Lecture Notes 1: Vector spaces

In this chapter we review certain basic concepts of linear algebra, highlighting their application to signal processing.

1 Vector spaces

Embedding signals in a vector space essentially means that we can add them up or scale them to produce new signals.

Definition 1.1 (Vector space)**.** *A vector space consists of a set $\mathcal{V}$, a scalar field that is usually either the real or the complex numbers and two operations $+$ and $\cdot$ satisfying the following conditions.*

1. *For any pair of elements $\vec{x}, \vec{y} \in \mathcal{V}$ the vector sum $\vec{x} + \vec{y}$ belongs to $\mathcal{V}$.*
2. *For any $\vec{x} \in \mathcal{V}$ and any scalar α, $\alpha \cdot \vec{x} \in \mathcal{V}$.*
3. *There exists a zero vector $\vec{0}$ such that $\vec{x} + \vec{0} = \vec{x}$ for any $\vec{x} \in \mathcal{V}$.*
4. *For any $\vec{x} \in \mathcal{V}$ there exists an additive inverse $\vec{y}$ such that $\vec{x} + \vec{y} = \vec{0}$, usually denoted by $-\vec{x}$.*
5. *The vector sum is commutative and associative, i.e. for all $\vec{x}, \vec{y}, \vec{z} \in \mathcal{V}$*
$$\vec{x} + \vec{y} = \vec{y} + \vec{x}, \quad (\vec{x} + \vec{y}) + \vec{z} = \vec{x} + (\vec{y} + \vec{z}). \tag{1}$$
6. *Scalar multiplication is associative, for any scalars α and β and any $\vec{x} \in \mathcal{V}$*
$$\alpha\,(\beta \cdot \vec{x}) = (\alpha\,\beta) \cdot \vec{x}. \tag{2}$$
7. *Scalar and vector sums are both distributive, i.e. for any scalars α and β and any $\vec{x}, \vec{y} \in \mathcal{V}$*
$$(\alpha + \beta) \cdot \vec{x} = \alpha \cdot \vec{x} + \beta \cdot \vec{x}, \quad \alpha \cdot (\vec{x} + \vec{y}) = \alpha \cdot \vec{x} + \alpha \cdot \vec{y}. \tag{3}$$

A subspace of a vector space $\mathcal{V}$ is any subset of $\mathcal{V}$ that is also itself a vector space.

From now on, for ease of notation we ignore the symbol for the scalar product $\cdot$, writing $\alpha \cdot \vec{x}$ as $\alpha\,\vec{x}$.

Depending on the signal of interest, we may want to represent it as an array of real or complex numbers, a matrix or a function. All of these mathematical objects can be represented as vectors in a vector space.

Example 1.2 (Real-valued and complex-valued vectors)**.** $\mathbb{R}^n$ with $\mathbb{R}$ as its associated scalar field is a vector space where each vector consists of a set of n real-valued numbers. This is by far the most useful vector space in data analysis. For example, we can represent images with n pixels as vectors in $\mathbb{R}^n$, where each pixel is assigned to an entry.

Similarly, $\mathbb{C}^n$ with $\mathbb{C}$ as its associated scalar field is a vector space where each vector consists of a set of n complex-valued numbers. In both $\mathbb{R}^n$ and $\mathbb{C}^n$, the zero vector is a vector containing zeros in every entry. △

Example 1.3 (Matrices)**.** Real-valued or complex-valued matrices of fixed dimensions form vector spaces with $\mathbb{R}$ and $\mathbb{C}$ respectively as their associated scalar fields. Adding matrices and multiplying matrices by scalars yields matrices of the same dimensions. In this case the zero vector corresponds to a matrix containing zeros in every entry. △

Example 1.4 (Functions)**.** Real-valued or complex-valued functions form a vector space (with $\mathbb{R}$ and $\mathbb{C}$ respectively as their associated scalar fields), since we can obtain new functions by adding functions or multiplying them by scalars. In this case the zero vector corresponds to a function that maps any number to zero. △

Linear dependence indicates when a vector can be represented in terms of other vectors.

Definition 1.5 (Linear dependence/independence)**.** *A set of m vectors $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_m$ is linearly dependent if there exist m scalar coefficients $\alpha_1, \alpha_2, \ldots, \alpha_m$ which are not all equal to zero and such that*

$$\sum_{i=1}^{m} \alpha_i \, \vec{x}_i = \vec{0}. \tag{4}$$

Otherwise, the vectors are linearly independent. Equivalently, any vector in a linearly dependent set can be expressed as a linear combination of the rest, which is not the case for linearly independent sets.

We define the span of a set of vectors $\{\vec{x}_1, \ldots, \vec{x}_m\}$ as the set of all possible linear combinations of the vectors:

$$\operatorname{span}(\vec{x}_1, \ldots, \vec{x}_m) := \left\{ \vec{y} \mid \vec{y} = \sum_{i=1}^{m} \alpha_i \, \vec{x}_i \quad \text{for some scalars } \alpha_1, \alpha_2, \ldots, \alpha_m \right\}. \tag{5}$$

It is not difficult to check that the span of any set of vectors belonging to a vector space $\mathcal{V}$ is a subspace of $\mathcal{V}$.

When working with a vector space, it is useful to consider the set of vectors with the smallest cardinality that spans the space. This is called a basis of the vector space.

Definition 1.6 (Basis)**.** *A basis of a vector space $\mathcal{V}$ is a set of independent vectors $\{\vec{x}_1, \ldots, \vec{x}_m\}$ such that*

$$\mathcal{V} = \operatorname{span}(\vec{x}_1, \ldots, \vec{x}_m) \tag{6}$$

An important property of all bases in a vector space is that they have the same cardinality.

Theorem 1.7 (Proof in Section 8.1)**.** *If a vector space $\mathcal{V}$ has a basis with finite cardinality then every basis of $\mathcal{V}$ contains the same number of vectors.*

This result allows us to define the dimension of a vector space.

Definition 1.8 (Dimension)**.** *The dimension $dim(\mathcal{V})$ of a vector space $\mathcal{V}$ is the cardinality of any of its bases, or equivalently the number of linearly independent vectors that span $\mathcal{V}$.*

This definition coincides with the usual geometric notion of dimension in $\mathbb{R}^2$ and $\mathbb{R}^3$: a line has dimension 1, whereas a plane has dimension 2 (as long as they contain the origin). Note that there exist infinite-dimensional vector spaces, such as the continuous real-valued functions defined on $[0,1]$ (we will define a basis for this space later on).

The vector space that we use to model a certain problem is usually called the ambient space and its dimension the ambient dimension. In the case of $\mathbb{R}^n$ the ambient dimension is n.

Lemma 1.9 (Dimension of $\mathbb{R}^n$)**.** *The dimension of $\mathbb{R}^n$ is n.*

Proof. Consider the set of vectors $\vec{e}_1, \ldots, \vec{e}_n \subseteq \mathbb{R}^n$ defined by

$$\vec{e}_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \vec{e}_2 = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \quad \ldots, \quad \vec{e}_n = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}. \tag{7}$$

One can easily check that this set is a basis. It is in fact the **standard basis** of $\mathbb{R}^n$. □

2 Inner product

Up to now, the only operations we have considered are addition and multiplication by a scalar. In this section, we introduce a third operation, the inner product between two vectors.

Definition 2.1 (Inner product)**.** *An inner product on a vector space $\mathcal{V}$ is an operation $\langle \cdot, \cdot \rangle$ that maps a pair of vectors to a scalar and satisfies the following conditions.*

- *If the scalar field associated to $\mathcal{V}$ is $\mathbb{R}$, it is symmetric. For any $\vec{x}, \vec{y} \in \mathcal{V}$*

  $$\langle \vec{x}, \vec{y} \rangle = \langle \vec{y}, \vec{x} \rangle . \tag{8}$$

  If the scalar field is $\mathbb{C}$, then for any $\vec{x}, \vec{y} \in \mathcal{V}$

  $$\langle \vec{x}, \vec{y} \rangle = \overline{\langle \vec{y}, \vec{x} \rangle}, \tag{9}$$

  where for any $\alpha \in \mathbb{C}$ $\overline{\alpha}$ is the complex conjugate of α.

- *It is linear in the first argument, i.e. for any* $\alpha \in \mathbb{R}$ *and any* $\vec{x}, \vec{y}, \vec{z} \in \mathcal{V}$

$$\langle \alpha\, \vec{x}, \vec{y} \rangle = \alpha \, \langle \vec{x}, \vec{y} \rangle \,, \tag{10}$$

$$\langle \vec{x} + \vec{y}, \vec{z} \rangle = \langle \vec{x}, \vec{z} \rangle + \langle \vec{y}, \vec{z} \rangle \,. \tag{11}$$

  Note that if the scalar field is $\mathbb{R}$*, it is also linear in the second argument by symmetry.*

- *It is positive definite:* $\langle \vec{x}, \vec{x} \rangle$ *is nonnegative for all* $\vec{x} \in \mathcal{V}$ *and if* $\langle \vec{x}, \vec{x} \rangle = 0$ *then* $\vec{x} = 0$*.*

Definition 2.2 (Dot product)**.** *The dot product between two vectors* $\vec{x}, \vec{y} \in \mathbb{R}^n$

$$\vec{x} \cdot \vec{y} := \sum_i \vec{x}\,[i] \; \vec{y}\,[i] \,, \tag{12}$$

where $\vec{x}\,[i]$ *is the* i*th entry of* $\vec{x}$*, is a valid inner product.* $\mathbb{R}^n$ *endowed with the dot product is usually called a Euclidean space of dimension* n*.*

Similarly, the dot product between two vectors $\vec{x}, \vec{y} \in \mathbb{C}^n$

$$\vec{x} \cdot \vec{y} := \sum_i \vec{x}\,[i] \; \overline{\vec{y}\,[i]} \tag{13}$$

is a valid inner product.

Definition 2.3 (Sample covariance)**.** *In statistics and data analysis, the sample covariance is used to quantify the joint fluctuations of two quantities or features. Let* (x_1, y_1)*,* (x_2, y_2)*, ...,* (x_n, y_n) *be a data set where each example consists of a measurement of the two features. The sample covariance is defined as*

$$\operatorname{cov}\left((x_1, y_1), \ldots, (x_n, y_n) \right) := \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \operatorname{av}\left(x_1, \ldots, x_n \right) \right) \left(y_i - \operatorname{av}\left(y_1, \ldots, y_n \right) \right) \tag{14}$$

where the average or sample mean of a set of n *numbers is defined by*

$$\operatorname{av}\left(a_1, \ldots, a_n \right) := \frac{1}{n} \sum_{i=1}^{n} a_i . \tag{15}$$

Geometrically the covariance is the scaled dot product of the two feature vectors after centering. The normalization constant is set so that if the measurements are modeled as independent samples $(\mathbf{x_1}, \mathbf{y_1})$*,* $(\mathbf{x_2}, \mathbf{y_2})$*, ...,* $(\mathbf{x_n}, \mathbf{y_n})$ *following the same distribution as two random variables* $\mathbf{x}$ *and* $\mathbf{y}$*, then the sample covariance of the sequence is an unbiased estimate of the covariance of* $\mathbf{x}$ *and* $\mathbf{y}$*,*

$$\mathrm{E}\left(\operatorname{cov}\left((\mathbf{x_1}, \mathbf{y_1}), \ldots, (\mathbf{x_n}, \mathbf{y_n}) \right) \right) = \operatorname{Cov}\left(\mathbf{x}, \mathbf{y} \right) := \mathrm{E}\left(\left(\mathbf{x} - \mathrm{E}\left(\mathbf{x} \right) \right) \left(\mathbf{y} - \mathrm{E}\left(\mathbf{y} \right) \right) \right) . \tag{16}$$

Definition 2.4 (Matrix inner product)**.** *The inner product between two $m \times n$ matrices A and B is*

$$\begin{align} \langle A, B \rangle &:= \operatorname{tr}\left(A^T B\right) \tag{17} \\ &= \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} B_{ij}, \tag{18} \end{align}$$

where the trace of an $n \times n$ matrix is defined as the sum of its diagonal

$$\operatorname{tr}\left(M\right) := \sum_{i=1}^{n} M_{ii}. \tag{19}$$

The following lemma shows a useful property of the matrix inner product.

Lemma 2.5. *For any pair of $m \times n$ matrices A and B*

$$\operatorname{tr}\left(B^T A\right) := \operatorname{tr}\left(A B^T\right). \tag{20}$$

Proof. Both sides are equal to (18). □

Note that the matrix inner product is equivalent to the inner product of the vectors with mn entries obtained by vectorizing the matrices.

Definition 2.6 (Function inner product)**.** *A valid inner product between two complex-valued square-integrable functions f, g defined in an interval $[a, b]$ of the real line is*

$$\vec{f} \cdot \vec{g} := \int_a^b f\left(x\right) \overline{g\left(x\right)} \, dx. \tag{21}$$

3 Norms

The norm of a vector is a generalization of the concept of *length* in Euclidean space.

Definition 3.1 (Norm)**.** *Let $\mathcal{V}$ be a vector space, a norm is a function $||\cdot||$ from $\mathcal{V}$ to $\mathbb{R}$ that satisfies the following conditions.*

- *It is homogeneous. For any scalar α and any $\vec{x} \in \mathcal{V}$*

$$||\alpha \, \vec{x}|| = |\alpha| \, ||\vec{x}||. \tag{22}$$

- *It satisfies the triangle inequality*

$$||\vec{x} + \vec{y}|| \leq ||\vec{x}|| + ||\vec{y}||. \tag{23}$$

  In particular, it is nonnegative (set $\vec{y} = -\vec{x}$).

- $||\vec{x}|| = 0$ *implies that* $\vec{x}$ *is the zero vector* $\vec{0}$.

A vector space equipped with a norm is called a normed space. Inner-product spaces are normed spaces because we can define a valid norm using the inner product.

Definition 3.2 (Inner-product norm)**.** *The norm induced by an inner product is obtained by taking the square root of the inner product of the vector with itself,*

$$||\vec{x}||_{\langle\cdot,\cdot\rangle} := \sqrt{\langle\vec{x},\vec{x}\rangle}. \tag{24}$$

Definition 3.3 (ℓ_2 norm)**.** *The* ℓ_2 *norm is the norm induced by the dot product in* $\mathbb{R}^n$ *or* $\mathbb{C}^n$*,*

$$||\vec{x}||_2 := \sqrt{\vec{x}\cdot\vec{x}} = \sqrt{\sum_{i=1}^{n} \vec{x}[i]^2}. \tag{25}$$

In the case of $\mathbb{R}^2$ *or* $\mathbb{R}^3$ *it is what we usually think of as the length of the vector.*

Definition 3.4 (Sample variance and standard deviation)**.** *Let* $\{x_1, x_2, \ldots, x_n\}$ *be a set of real-valued data. The sample variance is defined as*

$$\operatorname{var}(x_1, x_2, \ldots, x_n) := \frac{1}{n-1}\sum_{i=1}^{n} \left(x_i - \operatorname{av}(x_1, x_2, \ldots, x_n)\right)^2 \tag{26}$$

The sample standard deviation is the square root of the sample variance

$$\operatorname{std}(x_1, x_2, \ldots, x_n) := \sqrt{\operatorname{var}(x_1, x_2, \ldots, x_n)}. \tag{27}$$

Definition 3.5 (Sample variance and standard deviation)**.** *In statistics and data analysis, the sample variance is used to quantify the fluctuations of a quantity around its average. Assume that we have* n *real-valued measurements* x_1, x_2, $\ldots$, x_n. *The sample variance equals*

$$\operatorname{var}(x_1, x_2, \ldots, x_n) := \frac{1}{n-1}\sum_{i=1}^{n} \left(x_i - \operatorname{av}(x_1, x_2, \ldots, x_n)\right)^2 \tag{28}$$

The normalization constant is set so that if the measurements are modeled as independent samples $\mathbf{x_1}$, $\mathbf{x_2}$, $\ldots$, $\mathbf{x_n}$ *following the same distribution as a random variable* $\mathbf{x}$ *then the sample variance is an unbiased estimate of the variance of* $\mathbf{x}$*,*

$$\mathrm{E}\left(\operatorname{var}(\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n})\right) = \operatorname{Var}(\mathbf{x}) := \mathrm{E}\left((\mathbf{x} - \mathrm{E}(\mathbf{x}))^2\right). \tag{29}$$

The sample standard deviation is the square root of the sample variance

$$\operatorname{std}(x_1, x_2, \ldots, x_n) := \sqrt{\operatorname{var}(x_1, x_2, \ldots, x_n)}. \tag{30}$$

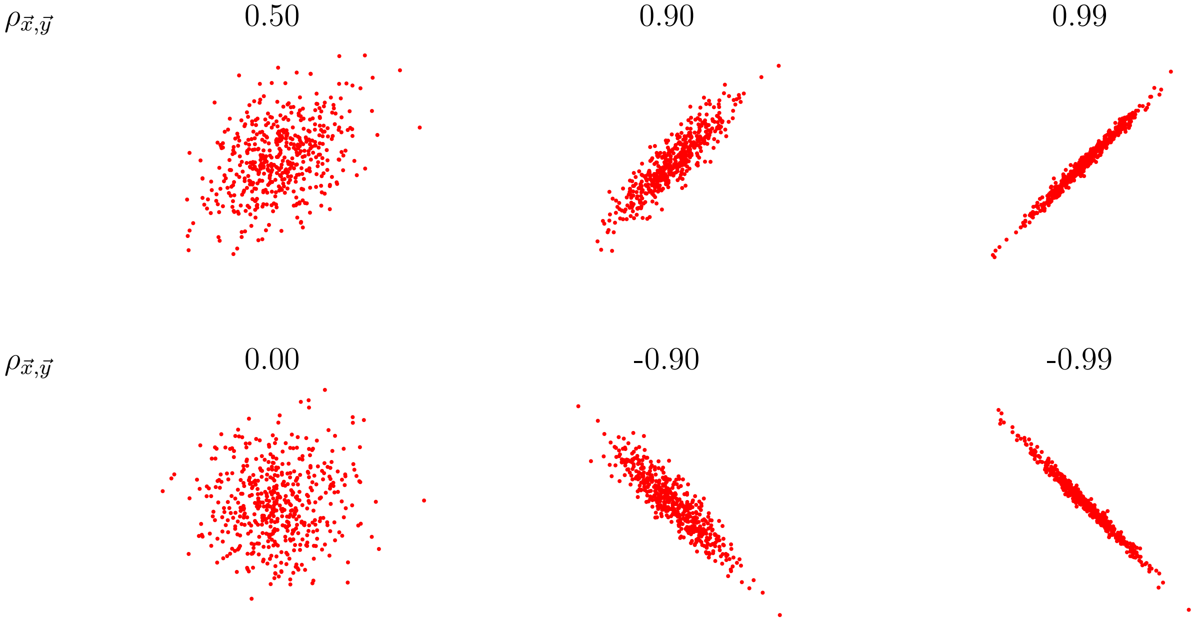


Figure 1: Scatter plot of the points $(\vec{x}_1, \vec{y}_1)$, $(\vec{x}_2, \vec{y}_2)$, ..., $(\vec{x}_n, \vec{y}_n)$ for vectors with different correlation coefficients.

Definition 3.6 (Correlation coefficient)**.** *When computing the sample covariance of two features the unit in which we express each quantity may severely affect the result. If one of the features is a distance, for example, expressing it in meters instead of kilometers increases the sample covariance by a factor of 1000! In order to obtain a measure of joint fluctuations that is invariant to scale, we normalize the covariance using the sample standard deviation of the features. This yields the correlation coefficient of the two quantities*

$$\rho_{(x_1,y_1),\ldots,(x_n,y_n)} := \frac{\operatorname{cov}\left((x_1, y_1), \ldots, (x_n, y_n)\right)}{\operatorname{std}(x_1, \ldots, x_n)\operatorname{std}(y_1, \ldots, y_n)}. \tag{31}$$

As illustrated in Figure 1 the correlation coefficient quantifies to what extent the entries of the two vectors are linearly related. Corollary 3.12 below shows that it is always between -1 and 1. If it is positive, we say that the two quantities are correlated. If it is negative, we say they are negatively correlated. If it is zero, we say that they are uncorrelated. In the following example we compute the correlation coefficient of some temperature data.

Example 3.7 (Correlation of temperature data)**.** In this example we analyze temperature data gathered at a weather station in Oxford over 150 years.[1] We first compute the correlation between the temperature in January and the temperature in August. The correlation coefficient is $\rho = 0.269$. This means that the two quantities are positively correlated: warmer temperatures in January tend to correspond to warmer temperatures

[1] The data is available at http://www.metoffice.gov.uk/pub/data/weather/uk/climate/stationdata/oxforddata.txt.

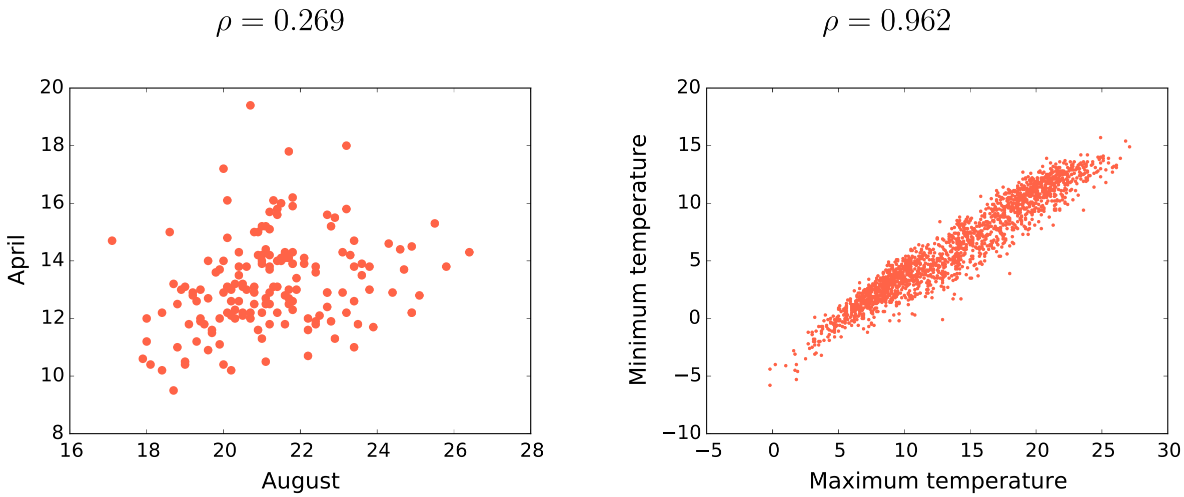


Figure 2: Scatterplot of the temperature in January and in August (left) and of the maximum and minimum monthly temperature (right) in Oxford over the last 150 years.

in August. The left image in Figure 2 shows a scatter plot where each point represents a different year. We repeat the experiment to compare the maximum and minimum temperature in the same month. The correlation coefficient between these two quantities is $\rho = 0.962$, indicating that the two quantities are extremely correlated. The right image in Figure 2 shows a scatter plot where each point represents a different month. $\triangle$

Definition 3.8 (Frobenius norm)**.** *The Frobenius norm is the norm induced by the matrix inner product. For any matrix $A \in \mathbb{R}^{m \times n}$*

$$||A||_{\mathrm{F}} := \sqrt{\operatorname{tr}\left(A^T A\right)} = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij}^2}. \tag{32}$$

It is equal to the ℓ_2 norm of the vectorized matrix.

Definition 3.9 ($\mathcal{L}_2$ norm)**.** *The $\mathcal{L}_2$ norm is the norm induced by the dot product in the inner-product space of square-integrable complex-valued functions defined on an interval $[a, b]$,*

$$||f||_{\mathcal{L}_2} := \sqrt{\langle f, f \rangle} = \sqrt{\int_a^b |f(x)|^2 \, dx}. \tag{33}$$

The inner-product norm is clearly homogeneous by linearity and symmetry of the inner product. $||\vec{x}||_{\langle \cdot, \cdot \rangle} = 0$ implies $\vec{x} = 0$ because the inner product is positive semidefinite. We only need to establish that the triangle inequality holds to ensure that the inner-product is a valid norm. This follows from a classic inequality in linear algebra, which is proved in Section 8.2.

Theorem 3.10 (Cauchy-Schwarz inequality)**.** *For any two vectors $\vec{x}$ and $\vec{y}$ in an inner-product space*

$$|\langle \vec{x}, \vec{y} \rangle| \leq ||\vec{x}||_{\langle \cdot,\cdot \rangle} \, ||\vec{y}||_{\langle \cdot,\cdot \rangle} \,. \tag{34}$$

Assume $||\vec{x}||_{\langle \cdot,\cdot \rangle} \neq 0$, then

$$\langle \vec{x}, \vec{y} \rangle = - ||\vec{x}||_{\langle \cdot,\cdot \rangle} \, ||\vec{y}||_{\langle \cdot,\cdot \rangle} \iff \vec{y} = - \frac{||\vec{y}||_{\langle \cdot,\cdot \rangle}}{||\vec{x}||_{\langle \cdot,\cdot \rangle}} \vec{x}, \tag{35}$$

$$\langle \vec{x}, \vec{y} \rangle = ||\vec{x}||_{\langle \cdot,\cdot \rangle} \, ||\vec{y}||_{\langle \cdot,\cdot \rangle} \iff \vec{y} = \frac{||\vec{y}||_{\langle \cdot,\cdot \rangle}}{||\vec{x}||_{\langle \cdot,\cdot \rangle}} \vec{x}. \tag{36}$$

Corollary 3.11. *The norm induced by an inner product satisfies the triangle inequality.*

Proof.

$$\begin{aligned} ||\vec{x} + \vec{y}||^2_{\langle \cdot,\cdot \rangle} &= ||\vec{x}||^2_{\langle \cdot,\cdot \rangle} + ||\vec{y}||^2_{\langle \cdot,\cdot \rangle} + 2 \langle \vec{x}, \vec{y} \rangle && (37) \\ &\leq ||\vec{x}||^2_{\langle \cdot,\cdot \rangle} + ||\vec{y}||^2_{\langle \cdot,\cdot \rangle} + 2 \, ||\vec{x}||_{\langle \cdot,\cdot \rangle} \, ||\vec{y}||_{\langle \cdot,\cdot \rangle} && \text{by the Cauchy-Schwarz inequality} \\ &= \left(||\vec{x}||_{\langle \cdot,\cdot \rangle} + ||\vec{y}||_{\langle \cdot,\cdot \rangle} \right)^2 . && (38) \end{aligned}$$

□

Another corollary of the Cauchy-Schwarz theorem is that the correlation coefficient is always between -1 and 1 and that if it equals either 1 or -1 then the two vectors are linearly dependent.

Corollary 3.12. *The correlation coefficient of two vectors $\vec{x}$ and $\vec{y}$ in $\mathbb{R}^n$ satisfies*

$$-1 \leq \rho_{(x_1,y_1),\ldots,(x_n,y_n)} \leq 1. \tag{39}$$

In addition,

$$\rho_{\vec{x},\vec{y}} = -1 \iff y_i = \operatorname{av}(y_1, \ldots, y_n) - \frac{\operatorname{std}(y_1, \ldots, y_n)}{\operatorname{std}(x_1, \ldots, x_n)} (x_i - \operatorname{av}(x_1, \ldots, x_n)), \tag{40}$$

$$\rho_{\vec{x},\vec{y}} = 1 \iff y_i = \operatorname{av}(y_1, \ldots, y_n) + \frac{\operatorname{std}(y_1, \ldots, y_n)}{\operatorname{std}(x_1, \ldots, x_n)} (x_i - \operatorname{av}(x_1, \ldots, x_n)). \tag{41}$$

Proof. The result follows from applying the Cauchy-Schwarz inequality to the vectors

$$\vec{a} := \begin{bmatrix} x_1 - \operatorname{av}(x_1, \ldots, x_n) & x_2 - \operatorname{av}(x_1, \ldots, x_n) & \cdots & x_n - \operatorname{av}(x_1, \ldots, x_n) \end{bmatrix}, \tag{42}$$

$$\vec{b} := \begin{bmatrix} y_1 - \operatorname{av}(y_1, \ldots, y_n) & y_2 - \operatorname{av}(y_1, \ldots, y_n) & \cdots & y_n - \operatorname{av}(y_1, \ldots, y_n) \end{bmatrix}, \tag{43}$$

since

$$\operatorname{std}(x_1, x_2, \ldots, x_n) = ||\vec{a}||_2 \,, \tag{44}$$

$$\operatorname{std}(y_1, y_2, \ldots, y_n) = ||\vec{b}||_2, \tag{45}$$

$$\operatorname{cov}((x_1, y_1), \ldots, (x_n, y_n)) = \left\langle \vec{a}, \vec{b} \right\rangle . \tag{46}$$

□

Norms are not always induced by an inner product. The parallelogram law provides a simple identity that allows to check whether this is the case.

Theorem 3.13 (Parallelogram law)**.** *A norm $\|\cdot\|$ on a vector space $\mathcal{V}$ is induced by an inner product if and only if*

$$2\|\vec{x}\|^2 + 2\|\vec{y}\|^2 = \|\vec{x}-\vec{y}\|^2 + \|\vec{x}+\vec{y}\|^2, \tag{47}$$

for any $\vec{x}, \vec{y} \in \mathcal{V}$.

Proof. If the norm is induced by an inner product then

$$\begin{aligned} \|\vec{x}-\vec{y}\|^2 + \|\vec{x}+\vec{y}\|^2 &= \langle \vec{x}-\vec{y}, \vec{x}-\vec{y}\rangle + \langle \vec{x}+\vec{y}, \vec{x}+\vec{y}\rangle && (48) \\ &= 2\|\vec{x}\|^2 + 2\|\vec{y}\|^2 - (\vec{x}, \vec{y}) - (\vec{y}, \vec{x}) + (\vec{x}, \vec{y}) + (\vec{y}, \vec{x}) && (49) \\ &= 2\|\vec{x}\|^2 + 2\|\vec{y}\|^2. && (50) \end{aligned}$$

If the identity holds then it can be shown that

$$\langle \vec{x}, \vec{y}\rangle := \frac{1}{4}\left(\|\vec{x}+\vec{y}\|^2 - \|\vec{x}-\vec{y}\|^2\right) \tag{51}$$

is a valid inner product for real scalars and

$$\langle \vec{x}, \vec{y}\rangle := \frac{1}{4}\left(\|\vec{x}+\vec{y}\|^2 - \|\vec{x}-\vec{y}\|^2 - i\left(\|\vec{x}+i\vec{y}\|^2 - \|\vec{x}-i\vec{y}\|^2\right)\right) \tag{52}$$

is a valid inner product for complex scalars. □

The following two norms do not satisfy the parallelogram identity and therefore are not induced by an inner product. Figure 3 compares their unit-norm balls with that of the ℓ_2 norm. Recall that the unit-norm ball of a norm $||\cdot||$ is the set of vectors $\vec{x}$ such that $||\vec{x}|| \leq 1$.

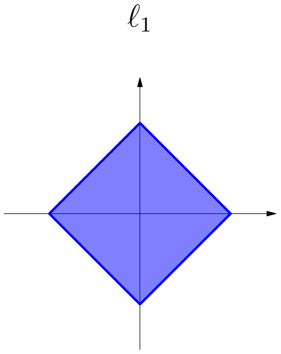
Definition 3.14 (ℓ_1 norm)**.** *The ℓ_1 norm of a vector in $\mathbb{R}^n$ or $\mathbb{C}^n$ is the sum of the absolute values of the entries,*

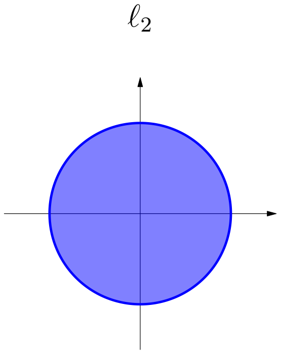
$$||\vec{x}||_1 := \sum_{i=1}^{n} |\vec{x}[i]|\,. \tag{53}$$

Definition 3.15 (ℓ_∞ norm)**.** *The ℓ_∞ norm of a vector in $\mathbb{R}^n$ or $\mathbb{C}^n$ is the maximum absolute value of its entries,*

$$||\vec{x}||_\infty := \max_i |\vec{x}[i]|\,. \tag{54}$$

Although they do not satisfy the Cauchy-Schwarz inequality, as they are not induced by any inner product, the ℓ_1 and ℓ_∞ norms can be used to bound the inner product between two vectors.





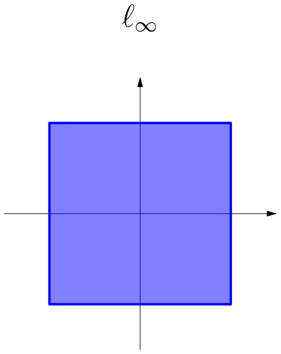


Figure 3: Unit ℓ_1, ℓ_2 and ℓ_∞ norm balls.

Theorem 3.16 (Hölder's inequality)**.** *For any two vectors $\vec{x}$ and $\vec{y}$ in $\mathbb{R}^n$ or $\mathbb{C}^n$*

$$|\langle \vec{x}, \vec{y} \rangle| \leq ||\vec{x}||_1 \, ||\vec{y}||_\infty . \tag{55}$$

Proof.

$$\begin{aligned} |\langle \vec{x}, \vec{y} \rangle| &\leq \sum_{i=1}^{n} |\vec{x}[i]| \, |\vec{y}[i]| && (56) \\ &\leq \max_i |\vec{y}[i]| \sum_{i=1}^{n} |\vec{x}[i]| && (57) \\ &= ||\vec{x}||_1 \, ||\vec{y}||_\infty . && (58) \end{aligned}$$

□

Distances in a normed space can be measured using the norm of the difference between vectors.

Definition 3.17 (Distance)**.** *The distance between two vectors $\vec{x}$ and $\vec{y}$ induced by a norm $||\cdot||$ is*

$$d(\vec{x}, \vec{y}) := ||\vec{x} - \vec{y}|| . \tag{59}$$

4 Nearest-neighbor classification

If we represent signals as vectors in a vector space, the distance between them quantifies their similarity. In this section we show how to exploit this to perform classification.

Definition 4.1 (Classification)**.** *Given a set of k predefined classes, the classification problem is to decide what class a signal belongs to. The assignment is done using a training set of examples, each of which consists of a signals and its corresponding label.*

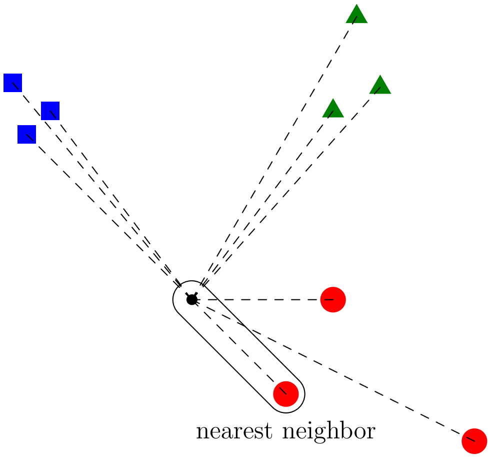


Figure 4: The nearest neighbor algorithm classifies points by assigning them the class of the closest point. In the diagram, the black point is assigned the *red circle* class because its nearest neighbor is a red circle.

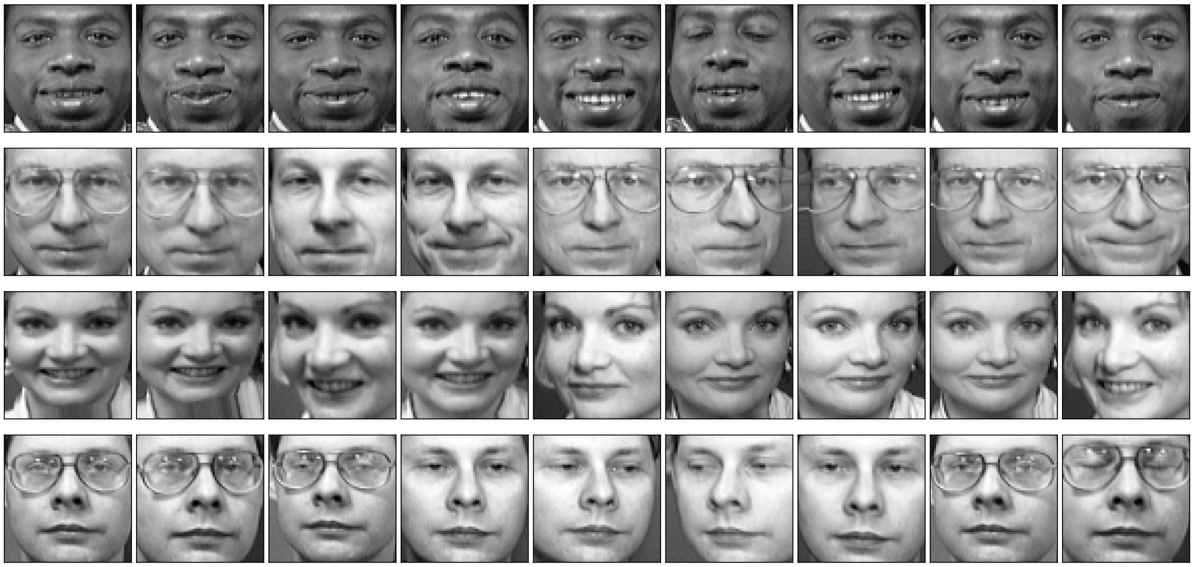


Figure 5: Training examples for four of the people in Example 4.3.

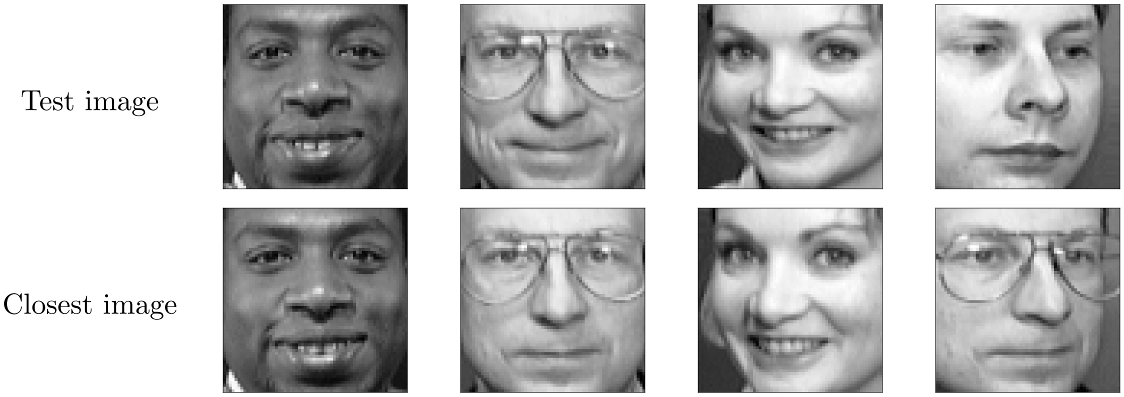


Figure 6: Results of nearest-neighbor classification for four of the people in Example 4.3. The assignments of the first three examples are correct, but the fourth is wrong.

The nearest-neighbor algorithm classifies signals by looking for the closest signal in the training set. Figure 4 shows a simple example.

Algorithm 4.2 (Nearest-neighbor classification)**.** *Assume that the signals of interest can be represented by vectors in a vector space endowed with a norm denoted by* $||\cdot||$*. The training set consequently consists of* n *pairs of vectors and labels:* $\{\vec{x}_1, l_1\}, \ldots, \{\vec{x}_n, l_n\}$*. To classify a test signal* $\vec{y}$ *we find the closest signal in the training set in terms of the distance induced by the norm,*

$$i^* := \arg \min_{1 \leq i \leq n} ||\vec{y} - \vec{x}_i|| \,, \tag{60}$$

and assign the corresponding label l_{i^*} *to* $\vec{y}$*.*

Example 4.3 (Face recognition)**.** The problem of face recognition consists of classifying images of faces to determine what person they correspond to. In this example we consider the Olivetti Faces data set[2]. The training set consists of 360 64×64 images taken from 40 different subjects (9 per subject). Figure 5 shows some of the faces in the training set. The test set consists of an image of each subject, which is different from the ones in the training set. We apply nearest-neighbor algorithm to classify the faces in the test set, modeling each image as a vector in $\mathbb{R}^{4096}$ and using the distance induced by the ℓ_2 norm. The algorithm classifies 36 of the 40 subjects correctly. Some of the results are shown in Figure 6. △

5 Orthogonality

When the inner product between two vectors is zero, we say that the vectors are orthogonal.

[2] Available at http://www.cs.nyu.edu/~roweis/data.html

Definition 5.1 (Orthogonality)**.** *Two vectors $\vec{x}$ and $\vec{y}$ are orthogonal if and only if*

$$\langle \vec{x}, \vec{y} \rangle = 0. \tag{61}$$

A vector $\vec{x}$ is orthogonal to a set $\mathcal{S}$, if

$$\langle \vec{x}, \vec{s} \rangle = 0, \quad \textit{for all } \vec{s} \in \mathcal{S}. \tag{62}$$

Two sets of $\mathcal{S}_1, \mathcal{S}_2$ are orthogonal if for any $\vec{x} \in \mathcal{S}_1, \vec{y} \in \mathcal{S}_2$

$$\langle \vec{x}, \vec{y} \rangle = 0. \tag{63}$$

The orthogonal complement of a subspace $\mathcal{S}$ is

$$\mathcal{S}^{\perp} := \{\vec{x} \mid \langle \vec{x}, \vec{y} \rangle = 0 \quad \textit{for all } \vec{y} \in \mathcal{S}\}. \tag{64}$$

Distances between orthogonal vectors measured in terms of the norm induced by the inner product are easy to compute.

Theorem 5.2 (Pythagorean theorem)**.** *If $\vec{x}$ and $\vec{y}$ are orthogonal vectors*

$$||\vec{x} + \vec{y}||^2_{\langle\cdot,\cdot\rangle} = ||\vec{x}||^2_{\langle\cdot,\cdot\rangle} + ||\vec{y}||^2_{\langle\cdot,\cdot\rangle}. \tag{65}$$

Proof. By linearity of the inner product

$$\begin{aligned} ||\vec{x} + \vec{y}||^2_{\langle\cdot,\cdot\rangle} &= ||\vec{x}||^2_{\langle\cdot,\cdot\rangle} + ||\vec{y}||^2_{\langle\cdot,\cdot\rangle} + 2\langle \vec{x}, \vec{y} \rangle \qquad (66) \\ &= ||\vec{x}||^2_{\langle\cdot,\cdot\rangle} + ||\vec{y}||^2_{\langle\cdot,\cdot\rangle}. \qquad (67) \end{aligned}$$

□

If we want to show that a vector is orthogonal to a certain subspace, it is enough to show that it is orthogonal to every vector in a basis of the subspace.

Lemma 5.3. *Let $\vec{x}$ be a vector and $\mathcal{S}$ a subspace of dimension n. If for any basis $\vec{b}_1, \vec{b}_2, \ldots, \vec{b}_n$ of $\mathcal{S}$,*

$$\left\langle \vec{x}, \vec{b}_i \right\rangle = 0, \quad 1 \leq i \leq n, \tag{68}$$

then $\vec{x}$ is orthogonal to $\mathcal{S}$.

Proof. Any vector $v \in \mathcal{S}$ can be represented as $v = \sum_i \alpha_{i=1}^n \vec{b}_i$ for $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, from (68)

$$\langle \vec{x}, v \rangle = \left\langle \vec{x}, \sum_i \alpha_{i=1}^n \vec{b}_i \right\rangle = \sum_i \alpha_{i=1}^n \left\langle \vec{x}, \vec{b}_i \right\rangle = 0. \tag{69}$$

□

If the vectors in a basis are normalized and mutually orthogonal, then the norm is said to be orthonormal.

Definition 5.4 (Orthonormal basis)**.** *A basis of mutually orthogonal vectors with inner-product norm equal to one is called an orthonormal basis.*

It is very easy to find the coefficients of a vector in an orthonormal basis: we just need to compute the dot products with the basis vectors.

Lemma 5.5 (Coefficients in an orthonormal basis)**.** *If $\{\vec{u}_1, \ldots, \vec{u}_n\}$ is an orthonormal basis of a vector space $\mathcal{V}$, for any vector $\vec{x} \in \mathcal{V}$*

$$\vec{x} = \sum_{i=1}^{n} \langle \vec{u}_i, \vec{x} \rangle \, \vec{u}_i. \tag{70}$$

Proof. Since $\{\vec{u}_1, \ldots, \vec{u}_n\}$ is a basis,

$$\vec{x} = \sum_{i=1}^{m} \alpha_i \, \vec{u}_i \quad \text{for some } \alpha_1, \alpha_2, \ldots, \alpha_m \in \mathbb{R}. \tag{71}$$

Immediately,

$$\langle \vec{u}_i, \vec{x} \rangle = \left\langle \vec{u}_i, \sum_{i=1}^{m} \alpha_i \, \vec{u}_i \right\rangle = \sum_{i=1}^{m} \alpha_i \, \langle \vec{u}_i, \vec{u}_i \rangle = \alpha_i \tag{72}$$

because $\langle \vec{u}_i, \vec{u}_i \rangle = 1$ and $\langle \vec{u}_i, \vec{u}_j \rangle = 0$ for $i \neq j$. □

We can construct an orthonormal basis for any subspace in a vector space by applying the Gram-Schmidt method to a set of linearly independent vectors spanning the subspace.

Algorithm 5.6 (Gram-Schmidt)**.** *Consider a set of linearly independent vectors $\vec{x}_1$, …, $\vec{x}_m$ in $\mathbb{R}^n$. To obtain an orthonormal basis of the span of these vectors we:*

1. *Set $\vec{u}_1 := \vec{x}_1 / \left|\left|\vec{x}_1\right|\right|_2$.*
2. *For $i = 1, \ldots, m$, compute*

$$\vec{v}_i := \vec{x}_i - \sum_{j=1}^{i-1} \langle \vec{u}_j, \vec{x}_i \rangle \, \vec{u}_j. \tag{73}$$

and set $\vec{u}_i := \vec{v}_i / \left|\left|\vec{v}_i\right|\right|_2$.

It is not difficult to show that the resulting set of vectors $\vec{u}_1$, …, $\vec{u}_m$ is an orthonormal basis for the span of $\vec{x}_1$, …, $\vec{x}_m$: they are orthonormal by construction and their span is the same as that of the original set of vectors.

6 Orthogonal projection

If two subspaces are disjoint, i.e. their only common point is the origin, then a vector that can be expressed as a sum of a vector from each subspace is said to belong to their direct sum.

Definition 6.1 (Direct sum)**.** *Let $\mathcal{V}$ be a vector space. For any subspaces $\mathcal{S}_1, \mathcal{S}_2 \subseteq \mathcal{V}$ such that*

$$\mathcal{S}_1 \cap \mathcal{S}_2 = \{0\} \tag{74}$$

the direct sum is defined as

$$\mathcal{S}_1 \oplus \mathcal{S}_2 := \{\vec{x} \mid \vec{x} = \vec{s}_1 + \vec{s}_2 \quad \vec{s}_1 \in \mathcal{S}_1, \vec{s}_2 \in \mathcal{S}_2\}\,. \tag{75}$$

The representation of a vector in the direct sum of two subspaces as the sum of vectors from the subspaces is unique.

Lemma 6.2. *Any vector $\vec{x} \in \mathcal{S}_1 \oplus \mathcal{S}_2$ has a unique representation*

$$\vec{x} = \vec{s}_1 + \vec{s}_2 \quad \vec{s}_1 \in \mathcal{S}_1, \vec{s}_2 \in \mathcal{S}_2. \tag{76}$$

Proof. If $\vec{x} \in \mathcal{S}_1 \oplus \mathcal{S}_2$ then by definition there exist $\vec{s}_1 \in \mathcal{S}_1, \vec{s}_2 \in \mathcal{S}_2$ such that $\vec{x} = \vec{s}_1 + \vec{s}_2$. Assume $\vec{x} = \vec{v}_1 + \vec{v}_2$, $\vec{v}_1 \in \mathcal{S}_1, \vec{v}_2 \in \mathcal{S}_2$, then $\vec{s}_1 - \vec{v}_1 = \vec{s}_2 - \vec{v}_2$. This implies that $\vec{s}_1 - \vec{v}_1$ and $\vec{s}_2 - \vec{v}_2$ are in $\mathcal{S}_1$ and also in $\mathcal{S}_2$. However, $\mathcal{S}_1 \cap \mathcal{S}_2 = \{0\}$, so we conclude $\vec{s}_1 = \vec{v}_1$ and $\vec{s}_2 = \vec{v}_2$. □

Given a vector x and a subspace $\mathcal{S}$, the orthogonal projection of $\vec{x}$ onto $\mathcal{S}$ is the vector that we reach when we go from x to $\mathcal{S}$ following a direction that is orthogonal to $\mathcal{S}$. This allows to express $\vec{x}$ as the sum of a component that belongs to $\mathcal{S}$ and another that belongs to its orthogonal complement. This is illustrated by a simple example in Figure 7.

Definition 6.3 (Orthogonal projection)**.** *Let $\mathcal{V}$ be a vector space. The orthogonal projection of a vector $\vec{x} \in \mathcal{V}$ onto a subspace $\mathcal{S} \subseteq \mathcal{V}$ is a vector denoted by $\mathcal{P}_{\mathcal{S}}\,\vec{x}$ such that $\vec{x} - \mathcal{P}_{\mathcal{S}}\,\vec{x} \in \mathcal{S}^{\perp}$.*

Theorem 6.4 (Properties of the orthogonal projection)**.** *Let $\mathcal{V}$ be a vector space. Every vector $\vec{x} \in \mathcal{V}$ has a unique orthogonal projection $\mathcal{P}_{\mathcal{S}}\,\vec{x}$ onto any subspace $\mathcal{S} \subseteq \mathcal{V}$ of finite dimension. In particular $\vec{x}$ can be expressed as*

$$\vec{x} = \mathcal{P}_{\mathcal{S}}\,\vec{x} + \mathcal{P}_{\mathcal{S}^{\perp}}\,\vec{x}. \tag{77}$$

For any vector $s \in \mathcal{S}$

$$\langle \vec{x}, s \rangle = \langle \mathcal{P}_{\mathcal{S}}\,\vec{x}, s \rangle\,. \tag{78}$$

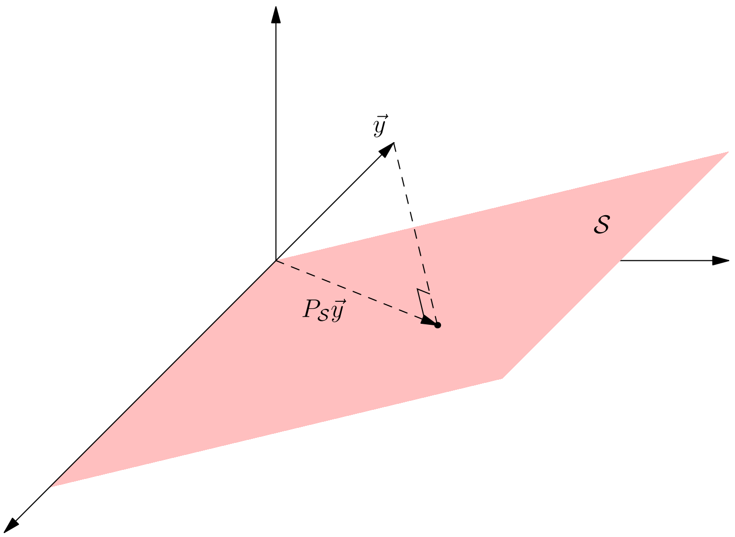


Figure 7: Orthogonal projection of a vector $\vec{x} \in \mathbb{R}^2$ on a two-dimensional subspace $\mathcal{S}$.

For any orthonormal basis $\vec{b}_1, \ldots, \vec{b}_m$ *of* $\mathcal{S}$,

$$\mathcal{P}_{\mathcal{S}}\, \vec{x} = \sum_{i=1}^{m} \left\langle \vec{x}, \vec{b}_i \right\rangle \vec{b}_i. \tag{79}$$

The orthogonal projection is a linear operation. For any vectors $\vec{x}$ *and* $\vec{y}$ *and any subspace* $\mathcal{S}$

$$\mathcal{P}_{\mathcal{S}}\, (\vec{x} + \vec{y}) = \mathcal{P}_{\mathcal{S}}\, \vec{x} + \mathcal{P}_{\mathcal{S}}\, \vec{y}. \tag{80}$$

Proof. Let us denote the dimension of $\mathcal{S}$ by m. Since m is finite, there exists an orthonormal basis of $\mathcal{S}$: $\vec{b}'_1, \ldots, \vec{b}'_m$. Consider the vector

$$\vec{p} := \sum_{i=1}^{m} \left\langle \vec{x}, \vec{b}'_i \right\rangle \vec{b}'_i. \tag{81}$$

It turns out that $\vec{x} - \vec{p}$ is orthogonal to every vector in the basis. For $1 \leq j \leq m$,

$$\left\langle \vec{x} - \vec{p}, \vec{b}'_j \right\rangle = \left\langle \vec{x} - \sum_{i=1}^{m} \left\langle \vec{x}, \vec{b}'_i \right\rangle \vec{b}'_i, \vec{b}'_j \right\rangle \tag{82}$$

$$= \left\langle \vec{x}, \vec{b}'_j \right\rangle - \sum_{i=1}^{m} \left\langle \vec{x}, \vec{b}'_i \right\rangle \left\langle \vec{b}'_i, \vec{b}'_j \right\rangle \tag{83}$$

$$= \left\langle \vec{x}, \vec{b}'_j \right\rangle - \left\langle \vec{x}, \vec{b}'_j \right\rangle = 0, \tag{84}$$

so $\vec{x} - \vec{p} \in \mathcal{S}^{\perp}$ and $\vec{p}$ is an orthogonal projection. Since $\mathcal{S} \cap \mathcal{S}^{\perp} = \{0\}$ [3] there cannot be two other vectors $\vec{x}_1 \in \mathcal{S}, \vec{x}_1 \in \mathcal{S}^{\perp}$ such that $\vec{x} = \vec{x}_1 + \vec{x}_2$ so the orthogonal projection is unique.

[3] For any vector $\vec{v}$ that belongs to both $\mathcal{S}$ and $\mathcal{S}^{\perp}$ $\langle \vec{v}, \vec{v} \rangle = ||\vec{v}||_2^2 = 0$, which implies $\vec{v} = 0$.

Notice that $\vec{o} := \vec{x} - \vec{p}$ is a vector in $\mathcal{S}^{\perp}$ such that $\vec{x} - \vec{o} = \vec{p}$ is in S and therefore in $\left(\mathcal{S}^{\perp}\right)^{\perp}$. This implies that $\vec{o}$ is the orthogonal projection of $\vec{x}$ onto $\mathcal{S}^{\perp}$ and establishes (77).

Equation (78) follows immediately from the orthogonality of any vector in $\mathcal{S}$ and $\mathcal{P}_{\mathcal{S}^{\perp}}\,\vec{x}$.

Equation (79) follows from (78).

Finally, linearity follows from (79) and linearity of the inner product

$$\begin{aligned}\mathcal{P}_{\mathcal{S}}\,(\vec{x} + \vec{y}) &= \sum_{i=1}^{m} \left\langle \vec{x} + \vec{y}, \vec{b}_i \right\rangle \vec{b}_i && (85)\\ &= \sum_{i=1}^{m} \left\langle \vec{x}, \vec{b}_i \right\rangle \vec{b}_i + \sum_{i=1}^{m} \left\langle \vec{y}, \vec{b}_i \right\rangle \vec{b}_i && (86)\\ &= \mathcal{P}_{\mathcal{S}}\,\vec{x} + \mathcal{P}_{\mathcal{S}}\,\vec{y}. && (87)\end{aligned}$$

□

The following corollary relates the dimensions of a subspace and its orthogonal complement within a finite-dimensional vector space.

Corollary 6.5 (Dimension of orthogonal complement)**.** *Let $\mathcal{V}$ be a finite-dimensional vector space, for any subspace $\mathcal{S} \subseteq \mathcal{V}$*

$$\dim\left(\mathcal{S}\right) + \dim\left(\mathcal{S}^{\perp}\right) = \dim\left(\mathcal{V}\right). \tag{88}$$

Proof. Consider a set of vectors $\mathcal{B}$ defined as the union of a basis of $\mathcal{S}$, which has $\dim\left(\mathcal{S}\right)$ elements, and a basis of $\mathcal{S}^{\perp}$, which has $\dim\left(\mathcal{S}^{\perp}\right)$ elements. Due to the orthogonality of $\mathcal{S}$ and $\mathcal{S}^{\perp}$ all the $\dim\left(\mathcal{S}\right) + \dim\left(\mathcal{S}^{\perp}\right)$ vectors in $\mathcal{B}$ are linearly independent and by (77) they span the whole space, which establishes the result. □

Computing the inner-product norm of the projection of a vector onto a subspace is easy if we have access to an orthonormal basis.

Lemma 6.6 (Norm of the projection)**.** *The norm of the projection of an arbitrary vector $\vec{x} \in \mathcal{V}$ onto a subspace $\mathcal{S} \subseteq \mathcal{V}$ of dimension d can be written as*

$$\left|\left|\mathcal{P}_{\mathcal{S}}\,\vec{x}\right|\right|_{\langle\cdot,\cdot\rangle} = \sqrt{\sum_{i}^{d} \left\langle \vec{b}_i, \vec{x} \right\rangle^2} \tag{89}$$

for any orthonormal basis $\vec{b}_1, \ldots, \vec{b}_d$ of $\mathcal{S}$.

Proof. By (79)

$$\begin{aligned}
\left|\left|\mathcal{P}_{\mathcal{S}}\,\vec{x}\right|\right|_{\langle\cdot,\cdot\rangle}^2 &= \langle \mathcal{P}_{\mathcal{S}}\,\vec{x}, \mathcal{P}_{\mathcal{S}}\,\vec{x} \rangle && (90)\\
&= \left\langle \sum_i^d \left\langle \vec{b}_i, \vec{x} \right\rangle \vec{b}_i, \sum_j^d \left\langle \vec{b}_j, \vec{x} \right\rangle \vec{b}_j \right\rangle && (91)\\
&= \sum_i^d \sum_j^d \left\langle \vec{b}_i, \vec{x} \right\rangle \left\langle \vec{b}_j, \vec{x} \right\rangle \left\langle \vec{b}_i, \vec{b}_j \right\rangle && (92)\\
&= \sum_i^d \left\langle \vec{b}_i, \vec{x} \right\rangle^2 . && (93)
\end{aligned}$$

□

The orthogonal projection of a vector $\vec{x}$ onto a subspace $\mathcal{S}$ has a very intuitive interpretation that generalizes to other sets: it is the vector in $\mathcal{S}$ that is closest to $\vec{x}$ in the distance associated to the inner-product norm.

Theorem 6.7 (The orthogonal projection is closest)**.** *The orthogonal projection $\mathcal{P}_{\mathcal{S}}\,\vec{x}$ of a vector $\vec{x}$ onto a subspace $\mathcal{S}$ is the solution to the optimization problem*

$$\begin{aligned}
&\underset{\vec{u}}{\text{minimize}} && \left|\left|\vec{x} - \vec{u}\right|\right|_{\langle\cdot,\cdot\rangle} && (94)\\
&\text{subject to} && \vec{u} \in \mathcal{S}. && (95)
\end{aligned}$$

Proof. Take any point $\vec{s} \in \mathcal{S}$ such that $\vec{s} \neq \mathcal{P}_{\mathcal{S}}\,\vec{x}$

$$\begin{aligned}
\left|\left|\vec{x} - \vec{s}\right|\right|_{\langle\cdot,\cdot\rangle}^2 &= \left|\left|\vec{x} - \mathcal{P}_{\mathcal{S}}\,\vec{x} + \mathcal{P}_{\mathcal{S}}\,\vec{x} - \vec{s}\right|\right|_{\langle\cdot,\cdot\rangle}^2 && (96)\\
&= \left|\left|\vec{x} - \mathcal{P}_{\mathcal{S}}\,\vec{x}\right|\right|_{\langle\cdot,\cdot\rangle}^2 + \left|\left|\mathcal{P}_{\mathcal{S}}\,\vec{x} - \vec{s}\right|\right|_{\langle\cdot,\cdot\rangle}^2 && (97)\\
&> \left|\left|\vec{x} - \mathcal{P}_{\mathcal{S}}\,\vec{x}\right|\right|_{\langle\cdot,\cdot\rangle}^2 \quad \text{because } \vec{s} \neq \mathcal{P}_{\mathcal{S}}\,\vec{x}, && (98)
\end{aligned}$$

where (97) follows from the Pythagorean theorem since because $\mathcal{P}_{\mathcal{S}^\perp}\,\vec{x} = \vec{x} - \mathcal{P}_{\mathcal{S}}\,\vec{x}$ belongs to $S^\perp$ and $\mathcal{P}_{\mathcal{S}}\,\vec{x} - \vec{s}$ to S. □

7 Denoising

In this section we consider the problem of denoising a signal that has been corrupted by an unknown perturbation.

Definition 7.1 (Denoising)**.** *The aim of denoising is to estimate a signal from perturbed measurements. If the noise is assumed to be additive, the data are modeled as the sum of the signal $\vec{x}$ and a perturbation $\vec{z}$*

$$\vec{y} := \vec{x} + \vec{z}. \qquad (99)$$

The goal is to estimate $\vec{x}$ from $\vec{y}$.

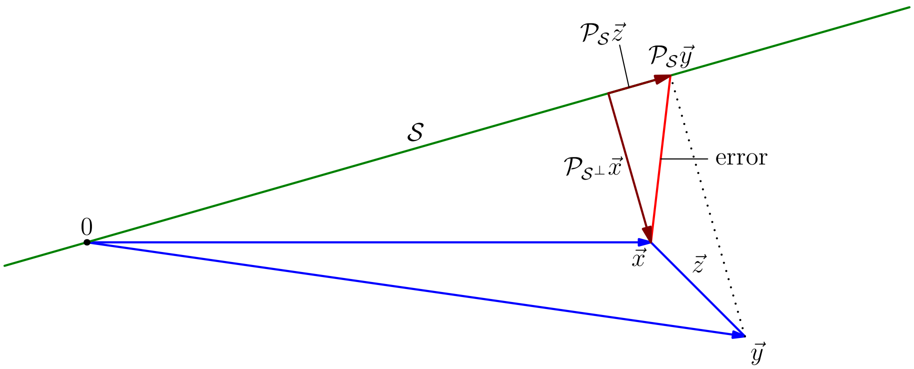

$\mathcal{P}_{\mathcal{S}}\vec{z}$
$\mathcal{P}_{\mathcal{S}}\vec{y}$
$\mathcal{S}$
$\mathcal{P}_{\mathcal{S}^\perp}\vec{x}$
error
0
$\vec{x}$
$\vec{z}$
$\vec{y}$

Figure 8: Illustration of the two terms in the error decomposition of Lemma 7.3 for a simple denoising example, where the data vector is denoised by projecting onto a 1D subspace.

In order to denoise a signal, we need to have some prior information about its structure. For instance, we may suspect that the signal is well approximated as belonging to a predefined subspace. This suggests estimating the signal by projecting the noisy data onto the subspace.

Algorithm 7.2 (Denoising via orthogonal projection)**.** *Denoising a data vector $\vec{y}$ via orthogonal projection onto a subspace $\mathcal{S}$, consists of setting the signal estimate to $\mathcal{P}_{\mathcal{S}}\,\vec{y}$, the projection of the noisy data onto $\mathcal{S}$.*

The following lemma gives a simple decomposition of the error incurred by this denoising technique, which is illustrated in Figure 8.

Lemma 7.3. *Let $\vec{y} := \vec{x} + \vec{z}$ and let $\mathcal{S}$ be an arbitrary subspace, then*

$$||\vec{x} - \mathcal{P}_{\mathcal{S}}\,\vec{y}||_2^2 = ||\mathcal{P}_{\mathcal{S}^\perp}\,\vec{x}||_2^2 + ||\mathcal{P}_{\mathcal{S}}\,\vec{z}||_2^2\,. \tag{100}$$

Proof. By linearity of the orthogonal projection

$$\begin{aligned}\vec{x} - \mathcal{P}_{\mathcal{S}}\,\vec{y} &= \vec{x} - \mathcal{P}_{\mathcal{S}}\,\vec{x} - \mathcal{P}_{\mathcal{S}}\,\vec{z} &(101)\\ &= \mathcal{P}_{\mathcal{S}^\perp}\,\vec{x} - \mathcal{P}_{\mathcal{S}}\,\vec{z}, &(102)\end{aligned}$$

so the result follows by the Pythagorean theorem. □

The error is divided into two terms. The first term is the projection of the signal onto the orthogonal complement of the chosen subspace $\mathcal{S}$. For this term to be small, the signal must be well represented by its projection onto $\mathcal{S}$. The second term is the projection of the noise onto $\mathcal{S}$. This term will be small if the noise is mostly orthogonal to $\mathcal{S}$. This makes sense: denoising via projection will only be effective if the projection preserves the signal but eliminates the noise.

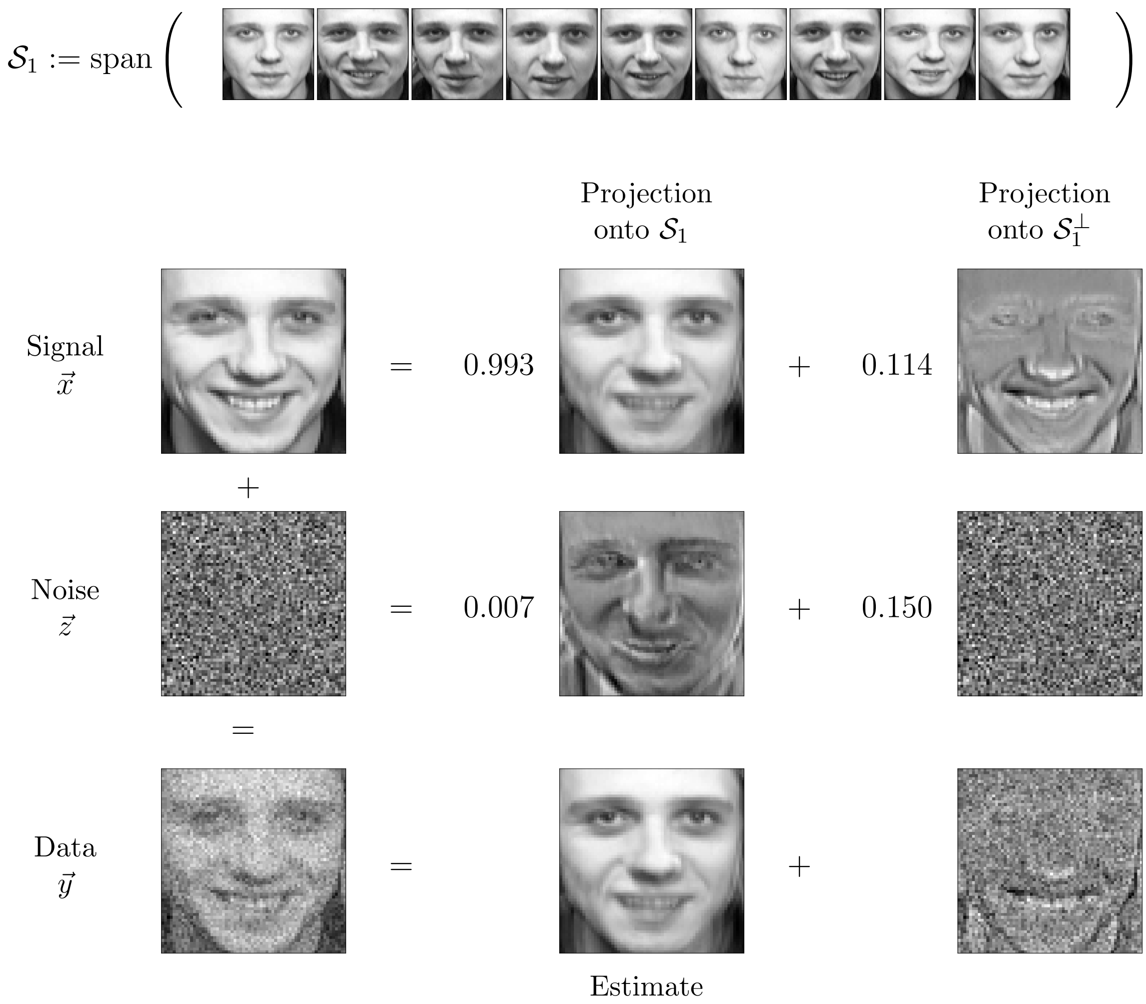


Figure 9: Denoising of the image of a face by projection onto the span of 9 other images of the same person, denoted by $\mathcal{S}_1$. The original image is normalized to have ℓ_2 norm equal to one. The noise has ℓ_2 norm equal to 0.1. The ℓ_2 norms of the projections of the original image and of the noise onto $\mathcal{S}_1$ and its orthogonal complement are indicated beside the corresponding images. The estimate is the projection of the noisy image onto $\mathcal{S}_1$.

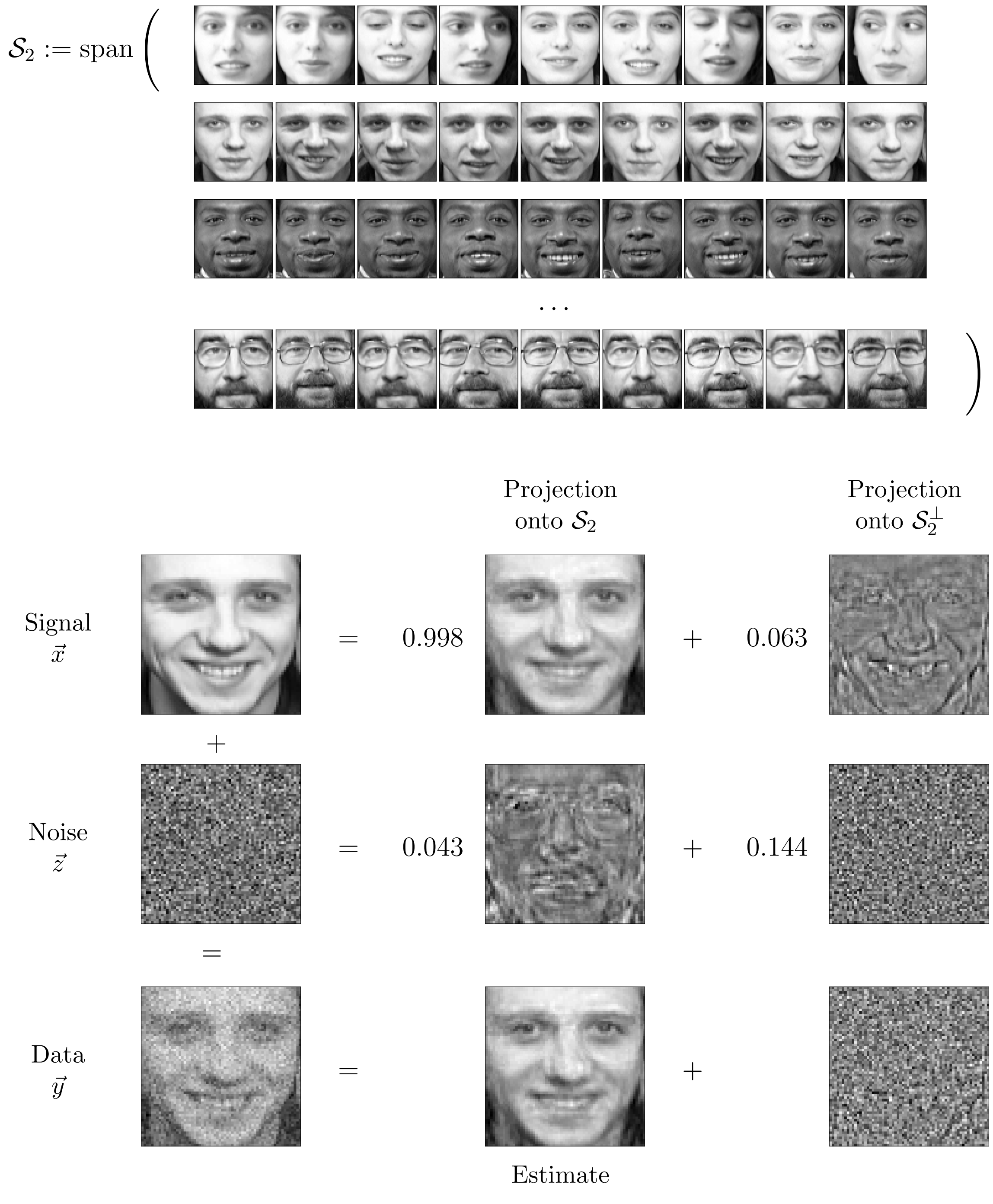

$\mathcal{S}_2 := \text{span}\Bigg($
$\dots$
$\Bigg)$
Projection onto $\mathcal{S}_2$
Projection onto $\mathcal{S}_2^\perp$
Signal $\vec{x}$
=
0.998
+
0.063
+
Noise $\vec{z}$
=
0.043
+
0.144
=
Data $\vec{y}$
=
+
Estimate

Figure 10: Denoising of the image of a face by projection onto the span of 360 other images of different people (including 9 of the same person), denoted by $\mathcal{S}_2$. The original image is normalized to have ℓ_2 norm equal to one. The noise has ℓ_2 norm equal to 0.1. The ℓ_2 norms of the projections of the original image and of the noise onto $\mathcal{S}_2$ and its orthogonal complement are indicated beside the corresponding images. The estimate is the projection of the noisy image onto $\mathcal{S}_2$.

Example 7.4 (Denoising of face images)**.** In this example we again consider the Olivetti Faces dataset[4], with a training set of 360 64 × 64 images taken from 40 different subjects (9 per subject). The goal is to denoise a test image $\vec{x}$ of the same dimensions that is not in the training set. The data $\vec{y}$ are obtained by adding noise to the test image. The entries of the noise vector z are sampled independently from a Gaussian distribution and scaled so that the signal-to-noise ratio equals 10,

$$\text{SNR} := \frac{||\vec{x}||_2}{||\vec{z}||_2} = 6.67. \tag{103}$$

We denoise the image by projecting onto two subspaces:

- $\mathcal{S}_1$: the span of the 9 images in the training set that correspond to the same subject.
- $\mathcal{S}_2$: the span of the 360 images in the training set.

Figure 9 and 10 show the results for $\mathcal{S}_1$ and $\mathcal{S}_2$ respectively. The relative ℓ_2-norm error of both estimates is:

$$\frac{||\vec{x} - \mathcal{P}_{\mathcal{S}_1}\,\vec{y}||_2}{||\vec{x}||_2} = 0.114, \tag{104}$$

$$\frac{||\vec{x} - \mathcal{P}_{\mathcal{S}_2}\,\vec{y}||_2}{||\vec{x}||_2} = 0.078. \tag{105}$$

The two estimates look very different. To interpret the results we separate the error into two components, as in Lemma 7.3. The norm of the projection of the noise onto $\mathcal{S}_1$ is smaller than its projection onto $\mathcal{S}_2$

$$0.007 = \frac{||\mathcal{P}_{\mathcal{S}_1}\,\vec{z}||_2}{||\vec{x}||_2} < \frac{||\mathcal{P}_{\mathcal{S}_2}\,\vec{z}||_2}{||\vec{x}||_2} = 0.043. \tag{106}$$

The reason is that $\mathcal{S}_1$ has lower dimension. The ratio between the two projections $(0.043/0.007 = 6.14)$ is close to the square root of the ratio of the dimensions of the subspaces (6.32). This is not a coincidence, as we will see later on. However, the projection of the signal onto $\mathcal{S}_1$ is not as close to $\vec{x}$ as the projection onto $\mathcal{S}_2$, which is particularly obvious in the lower half of the face,

$$0.063 = \frac{\left|\left|\mathcal{P}_{\mathcal{S}_2^\perp}\,\vec{x}\right|\right|_2}{||\vec{x}||_2} < \frac{\left|\left|\mathcal{P}_{\mathcal{S}_1^\perp}\,\vec{x}\right|\right|_2}{||\vec{x}||_2} = 0.114. \tag{107}$$

The conclusion is that the projection onto $\mathcal{S}_2$ produces a noisier looking image (because the noise-component of the error is larger), which nevertheless looks more similar to the original signal (because the signal-component of the error is smaller). This illustrates an important tradeoff when using projection-based denoising: subspaces with larger dimension approximate the signal better, but don't suppress the noise as much. △

[4] Available at http://www.cs.nyu.edu/~roweis/data.html

8 Proofs

8.1 Proof of Theorem 1.7

We prove the claim by contradiction. Assume that we have two bases $\{\vec{x}_1, \ldots, \vec{x}_m\}$ and $\{\vec{y}_1, \ldots, \vec{y}_n\}$ such that $m < n$ (or the second set has infinite cardinality). The proof follows from applying the following lemma m times (setting $r = 0, 1, \ldots, m-1$) to show that $\{\vec{y}_1, \ldots, \vec{y}_m\}$ spans $\mathcal{V}$ and hence $\{\vec{y}_1, \ldots, \vec{y}_n\}$ must be linearly dependent.

Lemma 8.1. *Under the assumptions of the theorem, if* $\{\vec{y}_1, \vec{y}_2, \ldots, \vec{y}_r, \vec{x}_{r+1}, \ldots, \vec{x}_m\}$ *spans* $\mathcal{V}$ *then* $\{\vec{y}_1, \ldots, \vec{y}_{r+1}, \vec{x}_{r+2}, \ldots, \vec{x}_m\}$ *also spans* $\mathcal{V}$ *(possibly after rearranging the indices* $r+1, \ldots, m$*) for* $r = 0, 1, \ldots, m-1$.

Proof. Since $\{\vec{y}_1, \vec{y}_2, \ldots, \vec{y}_r, \vec{x}_{r+1}, \ldots, \vec{x}_m\}$ spans $\mathcal{V}$

$$\vec{y}_{r+1} = \sum_{i=1}^{r} \beta_i \, \vec{y}_i + \sum_{i=r+1}^{m} \gamma_i \, \vec{x}_i, \quad \beta_1, \ldots, \beta_r, \gamma_{r+1}, \ldots, \gamma_m \in \mathbb{R}, \tag{108}$$

where at least one of the γ_j is non zero, as $\{\vec{y}_1, \ldots, \vec{y}_n\}$ is linearly independent by assumption. Without loss of generality (here is where we might need to rearrange the indices) we assume that $\gamma_{r+1} \neq 0$, so that

$$\vec{x}_{r+1} = \frac{1}{\gamma_{r+1}} \left(\sum_{i=1}^{r} \beta_i \, \vec{y}_i - \sum_{i=r+2}^{m} \gamma_i \vec{x}_i \right). \tag{109}$$

This implies that any vector in the span of $\{\vec{y}_1, \vec{y}_2, \ldots, \vec{y}_r, \vec{x}_{r+1}, \ldots, \vec{x}_m\}$, i.e. in $\mathcal{V}$, can be represented as a linear combination of vectors in $\{\vec{y}_1, \ldots, \vec{y}_{r+1}, \vec{x}_{r+2}, \ldots, \vec{x}_m\}$, which completes the proof. □

8.2 Proof of Theorem 3.10

If $||\vec{x}||_{\langle \cdot, \cdot \rangle} = 0$ then $\vec{x} = \vec{0}$ because the inner product is positive semidefinite, which implies $\langle \vec{x}, \vec{y} \rangle = 0$ and consequently that (34) holds with equality. The same is true if $||\vec{y}||_{\langle \cdot, \cdot \rangle} = 0$. Now assume that $||\vec{x}||_{\langle \cdot, \cdot \rangle} \neq 0$ and $||\vec{y}||_{\langle \cdot, \cdot \rangle} \neq 0$. By semidefiniteness of the inner product,

$$0 \leq \left|\left| \, ||\vec{y}||_{\langle \cdot, \cdot \rangle} \, \vec{x} + ||\vec{x}||_{\langle \cdot, \cdot \rangle} \, \vec{y} \right|\right|^2 = 2 \, ||\vec{x}||^2_{\langle \cdot, \cdot \rangle} \, ||\vec{y}||^2_{\langle \cdot, \cdot \rangle} + 2 \, ||\vec{x}||_{\langle \cdot, \cdot \rangle} \, ||\vec{y}||_{\langle \cdot, \cdot \rangle} \, \langle \vec{x}, \vec{y} \rangle \, , \tag{110}$$

$$0 \leq \left|\left| \, ||\vec{y}||_{\langle \cdot, \cdot \rangle} \, \vec{x} - ||\vec{x}||_{\langle \cdot, \cdot \rangle} \, \vec{y} \right|\right|^2 = 2 \, ||\vec{x}||^2_{\langle \cdot, \cdot \rangle} \, ||\vec{y}||^2_{\langle \cdot, \cdot \rangle} - 2 \, ||\vec{x}||_{\langle \cdot, \cdot \rangle} \, ||\vec{y}||_{\langle \cdot, \cdot \rangle} \, \langle \vec{x}, \vec{y} \rangle \, . \tag{111}$$

These inequalities establish (34).

Let us prove (40) by proving both implications.

($\Longrightarrow$) Assume $\langle \vec{x}, \vec{y} \rangle = - ||\vec{x}||_{\langle \cdot, \cdot \rangle} \, ||\vec{y}||_{\langle \cdot, \cdot \rangle}$. Then (110) equals zero, so $||\vec{y}||_{\langle \cdot, \cdot \rangle} \, \vec{x} = - ||\vec{x}||_{\langle \cdot, \cdot \rangle} \, \vec{y}$ because the inner product is positive semidefinite.

($\Longleftarrow$) Assume $||\vec{y}||_{\langle\cdot,\cdot\rangle}\,\vec{x} = -\,||\vec{x}||_{\langle\cdot,\cdot\rangle}\,\vec{y}$. Then one can easily check that (110) equals zero, which implies $\langle\vec{x},\vec{y}\rangle = -\,||\vec{x}||_{\langle\cdot,\cdot\rangle}\,||\vec{y}||_{\langle\cdot,\cdot\rangle}$.

The proof of (41) is identical (using (111) instead of (110)).