# Representation and classification of data

Wojciech Czaja and David Levermore

January 25, 2016

# Outline



Norbert Wiener Center for Harmonic Analysis and Applications

# Data representation

The phrase "**data representation**" usually refers to the methods used to represent information stored in a computer. Computers can store many different types of information, including, but not limited to: numbers, text, graphics (many varieties), sound, etc etc. Each such type of information must be formed using appropriate codes, because in the end, all of the information stored in a computer must be represented by finite sequences of 0s and 1s. This is how computer scientists and engineers think about data representation. Moreover, from this perspective, all processing of the stored information is performed on those binary sequences, and then appropriately interpreted through the use of aforementioned codes and other software techniques.

In mathematics, the concept of data representation is much broader: it is any mathematical form that data can be described by. This includes analytic, algebraic, statistical, or geometric representations. In this class we will narrow this scope to focus on analytic/geometric forms of data representation by means of **vector spaces**.

# Linear Representation by Vector Spaces

A **vector space** is a mathematical structure formed by two types of objects: a collection of elements called vectors, and two operations on pairs of vectors: addition and scalar multiplication. As such two vectors may be added together, and any vector can multiplied by numbers, called scalars. Thus, for a vector space $X$, for any two of its elements $x, y \in X$, and any numbers $\alpha, \beta \in \mathbb{F}$ (a field of numbers), we can form the linear combinations as new elements of X:

$$\alpha \cdot x + \beta \cdot y \in X.$$

# Axioms of Vector Spaces

The operations of addition and scalar multiplication in a vector space must satisfy a number of conditions called axioms:

- Associativity of addition
- Commutativity of addition
- Identity element of addition
- Inverse elements of addition
- Compatibility of scalar multiplication with field multiplication
- Identity element of scalar multiplication
- Distributivity of scalar multiplication with respect to vector addition
- Distributivity of scalar multiplication with respect to field addition

Giuseppe Peano, Geometrical Calculus, 1888

# Diversity of representations: from bases to frames

- A **basis** is a set of linearly independent vectors which can represent every vector in a given vector space through their linear combinations.
- An **orthogonal basis** for a vector space with an inner product, is a basis with vectors which are mutually orthogonal (perpendicular). If the vectors of an orthogonal basis are or length (norm) 1, the resulting basis is an **orthonormal basis (ONB)**.
- An ONB may cease to be an ONB after even a small perturbation, or when any of its elements is removed. We seek thus representation systems with more stability.
- **Frames** were introduced by Dunford and Schaeffer in 1952.
  R. J. Dunford and A. C. Schaeffer , "A class of nonharmonic Fourier series," Trans. Amer. Math. Soc., 1952, Vol. 72, pp. 341–366.

Norbert Wiener Center
for Harmonic Analysis and Applications

# Bases

The role of a basis is to allow us to **represent** elements of the vector space in terms of sequences of scalars (numbers), which are called vector coordinates. This is an important step, because thanks to this **representation**, abstract or complicated objects obtain a uniform mathematical format. The reason for this may not necessarily be clear when we think of the most typical example of a vector space: $d$-dimensional Euclidean vector space. This is because the Euclidean space is not just a good example of a vector space, it is also a prototypical example, and last but not least - a finite dimensional vector space.
Infinite dimensional vector spaces provide us with more intriguing examples of objects, and the role of a basis which allows us to replace these complicated objects by sequences of numbers becomes much more clear.

- Vector spaces of polynomials;
- Function spaces (Lipschitz, integrable, finite energy functions, etc.)

# Example: Polynomials of 1 variable with real coefficients

The set of polynomials with coefficients in $\mathbb{R}$ is a vector space over $\mathbb{R}$, denoted typically by $P(\mathbb{R})$. Vector addition and scalar multiplication are defined in the obvious manner. If the degree of the polynomials is unrestricted then the dimension of $P(\mathbb{R})$ is infinite. If instead one restricts the polynomials to those with degree less than or equal to an integer $n$, then we obtain a vector space of dimension $n+1$, commonly denoted by $P_n(\mathbb{R})$.
One possible basis for $P(\mathbb{R})$ is the monomial basis: $1, x, x^2, x^3, \ldots$.
The coordinates of any polynomial with respect to this basis are its coefficients. Say, for $f = 3 + x^2 + 17.1x^7$, the coefficients are $3, 0, 1, 0, 0, 0, 0, 17.1, 0, \ldots$.
Now, the true advantage of using basis representations for polynomials requires us to recognize polynomials as functions, rather than just abstract algebraic objects. For this, however, we need to introduce some new concepts allowing us to measure distances in vector spaces.

# Normed vector spaces

A **norm** in a vector space $X$ is a non-negative real-valued function $x \mapsto \|x\|$, which satisfies the axioms of sublinearlity (also known as Minkowski inequality or triangle inequality), and non-degeneracy.
A normed vector space is a pair $(X, \|\cdot\|)$ where $X$ is a vector space and $\|\cdot\|$ a norm on $X$.
Every finite ($N$) dimensional vector space $X$ can be equipped with a norm. Indeed, let $\{b_1, \ldots, b_N\}$ be a basis for $X$. Then, for any $x \in X$, we can write uniquely:

$$x = \sum_{n=1}^{N} x_n b_n.$$

With this notation we can define various norms, e.g.,:

- $\|x\|_p = (|x_1|^p + \ldots + |x_N|^p)^{\frac{1}{p}}$, $p \geq 1$,
- $\|x\|_\infty = \max\{|x_1|, \ldots, |x_N|\}$.

# Euclidean norm

When we fix $p = 2$ in the definition of the $p$-norm:

$$\|x\|_2 = \sqrt{|x_1|^2 + \ldots + |x_N|^2},$$

which is the standard **Euclidean norm** for the $N$-dimensional vector space. This is the same norm that is used in the least squares optimization problems.

# Euclidean space $\mathbb{R}^N$

The **Euclidean space** $\mathbb{R}^N$ is a normed vector space consisting of $N$-tuples of real numbers (which we call vectors), with the operations of standard (coordinate-wise) vector addition and (real) scalar multiplication, and equipped with the Euclidean norm $\|x\|_2$.
One of the major advantages of the Euclidean vector space over other normed spaces is the fact that we can measure distances between vectors using the concept of an **inner product**:

$$\langle x, y \rangle = x_1 y_1 + x_2 y_2 + \ldots + x_N y_N.$$

This inner product is naturally related to Euclidean norm via the following formula:

$$\|x\|_2^2 = \langle x, x \rangle.$$

# Example: Fourier Basis for $\mathbb{R}^N$

Given $N > 0$, define the following $N \times N$ matrix:

$$F(m,n) = \frac{1}{\sqrt{N}} e^{2\pi imn/N}, \quad m,n = 0,\ldots N-1.$$

The columns (or rows) of this matrix form an orthonormal basis for the space of $N$-dimensional complex vectors $\mathbb{R}^N$. This basis is called the **Fourier basis**. And the matrix $F$ is known as the Discrete Fourier Transform. Clearly $F$ is a unitary matrix, and as such it is, in particular, invertible.