Lecture Notes 2: Matrices

Matrices are rectangular arrays of numbers, which are extremely useful for data analysis. They can be interpreted as vectors in a vector space, linear functions or sets of vectors.

1 Basic properties

1.1 Column and row space

A matrix can be used to represent a set of vectors stored as columns or rows. The span of these vectors are called the column and row space of the matrix respectively.

Definition 1.1 (Column and row space)**.** *The column space* $\text{col}(A)$ *of a matrix* A *is the span of its columns. The row space* $\text{row}(A)$ *is the span of its rows.*

Interestingly, the row space and the column space of all matrices have the same dimension. We name this quantity the rank of the matrix.

Definition 1.2 (Rank)**.** *The rank of a matrix is the dimension of its column and of its row space.*

Theorem 1.3 (Proof in Section 5.1)**.** *The rank is well defined. For any matrix* A

$$\dim(\text{col}(A)) = \dim(\text{row}(A)). \tag{1}$$

If the dimension of the row and column space of an $m \times n$ matrix where $m < n$ is equal to m then the the rows are all linearly independent. Similarly, if $m > n$ and the rank is n then the columns are all linearly independent. In general, when the rank equals $\min\{n, m\}$ we say that the matrix is *full rank*.

Recall that the inner product between two matrices $A, B \in \mathbb{R}^{m \times n}$ is given by the trace of $A^T B$, and the norm induced by this inner product is the Frobenius norm. If the column spaces of two matrices are orthogonal, then the matrices are also orthogonal.

Lemma 1.4. *If the column spaces of any pair of matrices* $A, B \in \mathbb{R}^{m \times n}$ *are orthogonal then*

$$\langle A, B \rangle = 0. \tag{2}$$

Proof. We can write the inner product as a sum of products between the columns of A

and B, which are all zero under the assumption of the lemma

$$\langle A, B\rangle := \operatorname{tr}\left(A^T B\right) \tag{3}$$
$$= \sum_{i=1}^{n} \langle A_{:,i}, B_{:,i}\rangle \tag{4}$$
$$= 0. \tag{5}$$

□

The following corollary follows immediately from Lemma 1.4 and the Pythagorean theorem.

Corollary 1.5. *If the column spaces of any pair of matrices $A, B \in \mathbb{R}^{m\times n}$ are orthogonal*

$$||A + B||_{\mathrm{F}}^2 = ||A||_{\mathrm{F}}^2 + ||B||_{\mathrm{F}}^2 . \tag{6}$$

1.2 Linear maps

A map is a transformation that assigns a vector to another vector, possible belonging to a different vector space. The transformation is linear if it maps any linear combination of input vectors to the same linear combination of the corresponding outputs.

Definition 1.6 (Linear map)**.** *Given two vector spaces $\mathcal{V}$ and $\mathcal{R}$ associated to the same scalar field, a linear map $f : \mathcal{V} \to \mathcal{R}$ is a map from vectors in $\mathcal{V}$ to vectors in $\mathcal{R}$ such that for any scalar α and any vectors $\vec{x}_1, \vec{x}_2 \in \mathcal{V}$*

$$f\left(\vec{x}_1 + \vec{x}_2\right) = f\left(\vec{x}_1\right) + f\left(\vec{x}_2\right), \tag{7}$$
$$f\left(\alpha\, \vec{x}_1\right) = \alpha\, f\left(\vec{x}_1\right). \tag{8}$$

Every complex or real matrix of dimensions $m \times n$ defines a map from the space of n-dimensional vectors to the space of m-dimensional vectors through an operation called matrix-vector product. We denote the ith row of a matrix A by $A_{i:}$, the jth column by $A_{:j}$ and the (i, j) entry by A_{ij}.

Definition 1.7 (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*

$$\left(A\vec{x}\right)[i] = \sum_{j=1}^{n} A_{ij}\vec{x}\,[j]\,. \tag{9}$$

For real matrices, each entry in the matrix-vector product is the dot product between a row of the matrix and the vector,

$$\left(A\vec{x}\right)[i] = \langle A_{i:}, \vec{x}\rangle\,. \tag{10}$$

The matrix-vector product can also be interpreted in terms of the column of the matrix,

$$A\vec{x} = \sum_{j=1}^{n} \vec{x}\,[j]\, A_{:j}. \tag{11}$$

$A\vec{x}$ is a linear combination of the columns of A weighted by the entries in $\vec{x}$.

Matrix-vector multiplication is clearly linear. Perhaps surprisingly, the converse is also true: *any* linear map between $\mathbb{C}^n$ and $\mathbb{C}^m$ (or between $\mathbb{R}^n$ and $\mathbb{R}^m$) can be represented by a matrix.

Theorem 1.8 (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

$$F := \begin{bmatrix} f\left(\vec{e}_1\right) & f\left(\vec{e}_2\right) & \cdots & f\left(\vec{e}_n\right) \end{bmatrix}, \tag{12}$$

i.e., the columns of the matrix are the result of applying f to the standard basis. Indeed, for any vector $\vec{x} \in \mathbb{C}^n$

$$f\left(x\right) = f\left(\sum_{i=1}^{n} \vec{x}[i]\vec{e}_i\right) \tag{13}$$

$$= \sum_{i=1}^{n} \vec{x}[i] f\left(\vec{e}_i\right) \quad \text{by (7) and (8)} \tag{14}$$

$$= F\vec{x}. \tag{15}$$

The ith column of any matrix that represents the linear map must equal $f\left(\vec{e}_i\right)$ by (11), so the representation is unique. □

When a matrix $\mathbb{C}^{m \times n}$ is *fat*, i.e., $n > m$, we often say that it *projects* vectors onto a lower dimensional space. Note that such projections are not the same as the orthogonal projections we described in Lecture Notes 1. When a matrix is *tall*, i.e., $m > n$, we say that it *lifts* vectors to a higher-dimensional space.

1.3 Adjoint

The adjoint of a linear map f from an inner-product space $\mathcal{V}$ and another inner product space $\mathcal{R}$ maps elements of $\mathcal{R}$ back to $\mathcal{V}$ in a way that preserves their inner product with images of f.

Definition 1.9 (Adjoint)**.** *Given two vector spaces $\mathcal{V}$ and $\mathcal{R}$ associated to the same scalar field with inner products $\langle \cdot, \cdot \rangle_{\mathcal{V}}$ and $\langle \cdot, \cdot \rangle_{\mathcal{R}}$ respectively, the adjoint $f^* : \mathcal{R} \to \mathcal{V}$ of a linear map $f : \mathcal{V} \to \mathcal{R}$ satisfies*

$$\langle f\left(\vec{x}\right), \vec{y} \rangle_{\mathcal{R}} = \langle \vec{x}, f^*\left(\vec{y}\right) \rangle_{\mathcal{V}} \tag{16}$$

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

In the case of finite-dimensional spaces, the adjoint corresponds to the conjugate Hermitian transpose of the matrix associated to the linear map.

Definition 1.10 (Conjugate transpose)**.** *The entries of the conjugate transpose* $A^* \in \mathbb{C}^{n \times m}$ *of a matrix* $A \in \mathbb{C}^{m \times n}$ *are of the form*

$$(A^*)_{ij} = \overline{A_{ji}}, \quad 1 \leq i \leq n, \ 1 \leq j \leq m. \tag{17}$$

If the entries of the matrix are all real, this is just the transpose of the matrix.

Lemma 1.11 (Equivalence between conjugate transpose and adjoint)**.** *For finite* m, n *the adjoint* $f^* : \mathbb{C}^n \to \mathbb{C}^m$ *of a linear map* $f : \mathbb{C}^m \to \mathbb{C}^n$ *represented by a matrix* F *corresponds to the conjugate transpose of the matrix* F^**.*

Proof. For any $\vec{x} \in \mathbb{C}^n$ and $\vec{y} \in \mathbb{C}^m$,

$$\begin{aligned}
\langle f(\vec{x}), \vec{y} \rangle_{\mathbb{C}^m} &= \sum_{i=1}^{m} f(\vec{x})_i \, \overline{\vec{y}_i} && (18)\\
&= \sum_{i=1}^{m} \overline{\vec{y}_i} \sum_{j=1}^{n} F_{ij} \vec{x}_j && (19)\\
&= \sum_{j=1}^{n} \vec{x}_j \overline{\sum_{i=1}^{m} \overline{F_{ij}} \vec{y}_i} && (20)\\
&= \langle \vec{x}, F^* \vec{y} \rangle_{\mathbb{C}^n} . && (21)
\end{aligned}$$

By Theorem 1.8 a linear map is represented by a unique matrix (you can check that the adjoint map is linear), which completes the proof. □

A matrix that is equal to its adjoint is called self-adjoint. Self-adjoint real matrices are symmetric: they are equal to their transpose. Self-adjoint complex matrices are Hermitian: they are equal to their conjugate transpose.

1.4 Range and null space

The range of a linear map is the set of all possible vectors that can be reached by applying the map.

Definition 1.12 (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* $\mathcal{R}$ *that can be reached by applying* f *to a vector in* $\mathcal{V}$*:*

$$\text{range}(f) := \{\vec{y} \mid \vec{y} = f(\vec{x}) \quad \text{for some } \vec{x} \in \mathcal{V}\}. \tag{22}$$

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

The range of a matrix is the same as its column space.

Lemma 1.13 (The range is the column space)**.** *For any matrix* $A \in \mathbb{C}^{m \times n}$

$$\text{range}\,(A) = \text{col}\,(A)\,. \tag{23}$$

Proof. For any $\vec{x}$, $A\vec{x}$ is a linear combination of the columns of A, so the range is a subset of the column space. In addition, every column of A is in the range, since $A_{:i} = A\vec{e}_i$ for $1 \leq i \leq n$, so the column space is a subset of the range and both sets are equal. □

The null space of a map is the set of vectors that are mapped to zero.

Definition 1.14 (Null space)**.** *Let* $\mathcal{V}$ *and* $\mathcal{R}$ *be vector spaces associated to the same scalar field, the null space of a map* $f : \mathcal{V} \to \mathcal{R}$ *is the set of vectors in* $\mathcal{V}$ *that are mapped to the zero vector in* $\mathcal{R}$ *by* f*:*

$$\text{null}\,(f) := \left\{ \vec{x} \mid f\,(\vec{x}) = \vec{0} \right\}. \tag{24}$$

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

It is not difficult to prove that the null space of a map is a vector space, as long as the map is linear, since in that case scaling or adding elements of the null space yield vectors that are mapped to zero by the map.

The following lemma shows that for real matrices the null space is the orthogonal complement of the row space of the matrix.

Lemma 1.15. *For any matrix* $A \in \mathbb{R}^{m \times n}$

$$\text{null}\,(A) = \text{row}\,(A)^{\perp}\,. \tag{25}$$

Proof. Any vector $\vec{x}$ in the row space of A can be written as $\vec{x} = A^T\vec{z}$, for some vector $\vec{z} \in \mathbb{R}^m$. If $y \in \text{null}\,(A)$ then

$$\begin{aligned} \langle \vec{y}, \vec{x} \rangle &= \langle \vec{y}, A^T\vec{z} \rangle & (26)\\ &= \langle A\vec{y}, \vec{z} \rangle & (27)\\ &= 0. & (28) \end{aligned}$$

So $\text{null}\,(A) \subseteq \text{row}\,(A)^{\perp}$.

If $x \in \text{row}\,(A)^{\perp}$ then in particular it is orthogonal to every row of A, so $Ax = 0$ and $\text{row}\,(A)^{\perp} \subseteq \text{null}\,(A)$. □

An immediate corollary of Lemmas 1.13 and 1.15 is that the dimension of the range and the null space add up to the ambient dimension of the row space.

Corollary 1.16. *Let* $A \in \mathbb{R}^{m \times n}$

$$\dim\left(\text{range}\,(A)\right) + \dim\left(\text{null}\,(A)\right) = n. \tag{29}$$

This means that for every matrix $A \in \mathbb{R}^{m \times n}$ we can decompose any vector in $\mathbb{R}^n$ into two components: one is in the row space and is mapped to a nonzero vector in $\mathbb{C}^m$, the other is in the null space and is mapped to the zero vector.

1.5 Identity matrix and inverse

The identity matrix is a matrix that maps any vector to itself.

Definition 1.17 (Identity matrix)**.** *The identity matrix of dimensions $n \times n$ is*

$$I = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{30}$$

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

Square matrices have a unique inverse if they are full rank, since in that case the null space has dimension 0 and the associated linear map is a bijection. The inverse is a matrix that reverses the effect of the matrix on any vector.

Definition 1.18 (Matrix inverse)**.** *The inverse of a square matrix $A \in \mathbb{C}^{n \times n}$ is a matrix $A^{-1} \in \mathbb{C}^{n \times n}$ such that*

$$AA^{-1} = A^{-1}A = I. \tag{31}$$

1.6 Orthogonal and projection matrices

We often use the letters $U \in \mathbb{R}^{m \times n}$ or $V \in \mathbb{R}^{m \times n}$ for matrices with orthonormal columns. If such matrices are square then they are said to be orthogonal. Orthogonal matrices represent linear maps that do not affect the magnitude of a vector, just its direction.

Definition 1.19 (Orthogonal matrix)**.** *An orthogonal matrix is a real-valued square matrix such that its inverse is equal to its transpose,*

$$U^T U = UU^T = I. \tag{32}$$

By definition, the columns $U_{:1}, U_{:2}, \ldots, U_{:n}$ of any $n \times n$ orthogonal matrix have unit norm and orthogonal to each other, so they form an orthonormal basis (it's somewhat confusing that orthogonal matrices are not called orthonormal matrices instead). Applying U^T to a vector $\vec{x} \in \mathbb{R}^n$ is equivalent to computing the coefficients of its representation in the basis formed by the columns of U. Applying U to $U^T\vec{x}$ recovers $\vec{x}$ by scaling each basis vector with the corresponding coefficient:

$$\vec{x} = UU^T\vec{x} = \sum_{i=1}^{n} \langle U_{:i}, \vec{x} \rangle \, U_{:i}. \tag{33}$$

Since orthogonal matrices only rotate vectors, it is quite intuitive that the product of two orthogonal matrices yields another orthogonal matrix.

Lemma 1.20 (Product of orthogonal matrices)**.** *If $U, V \in \mathbb{R}^{n \times n}$ are orthogonal matrices, then UV is also an orthogonal matrix.*

Proof.

$$(UV)^T (UV) = V^T U^T U V = I. \tag{34}$$

□

The following lemma proves that orthogonal matrices preserve the ℓ_2 norms of vectors.

Lemma 1.21. *Let $U \in \mathbb{R}^{n \times n}$ be an orthogonal matrix. For any vector $\vec{x} \in \mathbb{R}^n$,*

$$||U\vec{x}||_2 = ||\vec{x}||_2 . \tag{35}$$

Proof. By the definition of orthogonal matrix

$$||U\vec{x}||_2^2 = \vec{x}^T U^T U \vec{x} \tag{36}$$

$$= \vec{x}^T \vec{x} \tag{37}$$

$$= ||\vec{x}||_2^2 . \tag{38}$$

□

Matrices with orthonormal columns can also be used to construct orthogonal-projection matrices, which represent orthogonal projections onto a subspace.

Lemma 1.22 (Orthogonal-projection matrix)**.** *Given a subspace $\mathcal{S} \subseteq \mathbb{R}^n$ of dimension d, the matrix*

$$P := UU^T, \tag{39}$$

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

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

$$P\vec{x} = UU^T \vec{x} \tag{40}$$

$$= \sum_{i=1}^{d} \langle U_{:i}, \vec{x} \rangle U_{:i} \tag{41}$$

$$= \mathcal{P}_{\mathcal{S}} \vec{x} \qquad \text{by (64) in the lecture notes on vector spaces.} \tag{42}$$

□

2 Singular-value decomposition

In this section we introduce the singular-value decomposition, a fundamental tool for manipulating matrices, and describe several applications in data analysis.

2.1 Definition

Every real matrix has a singular-value decomposition.

Theorem 2.1. *Every rank r real matrix $A \in R^{m \times n}$, has a singular-value decomposition (SVD) of the form*

$$A = \begin{bmatrix} \vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_r \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & \sigma_r \end{bmatrix} \begin{bmatrix} \vec{v}_1^T \\ \vec{v}_2^T \\ \vdots \\ \vec{v}_r^T \end{bmatrix} \tag{43}$$

$$= USV^T, \tag{44}$$

where 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, and 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 several singular values are the same, their left singular vectors can be replaced by any orthonormal basis of their span, and the same holds for the right singular vectors.

The SVD of an $m \times n$ matrix with $m \geq n$ can be computed in $\mathcal{O}(mn^2)$. We refer to any graduate linear algebra book for the proof of Theorem 2.1 and for the details on how to compute the SVD.

The SVD provides orthonormal bases for the column and row spaces of the matrix.

Lemma 2.2. *The left singular vectors are an orthonormal basis for the column space, whereas the right singular vectors are an orthonormal basis for the row space.*

Proof. We prove the statement for the column space, the proof for the row space is identical. All left singular vectors belong to the column space because $\vec{u}_i = A\left(\sigma_i^{-1}\vec{v}_i\right)$. In addition, every column of A is in their span because $A_{:i} = U\left(SV^T\vec{e}_i\right)$. Since they form an orthonormal set by Theorem 2.1, this completes the proof. □

The SVD presented in Theorem 2.1 can be augmented so that the number of singular values equals $\min(m, n)$. The additional singular values are all equal to zero. Their corresponding left and right singular vectors are orthonormal sets of vectors in the orthogonal complements of the column and row space respectively. If the matrix is tall or square, the additional right singular vectors are a basis of the null space of the matrix.

Corollary 2.3 (Singular-value decomposition)**.** *Every rank r real matrix $A \in R^{m \times n}$,*

where $m \geq n$, has a singular-value decomposition (SVD) of the form

$$A := [\underbrace{\vec{u}_1\ \vec{u}_2\ \cdots\ \vec{u}_r}_{\textit{Basis of } \operatorname{range}(A)}\ \ \vec{u}_{r+1}\ \cdots\ \vec{u}_n] \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 & 0 & \cdots & 0 \\ & & & \cdots & & & \\ 0 & 0 & \cdots & \sigma_r & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ & & & \cdots & & & \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix} [\underbrace{\vec{v}_1\ \vec{v}_2\ \cdots\ \vec{v}_r}_{\textit{Basis of } \operatorname{row}(A)}\ \underbrace{\vec{v}_{r+1}\ \cdots\ \vec{v}_n}_{\textit{Basis of } \operatorname{null}(A)}]^T, \tag{45}$$

where the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$ are positive real numbers, the left singular vectors $\vec{u}_1$, $\vec{u}_2$, ..., $\vec{u}_m$ form an orthonormal set in $\mathbb{R}^m$, and the right singular vectors $\vec{v}_1$, $\vec{v}_2$, ..., $\vec{v}_m$ form an orthonormal basis for $\mathbb{R}^n$.

If the matrix is fat, we can define a similar augmentation, where the additional left singular vectors form an orthonormal basis of the orthogonal complement of the range.

By the definition of rank and Lemma 2.2 the rank of a matrix is equal to the number of nonzero singular values.

Corollary 2.4. *The rank of a matrix is equal to the number of nonzero singular values.*

This interpretation of the rank allows to define an alternative definition that is very useful in practice, since matrices are often full rank due to numerical error, even if their columns or rows are *almost* linearly dependent.

Definition 2.5 (Numerical rank)**.** *Given a tolerance $\epsilon > 0$, the numerical rank of a matrix is the number of singular values that are greater than ϵ.*

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 three simple steps:

1. Rotation of $\vec{x}$ to align the component of $\vec{x}$ in the direction of the ith right singular vector $\vec{v}_i$ with the ith axis:

$$V^T \vec{x} = \sum_{i=1}^{n} \langle \vec{v}_i, \vec{x} \rangle \, \vec{e}_i. \tag{46}$$

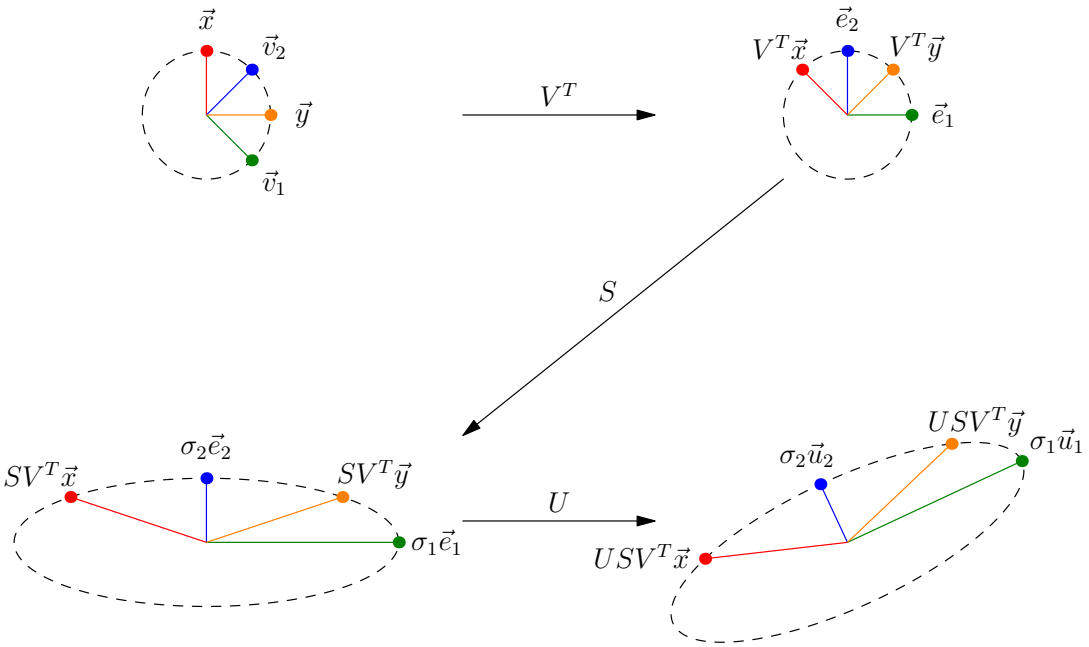
2. Scaling of each axis by the corresponding singular value

$$SV^T \vec{x} = \sum_{i=1}^{n} \sigma_i \, \langle \vec{v}_i, \vec{x} \rangle \, \vec{e}_i. \tag{47}$$

3. Rotation to align the ith axis with the ith left singular vector

$$USV^T \vec{x} = \sum_{i=1}^{n} \sigma_i \, \langle \vec{v}_i, \vec{x} \rangle \, \vec{u}_i. \tag{48}$$

(a) $\sigma_1 = 3$, $\sigma_2 = 1$.



(b) $\sigma_1 = 3$, $\sigma_2 = 0$.

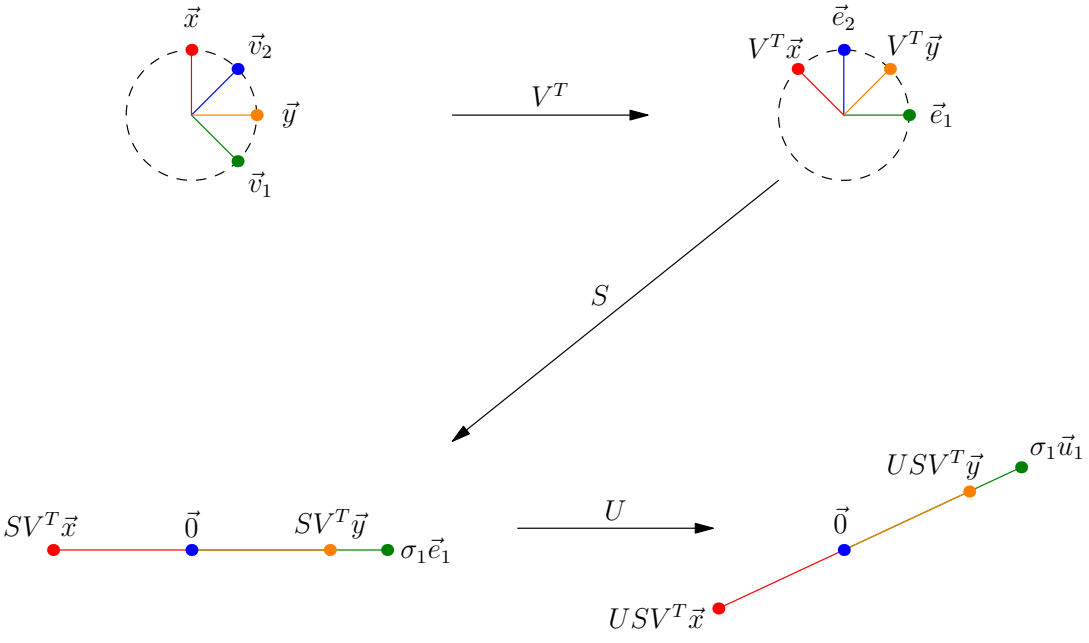


Figure 1: Any linear map can be decomposed into three steps: rotation to align the right singular vectors to the axes, scaling by the singular values and a final rotation to align the axes with the left singular vectors. In image (b) the second singular value is zero, so the linear map projects two-dimensional vectors onto a one-dimensional subspace.

Figure 1 illustrates this geometric analysis of the action of a linear map.

We end the section by showing that multiplying a matrix by an orthogonal matrix does not affect its singular values. This makes sense since it just modifies the rotation carried out by the left or right singular vectors.

Lemma 2.6. *For any matrix $A \in \mathbb{R}^{m\times n}$ and any orthogonal matrices $\widetilde{U} \in \mathbb{R}^{m\times m}$ and $\widetilde{V} \in \mathbb{R}^{m\times m}$ the singular values of $\widetilde{U}A$ and $A\widetilde{V}$ are the same as the singular values of A.*

Proof. Let $A = USV^T$ be the SVD of A. By Lemma 1.20 the matrices $\overline{U} := \widetilde{U}U$ and $\overline{V}^T := V^T\widetilde{V}$ are orthogonal matrices, so $\overline{U}SV^T$ and $US\overline{V}^T$ are valid SVDs for $\widetilde{U}A$ and $A\widetilde{V}$ respectively. The result follows by unicity of the SVD. □

2.2 Optimal approximations via the SVD

In the previous section, we show that linear maps rotate vectors, scale them according to the singular values and then rotate them again. This means that the maximum scaling possible is equal to the maximum singular value and occurs in the direction of the right singular vector $\vec{v}_1$. The following theorem makes this precise, showing that if we restrict our attention to the orthogonal complement of $\vec{v}_1$, then the maximum scaling is the second singular value, due to the orthogonality of the singular vectors. In general, the direction of maximum scaling orthogonal to the first $i-1$ left singular vectors is equal to the ith singular value and occurs in the direction of the ith singular vector.

Theorem 2.7. *For any matrix $A \in \mathbb{R}^{m\times n}$, with SVD given by (45), the singular values satisfy*

$$\sigma_1 = \max_{\left\{||\vec{x}||_2=1 \mid \vec{x}\in\mathbb{R}^n\right\}} ||A\vec{x}||_2 \tag{49}$$

$$= \max_{\left\{||\vec{y}||_2=1 \mid \vec{y}\in\mathbb{R}^m\right\}} \left|\left|A^T\vec{y}\right|\right|_2, \tag{50}$$

$$\sigma_i = \max_{\left\{||\vec{x}||_2=1 \mid \vec{x}\in\mathbb{R}^n, \vec{x}\perp\vec{v}_1,\ldots,\vec{v}_{i-1}\right\}} ||A\vec{x}||_2, \tag{51}$$

$$= \max_{\left\{||\vec{y}||_2=1 \mid \vec{y}\in\mathbb{R}^m, \vec{y}\perp\vec{u}_1,\ldots,\vec{u}_{i-1}\right\}} \left|\left|A^T\vec{y}\right|\right|_2, \qquad 2 \leq i \leq \min\{m,n\}, \tag{52}$$

the right singular vectors satisfy

$$\vec{v}_1 = \arg\max_{\left\{||\vec{x}||_2=1 \mid \vec{x}\in\mathbb{R}^n\right\}} ||A\vec{x}||_2, \tag{53}$$

$$\vec{v}_i = \arg\max_{\left\{||\vec{x}||_2=1 \mid \vec{x}\in\mathbb{R}^n, \vec{x}\perp\vec{v}_1,\ldots,\vec{v}_{i-1}\right\}} ||A\vec{x}||_2, \qquad 2 \leq i \leq m, \tag{54}$$

and the left singular vectors satisfy

$$\vec{u}_1 = \underset{\left\{||\vec{y}||_2=1 \mid \vec{y}\in\mathbb{R}^m\right\}}{\arg\max} \left|\left|A^T\vec{y}\right|\right|_2, \tag{55}$$

$$\vec{u}_i = \underset{\left\{||\vec{y}||_2=1 \mid \vec{y}\in\mathbb{R}^m, \vec{y}\perp\vec{u}_1,\ldots,\vec{u}_{i-1}\right\}}{\arg\max} \left|\left|A^T\vec{y}\right|\right|_2, \qquad 2 \leq i \leq n. \tag{56}$$

Proof. Consider a vector $\vec{x} \in \mathbb{R}^n$ with unit ℓ_2 norm that is orthogonal to $\vec{v}_1, \ldots, \vec{v}_{i-1}$, where $1 \leq i \leq n$ (if $i = 1$ then $\vec{x}$ is just an arbitrary vector). We express $\vec{x}$ in terms of the right singular vectors of A and a component that is orthogonal to their span

$$\vec{x} = \sum_{j=i}^{n} \alpha_j \vec{v}_j + \mathcal{P}_{\text{row}(A)^\perp}\, \vec{x} \tag{57}$$

where $1 = ||\vec{x}||_2^2 \geq \sum_{j=i}^{n} \alpha_j^2$. By the ordering of the singular values in Theorem 2.1

$$||A\vec{x}||_2^2 = \left\langle \sum_{k=1}^{n} \sigma_k \left\langle \vec{v}_k, \vec{x} \right\rangle \vec{u}_k, \sum_{k=1}^{n} \sigma_k \left\langle \vec{v}_k, \vec{x} \right\rangle \vec{u}_k \right\rangle \quad \text{by (48)} \tag{58}$$

$$= \sum_{k=1}^{n} \sigma_k^2 \left\langle \vec{v}_k, \vec{x} \right\rangle^2 \quad \text{because } \vec{u}_1, \ldots, \vec{u}_n \text{ are orthonormal} \tag{59}$$

$$= \sum_{k=1}^{n} \sigma_k^2 \left\langle \vec{v}_k, \sum_{j=i}^{n} \alpha_j \vec{v}_j + \mathcal{P}_{\text{row}(A)^\perp}\, \vec{x} \right\rangle^2 \tag{60}$$

$$= \sum_{j=i}^{n} \sigma_j^2 \alpha_j^2 \qquad \text{because } \vec{v}_1, \ldots, \vec{v}_n \text{ are orthonormal} \tag{61}$$

$$\leq \sigma_i^2 \sum_{j=i}^{n} \alpha_j^2 \quad \text{because } \sigma_i \geq \sigma_{i+1} \geq \ldots \geq \sigma_n \tag{62}$$

$$\leq \sigma_i^2 \quad \text{by (57)}. \tag{63}$$

This establishes (49) and (51). To prove (53) and (54) we show that $\vec{v}_i$ achieves the maximum

$$||A\vec{v}_i||_2^2 = \sum_{k=1}^{n} \sigma_k^2 \left\langle \vec{v}_k, \vec{v}_i \right\rangle^2 \tag{64}$$

$$= \sigma_i^2. \tag{65}$$

The same argument applied to A^T establishes (50), (55), (56) and (52). □

Given a set of vectors, it is often of interest to determine whether they are oriented in particular directions of the ambient space. This can be quantified in terms of the ℓ_2 norms of their projections on low-dimensional subspaces. The SVD provides an optimal k-dimensional subspace in this sense for *any value of* k.

Theorem 2.8 (Optimal subspace for orthogonal projection)**.** *For any matrix*

$$A := \begin{bmatrix} \vec{a}_1 & \vec{a}_2 & \cdots & \vec{a}_n \end{bmatrix} \in \mathbb{R}^{m \times n}, \tag{66}$$

with SVD given by (45)*, we have*

$$\sum_{i=1}^{n} \left|\left| \mathcal{P}_{\text{span}(\vec{u}_1, \vec{u}_2, \ldots, \vec{u}_k)} \, \vec{a}_i \right|\right|_2^2 \geq \sum_{i=1}^{n} ||\mathcal{P}_{\mathcal{S}} \, \vec{a}_i||_2^2 \,, \tag{67}$$

for any subspace $\mathcal{S}$ *of dimension* $k \leq \min\{m, n\}$*.*

Proof. Note that

$$\sum_{i=1}^{n} \left|\left| \mathcal{P}_{\text{span}(\vec{u}_1, \vec{u}_2, \ldots, \vec{u}_k)} \, \vec{a}_i \right|\right|_2^2 = \sum_{i=1}^{n} \sum_{j=1}^{k} \langle \vec{u}_j, \vec{a}_i \rangle^2 \tag{68}$$

$$= \sum_{j=1}^{k} \left|\left| A^T \vec{u}_j \right|\right|_2^2 . \tag{69}$$

We prove the result by induction on k. The base case $k = 1$ follows immediately from (55). To complete the proof we show that if the result is true for $k - 1 \geq 1$ (the induction hypothesis) then it also holds for k. Let $\mathcal{S}$ be an arbitrary subspace of dimension k. The intersection of $\mathcal{S}$ and the orthogonal complement to the span of $\vec{u}_1$, $\vec{u}_2$,..., $\vec{u}_{k-1}$ contains a nonzero vector $\vec{b}$ due to the following lemma.

Lemma 2.9 (Proof in Section 5.2)**.** *In a vector space of dimension* n*, the intersection of two subspaces with dimensions* d_1 *and* d_2 *such that* $d_1 + d_2 > n$ *has dimension at least one.*

We choose 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 $\vec{u}_1, \vec{u}_2, \ldots, \vec{u}_{k-1}$ (we can construct such a basis by Gram-Schmidt, starting with $\vec{b}$). By the induction hypothesis,

$$\sum_{i=1}^{k-1} \left|\left| A^T \vec{u}_i \right|\right|_2^2 = \sum_{i=1}^{n} \left|\left| \mathcal{P}_{\text{span}(\vec{u}_1, \vec{u}_2, \ldots, \vec{u}_{k-1})} \, \vec{a}_i \right|\right|_2^2 \tag{70}$$

$$\geq \sum_{i=1}^{n} \left|\left| \mathcal{P}_{\text{span}\left(\vec{b}_1, \vec{b}_2, \ldots, \vec{b}_{k-1}\right)} \, \vec{a}_i \right|\right|_2^2 \tag{71}$$

$$= \sum_{i=1}^{k-1} \left|\left| A^T \vec{b}_i \right|\right|_2^2 . \tag{72}$$

By (56)

$$\left|\left| A^T \vec{u}_k \right|\right|_2^2 \geq \left|\left| A^T \vec{b}_k \right|\right|_2^2 . \tag{73}$$

Combining (72) and (73) we conclude

$$\sum_{i=1}^{n} \left|\left|\mathcal{P}_{\text{span}(\vec{u}_1,\vec{u}_2,\ldots,\vec{u}_k)}\,\vec{a}_i\right|\right|_2^2 = \sum_{i=1}^{k} \left|\left|A^T\vec{u}_i\right|\right|_2^2 \tag{74}$$

$$\geq \sum_{i=1}^{k} \left|\left|A^T\vec{b}_i\right|\right|_2^2 \tag{75}$$

$$= \sum_{i=1}^{n} ||\mathcal{P}_{\mathcal{S}}\,\vec{a}_i||_2^2\,. \tag{76}$$

□

The SVD also allows to compute the optimal k-rank approximation to a matrix in Frobenius norm, for any value of k. For any matrix A, we denote by $A_{1:i,1:j}$ to denote the $i \times j$ submatrix formed by taking the entries that are both in the first i rows and the first j columns. Similarly, we denote by $A_{:,i:j}$ the matrix formed by columns i to j.

Theorem 2.10 (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 of A in the sense that*

$$U_{:,1:k}S_{1:k,1:k}V_{:,1:k}^T = \underset{\left\{\widetilde{A}\;|\;\text{rank}\left(\widetilde{A}\right)=k\right\}}{\arg\min} \left|\left|A-\widetilde{A}\right|\right|_{\text{F}}\,. \tag{77}$$

Proof. Let $\widetilde{A}$ be an arbitrary matrix in $\mathbb{R}^{m\times n}$ with $\text{rank}(\widetilde{A}) = k$, and let $\widetilde{U} \in \mathbb{R}^{m\times k}$ be a matrix with orthonormal columns such that $\text{col}(\widetilde{U}) = \text{col}(\widetilde{A})$. By Theorem 2.8,

$$\left|\left|U_{:,1:k}U_{:,1:k}^TA\right|\right|_{\text{F}}^2 = \sum_{i=1}^{n} \left|\left|\mathcal{P}_{\text{col}\left(U_{:,1:k}\right)}\,\vec{a}_i\right|\right|_2^2 \tag{78}$$

$$\geq \sum_{i=1}^{n} \left|\left|\mathcal{P}_{\text{col}\left(\widetilde{U}\right)}\,\vec{a}_i\right|\right|_2^2 \tag{79}$$

$$= \left|\left|\widetilde{U}\widetilde{U}^TA\right|\right|_{\text{F}}^2\,. \tag{80}$$

The column space of $A - \widetilde{U}\widetilde{U}^TA$ is orthogonal to the column space of $\widetilde{A}$ and $\widetilde{U}$, so by Corollary 1.5

$$\left|\left|A-\widetilde{A}\right|\right|_{\text{F}}^2 = \left|\left|A-\widetilde{U}\widetilde{U}^TA\right|\right|_{\text{F}}^2 + \left|\left|\widetilde{A}-\widetilde{U}\widetilde{U}^TA\right|\right|_{\text{F}}^2 \tag{81}$$

$$\geq \left|\left|A-\widetilde{U}\widetilde{U}^TA\right|\right|_{\text{F}}^2 \tag{82}$$

$$= ||A||_{\text{F}}^2 - \left|\left|\widetilde{U}\widetilde{U}^TA\right|\right|_{\text{F}}^2 \qquad \text{also by Corollary 1.5} \tag{83}$$

$$\geq ||A||_{\text{F}}^2 - \left|\left|U_{:,1:k}U_{:,1:k}^TA\right|\right|_{\text{F}}^2 \qquad \text{by (80)} \tag{84}$$

$$= \left|\left|A-U_{:,1:k}U_{:,1:k}^TA\right|\right|_{\text{F}}^2 \qquad \text{again by Corollary 1.5.} \tag{85}$$

□

2.3 Matrix norms

As we discussed in Lecture Notes 1, the inner-product norm in the vector space of matrices is the Frobenius norm. The following lemma establishes that the Frobenius norm of a matrix equals the ℓ_2 norm of its singular values.

Lemma 2.11. *For any matrix $A \in \mathbb{R}^{m \times n}$, with singular values $\sigma_1, \ldots, \sigma_{\min\{m,n\}}$*

$$||A||_{\mathrm{F}} = \sqrt{\sum_{i=1}^{\min\{m,n\}} \sigma_i^2}. \tag{86}$$

Proof. Let us denote the SVD of A by USV^T,

$$\begin{aligned}
||A||_{\mathrm{F}}^2 &= \operatorname{tr}\left(A^T A\right) && (87)\\
&= \operatorname{tr}\left(VSU^T USV^T\right) \qquad \text{by Lemma 2.5 in Lecture Notes 1} && (88)\\
&= \operatorname{tr}\left(VSSV^T\right) \qquad \text{because } U^T U = I && (89)\\
&= \operatorname{tr}\left(V^T VSS\right) && (90)\\
&= \operatorname{tr}\left(SS\right) \qquad \text{because } V^T V = I. && (91)
\end{aligned}$$

□

The operator norm quantifies how much a linear map can scale a vector in ℓ_2 norm.

Definition 2.12 (Operator norm)**.** *The operator norm of a linear map and of the corresponding matrix $A \in \mathbb{R}^{m \times n}$ is defined by*

$$||A|| := \max_{\left\{||\vec{x}||_2 = 1 \mid \vec{x} \in \mathbb{R}^n\right\}} ||A\,\vec{x}||_2 \,. \tag{92}$$

By Theorem 2.7 (see equation (49)) the operator norm is equal to the ℓ_∞ norm of the singular values, i.e. the largest one, is also a norm.

Corollary 2.13. *For any matrix $A \in \mathbb{R}^{m \times n}$, with singular values $\sigma_1, \ldots, \sigma_{\min\{m,n\}}$*

$$||A|| := \sigma_1. \tag{93}$$

We end the section by defining an additional matrix norm, this time directly in term of the singular values.

Definition 2.14 (Nuclear norm)**.** *The nuclear norm of a matrix $A \in \mathbb{R}^{m \times n}$ is equal to the ℓ_1 norm of its singular values $\sigma_1, \ldots, \sigma_{\min\{m,n\}}$*

$$||A||_* := \sum_{i=1}^{\min\{m,n\}} \sigma_i. \tag{94}$$

Any matrix norm that is a function of the singular values of a matrix is preserved after multiplication by an orthogonal matrix. This is a direct corollary of Lemma 2.6.

Corollary 2.15. *For any matrix $A \in \mathbb{R}^{m\times n}$ and any orthogonal matrices $\widetilde{U} \in \mathbb{R}^{m\times m}$ and $\widetilde{V} \in \mathbb{R}^{n\times n}$ the operator, Frobenius and nuclear norm of $\widetilde{U}A$ and $A\widetilde{V}$ are the same as those of A.*

The following theorem is analogous to Hölder's inequality for vectors.

Theorem 2.16 (Proof in Section 5.3)**.** *For any matrix $A \in \mathbb{R}^{m\times n}$,*

$$||A||_* = \sup_{\{||B||\leq 1 \mid B\in\mathbb{R}^{m\times n}\}} \langle A, B\rangle \,. \tag{95}$$

A direct consequence of the result is that the nuclear norm satisfies the triangle inequality. This implies that it is a norm, since it clearly satisfies the remaining properties.

Corollary 2.17. *For any $m \times n$ matrices A and B*

$$||A + B||_* \leq ||A||_* + ||B||_* \,. \tag{96}$$

Proof.

$$\begin{aligned} ||A + B||_* &= \sup_{\{||C||\leq 1 \mid C\in\mathbb{R}^{m\times n}\}} \langle A + B, C\rangle && (97)\\ &\leq \sup_{\{||C||\leq 1 \mid C\in\mathbb{R}^{m\times n}\}} \langle A, C\rangle + \sup_{\{||D||\leq 1 \mid D\in\mathbb{R}^{m\times n}\}} \langle B, D\rangle && (98)\\ &= ||A||_* + ||B||_* \,. && (99) \end{aligned}$$

□

2.4 Denoising via low-rank matrix estimation

In this section we consider the problem of denoising a set of n m-dimensional signals $\vec{x}_1$, $\vec{x}_2$, ..., $\vec{x}_n \in \mathbb{R}^m$. We model the noisy data as the sum between each signal and a noise vector

$$\vec{y}_i = \vec{x}_i + \vec{z}_i, \qquad 1 \leq i \leq n. \tag{100}$$

Our first assumption is that the signals are similar, in the sense that they approximately span a low-dimensional subspace due to the correlations between them. If this is the case, then the matrix

$$X := \begin{bmatrix} \vec{x}_1 & \vec{x}_2 & \cdots & \vec{x}_n \end{bmatrix} \tag{101}$$

obtained by stacking the signals as columns is approximately low rank. Note that in contrast to the subspace-projection denoising method described in Lecture Notes 1, we do *not* assume that the subspace is known.

Our second assumption is that the noise vectors are independent from each other, so that the noise matrix

$$Z := \begin{bmatrix} \vec{z}_1 & \vec{z}_2 & \cdots & \vec{z}_n \end{bmatrix} \tag{102}$$

is full rank. If the noise is not too large with respect to the signals, under these assumptions a low-rank approximation to the data matrix

$$\begin{aligned} Y &:= \begin{bmatrix} \vec{y}_1 & \vec{y}_2 & \cdots & \vec{y}_n \end{bmatrix} && (103) \\ &= X + Z && (104) \end{aligned}$$

should mostly suppress the noise and extract the component corresponding to the signals. Theorem 2.10 establishes that the best rank-k approximation to a matrix in Frobenius norm is achieved by truncating the SVD, for any value of k.

Algorithm 2.18 (Denoising via SVD truncation)**.** *Given n noisy data vectors $\vec{y}_1, \vec{y}_2, \ldots, \vec{y}_n \in \mathbb{R}^m$, we denoise the data by*

1. *Stacking the vectors as the columns of a matrix $Y \in \mathbb{R}^{m \times n}$.*
2. *Computing the SVD of $Y = USV^T$.*
3. *Truncating the SVD to produce the low-rank estimate L*

$$L := U_{:,1:k} S_{1:k,1:k} V^T_{:,1:k}, \tag{105}$$

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

An important decision is what rank k to choose. Higher ranks yield more accurate approximations to the original signals than lower-rank approximations, but they do not suppress the noise component in the data as much. The following example illustrates this tradeoff.

Example 2.19 (Denoising of digit images)**.** In this example we use the MNIST data set[1] to illustrate image denoising using SVD truncation. The signals consist of 6131 28×28 images of the number 3. The images are corrupted by noise sampled independently from a Gaussian distribution and scaled so that the signal-to-noise ratio (defined as the ratio between the ℓ_2 norms of the clean image and the noise) is 0.5 (there is more noise than signal!). Our assumption is that because of their similarities, the images interpreted as vectors in $\mathbb{R}^{784}$ form a low-dimensional (but unknown subspace), whereas the noise is uncorrelated and therefore is not restricted to a subspace. This assumption holds: Figure 11 shows the singular values of matrix formed by stacking the clean images, the noisy images and the noise.

We center each noisy image by subtracting the average of all the noisy images. Subtracting the average is a common preprocessing step in low-rank approximations (see Figure 5 for a geometric justification). We then apply SVD truncation to obtain a low-rank estimate of the image matrix. The noisy average is then added back to the images to produce the final estimate of the images. Figure 3 shows the results for rank-10 and rank-40 estimates. The lower-rank estimate suppresses the noise more, but does not approximate the original signals as effectively. △

[1] Available at http://yann.lecun.com/exdb/mnist/

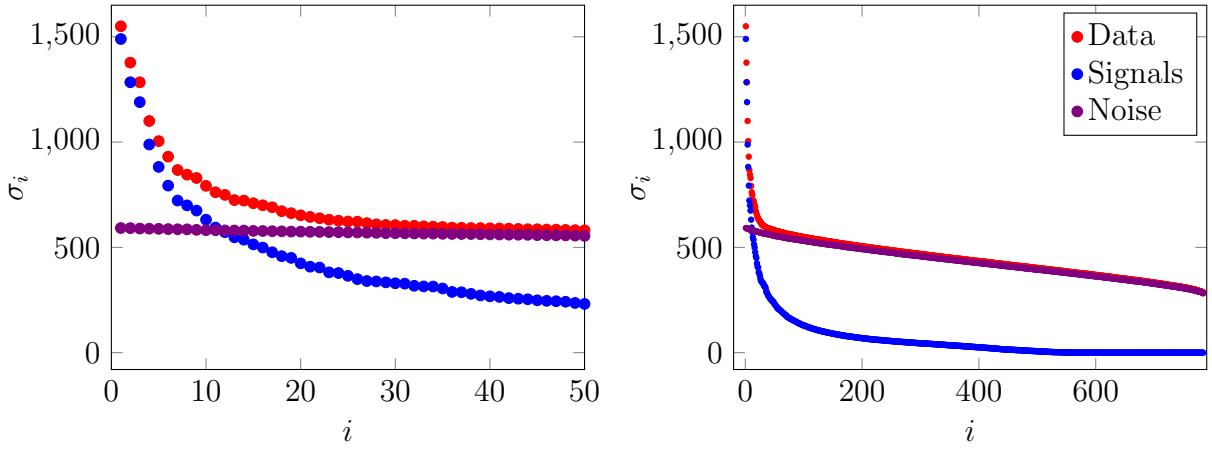


Figure 2: Plots of the singular values of the clean images, the noisy images and the noise in Example 2.19.

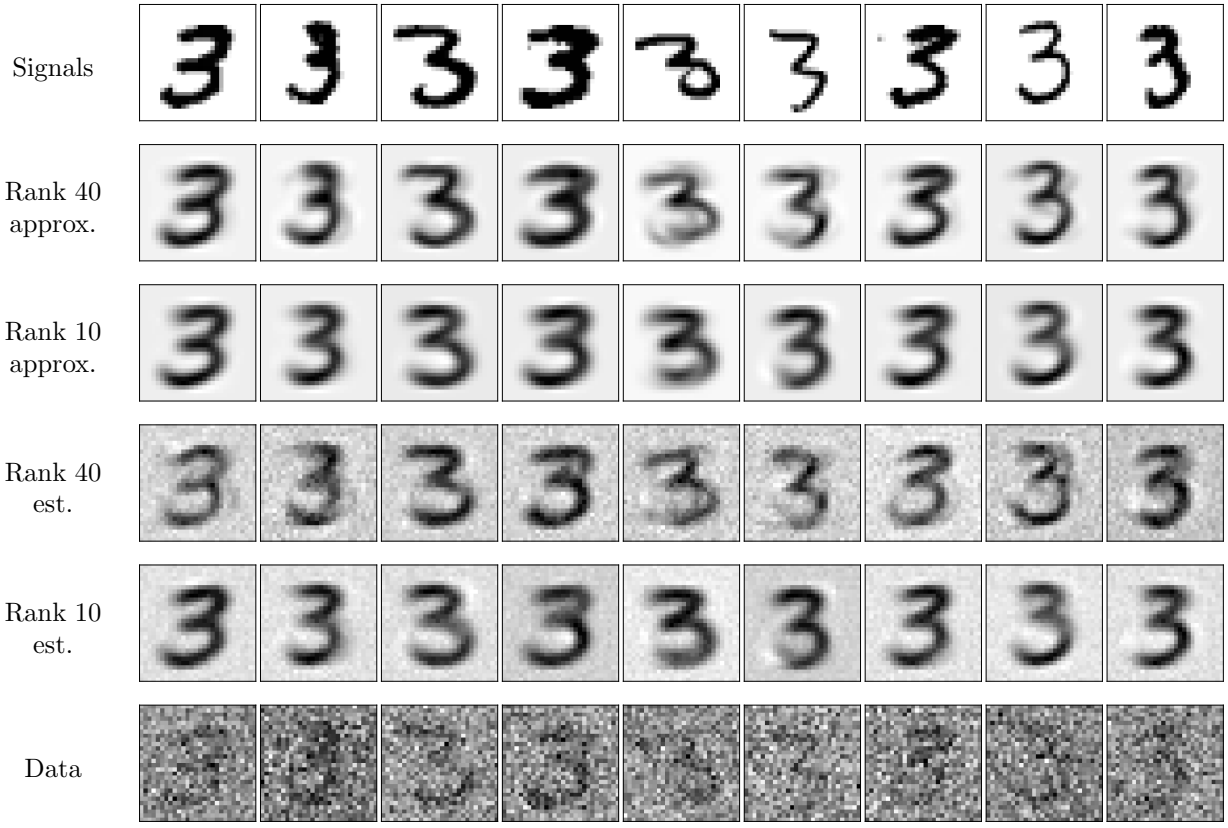


Figure 3: The images show 9 examples from the 6131 images used in Example 2.19. The top row shows the original clean images. The second and third rows show rank-40 and rank-10 approximations to the clean images. The fourth and fifth rows shows the results of applying SVD truncation to obtain rank-40 and rank-10 estimates respectively. The sixth row shows the noisy data.

2.5 Collaborative filtering

The aim of collaborative filtering is to pool together information from many users to obtain a model of their behavior. To illustrate the use of low-rank models in this application we consider a toy example. Bob, Molly, Mary and Larry rate the following six movies from 1 to 5,

$$A := \begin{array}{c} \begin{array}{cccc} \text{Bob} & \text{Molly} & \text{Mary} & \text{Larry} \end{array} \\ \begin{pmatrix} 1 & 1 & 5 & 4 \\ 2 & 1 & 4 & 5 \\ 4 & 5 & 2 & 1 \\ 5 & 4 & 2 & 1 \\ 4 & 5 & 1 & 2 \\ 1 & 2 & 5 & 5 \end{pmatrix} \end{array} \begin{array}{l} \text{The Dark Knight} \\ \text{Spiderman 3} \\ \text{Love Actually} \\ \text{Bridget Jones's Diary} \\ \text{Pretty Woman} \\ \text{Superman 2} \end{array} \tag{106}$$

A common assumption in collaborative filtering is that there are people that have similar tastes and hence produce similar ratings, and that there are movies that are similar and hence elicit similar reactions. Interestingly, this tends to induce low-rank structure in the matrix of ratings. To uncover this low-rank structure, we first subtract the average rating

$$\mu := \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij}, \tag{107}$$

from each entry in the matrix to obtain a centered matrix C and then compute its singular-value decomposition

$$A - \mu \vec{1}\,\vec{1}^{\,T} = USV^T = U \begin{bmatrix} 7.79 & 0 & 0 & 0 \\ 0 & 1.62 & 0 & 0 \\ 0 & 0 & 1.55 & 0 \\ 0 & 0 & 0 & 0.62 \end{bmatrix} V^T. \tag{108}$$

where $\vec{1} \in \mathbb{R}^4$ is a vector of ones. The fact that the first singular value is significantly larger than the rest suggests that the matrix may be well approximated by a rank-1 matrix. This is indeed the case:

$$\mu \vec{1}\,\vec{1}^{\,T} + \sigma_1 \vec{u}_1 \vec{v}_1^{\,T} = \begin{array}{c} \begin{array}{cccc} \text{Bob} & \text{Molly} & \text{Mary} & \text{Larry} \end{array} \\ \begin{pmatrix} 1.34\,(1) & 1.19\,(1) & 4.66\,(5) & 4.81\,(4) \\ 1.55\,(2) & 1.42\,(1) & 4.45\,(4) & 4.58\,(5) \\ 4.45\,(4) & 4.58\,(5) & 1.55\,(2) & 1.42\,(1) \\ 4.43\,(5) & 4.56\,(4) & 1.57\,(2) & 1.44\,(1) \\ 4.43\,(4) & 4.56\,(5) & 1.57\,(1) & 1.44\,(2) \\ 1.34\,(1) & 1.19\,(2) & 4.66\,(5) & 4.81\,(5) \end{pmatrix} \end{array} \begin{array}{l} \text{The Dark Knight} \\ \text{Spiderman 3} \\ \text{Love Actually} \\ \text{Bridget Jones's Diary} \\ \text{Pretty Woman} \\ \text{Superman 2} \end{array} \tag{109}$$

For ease of comparison the values of A are shown in brackets. The first left singular vector is equal to

$$\vec{u}_1 := \begin{pmatrix} \text{D. Knight} & \text{Spiderman 3} & \text{Love Act.} & \text{B.J.'s Diary} & \text{P. Woman} & \text{Superman 2} \\ -0.45 & -0.39 & 0.39 & 0.39 & 0.39 & -0.45 \end{pmatrix}.$$

This vector allows us to cluster the movies: movies with negative entries are similar (in this case they correspond to action movies) and movies with positive entries are similar (in this case they are romantic movies).

The first right singular vector is equal to

$$\vec{v}_1 = \begin{matrix} \text{Bob} & \text{Molly} & \text{Mary} & \text{Larry} \end{matrix} \\ \begin{pmatrix} 0.48 & 0.52 & -0.48 & -0.52 \end{pmatrix}. \tag{110}$$

This vector allows to cluster the users: negative entries indicate users that like action movies but hate romantic movies (Bob and Molly), whereas positive entries indicate the contrary (Mary and Larry).

For larger data sets, the model generalizes to a rank-k approximation, which approximates each ranking by a sum of k terms

$$\text{rating}\left(\text{movie } i, \text{user } j\right) = \sum_{l=1}^{k} \sigma_l \vec{u}_l\left[i\right] \vec{v}_l\left[j\right]. \tag{111}$$

The singular vectors cluster users and movies in different ways, whereas the singular values weight the importance of the different factors.

3 Principal component analysis

In Lecture Notes 1 we introduced the sample variance of a set of one-dimensional data, which measures the variation of measurements in a one-dimensional data set, as well as the sample covariance, which measures the joint fluctuations of two features. We now consider data sets where each example contains m features, and can therefore be interpreted as a vector in an m-dimensional ambient space. We are interested in analyzing the variation of the data in different directions of this space.

3.1 Sample covariance matrix

The sample covariance matrix of a data set contains the pairwise sample covariance between every pair of features in a data set. If the data are sampled from a multivariate distribution, then the sample covariance matrix can be interpreted as an estimate of the covariance matrix (see Section 3.3).

Definition 3.1 (Sample covariance matrix)**.** *Let $\{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\}$ be a set of m-dimensional real-valued data vectors, where each dimension corresponds to a different feature. The sample covariance matrix of these vectors is the $m \times m$ matrix*

$$\Sigma\left(\vec{x}_1, \ldots, \vec{x}_n\right) := \frac{1}{n-1} \sum_{i=1}^{n} \left(\vec{x}_i - \text{av}\left(\vec{x}_1, \ldots, \vec{x}_n\right)\right) \left(\vec{x}_i - \text{av}\left(\vec{x}_1, \ldots, \vec{x}_n\right)\right)^T, \tag{112}$$

where the center or average is defined as

$$\text{av}\left(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\right) := \frac{1}{n} \sum_{i=1}^{n} \vec{x}_i \tag{113}$$

contains the sample mean of each feature. The (i, j) entry of the covariance matrix, where $1 \leq i, j \leq d$, is given by

$$\Sigma\left(\vec{x}_1, \ldots, \vec{x}_n\right)_{ij} = \begin{cases} \text{var}\left(\vec{x}_1\left[i\right], \ldots, \vec{x}_n\left[i\right]\right) & \text{if } i = j, \\ \text{cov}\left(\left(\vec{x}_1\left[i\right], \vec{x}_1\left[j\right]\right), \ldots, \left(\vec{x}_n\left[i\right], \vec{x}_n\left[j\right]\right)\right) & \text{if } i \neq j. \end{cases} \tag{114}$$

In order to characterize the variation of a multidimensional data set around its center, we consider its variation in different directions. The average variation of the data in a certain direction is quantified by the sample variance of the projections of the data onto that direction. Let $\vec{d} \in \mathbb{R}^m$ be a unit-norm vector aligned with a direction of interest, the sample variance of the data set in the direction of $\vec{d}$ is given by

$$\text{var}\left(\vec{d}^T\vec{x}_1, \ldots, \vec{d}^T\vec{x}_n\right) = \frac{1}{n-1} \sum_{i=1}^{n} \left(\vec{d}^T\vec{x}_i - \text{av}\left(\vec{d}^T\vec{x}_1, \ldots, \vec{d}^T\vec{x}_n\right)\right)^2 \tag{115}$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} \left(\vec{d}^T\left(\vec{x}_i - \text{av}\left(\vec{x}_1, \ldots, \vec{x}_n\right)\right)\right)^2 \tag{116}$$

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

$$= \vec{d}^T \Sigma\left(\vec{x}_1, \ldots, \vec{x}_n\right) \vec{d}. \tag{117}$$

Using the sample covariance matrix we can express the variation in every direction! This is a deterministic analog of the fact that the covariance matrix of a random vector encodes its variance in every direction.

3.2 Principal component analysis

Principal-component analysis is a popular tool for data analysis, which consists of computing the singular-value decomposition of a set of vectors grouped as the columns of a matrix.

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

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

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

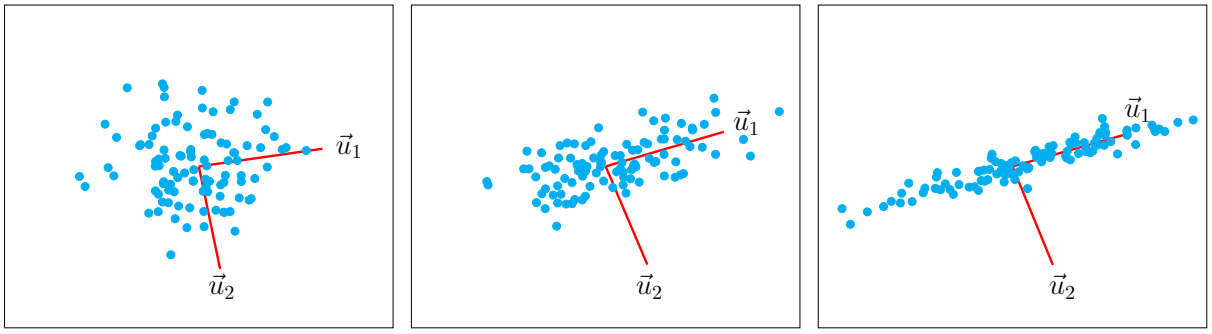


Figure 4: PCA of a dataset with $n = 100$ 2D vectors with different configurations. The two first singular values reflect how much energy is preserved by projecting onto the two first principal directions.

Algorithm 3.2 (Principal component analysis)**.** *Given n data vectors $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n \in \mathbb{R}^d$, we apply the following steps.*

1. *Center the data,*
$$\vec{c}_i = \vec{x}_i - \text{av}\left(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\right), \qquad 1 \leq i \leq n. \tag{118}$$
2. *Group the centered data as columns of a matrix*
$$C = \begin{bmatrix} \vec{c}_1 & \vec{c}_2 & \cdots & \vec{c}_n \end{bmatrix}. \tag{119}$$
3. *Compute the SVD of C. The left singular vectors are the principal directions. The principal values are the coefficients of the centered vectors when expressed in the basis of principal directions.*

The sample covariance matrix can be expressed in terms of the centered data matrix C

$$\Sigma\left(\vec{x}_1, \ldots, \vec{x}_n\right) = \frac{1}{n-1} C C^T. \tag{120}$$

This implies that by Theorem 2.7 the principal directions reveal the directions of maximum variation of the data.

Corollary 3.3. *Let $\vec{u}_1, \ldots, \vec{u}_k$ be the $k \leq \min\{m, n\}$ first principal directions obtained by applying Algorithm 3.2 to a set of vectors $\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^m$. Then the principal directions satisfy*

$$\vec{u}_1 = \underset{\left\{||\vec{d}||_2 = 1 \mid \vec{d} \in \mathbb{R}^n\right\}}{\arg\max} \text{var}\left(\vec{d}^T \vec{x}_1, \ldots, \vec{d}^T \vec{x}_n\right), \tag{121}$$

$$\vec{u}_i = \underset{\left\{||\vec{d}||_2 = 1 \mid \vec{d} \in \mathbb{R}^n, \vec{d} \perp \vec{u}_1, \ldots, \vec{u}_{i-1}\right\}}{\arg\max} \text{var}\left(\vec{d}^T \vec{x}_1, \ldots, \vec{d}^T \vec{x}_n\right), \qquad 2 \leq i \leq k, \tag{122}$$

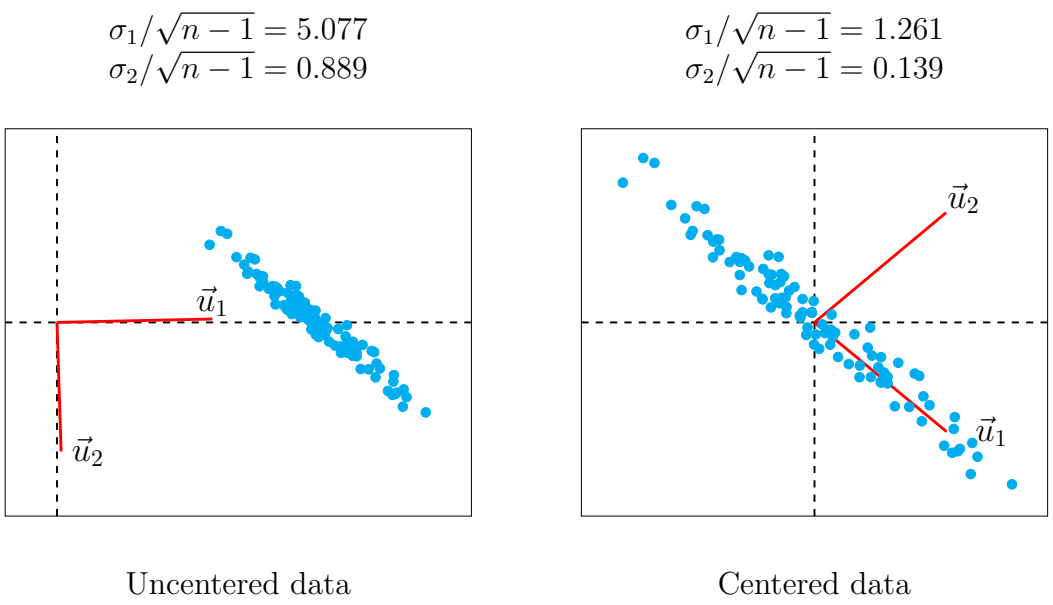


Figure 5: PCA applied to $n = 100$ 2D data points. On the left the data are not centered. As a result the dominant principal direction $\vec{u}_1$ lies in the direction of the mean of the data and PCA does not reflect the actual structure. Once we center, $\vec{u}_1$ becomes aligned with the direction of maximal variation.

and the associated singular values satisfy

$$\frac{\sigma_1}{\sqrt{n-1}} = \max_{\{||\vec{d}||_2=1 \,|\, \vec{d}\in\mathbb{R}^n\}} \text{std}\left(\vec{d}^T\vec{x}_1, \ldots, \vec{d}^T\vec{x}_n\right), \tag{123}$$

$$\frac{\sigma_i}{\sqrt{n-1}} = \max_{\{||\vec{d}||_2=1 \,|\, \vec{d}\in\mathbb{R}^n, \vec{d}\perp\vec{u}_1,\ldots,\vec{u}_{i-1}\}} \text{std}\left(\vec{d}^T\vec{x}_1, \ldots, \vec{d}^T\vec{x}_n\right), \qquad 2 \leq i \leq k. \tag{124}$$

Proof. For any vector $\vec{d}$

$$\text{var}\left(\vec{d}^T\vec{x}_1, \ldots, \vec{d}^T\vec{x}_n\right) = \vec{d}^T\Sigma\left(\vec{x}_1, \ldots, \vec{x}_n\right)\vec{d} \tag{125}$$

$$= \frac{1}{n-1}\vec{d}^T C C^T \vec{d} \tag{126}$$

$$= \frac{1}{n-1}\left|\left|C^T\vec{d}\right|\right|_2^2, \tag{127}$$

so the result follows from Theorem 2.7 applied to C. □

In words, $\vec{u}_1$ is the direction of maximum variation, $\vec{u}_2$ is the direction of maximum variation orthogonal to $\vec{u}_1$, and in general $\vec{u}_i$ is the direction of maximum variation that is orthogonal to $\vec{u}_1$, $\vec{u}_2$, ..., $\vec{u}_{i-1}$. Figure 4 shows the principal directions for several 2D examples.

Figure 5 illustrates the importance of centering, i.e., subtracting the sample mean of each feature, before applying PCA. Theorem 2.7 still holds if the data are not centered.

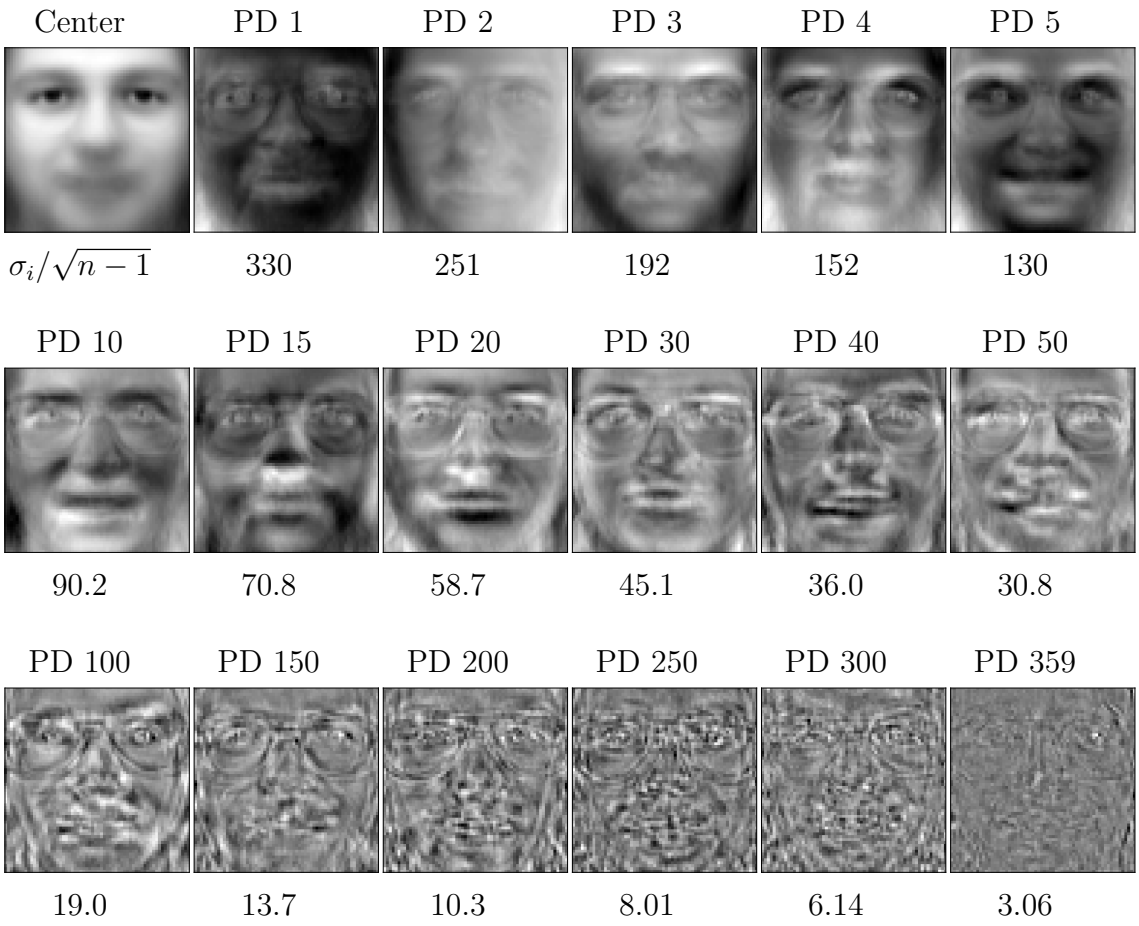


Figure 6: Average and principal directions (PD) of the faces data set in Example 3.4, along with their associated singular values.

However, the norm of the projection onto a certain direction no longer reflects the variation of the data. In fact, if the data are concentrated around a point that is far from the origin, the first principal direction tends be aligned with that point. This makes sense as projecting onto that direction captures more energy. As a result, the principal directions do not reflect the directions of maximum variation *within* the cloud of data. Centering the data set before applying PCA solves the issue.

In the following example, we apply PCA to a set of images.

Example 3.4 (PCA of faces)**.** In this example we consider the Olivetti Faces data set, which we described in Lecture Notes 1. We apply Algorithm 3.2 to a data set of 400 64×64 images taken from 40 different subjects (10 per subject). We vectorize each image so that each pixel is interpreted as a different feature. Figure 6 shows the center of the data and several principal directions, together with their associated singular values. The principal directions corresponding to the larger singular values seem to capture low-resolution structure, whereas the ones corresponding to the smallest singular values incorporate more intricate details.

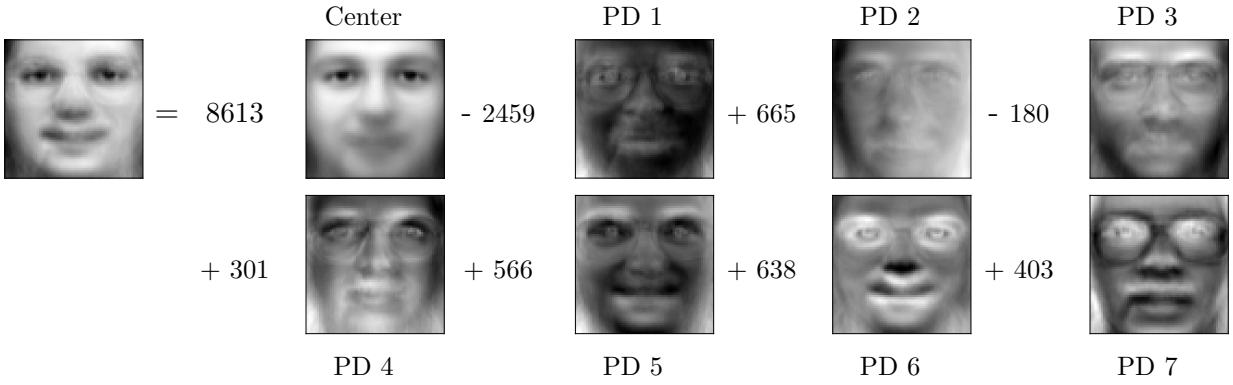


Figure 7: Projection of one of the faces $\vec{x}$ onto the first 7 principal directions and the corresponding decomposition into the 7 first principal components.

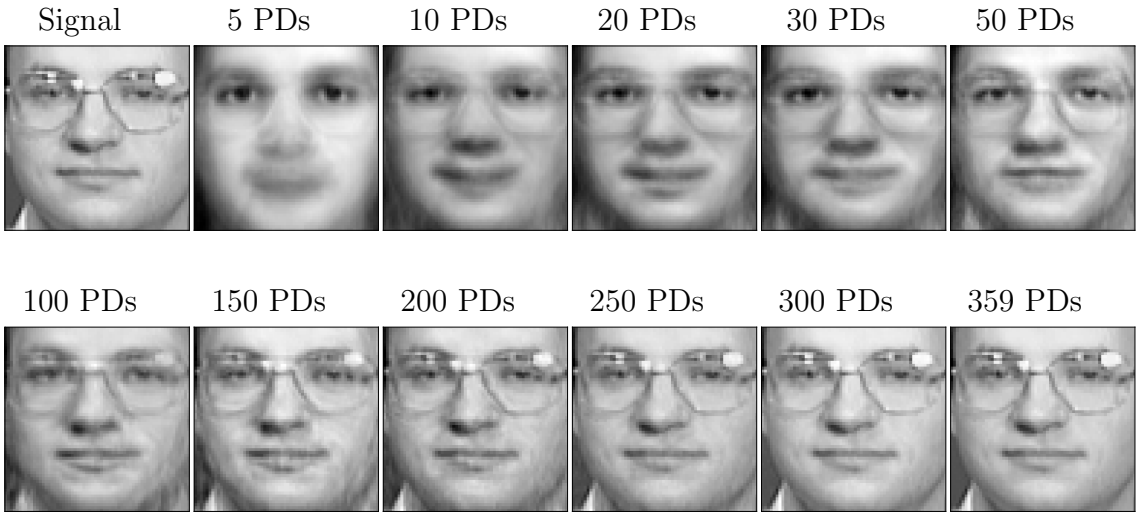


Figure 8: Projection of a face on different numbers of principal directions.

Figure 7 shows the projection of one of the faces onto the first 7 principal directions and the corresponding decomposition into its 7 first principal components. Figure 8 shows the projection of the same face onto increasing numbers of principal directions. As suggested by the visualization of the principal directions in Figure 6, the lower-dimensional projections produce blurry images.

△

3.3 Probabilistic interpretation

To provide a probabilistic interpretation of PCA, we first review some background on covariance matrices. The covariance matrix of a random vector captures the interaction between the components of the vector. It contains the variance of each component in the

diagonal and the covariances between different components in the off diagonals.

Definition 3.5. *The covariance matrix of a random vector* $\vec{\mathbf{x}}$ *is defined as*

$$\Sigma_{\vec{\mathbf{x}}} := \begin{bmatrix} \operatorname{Var}\left(\vec{\mathbf{x}}\left[1\right]\right) & \operatorname{Cov}\left(\vec{\mathbf{x}}\left[1\right], \vec{\mathbf{x}}\left[2\right]\right) & \cdots & \operatorname{Cov}\left(\vec{\mathbf{x}}\left[1\right], \vec{\mathbf{x}}\left[n\right]\right) \\ \operatorname{Cov}\left(\vec{\mathbf{x}}\left[2\right], \vec{\mathbf{x}}\left[1\right]\right) & \operatorname{Var}\left(\vec{\mathbf{x}}\left[2\right]\right) & \cdots & \operatorname{Cov}\left(\vec{\mathbf{x}}\left[2\right], \vec{\mathbf{x}}\left[n\right]\right) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{Cov}\left(\vec{\mathbf{x}}\left[n\right], \vec{\mathbf{x}}\left[1\right]\right) & \operatorname{Cov}\left(\vec{\mathbf{x}}\left[n\right], \vec{\mathbf{x}}\left[2\right]\right) & \cdots & \operatorname{Var}\left(\vec{\mathbf{x}}\left[n\right]\right) \end{bmatrix} \tag{128}$$

$$= \mathrm{E}\left(\vec{\mathbf{x}}\vec{\mathbf{x}}^T\right) - \mathrm{E}(\vec{\mathbf{x}})\mathrm{E}(\vec{\mathbf{x}})^T. \tag{129}$$

Note that if all the entries of a vector are uncorrelated, then its covariance matrix is diagonal. Using linearity of expectation, we obtain a simple expression for the covariance matrix of the linear transformation of a random vector.

Theorem 3.6 (Covariance matrix after a linear transformation)**.** *Let* $\vec{\mathbf{x}}$ *be a random vector of dimension* n *with covariance matrix* Σ*. For any matrix* $A \in \mathbb{R}^{m \times n}$ *,*

$$\Sigma_{A\vec{\mathbf{x}}} = A\Sigma_{\vec{\mathbf{x}}}A^T. \tag{130}$$

Proof. By linearity of expectation

$$\Sigma_{A\vec{\mathbf{x}}} = \mathrm{E}\left(\left(A\vec{\mathbf{x}}\right)\left(A\vec{\mathbf{x}}\right)^T\right) - \mathrm{E}\left(A\vec{\mathbf{x}}\right)\mathrm{E}\left(A\vec{\mathbf{x}}\right)^T \tag{131}$$

$$= A\left(\mathrm{E}\left(\vec{\mathbf{x}}\vec{\mathbf{x}}^T\right) - \mathrm{E}(\vec{\mathbf{x}})\mathrm{E}(\vec{\mathbf{x}})^T\right)A^T \tag{132}$$

$$= A\Sigma_{\vec{\mathbf{x}}}A^T. \tag{133}$$

□

An immediate corollary of this result is that we can easily decode the variance of the random vector *in any direction* from the covariance matrix. Mathematically, the variance of the random vector in the direction of a unit vector $\vec{v}$ is equal to the variance of its projection onto $\vec{v}$.

Corollary 3.7. *Let* $\vec{v}$ *be a unit-*ℓ_2*-norm vector,*

$$\operatorname{Var}\left(\vec{v}^T\vec{\mathbf{x}}\right) = \vec{v}^T\Sigma_{\vec{\mathbf{x}}}\vec{v}. \tag{134}$$

Consider the SVD of the covariance matrix of an n-dimensional random vector X

$$\Sigma_{\vec{\mathbf{x}}} = U\Lambda U^T \tag{135}$$

$$= \begin{bmatrix} \vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_n \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \sigma_n \end{bmatrix} \begin{bmatrix} \vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_n \end{bmatrix}^T. \tag{136}$$

Covariance matrices are symmetric by definition, so by Theorem 4.3 the eigenvectors $\vec{u}_1$, $\vec{u}_2$, ..., $\vec{u}_n$ can be chosen to be orthogonal. These singular vectors and singular values completely characterize the variance of the random vector in different directions. The theorem is a direct consequence of Corollary 3.7 and Theorem 2.7.

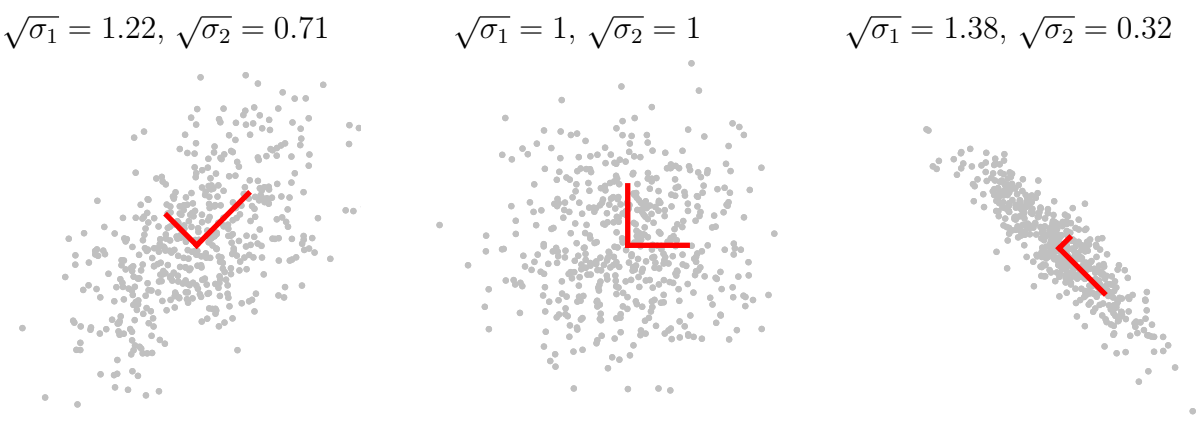


Figure 9: Samples from bivariate Gaussian random vectors with different covariance matrices are shown in gray. The eigenvectors of the covariance matrices are plotted in red. Each is scaled by the square roof of the corresponding singular value σ_1 or σ_2.

Theorem 3.8. *Let $\vec{\mathbf{x}}$ be a random vector of dimension n with covariance matrix $\Sigma_{\vec{\mathbf{x}}}$. The SVD of $\Sigma_{\vec{\mathbf{x}}}$ given by (136) satisfies*

$$\sigma_1 = \max_{||\vec{v}||_2=1} \text{Var}\left(\vec{v}^T\vec{\mathbf{x}}\right), \tag{137}$$

$$\vec{u}_1 = \arg\max_{||\vec{v}||_2=1} \text{Var}\left(\vec{v}^T\vec{\mathbf{x}}\right), \tag{138}$$

$$\sigma_k = \max_{||\vec{v}||_2=1, \vec{v}\perp\vec{u}_1,\ldots,\vec{u}_{k-1}} \text{Var}\left(\vec{v}^T\vec{\mathbf{x}}\right), \tag{139}$$

$$\vec{u}_k = \arg\max_{||\vec{v}||_2=1, \vec{v}\perp\vec{u}_1,\ldots,\vec{u}_{k-1}} \text{Var}\left(\vec{v}^T\vec{\mathbf{x}}\right). \tag{140}$$

In words, $\vec{u}_1$ is the *direction of maximum variance*. The second singular vector $\vec{u}_2$ is the direction of maximum variation that is orthogonal to $\vec{u}_1$. In general, the eigenvector $\vec{u}_k$ reveals the direction of maximum variation that is orthogonal to $\vec{u}_1$, $\vec{u}_2$, ..., $\vec{u}_{k-1}$. Finally, $\vec{u_n}$ is the direction of minimum variance. Figure 9 illustrates this with an example, where $n = 2$.

The sample variance and covariance are consistent estimators of the variance and covariance respectively, under certain assumptions on the higher moments of the underlying distributions. This provides an intuitive interpretation for principal component analysis under the assumption that the data are realizations of an iid sequence of random vectors: the principal components approximate the eigenvectors of the true covariance matrix, and hence the directions of maximum variance of the multidimensional distribution. Figure 10 illustrates this with a numerical example, where the principal directions indeed converge to the singular vectors as the number of data increases.

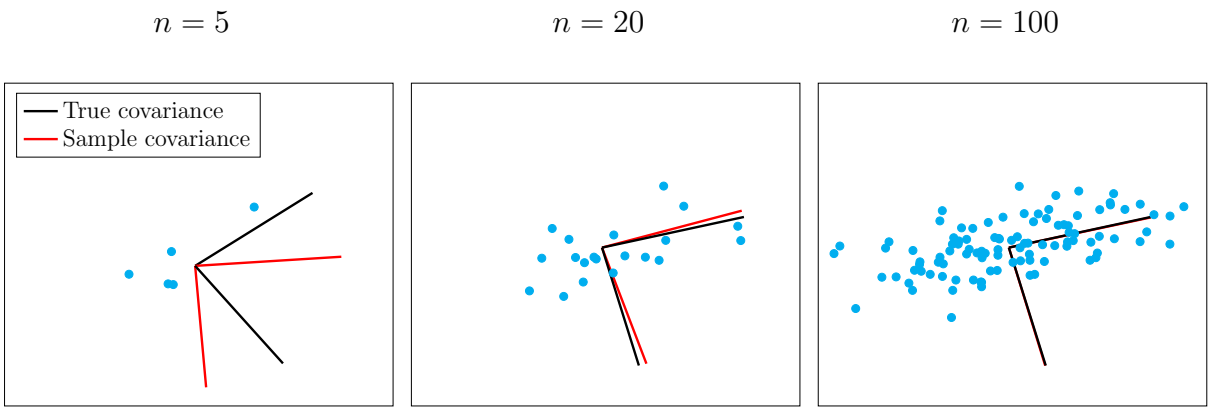


Figure 10: Principal directions of n samples from a bivariate Gaussian distribution (red) compared to the eigenvectors of the covariance matrix of the distribution (black).

3.4 Dimensionality reduction via PCA

Data containing a large number of features can be difficult to analyze or process. Dimensionality reduction is a useful preprocessing step for many data-analysis tasks, which consists of representing the data with a smaller number of variables. For data modeled as vectors in an ambient space $\mathbb{R}^m$ where each dimension corresponds to a feature, this can be achieved by projecting the vectors onto a lower-dimensional space $\mathbb{R}^k$, where $k < m$. If the projection is orthogonal, the new representation can be computed using an orthogonal basis for the lower-dimensional subspace $\vec{b}_1, \vec{b}_2, \ldots, \vec{b}_k$: each data vector $\vec{x} \in \mathbb{R}^m$ is described using the coefficients of its representation in the basis: $\langle \vec{b}_1, \vec{x} \rangle$, $\langle \vec{b}_2, \vec{x} \rangle$, $\ldots$, $\langle \vec{b}_k, \vec{x} \rangle$.

Given a data set of n vectors $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n \in \mathbb{R}^m$, the first k principal directions span the subspace that preserves the most energy (measured in ℓ_2 norm) in the centered data among all possible k-dimensional orthogonal projections by Theorem 2.8. This motivates the application of PCA for dimensionality reduction.

Example 3.9 (Nearest neighbors in principal-component space)**.** The nearest neighbors algorithm for classification (Algorithm 4.2 in Lecture Notes 1) requires computing n distances in an m-dimensional space (where m is the number of features) to classify each new example. The computational cost is $\mathcal{O}(nm)$, so if we need to classify p points the total cost is $\mathcal{O}(nmp)$. If we project each of the points onto a lower-dimensional space k computed via PCA before classifying them, then the computational cost is:

- $\mathcal{O}(mn \min\{m, n\})$ to compute the principal directions from the training data.
- kmn operations to project the training data onto the first k principal directions.
- kmp operations to project each point in the test set onto the first k principal directions.
- knp to perform nearest-neighbor classification in the lower-dimensional space.

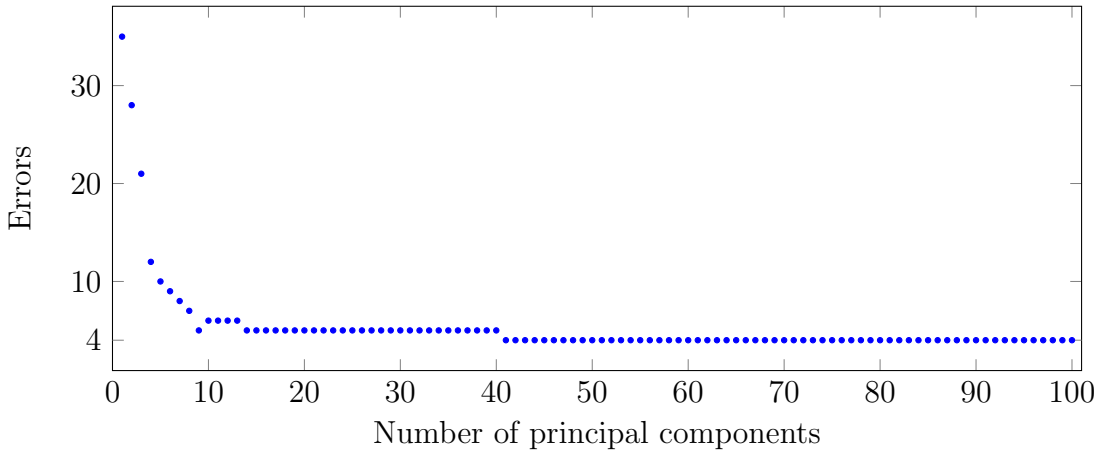


Figure 11: Errors for nearest-neighbor classification combined with PCA-based dimensionality reduction for different dimensions.

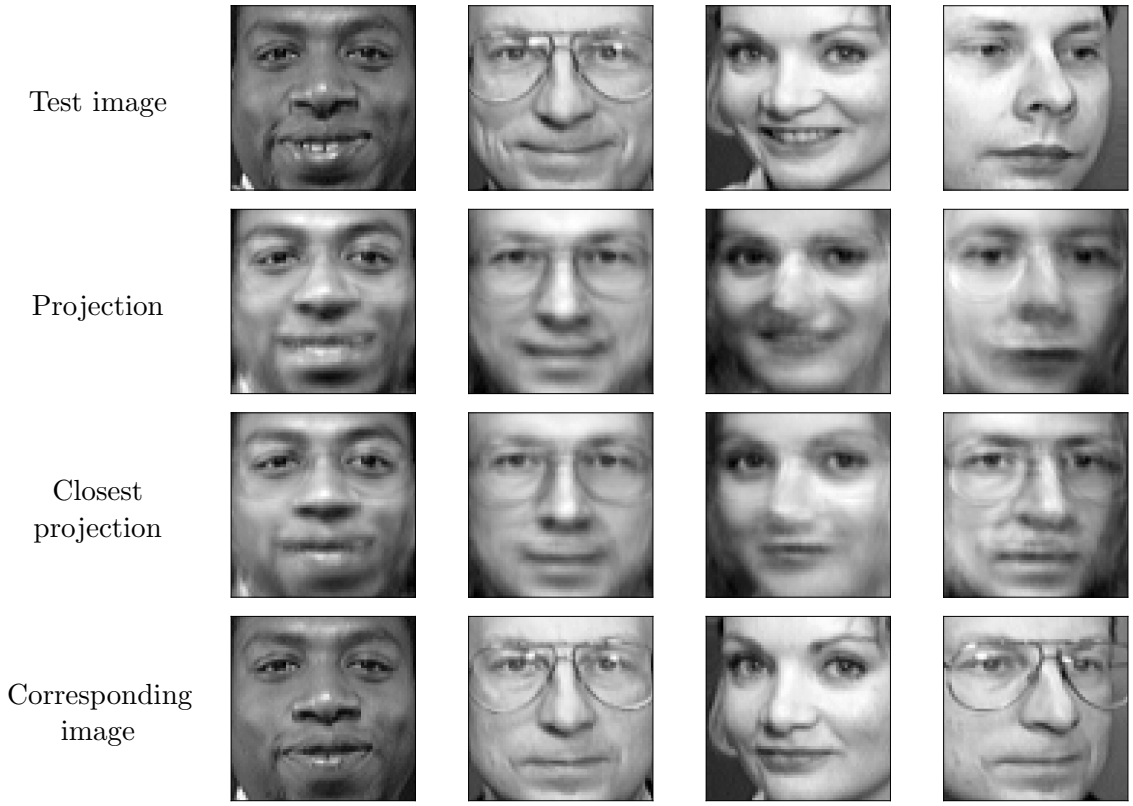


Figure 12: Results of nearest-neighbor classification combined with PCA-based dimensionality reduction of order 41 for four of the people in Example 3.9. The assignments of the first three examples are correct, but the fourth is wrong.

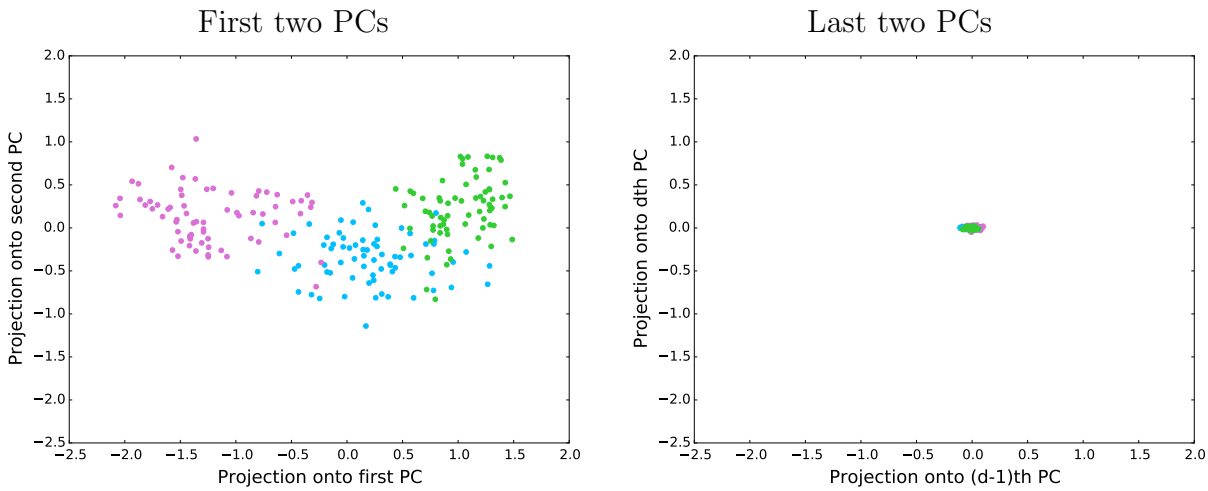


Figure 13: Projection of 7-dimensional vectors describing different wheat seeds onto the first two (left) and the last two (right) principal dimensions of the data set. Each color represents a variety of wheat.

If we have to classify a large number of points (i.e. $p \gg \max\{m, n\}$) the computational cost is reduced by operating in the lower-dimensional space.

Figure 11 shows the accuracy of the algorithm on the same data as Example 4.3 in Lecture Notes 1. The accuracy increases with the dimension at which the algorithm operates. This is not necessarily always the case because projections may actually be helpful for tasks such as classification (for example, factoring out small shifts and deformations). The same precision as in the ambient dimension (4 errors out of 40 test images) is achieved using just $k = 41$ principal components (in this example $n = 360$ and $m = 4096$). Figure 12 shows some examples of the projected data represented in the original m-dimensional space along with their nearest neighbors in the k-dimensional space. △

Example 3.10 (Dimensionality reduction for visualization)**.** Dimensionality reduction is often useful for visualization. The objective is to project the data onto 2D or 3D in a way that preserves its structure as much as possible. In this example, we consider a data set where each data point corresponds to a seed with seven features: area, perimeter, compactness, length of kernel, width of kernel, asymmetry coefficient and length of kernel groove. The seeds belong to three different varieties of wheat: Kama, Rosa and Canadian.[2] To visualize the data in 2D, we project each point onto the two first principal dimensions of the data set.

Figure 13 shows the projection of the data onto the first two and the last two principal directions. In the latter case, there is almost no discernible variation. As predicted by our theoretical analysis of PCA, the structure in the data is much better conserved by the two first directions, which allow to clearly visualize the difference between the three types of seeds. Note however that projection onto the first principal directions only ensures

[2]The data can be found at https://archive.ics.uci.edu/ml/datasets/seeds.

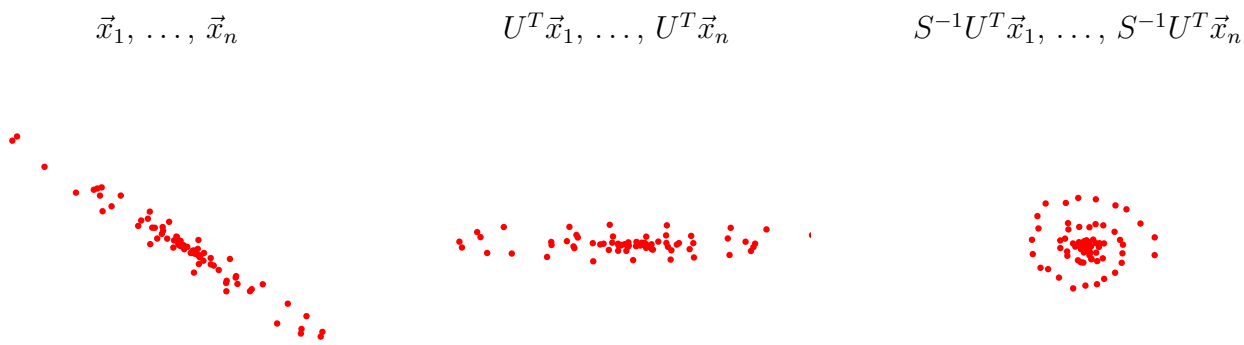


Figure 14: Effect of whitening a set of data. The original data are dominated by a linear skew (left). Applying U^T aligns the axes with the eigenvectors of the sample covariance matrix (center). Finally, S^{-1} reweights the data along those axes so that they have the same average variation, revealing the nonlinear structure that was obscured by the linear skew (right).

that we preserve as much variation as possible, but it does not necessarily preserve useful features for tasks such as clustering or classification. △

3.5 Whitening

The principal directions in a data set do not necessarily capture the most useful features for certain tasks. For instance, in the case of the faces data set in Example 3.4 the principal directions correspond to low-resolution images, so that the corresponding principal components capture low-resolution features. These features do not include important information contained in fine-scale details, which could be useful for tasks such as classification. Whitening is a preprocessing technique that reweights the principal components of every vector so every principal dimension has the same contribution.

Algorithm 3.11 (Whitening)**.** *Given n data vectors $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n \in \mathbb{R}^d$, we apply the following steps.*

1. *Center the data,*

$$\vec{c}_i = \vec{x}_i - \text{av}\left(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\right), \qquad 1 \leq i \leq n. \tag{141}$$

2. *Group the centered data as columns of a matrix*

$$C = \begin{bmatrix} \vec{c}_1 & \vec{c}_2 & \cdots & \vec{c}_n \end{bmatrix}. \tag{142}$$

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

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

$$\vec{w}_i := US^{-1}U^T\vec{c}_i. \tag{143}$$

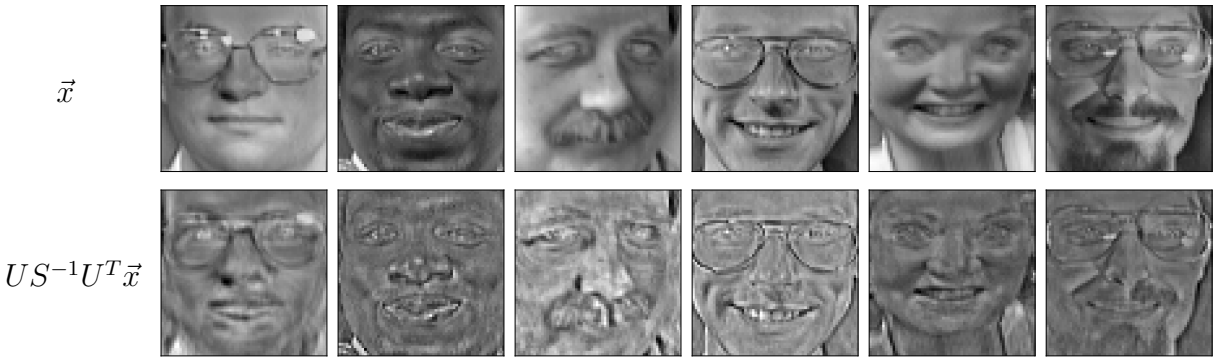


Figure 15: Centered faces in the data set from Example 3.4 before and after whitening.

The linear map $US^{-1}U^T$ scales the components of the centered vector in each principal direction by a factor inversely proportional to its corresponding singular value,

$$\vec{w}_i := \sum_{j=1}^{\min(m,n)} \frac{1}{\sigma_j} \langle \vec{u}_j, \vec{c}_i \rangle \, \vec{u}_j. \tag{144}$$

If we group the whitened vectors as columns of a matrix, the matrix can be expressed as

$$W = US^{-1}U^T C. \tag{145}$$

The covariance matrix of the whitened data is proportional to the identity

$$\Sigma\left(\vec{c}_1, \ldots, \vec{c}_n\right) = \frac{1}{n-1} WW^T \tag{146}$$

$$= \frac{1}{n-1} US^{-1}U^T CC^T US^{-1}U^T \tag{147}$$

$$= \frac{1}{n-1} US^{-1}U^T USV^T VSU^T US^{-1}U^T \tag{148}$$

$$= \frac{1}{n-1} I, \tag{149}$$

This means that the whitened data have no linear skews, there are no directions in space that contain more variation than others. As illustrated in Figure 14, this may reveal nonlinear structure in the data. Figure 15 shows some of the centered faces in the data set from Example 3.4 before and after whitening. Whitening enhances fine-detail features of the faces.

4 Eigendecomposition

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

$$A\vec{q} = \lambda \vec{q} \tag{150}$$

for a scalar λ which is the corresponding eigenvalue. Even if A is real, its eigenvectors and eigenvalues can be complex. If a matrix has n linearly independent eigenvectors then it is diagonalizable.

Lemma 4.1 (Eigendecomposition)**.** *If a square matrix $A \in \mathbb{R}^{n \times n}$ has n linearly independent eigenvectors $\vec{q}_1, \ldots, \vec{q}_n$ with eigenvalues $\lambda_1, \ldots, \lambda_n$ it can be expressed in terms of a matrix Q, whose columns are the eigenvectors, and a diagonal matrix containing the eigenvalues,*

$$A = \begin{bmatrix} \vec{q}_1 & \vec{q}_2 & \cdots & \vec{q}_n \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} \vec{q}_1 & \vec{q}_2 & \cdots & \vec{q}_n \end{bmatrix}^{-1} \tag{151}$$

$$= Q \Lambda Q^{-1} \tag{152}$$

Proof.

$$AQ = \begin{bmatrix} A\vec{q}_1 & A\vec{q}_2 & \cdots & A\vec{q}_n \end{bmatrix} \tag{153}$$

$$= \begin{bmatrix} \lambda_1 \vec{q}_1 & \lambda_2 \vec{q}_2 & \cdots & \lambda_2 \vec{q}_n \end{bmatrix} \tag{154}$$

$$= Q\Lambda. \tag{155}$$

If the columns of a square matrix are all linearly independent, then the matrix has an inverse, so multiplying the expression by Q^{-1} on both sides completes the proof. □

Lemma 4.2. *Not all matrices have an eigendecomposition.*

Proof. Consider the matrix

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}. \tag{156}$$

Assume an eigenvector $\vec{q}$ associated to an eigenvalue λ, then

$$\begin{bmatrix} \vec{q}[2] \\ 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \vec{q}[1] \\ \vec{q}[2] \end{bmatrix} = \begin{bmatrix} \lambda \vec{q}[1] \\ \lambda \vec{q}[2] \end{bmatrix}, \tag{157}$$

which implies that $\vec{q}[2] = 0$ and $\vec{q}[1] = 0$, so the matrix does not have eigenvectors associated to nonzero eigenvalues. □

Symmetric matrices are always diagonalizable.

Theorem 4.3 (Spectral theorem for symmetric matrices)**.** *If $A \in \mathbb{R}^n$ is symmetric, then it has an eigendecomposition of the form*

$$A = U \Lambda U^T \tag{158}$$

where the matrix of eigenvectors U is an orthogonal matrix.

This is a fundamental result in linear algebra that can be used to prove Theorem 2.1. We refer to any graduate-level linear-algebra text for the proof.

Together, Theorems 2.1 and 4.3 imply that the SVD $A = USV^T$ and the eigendecomposition $A = U\Lambda U^T$ of a symmetric matrix are almost the same. The left singular vectors can be taken to be equal to the eigenvectors. Nonnegative eigenvalues are equal to the singular vectors, and their right singular vectors are equal to the corresponding eigenvectors. The difference is that if an eigenvalue λ_i corresponding to an eigenvector $\vec{u}_i$ is negative, then $\sigma_i = -\lambda_i$ and the corresponding right-singular vector $\vec{v}_i = -\vec{u}_i$.

A useful application of the eigendecomposition is computing successive matrix products. Assume that we are interested in computing

$$AA \cdots A\vec{x} = A^k\vec{x}, \tag{159}$$

i.e., we want to apply A to $\vec{x}$ k times. A^k *cannot* be computed by taking the power of its entries (try out a simple example to convince yourself). However, if A has an eigendecomposition,

$$\begin{aligned} A^k &= Q\Lambda Q^{-1}Q\Lambda Q^{-1}\cdots Q\Lambda Q^{-1} && (160)\\ &= Q\Lambda^k Q^{-1} && (161)\\ &= Q \begin{bmatrix} \lambda_1^k & 0 & \cdots & 0 \\ 0 & \lambda_2^k & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \lambda_n^k \end{bmatrix} Q^{-1}, && (162) \end{aligned}$$

using the fact that for diagonal matrices applying the matrix repeatedly is equivalent to taking the power of the diagonal entries. This allows to compute the k matrix products using just 3 matrix products and taking the power of n numbers.

Let $A \in \mathbb{R}^{n\times n}$ be a matrix with eigendecomposition $Q\Lambda Q^{-1}$ and let $\vec{x}$ be an arbitrary vector in $\mathbb{R}^n$. Since the eigenvectors are linearly independent, they form a basis for $\mathbb{R}^n$, so we can represent $\vec{x}$ as

$$\vec{x} = \sum_{i=1}^{n} \alpha_i \vec{q}_i, \quad \alpha_i \in \mathbb{R},\ 1 \leq i \leq n. \tag{163}$$

Now let us apply A to $\vec{x}$ k times,

$$\begin{aligned} A^k\vec{x} &= \sum_{i=1}^{n} \alpha_i A^k \vec{q}_i && (164)\\ &= \sum_{i=1}^{n} \alpha_i \lambda_i^k \vec{q}_i. && (165) \end{aligned}$$

If we assume that the eigenvectors are ordered according to their magnitudes and that the magnitude of one of them is larger than the rest, $|\lambda_1| > |\lambda_2| \geq \ldots$, and that $\alpha_1 \neq 0$

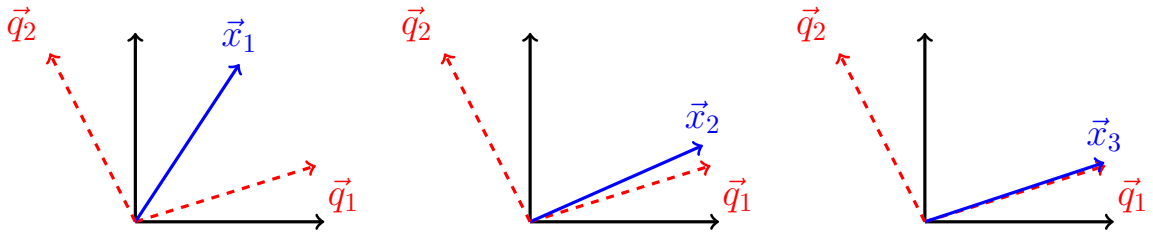


Figure 16: Illustration of the first three iterations of the power method for a matrix with eigenvectors $\vec{q}_1$ and $\vec{q}_2$, with corresponding eigenvalues $\lambda_1 = 1.05$ and $\lambda_2 = 0.1661$.

(which happens with high probability if we draw a random $\vec{x}$) then as k grows larger the term $\alpha_1 \lambda_1^k \vec{q}_1$ dominates. The term will blow up or tend to zero unless we normalize every time before applying A. Adding the normalization step to this procedure results in the power method or power iteration, an algorithm for estimating the eigenvector of a matrix that corresponds to the largest eigenvalue.

Algorithm 4.4 (Power method)**.**
Set $\vec{x}_1 := \vec{x} / ||\vec{x}||_2$*, where the entries of* $\vec{x}$ *are drawn at random. For* $i = 1, 2, 3, \ldots$*, compute*

$$\vec{x}_i := \frac{A\vec{x}_{i-1}}{||A\vec{x}_{i-1}||_2}. \tag{166}$$

Figure 16 illustrates the power method on a simple example, where the matrix is equal to

$$A = \begin{bmatrix} 0.930 & 0.388 \\ 0.237 & 0.286 \end{bmatrix}. \tag{167}$$

The convergence to the eigenvector corresponding to the eigenvalue with the largest magnitude is very fast.

We end this section with an example that applies a decomposition to analyze the evolution of the populations of two animals.

Example 4.5 (Deer and wolfs)**.** A biologist is studying the populations of deer and wolfs in Yellowstone. She concludes that a reasonable model for the populations in year $n+1$ is the linear system of equations

$$d_{n+1} = \frac{5}{4} d_n - \frac{3}{4} w_n, \tag{168}$$

$$w_{n+1} = \frac{1}{4} d_n + \frac{1}{4} w_n, \qquad n = 0, 1, 2, \ldots \tag{169}$$

where d_n and w_n denote the number of deer and wolfs in year n. She is interested in determining the evolution of the populations in the future so she computes an eigende-

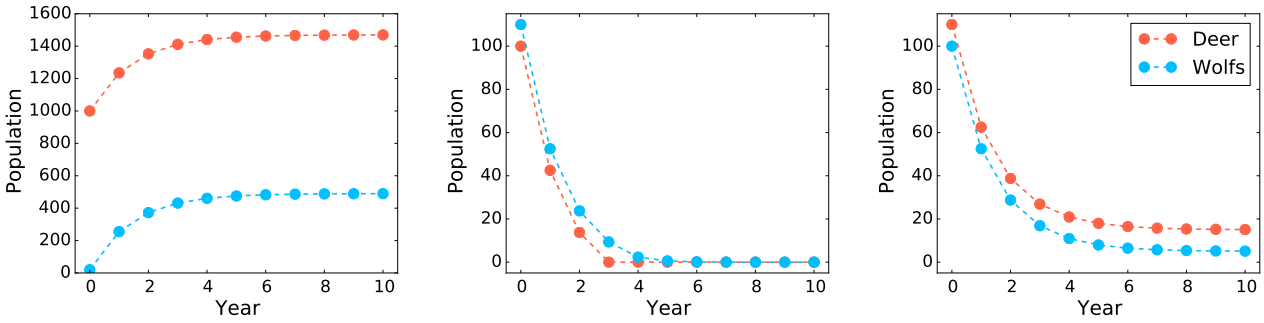


Figure 17: Evolution of the populations of deer and wolfs for different initial populations in Example 4.5.

composition of the matrix

$$A := \begin{bmatrix} 5/4 & -3/4 \\ 1/4 & 1/4 \end{bmatrix} \tag{170}$$

$$= \begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0.5 \end{bmatrix} \begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix}^{-1} := Q\Lambda Q^{-1}. \tag{171}$$

If we denote the initial populations of deer and wolfs as d_0 and w_0 respectively, the populations in year n are given by

$$\begin{bmatrix} d_n \\ w_n \end{bmatrix} = Q\Lambda^n Q^{-1} \begin{bmatrix} d_0 \\ w_0 \end{bmatrix} \tag{172}$$

$$= \begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0.5^n \end{bmatrix} \begin{bmatrix} 0.5 & -0.5 \\ -0.5 & 1.5 \end{bmatrix} \begin{bmatrix} d_0 \\ w_0 \end{bmatrix} \tag{173}$$

$$= \frac{d_0 - w_0}{2} \begin{bmatrix} 3 \\ 1 \end{bmatrix} + \frac{3w_0 - d_0}{8^n} \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \tag{174}$$

As $n \to \infty$, if the number of deer is larger than the number of wolfs, then the population of deer will converge to be three times the population of wolfs, which will converge to a half of the difference between their original populations. Since the populations cannot be negative, if the original population of wolfs is larger than that of deer, then both species will go extinct. This is confirmed by the simulations shown in Figure 17. △

5 Proofs

5.1 Proof of Theorem 1.3

It is sufficient to prove

$$\dim(\text{row}(A)) \leq \dim(\text{col}(A)) \tag{175}$$

for any arbitrary matrix A. Since the row space of A is equal to the column space of A^T and vice versa, applying (175) to A^T yields $\dim(\text{row}(A)) \geq \dim(\text{col}(A))$ which completes the proof.

To prove (175) let $r := \dim(\text{row}(A))$ and let $\vec{x}_1, \ldots, \vec{x}_r$ be a basis for $\text{row}(A)$. Consider the vectors $A\vec{x}_1, \ldots, A\vec{x}_r$. They belong to $\text{col}(A)$ by (11), so if they are linearly independent then $\dim(\text{col}(A)) \geq r$. We prove that this is the case by contradiction.

Assume that $A\vec{x}_1, \ldots, A\vec{x}_r$ are linearly dependent. Then there exist scalar coefficients $\alpha_1, \ldots, \alpha_r$ such that

$$\vec{0} = \sum_{i=1}^{r} \alpha_i A\vec{x}_i = A\left(\sum_{i=1}^{r} \alpha_i \vec{x}_i\right) \quad \text{by linearity of the matrix product,} \tag{176}$$

This implies that $\sum_{i=1}^{r} \alpha_i \vec{x}_i$ is orthogonal to every row of A and hence to every vector in $\text{row}(A)$. However it is in the span of a basis of $\text{row}(A)$ by construction! This is only possible if $\sum_{i=1}^{r} \alpha_i \vec{x}_i = 0$, which is a contradiction because $\vec{x}_1, \ldots, \vec{x}_r$ are assumed to be linearly independent.

5.2 Proof of Lemma 2.9

Let $\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_{d_1}$ be a basis for the first subspace and $\vec{b}_1, \vec{b}_2, \ldots, \vec{b}_{d_2}$ a basis for the second. Because the dimension of the vector space is n, the set of vectors $\vec{a}_1, \ldots, \vec{a}_{d_1}, \vec{b}_1, \ldots, \vec{b}_{d_2}$ are not linearly independent. There must exist scalars $\alpha_1, \ldots, \alpha_{d_1}, \beta_1, \ldots, \beta_{d_2}$, which are not all equal to zero, such that

$$\sum_{i=1}^{d_1} \alpha_i \vec{a}_i + \sum_{j=1}^{d_2} \beta_i \vec{b}_i = 0. \tag{177}$$

The vector

$$\vec{x} := \sum_{i=1}^{d_1} \alpha_i \vec{a}_i = -\sum_{j=1}^{d_2} \beta_i \vec{b}_i \tag{178}$$

cannot equal zero because both $\vec{a}_1, \ldots, \vec{a}_{d_1}$ and $\vec{b}_1, \ldots, \vec{b}_{d_2}$ are bases by assumption. $\vec{x}$ belongs to the intersection of the two subspaces, which completes the proof.

5.3 Proof of Theorem 2.16

The proof relies on the following lemma.

Lemma 5.1. *For any $Q \in \mathbb{R}^{n \times n}$*

$$\max_{1 \leq i \leq n} |Q_{ii}| \leq ||Q||. \tag{179}$$

Proof. Since $||\vec{e}_i||_2 = 1$,

$$\max_{1 \leq i \leq n} |Q_{ii}| \leq \max_{1 \leq i \leq n} \sqrt{\sum_{j=1}^{n} Q_{ji}^2} \tag{180}$$

$$= \max_{1 \leq i \leq n} ||Q\, \vec{e}_i||_2 \tag{181}$$

$$\leq ||Q|| \,. \tag{182}$$

□

We denote the SVD of A by USV^T,

$$\sup_{\{||B|| \leq 1 \mid B \in \mathbb{R}^{m \times n}\}} \text{tr}\left(A^T B\right) = \sup_{\{||B|| \leq 1 \mid B \in \mathbb{R}^{m \times n}\}} \text{tr}\left(V\, S\, U^T B\right) \tag{183}$$

$$= \sup_{\{||B|| \leq 1 \mid B \in \mathbb{R}^{m \times n}\}} \text{tr}\left(S\, B U^T V\right) \qquad \text{by Lemma 2.5 in Lecture Notes 1}$$

$$\leq \sup_{\{||M|| \leq 1 \mid M \in \mathbb{R}^{m \times n}\}} \text{tr}\left(S\, M\right) \qquad ||B|| = \left|\left|B U^T V\right|\right| \text{ by Corollary 2.15}$$

$$\leq \sup_{\left\{\max_{1 \leq i \leq n} |M_{ii}| \leq 1 \mid M \in \mathbb{R}^{m \times n}\right\}} \text{tr}\left(S\, M\right) \qquad \text{by Lemma 5.1}$$

$$\leq \sup_{\left\{\max_{1 \leq i \leq n} |M_{ii}| \leq 1 \mid M \in \mathbb{R}^{m \times n}\right\}} \sum_{i=1}^{n} M_{ii}\, \sigma_i \tag{184}$$

$$\leq \sum_{i=1}^{n} \sigma_i \tag{185}$$

$$= ||A||_* \,. \tag{186}$$

To complete the proof, we need to show that the equality holds. Note that UV^T has operator norm equal to one because its r singular values (recall that r is the rank of A) are equal to one. We have

$$\left\langle A, UV^T \right\rangle = \text{tr}\left(A^T U V^T\right) \tag{187}$$

$$= \text{tr}\left(V\, S\, U^T U V^T\right) \tag{188}$$

$$= \text{tr}\left(V^T V\, S\right) \qquad \text{by Lemma 2.5 in Lecture Notes 1} \tag{189}$$

$$= \text{tr}\left(S\right) \tag{190}$$

$$= ||A||_* \,. \tag{191}$$