

Week 1: Data Representation*

Sergei S. Pilyugin[†]

1 Introductory notes

Welcome to the new math course! We will learn some important mathematical concepts and methods used to analyze the data. We will also learn how to use MATLAB as a computer tool for implementing our algorithms and producing cool, important, realistic, and interesting results. Or so we hope.

2 Mathematical concepts

2.1 Vectors and matrices

A *vector* is an ordered collection of numerical values. In most applications, these values are real, and the vector is called a *real vector*. But we can also speak of integer, rational, and even complex (see further) vectors. The number of entries in a vector is called the *dimension*. The standard mathematical notation for a vector is

$$\vec{x} = \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \text{ or } \vec{x} = \mathbf{x} = (x_1, x_2, \dots, x_n)$$

for columns and rows respectively. The i th entry is denoted x_i .

Matrices are rectangular tables or arrays of values. Each column and each row of a matrix represents a vector. So we can associate a matrix with a collection of individual vectors arranged in a certain order. Just like vectors, matrices can be real, integer, rational, or complex. A matrix has two dimensions: n - the number of rows, and m -the number of columns. We will refer to it as an $n \times m$ matrix. The standard matrix notation is

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{pmatrix},$$

so that a_{ij} represents the entry located at the intersection of the i th row and the j th column.

What I have given you so far represents the bookkeeper's definition of a vector and/or matrix. There are geometric (visual) and algebraic (operational) concepts that are associated with these objects as well. For instance, one of the greatest mathematical inventions was the introduction of a coordinate system by a French mathematician *Rene Descartes*

*© Sergei S. Pilyugin, Department of Mathematics, University of Florida

[†]This course is made possible by the financial support from the Howard Hughes Medical Institute.

(a.k.a. *Renatus Cartesius*, if you happen to hail from Rome). A coordinate system allows to establish a unique correspondence between locations in the ambient space (points) P and triplet of real numbers of the form (x, y, z) representing the displacement of this point from the origin $(0, 0, 0)$. A similar concept is the of driving directions: to get from the origin to the point P you need to drive x blocks in the X -direction, then y blocks along the Y -direction, and then z blocks in the Z -direction. In a *Cartesian coordinate system*, the three directions are represented by three mutually perpendicular *coordinate axes* (see Fig.). Typically, the x - and y - axes are horizontal, while the z -axis is directed vertically upward, but it is a matter of choice and convenience.

Introducing a coordinate system is one convenient way to *visualize* the data. A typical collection of measurements at discrete time points (a.k.a. the time series data) can be conveniently recast as a $n \times 2$ matrix, where the first column represents the time points, and the second column represents the collected data.

Example 1

$$D = \begin{pmatrix} t_1 & x_1 \\ t_2 & x_2 \\ \vdots & \vdots \\ t_9 & x_9 \end{pmatrix} = \begin{pmatrix} 0.0 & 1.937 \times 10^3 \\ 6.0 & 1.170 \times 10^4 \\ 12.0 & 7.127 \times 10^4 \\ 18.0 & 4.289 \times 10^5 \\ 24.0 & 2.607 \times 10^6 \\ 30.0 & 1.565 \times 10^7 \\ 36.0 & 9.506 \times 10^7 \\ 42.0 & 5.749 \times 10^8 \\ 48.0 & 3.495 \times 10^9 \end{pmatrix} \quad (1)$$

for some hypothetical experimental data (e.g., cell count) collected every 6 hours over a 48 hour period. If the same experiment is repeated several times, the observed values may change! For instance, suppose that the above measurements were obtained in January of 2007. But when the experiment was repeated in March and then in June, the following measurements were obtained:¹

| hr | Jan. 07 | Mar. 07 | Jun. 07 |
|------|---------------------|---------------------|---------------------|
| 0.0 | 1.937×10^3 | 1.936×10^3 | 1.958×10^3 |
| 6.0 | 1.170×10^4 | 1.174×10^4 | 1.198×10^4 |
| 12.0 | 7.127×10^4 | 7.127×10^4 | 7.025×10^4 |
| 18.0 | 4.289×10^5 | 4.277×10^4 | 4.241×10^4 |
| 24.0 | 2.607×10^6 | 2.585×10^6 | 2.555×10^6 |
| 30.0 | 1.565×10^7 | 1.575×10^7 | 1.609×10^7 |
| 36.0 | 9.506×10^7 | 9.493×10^7 | 9.234×10^7 |
| 42.0 | 5.749×10^8 | 5.730×10^8 | 5.835×10^8 |
| 48.0 | 3.495×10^9 | 3.464×10^9 | 3.441×10^9 |

(2)

Clearly, there is small variation for different months. How significant is this variation? Is there a *trend* that would explain such variation? Could it be due to some stochastic (random) effects? Could it be due to some measurement error? Questions like these are addressed in *statistical analysis of data* and we will return to discuss this example later.

Example 2 This example is adapted from my own research on the rates of immune cell

¹Notice that I dropped the parentheses of the matrix to save room and added the header line to indicate which column corresponds to which experiment.

turnover. The following table was published in [1]

| | hours | | | |
|----------|--------------------|--------------------|--------------------|---------------------|
| $x_n(t)$ | 12 | 30 | 72 | 192 |
| 0 | 7.38×10^4 | 7.07×10^4 | 1.77×10^4 | 0.0 |
| 1 | 0.0 | 0.64×10^4 | 6.10×10^4 | 0.29×10^4 |
| 2 | 0.0 | 0.0 | 6.58×10^4 | 5.71×10^4 |
| 3 | 0.0 | 0.0 | 1.28×10^4 | 19.97×10^4 |
| 4 | 0.0 | 0.0 | 0.0 | 18.83×10^4 |
| 5 | 0.0 | 0.0 | 0.0 | 7.99×10^4 |
| 6+ | 0.0 | 0.0 | 0.0 | 4.00×10^4 |

(3)

It represents the result of the following experiment. A number of immune cells were labeled with a fluorescent dye (called CFSE) and transferred into a live animal (mouse). These cells continued to turn over (divide and die) in the new host. At discrete time points after transfer, the labeled cell counts were collected. The main property of CFSE is that it is distributed equally between two daughter cells upon cell division. Hence, the fluorescence decreases by a factor of 2 from one generation of cells to the next. This allows to determine how many times a given cell has divided, and count how many cells have undergone a given number of divisions n . In the table (3), $x_n(t)$ represents the number of cells that have undergone exactly n divisions by time t . Each of these data points represents averaged counts from three different mice, so clearly we would expect some statistical variation between different data points, and we will discuss it later. But the data represented in this example can be further manipulated to extract some additional information. For instance, we may calculate the total labeled cell counts at different times by adding the entries within each column. The result would be

| | | | | | |
|-----------------|--------------------|--------------------|---------------------|---------------------|-----|
| t | 12 | 30 | 72 | 192 | |
| $\sum_n x_n(t)$ | 7.38×10^4 | 7.71×10^4 | 1.573×10^5 | 5.679×10^5 | (4) |

which is yet another matrix.

2.2 Vector spaces

If we consider the set of all possible vectors of a given length, and describe to how add, subtract, and multiply these vectors by scalar quantities (numbers), the result is called a *vector space*.

Definition A vector space V over real numbers is a set of elements with operations of addition and scalar multiplication that obey the following rules: for any three vectors $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$ and any two real numbers $a, b \in \mathbb{R}$ ²

1. The vector sum $\mathbf{u} + \mathbf{v} \in V$;
2. The vector sum is commutative $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$;
3. The vector sum is associative $(\mathbf{u} + \mathbf{v}) + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w})$;
4. There exists a *zero vector* $\mathbf{0} \in V$ such that $\mathbf{0} + \mathbf{u} = \mathbf{u}$ for all $\mathbf{u} \in V$;

²This notation seems heavy only the first time you see it. The symbol \mathbb{R} traditionally denotes the set of all real numbers. The symbol \in reads "belongs to", so the statement $a, b \in \mathbb{R}$ in plain English reads "a and b are real numbers".

5. For every $\mathbf{u} \in V$, there exists its *negative* counterpart $-\mathbf{u} \in V$ such that $\mathbf{u} + (-\mathbf{u}) = \mathbf{0}$;
6. The scalar product $a\mathbf{u} \in V$;
7. $1\mathbf{u} = \mathbf{u}$ for all $\mathbf{u} \in V$;
8. Scalar multiplication is associative $a(b\mathbf{u}) = (ab)\mathbf{u}$;
9. Scalar multiplication obeys distributive laws $(a + b)\mathbf{u} = a\mathbf{u} + b\mathbf{u}$ and $a(\mathbf{u} + \mathbf{v}) = a\mathbf{u} + a\mathbf{v}$.

Sounds pretty stringent, does not it? It is a staple question on most math exams to recite the definition of the vector space. That's because it is one of the most fundamental mathematical concepts. Before we go further, let me give you an example of a vector space. It is called the n -dimensional real vector space denoted by $V = \mathbb{R}^n$. \mathbb{R}^n consists of all possible n -tuples of the form $\mathbf{x} = (x_1, x_2, \dots, x_n)$. Yes, we started by calling these guys vectors, and that's exactly what they are. The vector operations work as follows. For two vectors $\mathbf{x} = (x_1, x_2, \dots, x_n)$ and $\mathbf{y} = (y_1, y_2, \dots, y_n)$, their sum is calculated entry-wise

$$\mathbf{x} + \mathbf{y} = (x_1 + y_1, x_2 + y_2, \dots, x_n + y_n),$$

and so is scalar multiplication

$$k\mathbf{x} = (kx_1, kx_2, \dots, kx_n).$$

Given enough patience, one can check that all properties of the vector space are satisfied. What I want to do instead, is to translate these properties into plain English. For definiteness, let's talk about the space we live in - \mathbb{R}^3 .³ Recall our discussion of the Cartesian coordinates. A three-dimensional vector $\mathbf{x} = (x_1, x_2, x_3)$ is interpreted as a displacement from the origin of the coordinate system $\mathbf{0} = (0, 0, 0)$. Hence, it gives a simple directional instruction: follow the straight arrow to get from $\mathbf{0}$ to \mathbf{x} . The sum of two vectors $\mathbf{x} + \mathbf{y}$ represents the composite instruction: start at the origin, first follow \mathbf{x} , then follow \mathbf{y} . The place where you end up is the sum $\mathbf{x} + \mathbf{y}$ of two displacements. Now, it is only natural that $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$ because we follow two different sides of a parallelogram but the order is switched. Nonetheless, we always end up at the opposite vertex of the parallelogram. Now, we can go back and reinterpret all other properties of a vector space in this fashion. For instance, the instruction $\mathbf{0}$ is "don't move". The instruction $-\mathbf{x}$ is "retrace \mathbf{x} in the opposite direction". The instruction $2\mathbf{x}$ is "follow \mathbf{x} twice", etcetera.

Even the single vector instruction $\mathbf{x} = (x_1, x_2, x_3)$ can be thought of as a composite. It can be decomposed into the sum of three "orthogonal" instructions

$$\bullet \quad \mathbf{x} = (x_1, 0, 0) + (0, x_2, 0) + (0, 0, x_3),$$

or equivalently,

$$\mathbf{x} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + x_3\mathbf{e}_3,$$

where $\mathbf{e}_1 = (1, 0, 0)$, $\mathbf{e}_2 = (0, 1, 0)$, and $\mathbf{e}_3 = (0, 0, 1)$ are the vectors comprising the so-called *standard or canonical basis* of \mathbb{R}^3 . Each of these vectors carries the simple instruction "move one unit of length along the respective coordinate axis".

All these notions can be easily generalized to other dimensions of \mathbb{R}^n .

Exercise Can you list the elements of the canonical basis of \mathbb{R}^5 ?

³*Disclaimer:* Depending on who you talk to, people may claim that we live in spaces of higher dimension, that are "curved" and "convoluted", and aren't even vector spaces. They may as well be right, but we'll stick to the classical viewpoint for now.

2.3 Functions

As we have seen so far, a matrix can be interpreted as a collection of data points. In various situations, we know which entries represent the independent variable(s) (such as time, for instance) and which entries represent the dependent variable(s) (such as counts and other measurements). The logical distinction is that the independent variable should determine the value of dependent variable, and the other way around. For instance, it is possible to have the same cell count at two distinct time points, but it is impossible to have two distinct cell counts at the same time.⁴ By asking the question: *How does the number of cells change with time?*, we attempt to establish a correspondence between these quantities and ultimately draw some quantitative conclusions from the data. Ideally, we would like to be able to *extrapolate* the values of the dependent variable for *arbitrary values of the independent variable*. All experimental data consist of at most finite number of measurements. What we are asking here is some rule that would help us calculate the value of the dependent variable for any of the infinitely many possible values of the dependent variable. Such rules are called *functions*.

Definition A function is a rule that establishes a correspondence between dependent and independent variables. The independent variable is called the *argument*, and the dependent variable is called the *value* of the function. The notation is $y = f(x)$ where f is the function, x is the argument, and y is the value. The set of all admissible arguments x for $f(x)$ is called the *domain* of the function f , and the set of all values $y = f(x)$ produced by varying x over the entire domain of f , is called the *range* of the function f . Each argument x uniquely determines the corresponding value $y = f(x)$, but the same value can be produced by more than one argument.

2.4 Distance functions

By inspecting two vectors, we can easily determine if they are identical or not. And if they are distinct, they must correspond to distinct points in space. How far are these points from each other? Can we talk about two points (or vectors) being "closer" in some sense than some prescribed value? Can we talk about a length of a vector? For instance, if we treat the data in Table 2 as vectors, can we make a statement of the sort: two experiments are "really close" and the third one is an "outlier"?

Definition. A *distance function* or a *metric* on a set A is a quantitative measure of the *distance* $d(a, b)$ between two different objects a and b that are elements of the set A . The distance function must satisfy the following rules:

1. The distance between any two objects a and b must be nonnegative, that is $d(a, b) \geq 0$;
2. The distance between a and b is zero if and only if $a = b$, that is $d(a, b) = 0$ iff $a = b$;
3. For any two objects a and b , the distances from a to b and from b to a must be identical, that is $d(a, b) = d(b, a)$;
4. For any three objects a, b, c , the distance from a to c cannot exceed the sum of the distances from a to b and from b to c , that is

$$d(a, c) \leq d(a, b) + d(b, c).$$

⁴I should point out that a priori we cannot decide which variables must be designated as independent, and which as dependent, without knowing what these variables actually represent.

The standard distance function in the n -dimensional space \mathbb{R}^n is called the Euclidean distance. For two points $X = (x_1, x_2, \dots, x_n)$ and $Y = (y_1, y_2, \dots, y_n)$ the Euclidean distance between X and Y is defined as

$$d(X, Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2}.$$

In the case of real numbers ($n = 1$), this distance function simplifies to

$$d(x, y) = \sqrt{(x - y)^2} = |x - y|.$$

Euclidean distance may be modified by assigning positive weights $w_i > 0$ to each of the squared differences, and the result is another distance function

$$d_w(X, Y) = \sqrt{w_1(x_1 - y_1)^2 + w_2(x_2 - y_2)^2 + \dots + w_n(x_n - y_n)^2}.$$

A very nice property of the Euclidean distance is its "compatibility" with the structure of the vector space \mathbb{R}^n . Euclidean distance $d(X, Y)$ is really the length of the straight segment connecting the points X and Y . We can translate it into the notion of the *length* or the *magnitude* or the *norm* for vectors by defining the Euclidean vector norm

$$\|\mathbf{x}\| = d(\mathbf{0}, \mathbf{x}) = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2}.$$

Definition. A norm $|\cdot|$ on the vector space V is a scalar function such that

1. $\|\mathbf{x}\| \geq 0$ for all $\mathbf{x} \in V$;
2. $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$;
3. $\|k\mathbf{x}\| = |k|\|\mathbf{x}\|$ for all $\mathbf{x} \in V$ and all scalars k ;
4. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in V$.

The last property is conventionally called *the triangle inequality* because it states exactly that: the length of any side of a triangle cannot exceed the sum of the lengths of the other two sides.

2.4.1 Triangle inequality for the Euclidean distance

To show that the Euclidean distance satisfies the triangle inequality, we will first prove the *Cauchy-Schwartz inequality*: for any two vectors \mathbf{x} and \mathbf{y} , the absolute value of their dot product does not exceed the product of their Euclidean norms:

$$|\mathbf{x} \cdot \mathbf{y}| \leq \|\mathbf{x}\|\|\mathbf{y}\|,$$

where the dot product $\mathbf{x} \cdot \mathbf{y}$ is defined as

$$\mathbf{x} \cdot \mathbf{y} = x_1y_1 + x_2y_2 + \dots + x_ny_n.$$

Proof: The dot product of any vector with itself is always non-negative, hence the inequality

$$0 \leq (\mathbf{x} + t\mathbf{y})(\mathbf{x} + t\mathbf{y}) = \|\mathbf{x}\|^2 + 2t(\mathbf{x} \cdot \mathbf{y}) + t^2\|\mathbf{y}\|^2$$

holds for all values of t . This implies that the discriminant of this quadratic in t is non-positive, that is,

$$4(\mathbf{x} \cdot \mathbf{y})^2 - 4\|\mathbf{x}\|^2\|\mathbf{y}\|^2 \leq 0.$$

Rewriting this inequality as

$$(\mathbf{x} \cdot \mathbf{y})^2 \leq \|\mathbf{x}\|^2 \|\mathbf{y}\|^2$$

and taking square roots on both sides, we arrive at the Cauchy-Schwartz inequality. \diamond

For any two vectors \mathbf{x} and \mathbf{y} , we have that

$$\|\mathbf{x} + \mathbf{y}\|^2 = (\mathbf{x} + \mathbf{y}) \cdot (\mathbf{x} + \mathbf{y}) = \mathbf{x} \cdot \mathbf{x} + 2\mathbf{x} \cdot \mathbf{y} + \mathbf{y} \cdot \mathbf{y}.$$

By the Cauchy-Schwartz inequality, the middle term satisfies

$$2\mathbf{x} \cdot \mathbf{y} \leq 2|\mathbf{x} \cdot \mathbf{y}| \leq 2\|\mathbf{x}\|\|\mathbf{y}\|,$$

hence

$$\|\mathbf{x} + \mathbf{y}\|^2 \leq \|\mathbf{x}\|^2 + 2\|\mathbf{x}\|\|\mathbf{y}\| + \|\mathbf{y}\|^2 = (\|\mathbf{x}\| + \|\mathbf{y}\|)^2.$$

Taking square roots of both sides, we obtain the triangle inequality

$$\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|.$$

2.4.2 Other examples of distance functions

There are other important examples of distances. Some of these do not have a clear geometric interpretation. For example, in information theory and linguistics, the *Hamming distance* between two symbolic strings (words) of equal length is the number of positions for which the corresponding symbols are different [3]. The Hamming distance between the words "biology" and "zoology" is 2 (because they only differ in the first two letters), while the distance between "biology" and "physics" is 7 (the maximal possible ever: the letters in all positions differ!). In the world of the Hamming distance, one would have to conclude that biology is much closer to zoology than it is to physics.

The notion of the Hamming distance is limited to the strings of the same length. There is a natural extension of this distance function to the strings of arbitrary lengths which is called the *Levenshtein* or *edit distance* [4]. It is used in spell checkers (e.g. to suggest the closest words that are grammatically correct) and in the analysis of DNA strings. The edit distance between two strings is the minimal number of elementary editing operations required to transform one string into the other. The elementary editing operations include deletion, insertion, and substitution of a single character. If two strings have equal length, then the edit distance is the same as the Hamming distance (if we allow only substitutions to be used in the editing process).

Here is an algorithm for computing the edit distance. Suppose we wish to calculate the edit distance between the strings $S = s_1s_2\dots s_n$ and $T = t_1t_2\dots t_m$.

1. We begin by forming an $(m + 1) \times (n + 1)$ matrix D initially containing all zeros, that is, $d_{ij} = 0$ for $i = 0, 1, 2, \dots, m$ and $j = 0, 1, 2, \dots, n$.
2. Assign values $d_{0j} = j$, $j = 1, 2, \dots, m$, and $d_{i0} = i$, $i = 1, 2, \dots, n$.
3. Starting from the second top row and going from left to right, we fill in the values d_{ij} according to the following rule:

$$\begin{array}{ccc} A = d_{i-1,j-1} + cost & & B = d_{i-1,j} + 1 \\ & \searrow \downarrow & \\ C = d_{i,j-1} + 1 & \rightarrow & d_{ij} = \min(A, B, C) \end{array},$$

where $cost = 0$ if $t_i = s_j$ and $cost = 1$ if $t_i \neq s_j$.

4. After completing a row, move to the row below, until the bottom row is reached.
5. The value d_{mn} is the edit distance between the strings S and T .

Example Here is an example of computing the edit distance between the strings *mathematics* and *physics*.

| | | | | | | | | | | | | |
|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|-----------|
| | s_j | <i>m</i> | <i>a</i> | <i>t</i> | <i>h</i> | <i>e</i> | <i>m</i> | <i>a</i> | <i>t</i> | <i>i</i> | <i>c</i> | <i>s</i> |
| t_i | <u>0</u> | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| <i>p</i> | 1 | <u>1</u> | <u>2</u> | <u>3</u> | <u>4</u> | <u>5</u> | 6 | 7 | 8 | 9 | 10 | 11 |
| <i>h</i> | 2 | 2 | 2 | 3 | 3 | 4 | <u>5</u> | 6 | 7 | 8 | 9 | <u>10</u> |
| <i>y</i> | 3 | 3 | 3 | 3 | 4 | 4 | 5 | <u>6</u> | 7 | 8 | 9 | <u>10</u> |
| <i>s</i> | 4 | 4 | 4 | 4 | 4 | 5 | 5 | 6 | <u>7</u> | 8 | 9 | 9 |
| <i>i</i> | 5 | 5 | 5 | 5 | 5 | 5 | 6 | 6 | 7 | <u>7</u> | 8 | 9 |
| <i>c</i> | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 7 | 7 | 8 | <u>7</u> | 8 |
| <i>s</i> | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 8 | 8 | 8 | <u>7</u> |

The main idea behind the algorithm described above is that each entry d_{ij} corresponds to the minimal number of editing operations required to transform the substring $T_i = t_1 t_2 \dots t_i$ into the substring $S_j = s_1 s_2 \dots s_j$. Initially, an empty string can be transformed into a string of k characters by using exactly k additions (Step 2). Explanation of Step 3:

- If we can transform T_i into S_{j-1} in $d_{i,j-1}$ operations, then we can transform T_i into S_j in $C = d_{i,j-1} + 1$ operations by simply adding the character s_j to S_{j-1} .
- If we can transform T_{i-1} into S_j in $d_{i-1,j}$ operations, then we can transform T_i into S_j in $B = d_{i-1,j} + 1$ operations by simply deleting the character t_i from T_i .
- If we can transform T_{i-1} into S_{j-1} in $d_{i-1,j-1}$ operations, then we can transform T_i into S_j in $A = d_{i-1,j-1} + cost$ operations by replacing the character t_i with s_j if they are different ($cost = 1$).
- The minimal number of operations required to transform T_i into S_j is the minimum of the three quantities: $d_{ij} = \min(A, B, C)$.

Exercises

1. (less serious): Compute the pairwise edit distances between the words "zoology", "physics", and "chemistry", and compare the relative ratios of these distances to the Euclidean distances between the zoology, physics, and chemistry buildings at UF.
2. (more serious): Write a MATLAB code that would compute the edit distance between two given strings.
3. (serious) In the definition of the edit distance, we assume that all three editing operations have the same unit cost. Suppose that performing a single addition is associated with cost $\alpha > 0$, the cost of a single deletion is $\delta > 0$, and the cost of a single replacement is $\rho > 0$. Define the *weighted edit distance* as the minimal cost of transforming one string into another. How should you modify the algorithm to calculate this new distance?
4. Suppose that $\alpha = 1.5$, $\delta = 1$, and $\rho = 2.5$. Calculate the weighted edit distance between the words "physics" and "mathematics".

2.4.3 Set-theoretic distance between finite sets

In mathematics, a set is defined as a collection of elements. For instance, the set $A = \{a, b, c\}$ consists of exactly three elements a, b, c . Moreover, as a collection, the set $A' = \{b, c, a\}$ is the same as A , that is *the order of elements in the set is irrelevant*. We use the notation $a \in A$ to indicate that *the element a belongs to the set A* . A set is finite if it consists of finitely many elements. The *cardinality* of a finite set is defined as the number of elements contained in this set. For example, if $A = \{a, b, c\}$ then $|A| = 3$. The empty set $\emptyset = \{\}$ containing no elements has cardinality 0.

We say that A is a subset of B if B contains all elements of A (and possibly more). The standard notation is $A \subset B$. Clearly, $A \subset B$ implies that $|A| \leq |B|$.

We define the following operations on sets:

- The union $A \cup B$ of the sets A and B is the set containing all elements that belong to either A or B .
- The intersection $A \cap B$ of the sets A and B is the set containing all elements that belong to both A and B .
- The set difference $A \setminus B$ of the sets A and B is the set containing all elements that belong to A but do not belong to B .
- Finally, the symmetric set difference $A \Delta B$ is defined as $A \Delta B := (A \setminus B) \cup (B \setminus A)$. In other words, $A \Delta B$ consists of elements that belong to either A or B , but not both.

Observe that the set difference $A \Delta B$ is indeed symmetric, that is, $A \Delta B = B \Delta A$.

Example If $A = \{1, 2, 3, 4\}$ and $B = \{1, 2, 4, 8, 16\}$ then

$$\begin{aligned} A \cup B &= \{1, 2, 3, 4, 8, 16\}, \\ A \cap B &= \{1, 2, 4\}, \\ A \setminus B &= \{3\}, \\ B \setminus A &= \{8, 16\}, \\ A \Delta B &= \{3, 8, 16\}. \end{aligned}$$

Definition. If the sets A and B are finite, then so is the set $A \Delta B$. We define the set-theoretic distance between A and B as $d(A, B) := |A \Delta B|$.

Does this function fit the definition of a distance? Yes, it does. The first three properties are more or less immediate. The quantity $d(A, B)$ is certainly nonnegative. Since $A \Delta A = \emptyset$, we have $d(A, A) = 0$. On the other hand if $d(A, B) = 0$ that means that each element of A belongs to B , and each element of B belongs to A , hence the sets A and B are equal. We also have that $d(A, B) = d(B, A)$ because $A \Delta B = B \Delta A$.

Verifying the triangle inequality is a bit more interesting. To show that

$$d(A, B) \leq d(A, C) + d(C, B)$$

amounts to proving that

$$|A \Delta B| \leq |A \Delta C| + |C \Delta B|.$$

Consider any element $x \in A \setminus B$, that is, x belongs to A but does not belong to B . If $x \in C$, then $x \in C \Delta B$. If $x \notin C$, then $x \in A \Delta C$. We are forced to conclude that each element

of $A \setminus B$ is counted exactly once in either $|A \Delta C|$ or in $|C \Delta B|$. Similarly, each element of $B \setminus A$ is counted exactly once in either $|A \Delta C|$ or in $|C \Delta B|$. Hence, the triangle inequality holds. \diamond

Exercises.

1. Construct an example of three distinct sets A, B, C such that $d(A, B) < d(A, C) + d(C, B)$.
2. Construct an example of three distinct sets A, B, C such that $d(A, B) = d(A, C) + d(C, B)$.
3. Verify that $A \Delta B = (A \cup B) \setminus (A \cap B)$.
4. Verify the following *counting formula*

$$|A \cup B| = |A| + |B| - |A \cap B| = |A \cap B| + d(A, B).$$

Interestingly, it turns out that the Hamming distance is a special case of the set-theoretic distance. Indeed, each string of characters can be identified with a special type of set. Such set consists of pairs (position, character). For instance, the string "florida" corresponds to the set

$$F = \{(1, f), (2, l), (3, o), (4, r), (5, i), (6, d), (7, a)\},$$

that is, at the first position is occupied by "f", the second by "l", etcetera. Similarly, the string "georgia" corresponds to the set

$$G = \{(1, g), (2, e), (3, o), (4, r), (5, g), (6, i), (7, a)\}.$$

The Hamming distance between the strings "florida" and "georgia" equals 4. The set-theoretic distance $d(F, G) = 8$, that is exactly twice as much!

Exercises.

1. Compute the set-theoretic distance between the strings "physics" and "mathematics" and compare it to the edit distance between these strings.
2. Argue that for two strings of the same length, their set-theoretic distance equals twice their Hamming distance.
3. In general, the set-theoretic distance is not necessarily even. For instance, calculate $d(A, B)$ for $A = \{1, 2, 3, 4\}$ and $B = \{1, 2, 3, 4, 5, 6, 7\}$.
4. Using the counting formula (see above), show that the set-theoretic distance between two finite sets of the same cardinality (i.e. $|A| = |B|$) is always even. Hint: $|A \Delta B| = 2 * |A| - 2 * |\cap B|$.

2.5 Complex numbers

Contrary to the common student myth, complex numbers were introduced to make our life easier. It turns out that many calculations can be performed much smoother and almost painlessly if we use complex numbers instead of reals. As we shall see, the real numbers actually live inside the set of complex as one big happy family.

Definition The set of complex numbers \mathbb{C} consists of all elements of the form $z = a + ib$ where $a, b \in \mathbb{R}$. Here, the symbol i represents the *imaginary unit*, that is, a number such that $i^2 = -1$. The real coefficients $a = \operatorname{Re}(z)$ and $b = \operatorname{Im}(z)$ are called the real and imaginary parts of z respectively. The arithmetic operations on the set \mathbb{C} are defined as follows

$$\begin{aligned} \text{addition} & \quad (a + ib) + (c + id) = (a + b) + i(c + d); \\ \text{multiplication} & \quad (a + ib)(c + id) = (ac - bd) + i(ad + bc). \end{aligned}$$

Equivalently, we can carry out the symbolic operation, replace all i^2 with -1 , and collect the i terms. Just like with nonzero reals, each nonzero complex number z has a unique inverse $\frac{1}{z}$ such that $z\frac{1}{z} = \frac{1}{z}z = 1$. Multiplication by the inverse is the same as division. It is easy to check that

$$\frac{1}{z} = \frac{a - ib}{a^2 + b^2}.$$

For ease of notation, we call $\bar{z} = a - ib$ the *complex conjugate* of z . Complex numbers are graphically represented as points in (a, b) plane. Since both the addition and multiplication by a real number are performed entry-wise, the set of complex numbers \mathbb{C} is a two dimensional real vector space with the basis 1 and i . The multiplication of complex numbers corresponds closely to the cross-product of three-dimensional real vectors. The Euclidean norm in this space is used to define the absolute value of a complex number

$$|z| = \sqrt{a^2 + b^2}.$$

It is easy to check that $|z_1 z_2| = |z_1| |z_2|$ and that $|z| = |\bar{z}|$.

Now, let me illustrate the use of complex numbers by solving a quadratic equation $z^2 + 2z + 2 = 0$. Substituting the coefficients $a = 1$, $b = 2$, $c = 2$ into the formula⁵

$$z_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a},$$

we find that

$$z_{1,2} = \frac{-2 \pm \sqrt{4 - 8}}{2} = -1 \pm \sqrt{-1} = -1 \pm i.$$

Clearly, this quadratic has no real roots, but it does have two complex roots. According to the Fundamental Theorem of Algebra, any polynomial equation of degree n with complex coefficients is *guaranteed* n roots if we count their multiplicities. In particular, this applies to polynomial equations with real coefficients. We will recall this fact every time we search for *eigenvalues*.

Another way to represent a complex number, is to express it in *polar form*. Polar coordinates in the (x, y) plane are introduced as follows

$$x = r \cos \theta, \quad y = r \sin \theta,$$

where $r = \sqrt{x^2 + y^2} \geq 0$ and $\tan \theta = y/x$ with $0 \leq \theta < 2\pi$. In polar coordinates, the points in the plane are located according to the *azimuthal angle* θ measured counterclockwise from the positive x -axis, and the distance r from the origin. If we transfer this idea to complex numbers, we get

$$a = \operatorname{Re}(z) = r \cos \theta, \quad b = \operatorname{Im}(z) = r \sin \theta,$$

which implies

$$r = \sqrt{a^2 + b^2} = |z|, \quad z = |z|(\cos \theta + i \sin \theta).$$

The latter expression is called the polar form of z , and the angle θ is called the argument of z denoted $\theta = \arg(z)$.

⁵All must remember!

2.5.1 Euler's formula

The Euler's formula

$$e^{\alpha+i\beta} = e^{\alpha}(\cos \beta + i \sin \beta).$$

is of fundamental importance. It relates trigonometric functions to the exponential function via complex numbers. It would not be far from the truth to say that Euler's formula "has it all". The formula is derived by combining two facts:

Fact 1. Using the power series definition of the exponential function

$$e^z := \sum_{n=0}^{\infty} \frac{1}{n!} z^n,$$

and the binomial formula

$$(x + y)^n = \sum_{k=0}^n \binom{n}{k} x^k y^{n-k} = \sum_{k=0}^n \frac{n!}{k!(n-k)!} x^k y^{n-k},$$

we can write

$$e^{x+y} = \sum_{n=0}^{\infty} \frac{1}{n!} (x + y)^n = \sum_{n=0}^{\infty} \sum_{k=0}^n \frac{1}{k!(n-k)!} x^k y^{n-k}.$$

Observe that each product of the form $x^n y^m$ has the coefficient $\frac{1}{n!m!}$. On the other hand, we have that

$$e^x e^y = \left(\sum_{n=0}^{\infty} \frac{1}{n!} x^n \right) \left(\sum_{m=0}^{\infty} \frac{1}{m!} y^m \right) = \sum_{n,m} \frac{1}{n!m!} x^n y^m.$$

Hence, the coefficients of the power series for e^{x+y} and $e^x e^y$ coincide. The convergence argument allows to establish the identity

$$e^{x+y} = e^x e^y.$$

Fact 2. Using the same power series definition of the exponential function, we have that

$$e^{i\beta} = \sum_{n=0}^{\infty} \frac{1}{n!} (i\beta)^n = 1 + i\beta + \frac{1}{2!} i^2 \beta^2 + \frac{1}{3!} i^3 \beta^3 + \frac{1}{4!} i^4 \beta^4 + \frac{1}{5!} i^5 \beta^5 + \dots$$

Since

$$i^2 = i^6 = i^1 0 = \dots = -1, \quad i^3 = i^7 = \dots = -i, \quad i^4 = i^8 = \dots = 1,$$

the following pattern emerges

$$e^{i\beta} = \left(1 - \frac{1}{2!} \beta^2 + \frac{1}{4!} \beta^4 - \dots \right) + i \left(\beta - \frac{1}{3!} \beta^3 + \frac{1}{5!} \beta^5 - \dots \right).$$

Recognizing the real part as the power series for $\cos \beta$ and the imaginary part as the power series for $\sin \beta$, we have the identity

$$e^{i\beta} = \cos \beta + i \sin \beta.$$

Combining this identity with Fact 1, we obtain the Euler's formula.

Among other things, Euler's formula includes all you ever need to remember about trigonometry. Using Euler's formula, we can express complex numbers in *exponential form*

$$z = |z|e^{i\theta}.$$

Let me illustrate how Euler's formula works with complex multiplication. Suppose that we want to multiply two complex numbers z_1 and z_2 that are given in polar form. Here is what we do

$$z_1 z_2 = |z_1|e^{i\theta_1}|z_2|e^{i\theta_2} = |z_1||z_2|e^{i\theta_1}e^{i\theta_2} = |z_1||z_2|e^{i(\theta_1+\theta_2)}.$$

I used the main property of the exponential function here. The end result is that

$$|z_1 z_2| = |z_1||z_2|, \quad \arg(z_1 z_2) = \arg(z_1) + \arg(z_2).$$

As far as the division is concerned, we have that

$$|z_1/z_2| = |z_1|/|z_2|, \quad \arg(z_1/z_2) = \arg(z_1) - \arg(z_2).$$

There is one thing I need to clarify here: when we add/subtract two arguments in the above expressions, we may end up with a number which is either greater than 2π or less than 0. In this case, we need to add/subtract 2π to bring the argument back into the interval $0 \leq \theta < 2\pi$.

Exercise

- Calculate $(2 - i)(3 + 4i)(1 + i) - (5 - 7i)$ and $\frac{2+i}{1+2i}$;
- Express the numbers i and $1 + i$ in polar and exponential form
- Using the exponential form to compute $(1 + i)^8$.

2.6 Graphical representation of data (graphs, plots, histograms)

A graph is used to depict the shape of a given function $y = f(x)$. It is the collection of all points of the form $(x, f(x))$ over the domain of the function f . The graphing procedure is implemented in MATLAB by the command `plot`. Although the graph of a function f may contain infinitely many points, MATLAB can only plot a finite subset of those points and connect them by straight segments. The basic syntax of the `plot` command is as follows:

```
>>x=xmin:xstep:xmax; % x varies from xmin to xmax with a step xstep
y=f(x); % for each value of x, the value of y is determined by y=f(x)
plot(x,y) % all pairs (x,y) are plotted and joined by straight segments
```

Histograms (bar charts) and pie charts are used to visualize proportions. A histogram of a vector (x_1, x_2, \dots, x_n) consists of n vertical bars of same width whose heights are given by x_i . A pie chart of a nonnegative vector (x_1, x_2, \dots, x_n) is given by a disk partitioned into concentric sectors whose relative areas are given by fractions

$$f_i = \frac{x_i}{x_1 + x_2 + \dots + x_n}.$$

Larger fractions correspond to larger slices of the pie. Since MATLAB cannot produce sectors of zero area, all zero entries in the vector are discarded. The basic usage of `bar` and `pie` is illustrated below.

3 MATLAB notes

3.1 Entering vectors and matrices

To enter a vector in MATLAB, we can simply type

```
>>x=[1 2 3]
```

or

```
>>x=[1; 2; 3]
```

The difference is that the first sequence creates a row vector, and the second sequence creates a column vector. MATLAB is kind enough to tell us that by responding with

```
x=
    1    2    3
```

or

```
x=
     1
     2
     3
```

respectively.

There are various ways to enter and/or create matrices (and vectors) in MATLAB. Matrices can be entered in MATLAB manually. For instance, typing

```
>>A=[1 2 3; 4 5 6; 7 8 9]
```

in the command line will store a new matrix A in MATLAB memory. The stored matrix has the form

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}.$$

Matrices can also be imported from an external file using the `load` command. For instance, we can create a text file `A.dat` in the current working directory that contains the following text:

```
1 2 3
4 5 6
7 8 9
```

generated with spaces and line breaks. Then, typing

```
>>load A.dat
```

in the command line would instruct MATLAB to import the contents of that file and save it as a matrix called A . Matrices can also be created using matrix functions and M-files (more about these later).

3.2 Producing plots

In its simplest form, the `plot(y)` command produces a plot of the entries of a vector `y` against the indices of these entries. If two vectors are specified as arguments, `plot(x,y)` produces a plot of the entries of `y` against the entries of `x`. Try the following:

```
>>x=[1 2 3 4 5]; plot(x)
```

and compare the results with

```
>>x=[5 3 4 2 1]; plot(x)
```

The results are shown in Figure~\ref{fig1}.

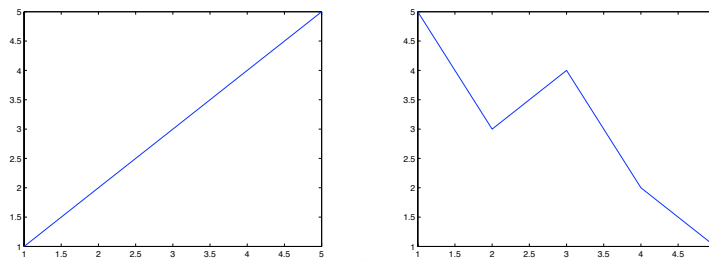


Figure 1: Compare the two `x` plots

Now try this

```
>>x=0:0.1:1; y=x.^2; plot(x,y)
```

The result is shown in Figure 2. To produce multiple plots, you simply list all pairs of

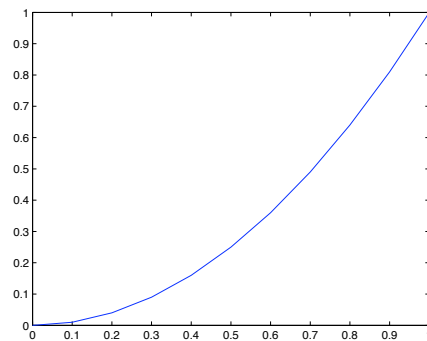


Figure 2: The `(x,y)` plot

vectors to be plotted:

```
>>x1=0:0.1:1; y1=x1.^2; x2=-1:0.1:1; y2=2-abs(x2); plot(x1,y1,x2,y2)
```

To add a legend to this figure, follow up with the command

```
legend('y1','y2')
```

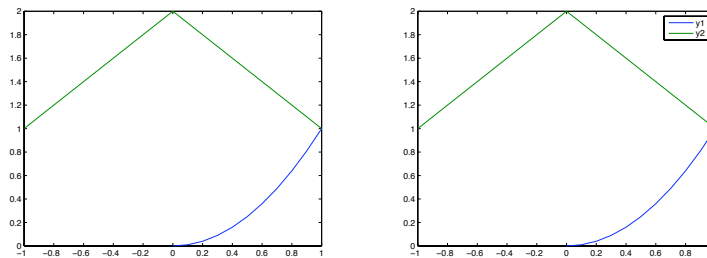


Figure 3: The multiple plot with and without the legend

The results are shown in Figure 3. You may have noticed a peculiar dot I used in the definition $y=x.^2$. This is because I am instructing MATLAB to square the variable x which is technically a matrix. Although we have not discussed it yet, but there are different way to square matrices. Here, I am specifically instructing MATLAB to take squares of *individual entries* of x , that is, perform the indicated operation element-by-element.

3.3 Producing histograms and bar charts

To produce a histogram/bar chart, type the following

```
>>x=[1 3 4 2 3]; bar(x)
```

and to produce a pie chart, type

```
>>x=[1 3 4 2 3]; pie(x)
```

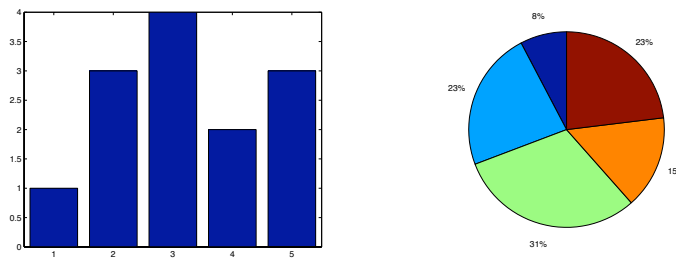


Figure 4: The bar chart (histogram) and the pie chart

The results are shown in Figure 4. To improve the appearance of these diagrams, you can create similar 3D plots by using `bar3` and `pie3`.

Exercise 3 Create a .dat file containing the information in Table (3). Use `load` to import these data to MATLAB, and create a 3D bar chart from the imported data.

References

- [1] S. S. Pilyugin, V. V. Ganusov, K. Murali-Krishna, R. Ahmed, and R. Antia (2003), *The rescaling method for quantifying the turnover of cell populations*, J. Theor. Biol., 225, 275-283. [PDF Source](#)

- [2] See more at http://en.wikipedia.org/wiki/Complex_number
- [3] See more at http://en.wikipedia.org/wiki/Hamming_distance
- [4] See more at http://en.wikipedia.org/wiki/Levenshtein_distance

www.math.ufl.edu/~pilyugin