langle A, B \rangle .$$

Consequence: nuclear norm satisfies triangle inequality

$$||A + B||_*$$

Hölder's inequality for matrices

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

$$||A||_* = \sup_{\{||B|| \leq 1 \mid B \in \mathbb{R}^{m \times n}\}} \langle A, B \rangle .$$

Consequence: nuclear norm satisfies triangle inequality

$$||A + B||_* = \sup_{\{||C|| \leq 1 \mid C \in \mathbb{R}^{m \times n}\}} \langle A + B, C \rangle$$

Hölder's inequality for matrices

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

$$||A||_* = \sup_{\{||B|| \leq 1 \mid B \in \mathbb{R}^{m \times n}\}} \langle A, B \rangle .$$

Consequence: nuclear norm satisfies triangle inequality

$$\begin{aligned} ||A + B||_* &= \sup_{\{||C|| \leq 1 \mid C \in \mathbb{R}^{m \times n}\}} \langle A + B, C \rangle \\ &\leq \sup_{\{||C|| \leq 1 \mid C \in \mathbb{R}^{m \times n}\}} \langle A, C \rangle + \sup_{\{||D|| \leq 1 \mid D \in \mathbb{R}^{m \times n}\}} \langle B, D \rangle \end{aligned}$$

Hölder's inequality for matrices

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

$$||A||_* = \sup_{\{||B|| \leq 1 \mid B \in \mathbb{R}^{m \times n}\}} \langle A, B \rangle .$$

Consequence: nuclear norm satisfies triangle inequality

$$\begin{aligned}
||A + B||_* &= \sup_{\{||C|| \leq 1 \mid C \in \mathbb{R}^{m \times n}\}} \langle A + B, C \rangle \\
&\leq \sup_{\{||C|| \leq 1 \mid C \in \mathbb{R}^{m \times n}\}} \langle A, C \rangle + \sup_{\{||D|| \leq 1 \mid D \in \mathbb{R}^{m \times n}\}} \langle B, D \rangle \\
&= ||A||_* + ||B||_*
\end{aligned}$$

Proof

The proof relies on the following lemma:

For any $Q \in \mathbb{R}^{n \times n}$

$$\max_{1 \le i \le n} |Q_{ii}| \le ||Q||$$

Proof:

Since $||\vec{e_i}||_2 = 1$,

$$\begin{aligned}
\max_{1 \le i \le n} |Q_{ii}| &\le \max_{1 \le i \le n} \sqrt{\sum_{j=1}^{n} Q_{ji}^2} \\
&= \max_{1 \le i \le n} ||Q\,\vec{e_i}||_2 \\
&\le ||Q||
\end{aligned}$$

Proof

Let $A := USV^T$,

$$\sup_{||B|| \leq 1} \text{Trace}\left(A^T B\right)$$

Proof

Let $A := USV^T$,

$$\sup_{||B|| \leq 1} \text{Trace}\left(A^T B\right) = \sup_{||B|| \leq 1} \text{Trace}\left(V\, S\, U^T B\right)$$

Proof

Let $A := USV^T$,

$$\begin{aligned}\sup_{\|B\| \leq 1} \text{Trace}\left(A^T B\right) &= \sup_{\|B\| \leq 1} \text{Trace}\left(V\, S\, U^T B\right) \\ &= \sup_{\|B\| \leq 1} \text{Trace}\left(S\, B U^T V\right)\end{aligned}$$

Proof

Let $A := USV^T$,

$$\begin{aligned}
\sup_{||B|| \leq 1} \text{Trace}\left(A^T B\right) &= \sup_{||B|| \leq 1} \text{Trace}\left(V\,S\,U^T B\right) \\
&= \sup_{||B|| \leq 1} \text{Trace}\left(S\,BU^T V\right) \\
&\leq \sup_{||M|| \leq 1} \text{Trace}\,(S\,M) \quad \text{since } ||B|| = \left|\left|BU^T V\right|\right| \\
&\leq \sup_{\max_{1 \leq i \leq n} |M_{ii}| \leq 1} \text{Trace}\,(S\,M) \qquad \text{by the lemma}
\end{aligned}$$

Proof

Let $A := USV^T$,

$$\begin{aligned}
\sup_{||B|| \leq 1} \text{Trace}\left(A^T B\right) &= \sup_{||B|| \leq 1} \text{Trace}\left(V\, S\, U^T B\right) \\
&= \sup_{||B|| \leq 1} \text{Trace}\left(S\, B U^T V\right) \\
&\leq \sup_{||M|| \leq 1} \text{Trace}\left(S\, M\right) \quad \text{since } ||B|| = \left|\left|BU^T V\right|\right| \\
&\leq \sup_{\max_{1 \leq i \leq n} |M_{ii}| \leq 1} \text{Trace}\left(S\, M\right) \qquad \text{by the lemma} \\
&\leq \sup_{\max_{1 \leq i \leq n} |M_{ii}| \leq 1} \sum_{i=1}^{n} M_{ii}\, \sigma_i
\end{aligned}$$

Proof

Let $A := USV^T$,

$$\begin{aligned}
\sup_{||B|| \leq 1} \text{Trace}\left(A^T B\right) &= \sup_{||B|| \leq 1} \text{Trace}\left(V\, S\, U^T B\right) \\
&= \sup_{||B|| \leq 1} \text{Trace}\left(S\, B U^T V\right) \\
&\leq \sup_{||M|| \leq 1} \text{Trace}\left(S\, M\right) \quad \text{since } ||B|| = \left|\left|B U^T V\right|\right| \\
&\leq \sup_{\max_{1 \leq i \leq n} |M_{ii}| \leq 1} \text{Trace}\left(S\, M\right) \qquad \text{by the lemma} \\
&\leq \sup_{\max_{1 \leq i \leq n} |M_{ii}| \leq 1} \sum_{i=1}^{n} M_{ii}\, \sigma_i \\
&\leq \sum_{i=1}^{n} \sigma_i
\end{aligned}$$

Proof

Let $A := USV^T$,

$$
\begin{aligned}
\sup_{||B|| \leq 1} \operatorname{Trace}\left(A^T B\right) &= \sup_{||B|| \leq 1} \operatorname{Trace}\left(V\, S\, U^T B\right) \\
&= \sup_{||B|| \leq 1} \operatorname{Trace}\left(S\, B U^T V\right) \\
&\leq \sup_{||M|| \leq 1} \operatorname{Trace}\left(S\, M\right) \quad \text{since } ||B|| = \left|\left|B U^T V\right|\right| \\
&\leq \sup_{\max_{1 \leq i \leq n} |M_{ii}| \leq 1} \operatorname{Trace}\left(S\, M\right) \qquad \text{by the lemma} \\
&\leq \sup_{\max_{1 \leq i \leq n} |M_{ii}| \leq 1} \sum_{i=1}^{n} M_{ii}\, \sigma_i \\
&\leq \sum_{i=1}^{n} \sigma_i \\
&= ||A||_*
\end{aligned}
$$

Proof

To complete the proof, we show that the equality holds

UV^T has operator norm equal to one

$$\left\langle A, UV^T \right\rangle$$

Proof

To complete the proof, we show that the equality holds

UV^T has operator norm equal to one

$$\left\langle A, UV^T \right\rangle = \operatorname{tr}\left(A^T UV^T\right)$$

Proof

To complete the proof, we show that the equality holds

UV^T has operator norm equal to one

$$\left\langle A, UV^T \right\rangle = \operatorname{tr}\left(A^T UV^T\right)$$
$$= \operatorname{tr}\left(V\, S\, U^T UV^T\right)$$

Proof

To complete the proof, we show that the equality holds

UV^T has operator norm equal to one

$$\begin{aligned}\left\langle A, UV^T \right\rangle &= \operatorname{tr}\left(A^T UV^T\right) \\ &= \operatorname{tr}\left(V\, S\, U^T UV^T\right) \\ &= \operatorname{tr}\left(V^T V\, S\right)\end{aligned}$$

Proof

To complete the proof, we show that the equality holds

UV^T has operator norm equal to one

$$\begin{aligned} \left\langle A, UV^T \right\rangle &= \operatorname{tr}\left(A^T UV^T\right) \\ &= \operatorname{tr}\left(V\, S\, U^T UV^T\right) \\ &= \operatorname{tr}\left(V^T V\, S\right) \\ &= \operatorname{tr}\left(S\right) \end{aligned}$$

Proof

To complete the proof, we show that the equality holds

UV^T has operator norm equal to one

$$\begin{aligned}\left\langle A, UV^T \right\rangle &= \operatorname{tr}\left(A^T UV^T\right) \\ &= \operatorname{tr}\left(V\, S\, U^T UV^T\right) \\ &= \operatorname{tr}\left(V^T V\, S\right) \\ &= \operatorname{tr}\left(S\right) \\ &= ||A||_*\end{aligned}$$

Denoising correlated signals

Aim: Estimating n m-dimensional signals $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n \in \mathbb{R}^m$ from

$$\vec{y}_i = \vec{x}_i + \vec{z}_i, \qquad 1 \leq i \leq n$$

Assumption 1: Signals are similar and approximately span an unknown low-dimensional subspace

Assumption 2: Noisy perturbations are independent / uncorrelated

Denoising correlated signals

Aim: Estimating n m-dimensional signals $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n \in \mathbb{R}^m$ from

$$\vec{y}_i = \vec{x}_i + \vec{z}_i, \qquad 1 \leq i \leq n$$

Assumption 1: Signals are similar and approximately span an unknown low-dimensional subspace

Assumption 2: Noisy perturbations are independent / uncorrelated

Denoising correlated signals

By the assumptions

$$X := \begin{bmatrix} \vec{x}_1 & \vec{x}_2 & \cdots & \vec{x}_n \end{bmatrix}$$

is approximately low rank, whereas

$$Z := \begin{bmatrix} \vec{z}_1 & \vec{z}_2 & \cdots & \vec{z}_n \end{bmatrix}$$

is full rank

If Z is not too large, low-rank approximation to

$$\begin{aligned} Y &:= \begin{bmatrix} \vec{y}_1 & \vec{y}_2 & \cdots & \vec{y}_n \end{bmatrix} \\ &= X + Z \end{aligned}$$

should correspond mostly to X

Denoising via SVD truncation

1. Stack the vectors as the columns of a matrix $Y \in \mathbb{R}^{m \times n}$

2. Compute the SVD of $Y = USV^T$

3. Truncate the SVD to produce the low-rank estimate L

$$L := U_{:,1:k} S_{1:k,1:k} V_{:,1:k}^T,$$

   for a fixed value of $k \leq \min\{m, n\}$

Important decision

What rank k to choose?

- *Large k*
- *Small k*

Important decision

What rank k to choose?

- *Large* k will approximate signals well but not suppress noise
- *Small* k

Important decision

What rank k to choose?

- *Large* k will approximate signals well but not suppress noise
- *Small* k will suppress noise but may not approximate signals well

Denoising of digit images

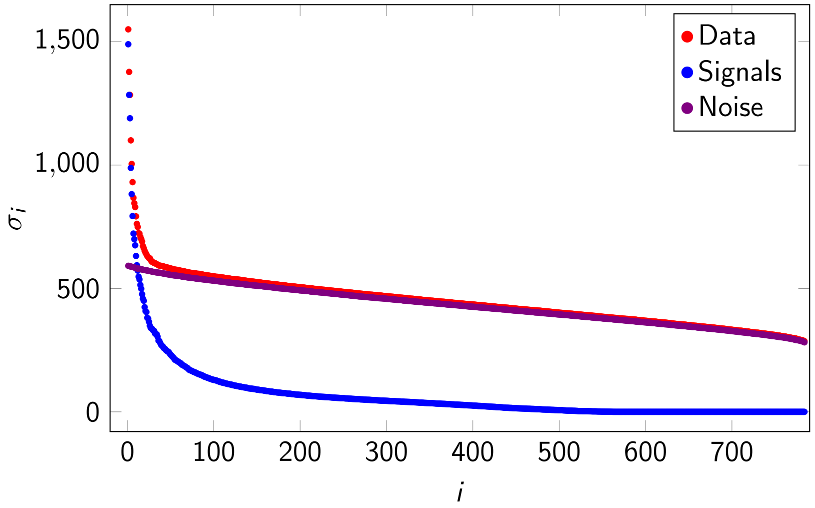
MNIST data

Signals: 6131 28×28 images of the number 3

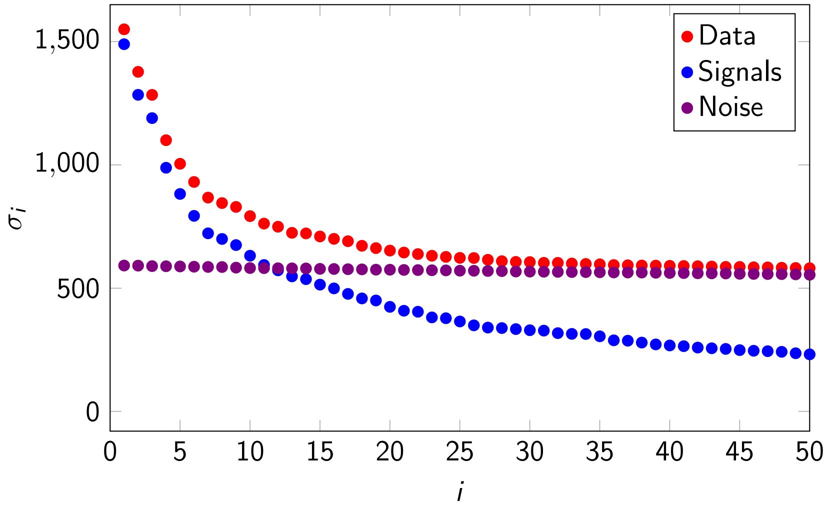
Noise: iid Gaussian so that SNR is 0.5 in ℓ_2 norm

More noise than signal!

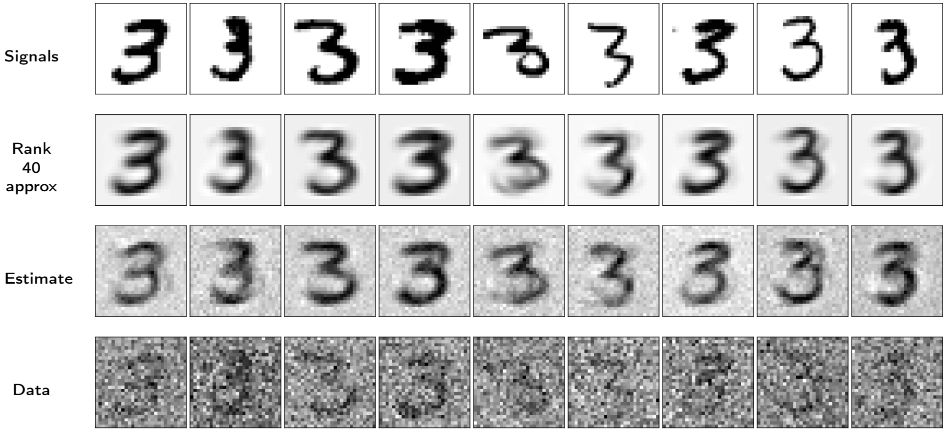
Denoising of digit images



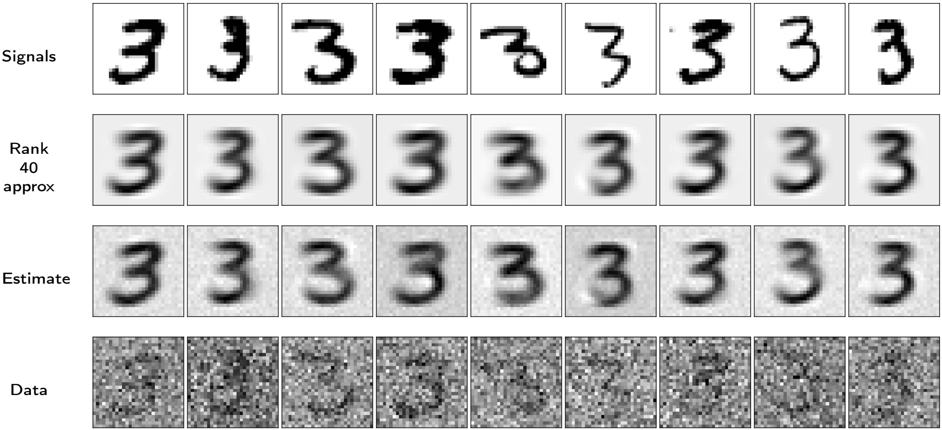
Denoising of digit images



$k = 40$



$k = 10$



Collaborative filtering

$$A := \begin{pmatrix} 1 & 1 & 5 & 4 \\ 2 & 1 & 4 & 5 \\ 4 & 5 & 2 & 1 \\ 5 & 4 & 2 & 1 \\ 4 & 5 & 1 & 2 \\ 1 & 2 & 5 & 5 \end{pmatrix}$$

Bob	Molly	Mary	Larry	
1	1	5	4	The Dark Knight
2	1	4	5	Spiderman 3
4	5	2	1	Love Actually
5	4	2	1	Bridget Jones's Diary
4	5	1	2	Pretty Woman
1	2	5	5	Superman 2

Intuition

Some people have similar tastes and hence produce similar ratings

Some movies are similar and hence elicit similar reactions

This tends to induce low-rank structure in the matrix of ratings

SVD

$$A - \mu\vec{1}\,\vec{1}^{\,T} = U\Sigma V^T = U \begin{bmatrix} 7.79 & 0 & 0 & 0 \\ 0 & 1.62 & 0 & 0 \\ 0 & 0 & 1.55 & 0 \\ 0 & 0 & 0 & 0.62 \end{bmatrix} V^T$$

$$\mu := \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij}$$

Rank 1 model

$$\bar{A} + \sigma_1 \vec{u}_1 \vec{v}_1^T =$$

	Bob	Molly	Mary	Larry	
	1.34 (1)	1.19 (1)	4.66 (5)	4.81 (4)	The Dark Knight
	1.55 (2)	1.42 (1)	4.45 (4)	4.58 (5)	Spiderman 3
	4.45 (4)	4.58 (5)	1.55 (2)	1.42 (1)	Love Actually
	4.43 (5)	4.56 (4)	1.57 (2)	1.44 (1)	B.J.'s Diary
	4.43 (4)	4.56 (5)	1.57 (1)	1.44 (2)	Pretty Woman
	1.34 (1)	1.19 (2)	4.66 (5)	4.81 (5)	Superman 2

First left singular vector

	D. Knight	Sp. 3	Love Act.	B.J.'s Diary	P. Woman	Sup. 2
$\vec{u}_1 =$ (	-0.45	-0.39	0.39	0.39	0.39	-0.45)

Coefficients cluster movies into action (+) and romantic (-)

First right singular vector

$$\begin{array}{ccccc} & \text{Bob} & \text{Molly} & \text{Mary} & \text{Larry} \\ \vec{v}_1 = & (0.48 & 0.52 & -0.48 & -0.52) \end{array}$$

Coefficients cluster people into action (-) and romantic (+)

Generalization

Each rating is a sum of k terms

$$\text{rating}\left(\text{movie } i, \text{user } j\right) = \sum_{l=1}^{k} \sigma_l \vec{u}_l\left[i\right] \vec{v}_l\left[j\right]$$

Singular vectors cluster users and movies in different ways

Singular values weight the importance of the different factors.

Sample covariance matrix

The sample covariance matrix of $\{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\} \in \mathbb{R}^m$

$$\Sigma\left(\vec{x}_1, \ldots, \vec{x}_n\right) := \frac{1}{n-1} \sum_{i=1}^{n} \left(\vec{x}_i - \operatorname{av}\left(\vec{x}_1, \ldots, \vec{x}_n\right)\right)\left(\vec{x}_i - \operatorname{av}\left(\vec{x}_1, \ldots, \vec{x}_n\right)\right)^T$$

$$\operatorname{av}\left(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\right) := \frac{1}{n} \sum_{i=1}^{n} \vec{x}_i$$

$$\Sigma\left(\vec{x}_1, \ldots, \vec{x}_n\right)_{ij} = \begin{cases} \operatorname{var}\left(\vec{x}_1[i], \ldots, \vec{x}_n[i]\right) & \text{if } i = j, \\ \operatorname{cov}\left(\left(\vec{x}_1[i], \vec{x}_1[j]\right), \ldots, \left(\vec{x}_n[i], \vec{x}_n[j]\right)\right) & \text{if } i \neq j \end{cases}$$

Variation in a certain direction

For a unit vector $\vec{d} \in \mathbb{R}^m$

$$\text{var}\left(\vec{d}^{\,T}\vec{x}_1, \ldots, \vec{d}^{\,T}\vec{x}_n\right)$$

Variation in a certain direction

For a unit vector $\vec{d} \in \mathbb{R}^m$

$$\operatorname{var}\left(\vec{d}^{\,T}\vec{x}_1, \ldots, \vec{d}^{\,T}\vec{x}_n\right)$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(\vec{d}^{\,T}\vec{x}_i - \operatorname{av}\left(\vec{d}^{\,T}\vec{x}_1, \ldots, \vec{d}^{\,T}\vec{x}_n\right)\right)^2$$

Variation in a certain direction

For a unit vector $\vec{d} \in \mathbb{R}^m$

$$\operatorname{var}\left(\vec{d}^{\,T}\vec{x}_1, \ldots, \vec{d}^{\,T}\vec{x}_n\right)$$
$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(\vec{d}^{\,T}\vec{x}_i - \operatorname{av}\left(\vec{d}^{\,T}\vec{x}_1, \ldots, \vec{d}^{\,T}\vec{x}_n\right)\right)^2$$
$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(\vec{d}^{\,T}\left(\vec{x}_i - \operatorname{av}\left(\vec{x}_1, \ldots, \vec{x}_n\right)\right)\right)^2$$

Variation in a certain direction

For a unit vector $\vec{d} \in \mathbb{R}^m$

$$
\begin{aligned}
&\operatorname{var}\left(\vec{d}^{\,T}\vec{x}_1, \ldots, \vec{d}^{\,T}\vec{x}_n\right) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(\vec{d}^{\,T}\vec{x}_i - \operatorname{av}\left(\vec{d}^{\,T}\vec{x}_1, \ldots, \vec{d}^{\,T}\vec{x}_n\right)\right)^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(\vec{d}^{\,T}\left(\vec{x}_i - \operatorname{av}\left(\vec{x}_1, \ldots, \vec{x}_n\right)\right)\right)^2 \\
&= \vec{d}^{\,T}\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(\vec{x}_i - \operatorname{av}\left(\vec{x}_1, \ldots, \vec{x}_n\right)\right)\left(\vec{x}_i - \operatorname{av}\left(\vec{x}_1, \ldots, \vec{x}_n\right)\right)^T\right)\vec{d}
\end{aligned}
$$

Variation in a certain direction

For a unit vector $\vec{d} \in \mathbb{R}^m$

$$
\begin{aligned}
&\operatorname{var}\left(\vec{d}^{\,T}\vec{x}_1, \ldots, \vec{d}^{\,T}\vec{x}_n\right) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(\vec{d}^{\,T}\vec{x}_i - \operatorname{av}\left(\vec{d}^{\,T}\vec{x}_1, \ldots, \vec{d}^{\,T}\vec{x}_n\right)\right)^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(\vec{d}^{\,T}\left(\vec{x}_i - \operatorname{av}\left(\vec{x}_1, \ldots, \vec{x}_n\right)\right)\right)^2 \\
&= \vec{d}^{\,T}\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(\vec{x}_i - \operatorname{av}\left(\vec{x}_1, \ldots, \vec{x}_n\right)\right)\left(\vec{x}_i - \operatorname{av}\left(\vec{x}_1, \ldots, \vec{x}_n\right)\right)^T\right)\vec{d} \\
&= \vec{d}^{\,T}\Sigma\left(\vec{x}_1, \ldots, \vec{x}_n\right)\vec{d}
\end{aligned}
$$

Covariance matrix captures variance in every direction!

Principal component analysis

Given n data vectors $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n \in \mathbb{R}^d$,

1. Center the data,

$$\vec{c}_i = \vec{x}_i - \operatorname{av}\left(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\right), \qquad 1 \leq i \leq n$$

2. Group the centered data as columns of a matrix

$$C = \begin{bmatrix} \vec{c}_1 & \vec{c}_2 & \cdots & \vec{c}_n \end{bmatrix}.$$

3. Compute the SVD of C

The left singular vectors are the principal directions

The principal values are the coefficients of the centered vectors in the basis of principal directions.

Directions of maximum variance

The principal directions satisfy

$$\vec{u}_1 = \underset{\left\{ \left|\left|\vec{d}\right|\right|_2 = 1 \;|\; \vec{d} \in \mathbb{R}^n \right\}}{\arg\max} \operatorname{var}\left(\vec{d}^{\,T} \vec{x}_1, \ldots, \vec{d}^{\,T} \vec{x}_n \right)$$

$$\vec{u}_i = \underset{\left\{ \left|\left|\vec{d}\right|\right|_2 = 1 \;|\; \vec{d} \in \mathbb{R}^n,\, \vec{d} \perp \vec{u}_1, \ldots, \vec{u}_{i-1} \right\}}{\arg\max} \operatorname{var}\left(\vec{d}^{\,T} \vec{x}_1, \ldots, \vec{d}^{\,T} \vec{x}_n \right), \qquad 2 \leq i \leq k$$

Directions of maximum variance

The associated singular values satisfy

$$\frac{\sigma_1}{\sqrt{n-1}} = \max_{\left\{ ||\vec{d}||_2 = 1 \mid \vec{d} \in \mathbb{R}^n \right\}} \text{std}\left(\vec{d}^{\,T} \vec{x}_1, \ldots, \vec{d}^{\,T} \vec{x}_n \right)$$

$$\frac{\sigma_i}{\sqrt{n-1}} = \max_{\left\{ ||\vec{d}||_2 = 1 \mid \vec{d} \in \mathbb{R}^n,\, \vec{d} \perp \vec{u}_1, \ldots, \vec{u}_{i-1} \right\}} \text{std}\left(\vec{d}^{\,T} \vec{x}_1, \ldots, \vec{d}^{\,T} \vec{x}_n \right), \quad 2 \leq i \leq k$$

Proof

$$\begin{aligned}\Sigma(\vec{x}_1, \ldots, \vec{x}_n) &= \frac{1}{n-1} \sum_{i=1}^{n} (\vec{x}_i - \mathrm{av}(\vec{x}_1, \ldots, \vec{x}_n)) (\vec{x}_i - \mathrm{av}(\vec{x}_1, \ldots, \vec{x}_n))^T \\ &= \frac{1}{n-1} C C^T\end{aligned}$$

Proof

$$\Sigma\left(\vec{x}_1,\ldots,\vec{x}_n\right) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\vec{x}_i - \operatorname{av}\left(\vec{x}_1,\ldots,\vec{x}_n\right)\right)\left(\vec{x}_i - \operatorname{av}\left(\vec{x}_1,\ldots,\vec{x}_n\right)\right)^T$$

$$= \frac{1}{n-1} C C^T$$

For any vector $\vec{d}$

$$\operatorname{var}\left(\vec{d}^{\,T}\vec{x}_1,\ldots,\vec{d}^{\,T}\vec{x}_n\right) = \vec{d}^{\,T}\Sigma\left(\vec{x}_1,\ldots,\vec{x}_n\right)\vec{d}$$

Proof

$$\Sigma\left(\vec{x}_1,\ldots,\vec{x}_n\right) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\vec{x}_i - \operatorname{av}\left(\vec{x}_1,\ldots,\vec{x}_n\right)\right)\left(\vec{x}_i - \operatorname{av}\left(\vec{x}_1,\ldots,\vec{x}_n\right)\right)^T$$

$$= \frac{1}{n-1} C C^T$$

For any vector $\vec{d}$

$$\operatorname{var}\left(\vec{d}^{\,T}\vec{x}_1,\ldots,\vec{d}^{\,T}\vec{x}_n\right) = \vec{d}^{\,T}\Sigma\left(\vec{x}_1,\ldots,\vec{x}_n\right)\vec{d}$$

$$= \frac{1}{n-1}\vec{d}^{\,T} C C^T \vec{d}$$

Proof

$$\Sigma\left(\vec{x}_1,\ldots,\vec{x}_n\right)=\frac{1}{n-1}\sum_{i=1}^{n}\left(\vec{x}_i-\operatorname{av}\left(\vec{x}_1,\ldots,\vec{x}_n\right)\right)\left(\vec{x}_i-\operatorname{av}\left(\vec{x}_1,\ldots,\vec{x}_n\right)\right)^T$$
$$=\frac{1}{n-1}CC^T$$

For any vector $\vec{d}$

$$\begin{aligned}\operatorname{var}\left(\vec{d}^{\,T}\vec{x}_1,\ldots,\vec{d}^{\,T}\vec{x}_n\right)&=\vec{d}^{\,T}\Sigma\left(\vec{x}_1,\ldots,\vec{x}_n\right)\vec{d}\\&=\frac{1}{n-1}\vec{d}^{\,T}CC^T\vec{d}\\&=\frac{1}{n-1}\left|\left|C^T\vec{d}\right|\right|_2^2\end{aligned}$$

Singular values

The singular values of A satisfy

$$
\begin{aligned}
\sigma_1 &= \max_{\left\{||\vec{x}||_2=1 \ | \ \vec{x}\in\mathbb{R}^n\right\}} ||A\vec{x}||_2 \\
&= \max_{\left\{||\vec{y}||_2=1 \ | \ \vec{y}\in\mathbb{R}^m\right\}} \left|\left|A^T\vec{y}\right|\right|_2 \\
\sigma_i &= \max_{\left\{||\vec{x}||_2=1 \ | \ \vec{x}\in\mathbb{R}^n, \vec{x}\perp\vec{u}_1,\dots,\vec{u}_{i-1}\right\}} ||A\vec{x}||_2 \\
&= \max_{\left\{||\vec{y}||_2=1 \ | \ \vec{y}\in\mathbb{R}^m, \vec{y}\perp\vec{v}_1,\dots,\vec{v}_{i-1}\right\}} \left|\left|A^T\vec{y}\right|\right|_2, \qquad 2 \leq i \leq \min\{m, n\}
\end{aligned}
$$

Singular vectors

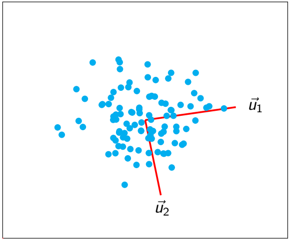
The left singular vectors of A satisfy

$$\vec{u}_1 = \arg\max_{\left\{||\vec{y}||_2=1 \ | \ \vec{y}\in\mathbb{R}^m\right\}} \left|\left|A^T\vec{y}\right|\right|_2$$

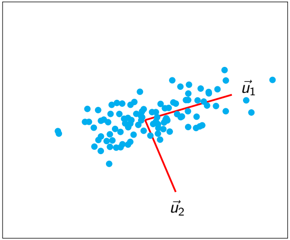
$$\vec{u}_i = \arg\max_{\left\{||\vec{y}||_2=1 \ | \ \vec{y}\in\mathbb{R}^m, \vec{y}\perp\vec{v}_1,\ldots,\vec{v}_{i-1}\right\}} \left|\left|A^T\vec{y}\right|\right|_2, \qquad 2 \leq i \leq n$$

PCA in 2D

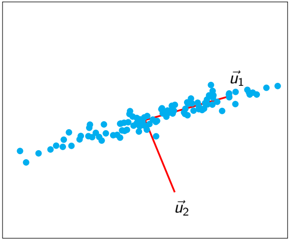
$$\sigma_1/\sqrt{n-1} = 0.705, \quad \sigma_2/\sqrt{n-1} = 0.690$$



$$\sigma_1/\sqrt{n-1} = 0.983, \quad \sigma_2/\sqrt{n-1} = 0.356$$

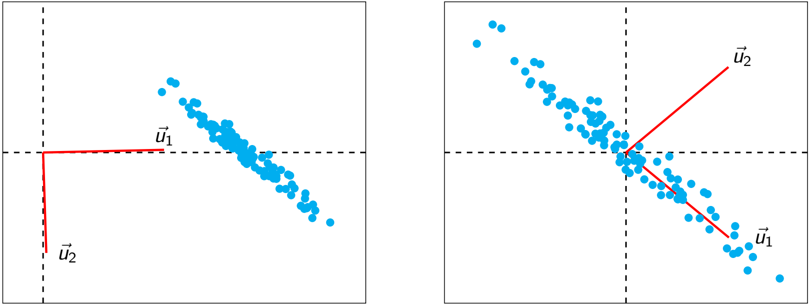


$$\sigma_1/\sqrt{n-1} = 1.349, \quad \sigma_2/\sqrt{n-1} = 0.144$$



Centering

$$\sigma_1/\sqrt{n-1} = 5.077 \qquad \sigma_1/\sqrt{n-1} = 1.261$$
$$\sigma_2/\sqrt{n-1} = 0.889 \qquad \sigma_2/\sqrt{n-1} = 0.139$$



Uncentered data   Centered data

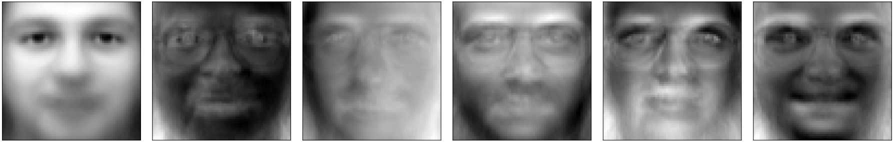
PCA of faces

Data set of 400 64×64 images from 40 subjects (10 per subject)

Each face is vectorized and interpreted as a vector in $\mathbb{R}^{4096}$

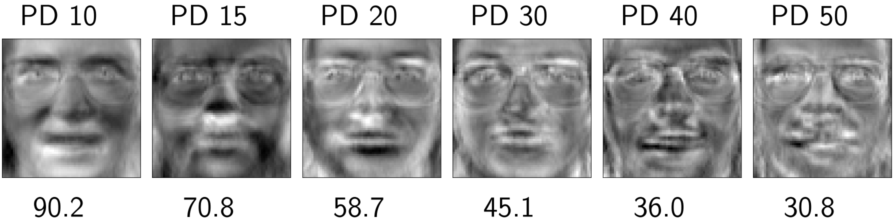
PCA of faces

Center	PD 1	PD 2	PD 3	PD 4	PD 5

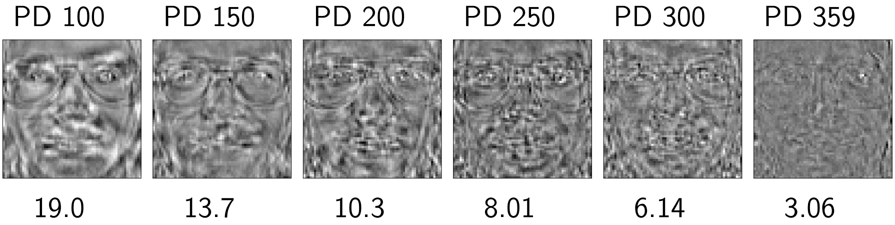


$\sigma_i/\sqrt{n-1}$	330	251	192	152	130

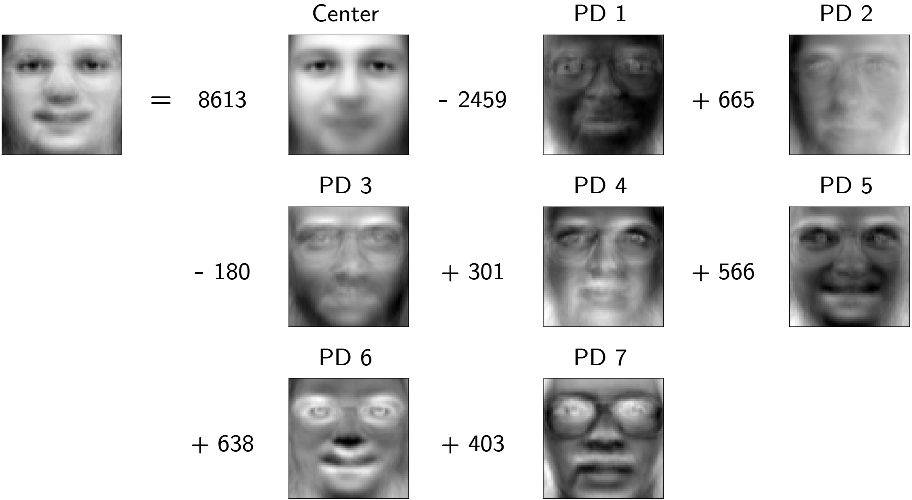
PCA of faces



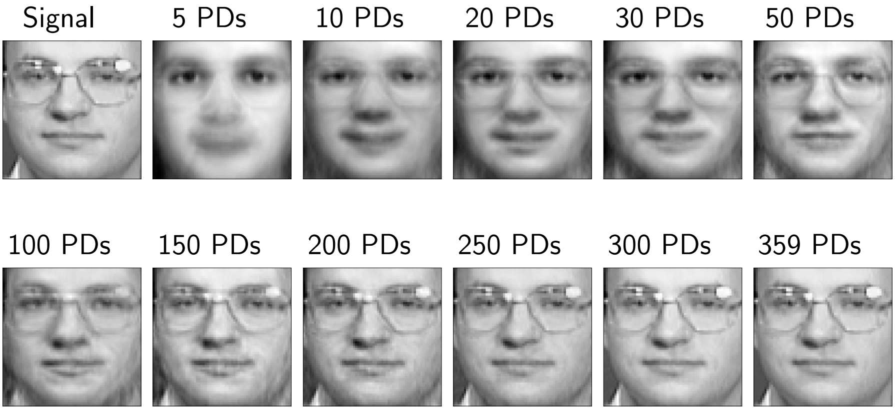
PCA of faces



Projection onto first 7 principal directions



Projection onto first k principal directions



Covariance matrix

The covariance matrix of a random vector $\vec{x}$ is defined as

$$\Sigma_{\vec{x}} := \begin{bmatrix} \mathrm{Var}\left(\vec{x}[1]\right) & \mathrm{Cov}\left(\vec{x}[1], \vec{x}[2]\right) & \cdots & \mathrm{Cov}\left(\vec{x}[1], \vec{x}[n]\right) \\ \mathrm{Cov}\left(\vec{x}[2], \vec{x}[1]\right) & \mathrm{Var}\left(\vec{x}[2]\right) & \cdots & \mathrm{Cov}\left(\vec{x}[2], \vec{x}[n]\right) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}\left(\vec{x}[n], \vec{x}[1]\right) & \mathrm{Cov}\left(\vec{x}[n], \vec{x}[2]\right) & \cdots & \mathrm{Var}\left(\vec{x}[n]\right) \end{bmatrix}$$

$$= \mathrm{E}\left(\vec{x}\vec{x}^T\right) - \mathrm{E}(\vec{x})\mathrm{E}(\vec{x})^T$$

If the covariance matrix is diagonal, the entries are uncorrelated

Covariance matrix after a linear transformation

Let $\Sigma \in \mathbb{R}^{n\times n}$ be the covariance matrix of $\vec{\mathbf{x}}$. For any matrix $A \in \mathbb{R}^{m\times n}$

$$\Sigma_{A\vec{\mathbf{x}}+\vec{b}} = A\Sigma_{\vec{\mathbf{x}}}A^T$$

Proof:

$$\Sigma_{A\vec{\mathbf{x}}}$$

Covariance matrix after a linear transformation

Let $\Sigma \in \mathbb{R}^{n \times n}$ be the covariance matrix of $\vec{x}$. For any matrix $A \in \mathbb{R}^{m \times n}$

$$\Sigma_{A\vec{x}+\vec{b}} = A \Sigma_{\vec{x}} A^T$$

Proof:

$$\Sigma_{A\vec{x}} = \mathrm{E}\left((A\vec{x}) (A\vec{x})^T \right) - \mathrm{E}(A\vec{x}) \, \mathrm{E}(A\vec{x})^T$$

Covariance matrix after a linear transformation

Let $\Sigma \in \mathbb{R}^{n \times n}$ be the covariance matrix of $\vec{\mathbf{x}}$. For any matrix $A \in \mathbb{R}^{m \times n}$

$$\Sigma_{A\vec{\mathbf{x}}+\vec{b}} = A\Sigma_{\vec{\mathbf{x}}}A^T$$

Proof:

$$\begin{aligned}\Sigma_{A\vec{\mathbf{x}}} &= \mathrm{E}\left((A\vec{\mathbf{x}})(A\vec{\mathbf{x}})^T\right) - \mathrm{E}(A\vec{\mathbf{x}})\,\mathrm{E}(A\vec{\mathbf{x}})^T \\ &= A\left(\mathrm{E}\left(\vec{\mathbf{x}}\vec{\mathbf{x}}^T\right) - \mathrm{E}(\vec{\mathbf{x}})\mathrm{E}(\vec{\mathbf{x}})^T\right)A^T\end{aligned}$$

Covariance matrix after a linear transformation

Let $\Sigma \in \mathbb{R}^{n \times n}$ be the covariance matrix of $\vec{\mathbf{x}}$. For any matrix $A \in \mathbb{R}^{m \times n}$

$$\Sigma_{A\vec{\mathbf{x}}+\vec{b}} = A\Sigma_{\vec{\mathbf{x}}}A^T$$

Proof:

$$\begin{aligned}
\Sigma_{A\vec{\mathbf{x}}} &= \mathrm{E}\left((A\vec{\mathbf{x}})\,(A\vec{\mathbf{x}})^T\right) - \mathrm{E}\,(A\vec{\mathbf{x}})\,\mathrm{E}\,(A\vec{\mathbf{x}})^T \\
&= A\left(\mathrm{E}\left(\vec{\mathbf{x}}\vec{\mathbf{x}}^T\right) - \mathrm{E}(\vec{\mathbf{x}})\mathrm{E}(\vec{\mathbf{x}})^T\right)A^T \\
&= A\Sigma_{\vec{\mathbf{x}}}A^T
\end{aligned}$$

Variance in a fixed direction

The variance of $\vec{\mathbf{x}}$ in the direction of a unit-norm vector $\vec{v}$ equals

$$\mathrm{Var}\left(\vec{v}^{\,T}\vec{\mathbf{x}}\right) = \vec{v}^{\,T}\Sigma_{\vec{\mathbf{x}}}\vec{v}$$

SVD of covariance matrix

$$\Sigma_{\vec{\mathbf{x}}} = U\Lambda U^T$$

$$= \begin{bmatrix} \vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_n \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \sigma_n \end{bmatrix} \begin{bmatrix} \vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_n \end{bmatrix}^T$$

Directions of maximum variance

The SVD of the covariance matrix $\Sigma_{\vec{\mathbf{x}}}$ of a random vector $\vec{\mathbf{x}}$ satisfies

$$\sigma_1 = \max_{||\vec{v}||_2=1} \mathrm{Var}\left(\vec{v}^{\,T}\vec{\mathbf{x}}\right)$$

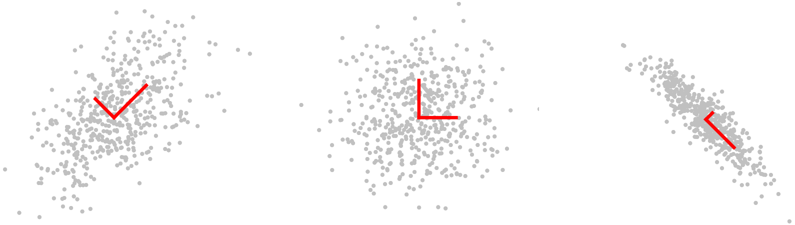
$$\vec{u}_1 = \arg\max_{||\vec{v}||_2=1} \mathrm{Var}\left(\vec{v}^{\,T}\vec{\mathbf{x}}\right)$$

$$\sigma_k = \max_{||\vec{v}||_2=1, \vec{v}\perp\vec{u}_1,\ldots,\vec{u}_{k-1}} \mathrm{Var}\left(\vec{v}^{\,T}\vec{\mathbf{x}}\right)$$

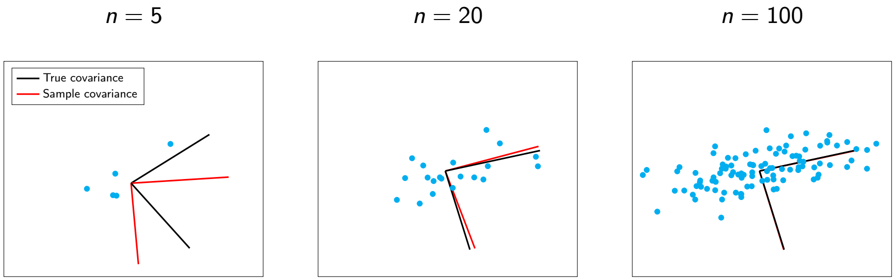
$$\vec{u}_k = \arg\max_{||\vec{v}||_2=1, \vec{v}\perp\vec{u}_1,\ldots,\vec{u}_{k-1}} \mathrm{Var}\left(\vec{v}^{\,T}\vec{\mathbf{x}}\right)$$

Directions of maximum variance

$\sqrt{\sigma_1} = 1.22,\ \sqrt{\sigma_2} = 0.71$ $\sqrt{\sigma_1} = 1,\ \sqrt{\sigma_2} = 1$ $\sqrt{\sigma_1} = 1.38,\ \sqrt{\sigma_2} = 0.32$



Probabilistic interpretation of PCA



Dimensionality reduction

Data with a large number of features can be difficult to analyze or process

Dimensionality reduction is a useful preprocessing step

If data modeled are vectors in $\mathbb{R}^m$ we can reduce the dimension by projecting onto $\mathbb{R}^k$, where $k < m$

For orthogonal projections, the new representation is $\langle \vec{b}_1, \vec{x} \rangle$, $\langle \vec{b}_2, \vec{x} \rangle$, ..., $\langle \vec{b}_k, \vec{x} \rangle$ for a basis $\vec{b}_1$, ..., $\vec{b}_k$ of the subspace that we project on

Problem: How do we choose the subspace?

Optimal subspace for orthogonal projection

Given a set of vectors $\vec{a}_1, \vec{a}_2, \ldots \vec{a}_n$ and a fixed dimension $k \leq n$, the SVD of

$$A := \begin{bmatrix} \vec{a}_1 & \vec{a}_2 & \cdots & \vec{a}_n \end{bmatrix} \in \mathbb{R}^{m \times n}$$

provides the k-dimensional subspace that captures the most energy

$$\sum_{i=1}^{n} \left|\left|\mathcal{P}_{\text{span}(\vec{u}_1, \vec{u}_2, \ldots, \vec{u}_k)} \, \vec{a}_i\right|\right|_2^2 \geq \sum_{i=1}^{n} \left|\left|\mathcal{P}_{\mathcal{S}} \, \vec{a}_i\right|\right|_2^2$$

for any subspace $\mathcal{S}$ of dimension k

Nearest neighbors in principal-component space

Nearest neighbors classification (Algorithm 4.2 in Lecture Notes 1) computes n distances in $\mathbb{R}^m$ for each new example

Cost: $\mathcal{O}(nmp)$ for p examples

Idea: Project onto first k main principal directions beforehand

Cost:

- $\mathcal{O}\left(m^2 n\right)$, if $m < n$, to compute principal dimensions
- kmn operations to project training set
- kmp operations to project test set
- knp to perform nearest-neighbor classification

Faster if $p > m$

Face recognition

Training set: 360 64×64 images from 40 different subjects (9 each)

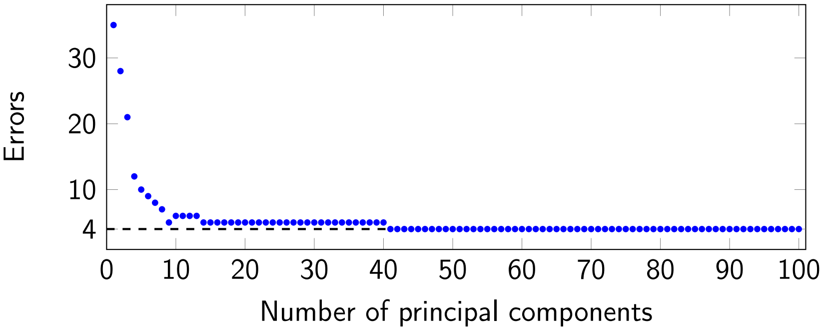
Test set: 1 new image from each subject

We model each image as a vector in $\mathbb{R}^{4096}$ ($m = 4096$)

To classify we:

1. Project onto first k principal directions
2. Apply nearest-neighbor classification using the ℓ_2-norm distance in $\mathbb{R}^k$

Performance



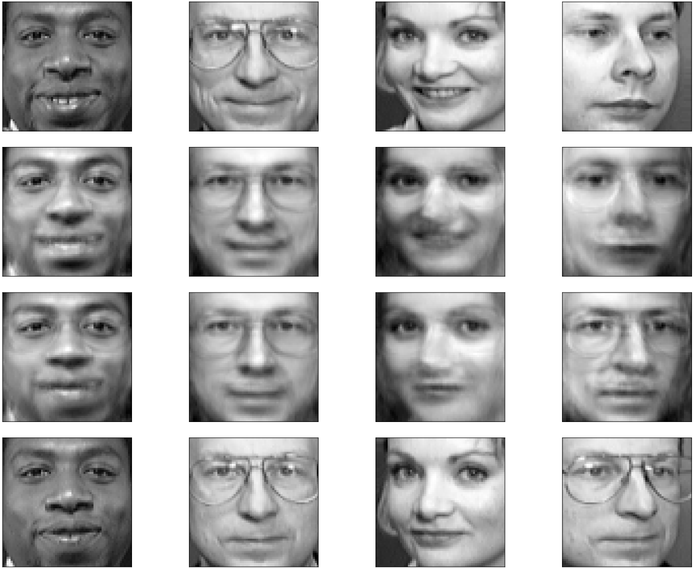
Nearest neighbor in $\mathbb{R}^{41}$

Test image

Projection

Closest projection

Corresponding image



Dimensionality reduction for visualization

Motivation: Visualize high-dimensional features projected onto 2D or 3D

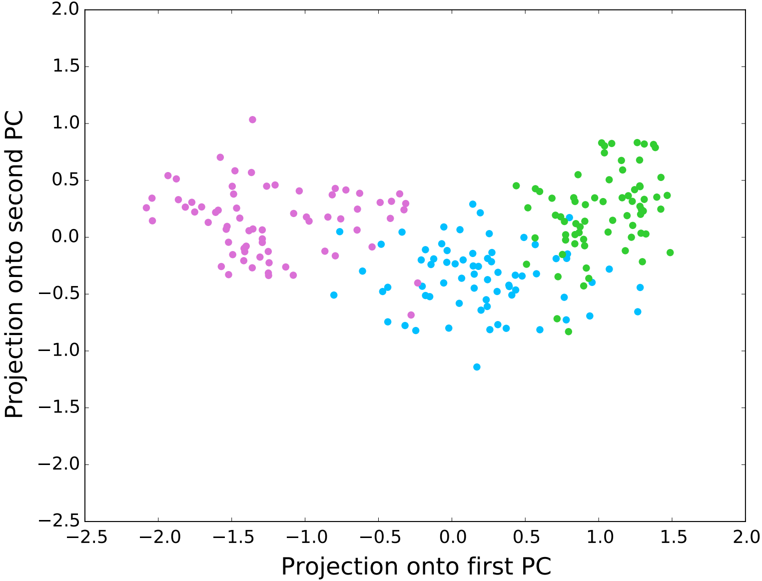
Example:

Seeds from three different varieties of wheat: Kama, Rosa and Canadian

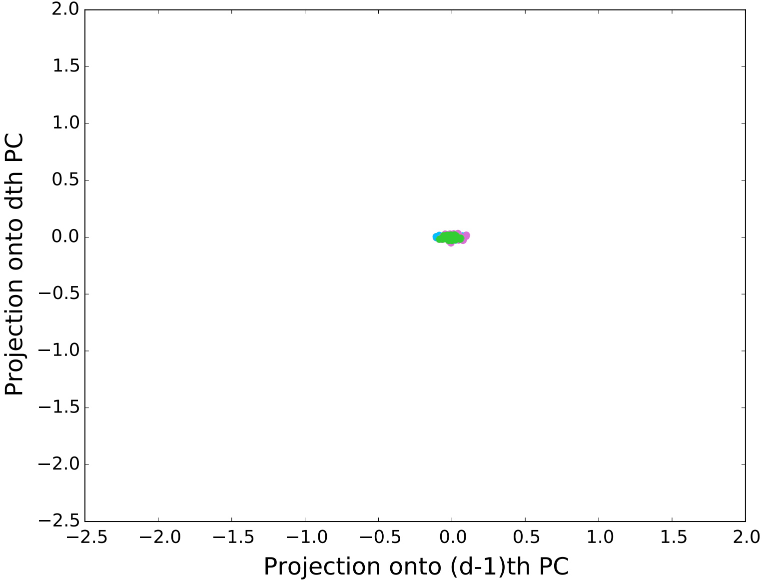
Features:

- Area
- Perimeter
- Compactness
- Length of kernel
- Width of kernel
- Asymmetry coefficient
- Length of kernel groove

Projection onto two first PCs



Projection onto two last PCs



Whitening

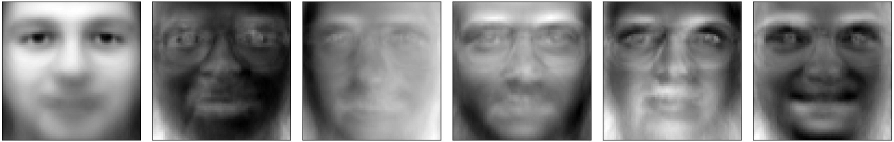
Motivation: Dominating principal directions are not necessarily the most informative

Principal directions corresponding to small singular values may contain information that is *drowned* by main directions

Intuitively, linear skew obscures useful structure

PCA of faces

Center	PD 1	PD 2	PD 3	PD 4	PD 5



$\sigma_i/\sqrt{n-1}$	330	251	192	152	130

Whitening

Given $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n \in \mathbb{R}^d$ we:

1. Center the data,

$$\vec{c}_i = \vec{x}_i - \operatorname{av}\left(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\right), \qquad 1 \leq i \leq n$$

2. Group the centered data as columns of a matrix

$$C = \begin{bmatrix} \vec{c}_1 & \vec{c}_2 & \cdots & \vec{c}_n \end{bmatrix}$$

3. Compute the SVD of $C = USV^T$

4. Whiten by applying the linear map $US^{-1}U^T$

$$\vec{w}_i := US^{-1}U^T \vec{c}_i$$

Covariance of whitened data

Matrix of whitened vectors

$$W = US^{-1}U^{T}C$$

The covariance matrix of the whitened data is

$$\Sigma\left(\vec{c}_1, \ldots, \vec{c}_n\right)$$

Covariance of whitened data

Matrix of whitened vectors

$$W = US^{-1}U^{T}C$$

The covariance matrix of the whitened data is

$$\Sigma\left(\vec{c}_1, \ldots, \vec{c}_n\right) = \frac{1}{n-1} W W^{T}$$

Covariance of whitened data

Matrix of whitened vectors

$$W = US^{-1}U^T C$$

The covariance matrix of the whitened data is

$$\begin{aligned}\Sigma\left(\vec{c}_1, \ldots, \vec{c}_n\right) &= \frac{1}{n-1} W W^T \\ &= \frac{1}{n-1} US^{-1}U^T C C^T US^{-1}U^T\end{aligned}$$

Covariance of whitened data

Matrix of whitened vectors

$$W = US^{-1}U^T C$$

The covariance matrix of the whitened data is

$$\begin{aligned}
\Sigma(\vec{c}_1, \ldots, \vec{c}_n) &= \frac{1}{n-1} WW^T \\
&= \frac{1}{n-1} US^{-1}U^T CC^T US^{-1}U^T \\
&= \frac{1}{n-1} US^{-1}U^T USV^T VSU^T US^{-1}U^T
\end{aligned}$$

Covariance of whitened data

Matrix of whitened vectors

$$W = US^{-1}U^T C$$

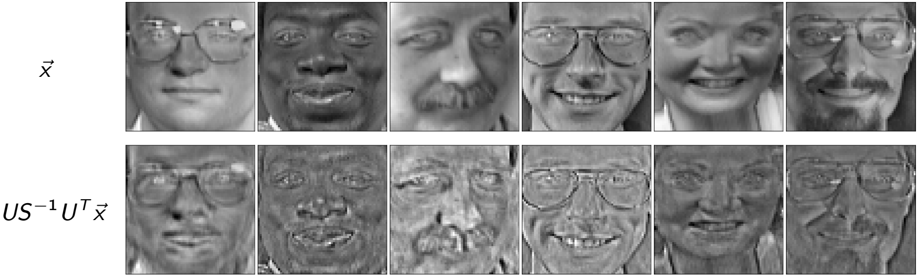
The covariance matrix of the whitened data is

$$\begin{aligned}
\Sigma(\vec{c}_1, \ldots, \vec{c}_n) &= \frac{1}{n-1} WW^T \\
&= \frac{1}{n-1} US^{-1}U^T CC^T US^{-1}U^T \\
&= \frac{1}{n-1} US^{-1}U^T USV^T VSU^T US^{-1}U^T \\
&= \frac{1}{n-1} I
\end{aligned}$$

Whitening

$\vec{x}_1, \ldots, \vec{x}_n$ $\quad U^T\vec{x}_1, \ldots, U^T\vec{x}_n$ $\quad S^{-1}U^T\vec{x}_1, \ldots, S^{-1}U^T\vec{x}_n$

Whitened faces



Eigenvectors

An eigenvector $\vec{q}$ of a square matrix $A \in \mathbb{R}^{n \times n}$ satisfies

$$A\vec{q} = \lambda \vec{q}$$

for a scalar λ which is the corresponding eigenvalue

Even if A is real, its eigenvectors and eigenvalues can be complex

Eigendecomposition

If a matrix has n linearly independent eigenvectors then it is diagonalizable

Let $\vec{q}_1, \ldots, \vec{q}_n$ be lin. indep. eigenvectors of $A \in \mathbb{R}^{n \times n}$ with eigenvalues $\lambda_1, \ldots, \lambda_n$

$$
\begin{aligned}
A &= \begin{bmatrix} \vec{q}_1 & \vec{q}_2 & \cdots & \vec{q}_n \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} \vec{q}_1 & \vec{q}_2 & \cdots & \vec{q}_n \end{bmatrix}^{-1} \\
&= Q \Lambda Q^{-1}
\end{aligned}
$$

Proof

AQ

Proof

$$AQ = \begin{bmatrix} A\vec{q}_1 & A\vec{q}_2 & \cdots & A\vec{q}_n \end{bmatrix}$$

Proof

$$
\begin{aligned}
AQ &= \begin{bmatrix} A\vec{q}_1 & A\vec{q}_2 & \cdots & A\vec{q}_n \end{bmatrix} \\
&= \begin{bmatrix} \lambda_1\vec{q}_1 & \lambda_2\vec{q}_2 & \cdots & \lambda_2\vec{q}_n \end{bmatrix}
\end{aligned}
$$

Proof

$$
\begin{aligned}
AQ &= \begin{bmatrix} A\vec{q}_1 & A\vec{q}_2 & \cdots & A\vec{q}_n \end{bmatrix} \\
&= \begin{bmatrix} \lambda_1\vec{q}_1 & \lambda_2\vec{q}_2 & \cdots & \lambda_2\vec{q}_n \end{bmatrix} \\
&= Q\Lambda
\end{aligned}
$$

Not all matrices have an eigendecomposition

$$A := \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

Assume A has an eigenvector $\vec{q}$ associated to an eigenvalue λ

$$\begin{bmatrix} \vec{q}\,[2] \\ 0 \end{bmatrix}$$

Not all matrices have an eigendecomposition

$$A := \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

Assume A has an eigenvector $\vec{q}$ associated to an eigenvalue λ

$$\begin{bmatrix} \vec{q}\,[2] \\ 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \vec{q}\,[1] \\ \vec{q}\,[2] \end{bmatrix}$$

Not all matrices have an eigendecomposition

$$A := \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

Assume A has an eigenvector $\vec{q}$ associated to an eigenvalue λ

$$\begin{aligned} \begin{bmatrix} \vec{q}\,[2] \\ 0 \end{bmatrix} &= \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \vec{q}\,[1] \\ \vec{q}\,[2] \end{bmatrix} \\ &= \begin{bmatrix} \lambda \vec{q}\,[1] \\ \lambda \vec{q}\,[2] \end{bmatrix} \end{aligned}$$

Spectral theorem for symmetric matrices

If $A \in \mathbb{R}^n$ is symmetric, then it has an eigendecomposition of the form

$$A = U \Lambda U^T$$

where the matrix of eigenvectors U is an orthogonal matrix

Eigendecomposition vs SVD

Symmetric matrices also have singular value decomposition

$$A = USV^T$$

Left singular vectors = eigenvectors

Singular values = magnitude of eigenvalues

$\vec{v}_i = \vec{u}_i$ if $\lambda_i \geq 0$

$\vec{v}_i = -\vec{u}_i$ if $\lambda_i < 0$

Application of eigendecomposition

Fast computation of

$$AA \cdots A\vec{x} = A^k \vec{x}$$

A^k

Application of eigendecomposition

Fast computation of

$$AA \cdots A\vec{x} = A^k \vec{x}$$

$$A^k = Q \Lambda Q^{-1} Q \Lambda Q^{-1} \cdots Q \Lambda Q^{-1}$$

Application of eigendecomposition

Fast computation of

$$AA \cdots A\vec{x} = A^k \vec{x}$$

$$\begin{aligned} A^k &= Q\Lambda Q^{-1} Q\Lambda Q^{-1} \cdots Q\Lambda Q^{-1} \\ &= Q\Lambda^k Q^{-1} \end{aligned}$$

Application of eigendecomposition

Fast computation of

$$AA \cdots A\vec{x} = A^k \vec{x}$$

$$\begin{aligned} A^k &= Q\Lambda Q^{-1} Q\Lambda Q^{-1} \cdots Q\Lambda Q^{-1} \\ &= Q\Lambda^k Q^{-1} \\ &= Q \begin{bmatrix} \lambda_1^k & 0 & \cdots & 0 \\ 0 & \lambda_2^k & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \lambda_n^k \end{bmatrix} Q^{-1} \end{aligned}$$

Application of eigendecomposition

For any vector $\vec{x}$

$$A^k \vec{x}$$

Application of eigendecomposition

For any vector $\vec{x}$

$$A^k \vec{x} = \sum_{i=1}^{n} \alpha_i A^k \vec{q}_i$$

Application of eigendecomposition

For any vector $\vec{x}$

$$A^k \vec{x} = \sum_{i=1}^{n} \alpha_i A^k \vec{q}_i$$
$$= \sum_{i=1}^{n} \alpha_i \lambda_i^k \vec{q}_i$$

Application of eigendecomposition

For any vector $\vec{x}$

$$A^k \vec{x} = \sum_{i=1}^{n} \alpha_i A^k \vec{q}_i$$
$$= \sum_{i=1}^{n} \alpha_i \lambda_i^k \vec{q}_i$$

If $|\lambda_1| > |\lambda_2| \geq \ldots$ then $\vec{q}_1$ will eventually dominate

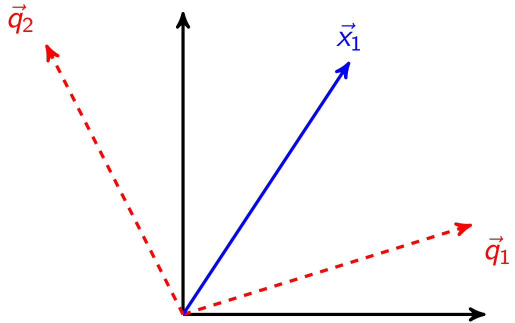
Power method

Set $\vec{x}_1 := \vec{x} / \left|\left|\vec{x}\right|\right|_2$, where $\vec{x}$ is a randomly generated.

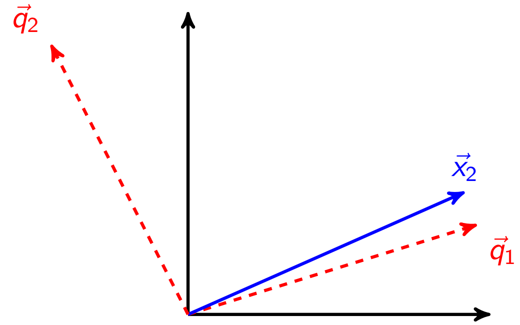
For $i = 1, 2, 3, \ldots$, compute

$$\vec{x}_i := \frac{A\vec{x}_{i-1}}{\left|\left|A\vec{x}_{i-1}\right|\right|_2}$$

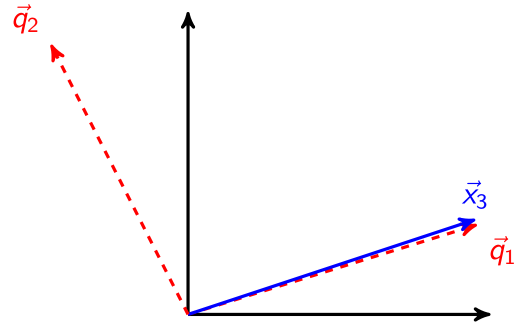
Power method



Power method



Power method



Deer and wolfs

Model for deer d_{n+1} and wolf w_{n+1} population in year $n+1$

$$\begin{aligned} d_{n+1} &= \frac{5}{4}d_n - \frac{3}{4}w_n \\ w_{n+1} &= \frac{1}{4}d_n + \frac{1}{4}w_n \qquad n = 0, 1, 2, \ldots \end{aligned}$$

Deer and wolfs

$$\begin{bmatrix} d_n \\ w_n \end{bmatrix}$$

Deer and wolfs

$$\begin{bmatrix} d_n \\ w_n \end{bmatrix} = Q\Lambda^n Q^{-1} \begin{bmatrix} d_0 \\ w_0 \end{bmatrix}$$

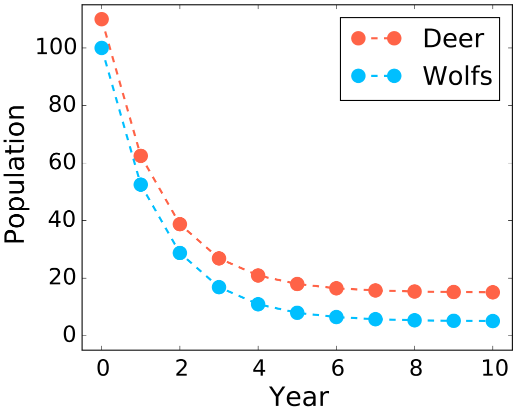
Deer and wolfs

$$
\begin{aligned}
\begin{bmatrix} d_n \\ w_n \end{bmatrix} &= Q\Lambda^n Q^{-1} \begin{bmatrix} d_0 \\ w_0 \end{bmatrix} \\
&= \begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0.5^n \end{bmatrix} \begin{bmatrix} 0.5 & -0.5 \\ -0.5 & 1.5 \end{bmatrix} \begin{bmatrix} d_0 \\ w_0 \end{bmatrix}
\end{aligned}
$$

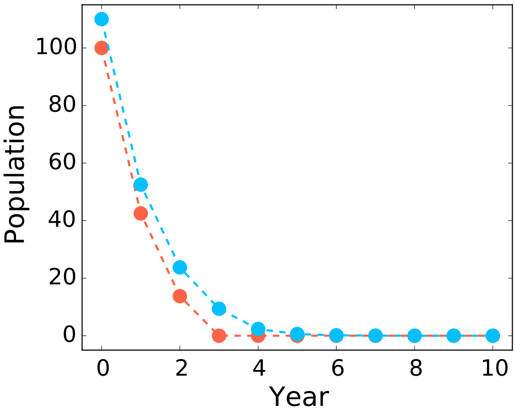
Deer and wolfs

$$
\begin{aligned}
\begin{bmatrix} d_n \\ w_n \end{bmatrix} &= Q\Lambda^n Q^{-1} \begin{bmatrix} d_0 \\ w_0 \end{bmatrix} \\
&= \begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0.5^n \end{bmatrix} \begin{bmatrix} 0.5 & -0.5 \\ -0.5 & 1.5 \end{bmatrix} \begin{bmatrix} d_0 \\ w_0 \end{bmatrix} \\
&= \frac{d_0 - w_0}{2} \begin{bmatrix} 3 \\ 1 \end{bmatrix} + \frac{3w_0 - d_0}{8^n} \begin{bmatrix} 1 \\ 1 \end{bmatrix}
\end{aligned}
$$

Deer and wolfs



Deer and wolfs



Deer and wolfs