Vector

- Consider an n-dimensional vector, x

$$x = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} \in R^n, \text{ with } v_i \in R$$

- x is usually referred to as a length-n finite duration real-valued signal.

Basis vectors

- Given a signal $x \in R^n$, we often want to express it as a linear combination of basis vectors (or bases).

$$x \cong c_1 a_1 + c_2 a_2 + \cdots + c_m a_m,$$

where $c_i \in R$ are the coefficients and $a_i \in R^n$ are the basis vectors.

Basis vectors

- Choosing the bases or learning them from data is a core problem in many fields:
  - such as signal processing, pattern recognition, and machine learning.

Matrix form

- In matrix form, the basis vectors can be combined in a matrix A, where

  $A = [a_1, a_2, \dots, a_m]$ is an $n \times m$ matrix.
- The problem of representing a signal x in terms of the bases is therefore written as

$$x = Ac,$$

where $c = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{bmatrix}$ is the coefficient vector.

The case of $m = n$

- A typical case is $m = n$. That is, to represent an n-dimensional signal, we use n basis vectors in the same-dimensional space.
- Finding the coefficients c is equivalent to solving the linear equation systems,

$$x = Ac,$$

and the solution is simply $c = A^{-1}x$.

Orthonormal Bases

- Moreover, the bases matrix A is often to be orthonormal (for real-valued matrix), i.e.,

$$A^T A = AA^T = I,$$

where 'T' denotes the transpose and I is the identity matrix.

- or A is unitary (for complex-valued matrix), i.e., $A^* A = AA^* = I$, where * denotes the conjugate transpose.
- For orthonormal and unitary matrices, $A^{-1} = A^T$ and $A^{-1} = A^*$, respectively.

Orthonormal or unitary bases

- When the bases A is orthonormal or unitary, given an input signal x, finding the linear combination coefficients to recover x from the bases thus becomes simple,

$$x = Ac \;\; \Rightarrow \;\; c = A^T x$$

or

$$x = Ac \;\; \Rightarrow \;\; c = A^* x$$

Example of common orthonormal/unitary bases

- Fourier bases (that is the main and fundamental part of this course), useful for frequency representation.
- $A =$

$$\begin{pmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ 1 & e^{-2\pi j \frac{1}{n}} & e^{-2\pi j \frac{2}{n}} & e^{-2\pi j \frac{3}{n}} & \cdots & e^{-2\pi j \frac{n-1}{n}} \\ 1 & e^{-2\pi j \frac{2}{n}} & e^{-2\pi j \frac{4}{n}} & e^{-2\pi j \frac{6}{n}} & \cdots & e^{-2\pi j \frac{2(n-1)}{n}} \\ 1 & e^{-2\pi j \frac{3}{n}} & e^{-2\pi j \frac{6}{n}} & e^{-2\pi j \frac{9}{n}} & \cdots & e^{-2\pi j \frac{3(n-1)}{n}} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & e^{-2\pi j \frac{n-1}{n}} & e^{-2\pi j \frac{2(n-1)}{n}} & e^{-2\pi j \frac{3(n-1)}{n}} & \cdots & e^{-2\pi j \frac{(n-1)(n-1)}{n}} \end{pmatrix}$$

Example of common orthonormal/unitary bases

- Discrete Cosine Transform (DCT) bases, useful for image data compression.
- $A^T =$

$$T = \begin{bmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
\cos\frac{\pi}{16} & \cos\frac{3\pi}{16} & \cos\frac{5\pi}{16} & \cos\frac{7\pi}{16} & -\cos\frac{7\pi}{16} & -\cos\frac{5\pi}{16} & -\cos\frac{3\pi}{16} & -\cos\frac{\pi}{16} \\
\cos\frac{\pi}{8} & \cos\frac{3\pi}{8} & -\cos\frac{3\pi}{8} & -\cos\frac{\pi}{8} & -\cos\frac{\pi}{8} & -\cos\frac{3\pi}{8} & \cos\frac{3\pi}{8} & \cos\frac{\pi}{8} \\
\cos\frac{3\pi}{16} & -\cos\frac{7\pi}{16} & -\cos\frac{\pi}{16} & -\cos\frac{5\pi}{16} & \cos\frac{5\pi}{16} & \cos\frac{\pi}{16} & \cos\frac{7\pi}{16} & -\cos\frac{3\pi}{16} \\
\frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\
\cos\frac{5\pi}{16} & -\cos\frac{\pi}{16} & \cos\frac{7\pi}{16} & \cos\frac{3\pi}{16} & -\cos\frac{3\pi}{16} & -\cos\frac{7\pi}{16} & \cos\frac{\pi}{16} & -\cos\frac{5\pi}{16} \\
\cos\frac{3\pi}{8} & -\cos\frac{\pi}{8} & \cos\frac{\pi}{8} & -\cos\frac{3\pi}{8} & -\cos\frac{3\pi}{8} & \cos\frac{\pi}{8} & -\cos\frac{\pi}{8} & \cos\frac{3\pi}{8} \\
\cos\frac{7\pi}{16} & -\cos\frac{5\pi}{16} & \cos\frac{3\pi}{16} & -\cos\frac{\pi}{16} & \cos\frac{\pi}{16} & -\cos\frac{3\pi}{16} & \cos\frac{5\pi}{16} & -\cos\frac{7\pi}{16}
\end{bmatrix}$$

Example of common orthonormal/unitary bases

- Discrete Wavelet Transform (DWT) bases, useful for feature extraction.
- $A^T =$

$$
\begin{bmatrix}
c_0 & c_1 & c_2 & c_3 & & & & & & & \\
c_3 & -c_2 & c_1 & -c_0 & & & & & & & \\
 & & c_0 & c_1 & c_2 & c_3 & & & & & \\
 & & c_3 & -c_2 & c_1 & -c_0 & & & & & \\
\vdots & \vdots & & & & & \ddots & & & & \\
 & & & & & & & c_0 & c_1 & c_2 & c_3 \\
 & & & & & & & c_3 & -c_2 & c_1 & -c_0 \\
c_2 & c_3 & & & & & & & & c_0 & c_1 \\
c_1 & -c_0 & & & & & & & & c_3 & -c_2
\end{bmatrix}, \quad \text{(A1)}
$$

Example: Daubechies-4 wavelet

The case of $m < n$

- When $m < n$, the number of bases m is smaller than the signal dimension n.
- In this case, it is impossible to find the coefficients c such that $x = Ac$.
- We usually find $\hat{c}$ instead that minimizes the squared loss,

$$\hat{c} = \underset{c}{\operatorname{argmin}} \|x - Ac\|^2 .$$

Subspace

- That is,

$$\|x - (c_1 a_1 + c_2 a_2 + \cdots + c_m a_m)\|^2$$

  is minimized.
- Let the m-dimensional subspace (or hyperplane) spanned by the vectors $a_1, a_2, \dots, a_m$ be denoted as Π_{m}.
- The problem is equivalently to finding a point $\hat{p}$ lying on the subspace Π_{m}, such that $\|x - \hat{p}\| \leq \|x - p\|, \forall p \in \Pi_{\mathrm{m}}$.

Orthogonality principle

Eg., in 3D space,

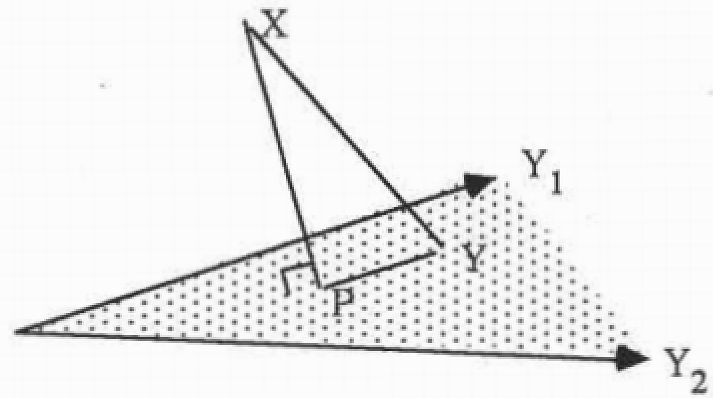


Figure 4.1: P is the projection of the vector X onto the subspace spanned by Y_1 and Y_2. That is, $X - P$ is orthogonal to all points in the subspace, and, as seen geometrically, the distance from X to P is smaller than the distance from X to any other point Y in the subspace.

- For the point X, its orthogonal projection P on the plane spanned by Y_1 and Y_2 is also its closest point in that plane.

Orthogonal projection

- Hence, the solution of

$$\underset{p \in \Pi_m}{\operatorname{argmin}} \|x - p\|^2$$

is $\hat{p} = proj_{\Pi_m}(x)$, i.e., the orthogonal projection of x on the hyperplane Π_m.

- How to find the orthogonal projection? The projection operator is computed via the pseudo inverse of A (denoted as $A^+ = \left(A^T A\right)^{-1} A^T$)

$$proj_{\Pi_m}(x) = AA^+x = A\left(A^T A\right)^{-1} A^T x$$

- The optimal coefficient vector is given by

$$\hat{c} = \left(A^T A\right)^{-1} A^T x = A^+ x$$

Orthonormal bases

- For orthonormal bases, i.e., $A^T A = I_m$ (the m-dimensional identity matrix), the pseudo inverse $A^+ = A^T$.
- So, the solution of

$$\underset{c}{\operatorname{argmin}} ||x - Ac||^2$$

is simply $\hat{c} = A^T x$ when A is orthonormal.

Learning bases from data

- In the above, we consider the problem of **finding the coefficients**, where the bases are given.
- Learning the bases from data:
  - If the bases are unknown, they could be learned from data.
  - Assume a set of data, called the training data, are available. We learn both the bases and coefficients that represent the training data.
  - Here, we do not assume that the training data have labels; thus it is a kind of **unsupervised learning**

Dictionary Learning

- Problem formulation of **dictionary learning**:

$$X \cong BC,$$

where $X \in R^{n \times N}$ is the **data matrix** consisting of N training samples with each sample an n-dimensional column vector.

- $X = [x_1\ x_2 \cdots x_N], x_i \in R^n, i = 1 \cdots N$.
- $B \in R^{n \times m}$ is the dictionary (i.e., bases matrix), $m < n$.
- $C \in R^{m \times N}$ is the coefficient matrix, $m < n$.
- Both the dictionary B and the coefficients C are **unknown**.

Matrix Factorization

$$X \cong BC$$

- The rank of BC is m and the data matrix X is of rank n for the general setting, where $m < n$.
- The above dictionary learning problem is thus analogous to the problem of low-rank matrix factorization, or simply **matrix factorization**.
- That is, factorizing the data matrix X (rank n) into two low-rank matrices, B and C (rank m).

Squared-error loss

- We often consider the squared-error loss, i.e., find both the dictionary and coefficients satisfying

$$\underset{B,C}{\operatorname{argmin}} \|X - BC\|_F^2 = \sum_{i=1}^{N} \|x_i - Bc_i\|^2 ,$$

where 'F' denotes the Frobenius norm, x_i is the i-th column of X and also the i-th training sample; c_i consists of the coefficients w.r.t. the bases B.

- The dictionary learned from the training data can then serve as the bases for future data representation.

Infinite solutions of the bases (cont.)

$$\underset{B,C}{\operatorname{argmin}} \|X - BC\|_F^2 \ \text{ with } m < n$$

- If B, C are solutions, then $\hat{B} = BQ$, and $\hat{C} = Q^{-1}C$ are also solutions, where $Q \in R^{m\times m}$ is an arbitrary matrix satisfying $QQ^{-1} = I$.

- So, the matrix factorization problem has infinite solutions.

Finding the bases: the case of $m < n$

- When the number of bases m is smaller than the data dimension n, consider the equation

$$\underset{B,C}{\operatorname{argmin}} \sum_{i=1}^{N} \|x_i - Bc_i\|^2 .$$

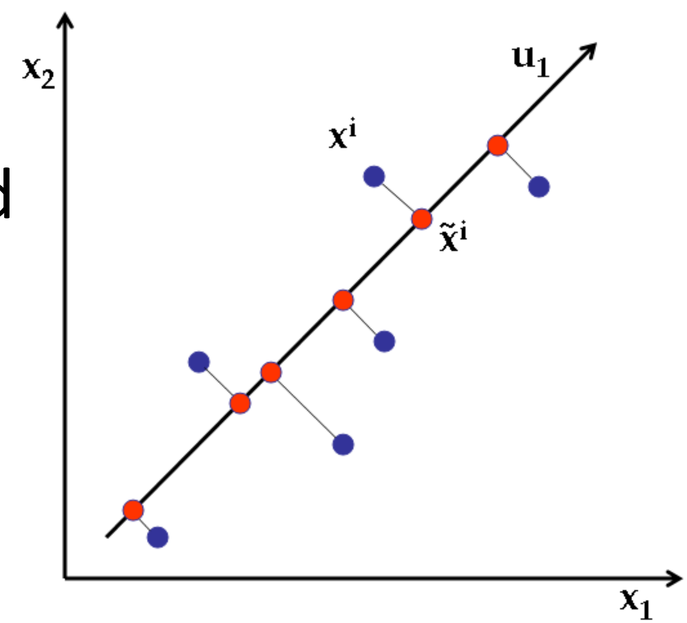
- Let $b_1, b_2, \ldots, b_m$ be the columns of B.
- Denote the m-dimensional subspace spanned by the bases $b_1, b_2, \ldots, b_m$ (B's columns) as $\Pi_{\mathrm{B;m}}$.

- According to the orthogonal principle, the problem becomes

$$\underset{B}{\operatorname{argmin}} \sum_{i=1}^{N} \left\| x_i - proj(x_i, \Pi_{B;m}) \right\|^2,$$

i.e., find the m-dimensional hyperplane $(m < n)$ such that the sum of squared distances between the training samples and their projections to that hyperplane is minimized.

Eg., the case of $n = 2$ and $m = 1$ is illustrated as



One solution of this problem can be found by **principle component analysis (PCA)**