Unit 2: Matrices and vectors

Dr.P.M.Bajracharya

September 17, 2021

Matrices and vectors are the main objects of study in Linear Algebra which, in turn, is widely-used in machine learning, data analysis, and data science. We said that a matrix is more than just a thing A_{ij} with two subscripts. Essentially, it is a thing that is characterized by a certain set operations that can be performed on it. It is these operations that make matrices so useful.

1 Ways of viewing a matrix

There are **four** major ways to view a matrix. They are

As a set of points. In this case, m points, each of which are in $\mathbb{R}^n$ can be written as an $m \times n$ matrix.

A very simple example of this is one point on the real line, i.e., on $\mathbb{R}$.

Slightly less simple examples of this are one point on the plane, two points on the line, two points on the plane. All of these simple examples can be written as $1 \times 1, 2 \times 1, 1 \times 2$, and 2×2 matrices, but the examples are so simple that this is not usually done.

As a linear transformation. Matrix-vector products, and matrix-matrix products, e.g., the functions

$$Ax \to y,\ Ay \to z,\ \text{etc.}$$

can be viewed as defining linear functions taking vectors x as inputs and returning vectors y as outputs. If the matrix is $m \times n$, we will see that this involves linear functions from $\mathbb{R}^n$ to $\mathbb{R}^m$.

This is a generalization of the familiar $y = ax$, the equation of a line with slope a going through the origin, where y, a, x are all real numbers, and we are thinking of y as a simple function of x. In this case, given the value of a, one may want to compute y as a function of the input x.

As a quadratic form. This is s a generalization of $y = ax^2$, the familiar equation for a simple quadratic form involving one variable $x \in \mathbb{R}$. For example, if we have a symmetric matrix A and a vector x, then we can define the transpose of x^T and then extend the matrix-vector multiplication to write $y = x^T Ax$, which is a quadratic form in $x = (x_1, ..., x_n)^T \in \mathbb{R}^n$.

We know that the one-dimensional quadratic form $y = ax^2$ has very different properties depending on whether $a > 0$ or $a = 0$ or $a < 0$. We will see that for higher-dimensional quadratic forms, i.e., involving vectors $x \in \mathbb{R}^n$ rather than numbers $x \in \mathbb{R}^2$, there is a rich array of properties that are possible.

In terms of linear equations. Given a vector y and a matrix A, we might want to find a vector x such that $y = Ax$. This could be a vector where the random walk process leaves the vectors unchanged, if such a vector exists, but it could also be all sorts of other things. This has similarities with the linear transformation perspective, except when one thinks about matrices in terms of linear equation solving one typically knows y and A and one wants to find x.

When restricted to the familiar case, one wants to solve for x in the equation $y = ax$, and we know that the solution is $x = a^{-1}y = y/a$, assuming that $a \neq 0$. Again, when we generalize to higher-dimensional linear equations, a rich array of properties are possible.

While the linear transformation and linear equation perspectives both involve expressions of the form $y = Ax$, there is an important difference in terms of what we assume we know.

- For the linear transformation perspective, we know A and x and we are interested in computing y. This is sometimes known as a **forward process**.
- For the linear equation perspective, we know A and y and we are interested in computing x. This is sometimes known as a **backward** or **inverse process**, since at least formally this involves the inverse of A.

Inverse processes tend to be more difficult. Most linear algebra classes start with this inverse process of solving linear equations as the main motivation. We will get to that eventually as an important application, but we won't start with it.

Viewing matrices from each of these perspectives has its advantages and disadvantages. In particular, the last perspective (i.e., in terms of linear equations) is most common and is the most important for many traditional applications of linear algebra, e.g., in engineering, physics, and scientific computation, and so it gets the greatest emphasis in many traditional linear algebra classes. This approach leads to an emphasis on things like *Reduced Row Echelon Forms*, *QR decompositions*, and so on.

The other perspectives are more useful in *data science*, *data analysis*, and *machine learning*, and so in this class we will be much more interested in those perspectives. In particular, we will start by viewing matrices as consisting of a bunch of points in a high-dimensional vector space, in which case the linear transformation and quadratic form perspectives will arise naturally.

2 Matrices, vectors, and $\mathbb{R}^n$

2.1 What is $\mathbb{R}^n$?

- **One-dimensional Euclidean space** $\mathbb{R}$: This is just the set of real numbers. This is typically visualized as the **real number line**. The elements of $\mathbb{R}$ can be added and multiplied, and both of these operations have an inverse, except for multiplication by 0 which doesn't have an inverse, etc. By that last comment, we mean that $y = ax$ has the solution $x = a^{-1}y$ for all $a \neq 0$; and if $a = 0$, then there is in general no solution.

- **Two-dimensional Euclidean space** $\mathbb{R}^2$: This is just the set of ordered pairs of real numbers. That is,

  $$x \in \mathbb{R}^2 \Leftrightarrow x = (x_1, x_2),$$

  where $x_1, x_2 \in \mathbb{R}$.

  This is visualized as the **plane**. We will explore operations like *addition* and *multiplication* for points of the plane.

- **Three-dimensional Euclidean space** $\mathbb{R}^3$: This is the set of ordered triples of real numbers. That is,

  $$x \in \mathbb{R}^3 \Leftrightarrow x = (x_1, x_2, x_3),$$

  where $x_1, x_2, x_3 \in \mathbb{R}$.

  This is visualized as the **space**. We will explore operations like *addition* and *multiplication* for points of the space.

- **Four-dimensional Euclidean space** $\mathbb{R}^4$: This is the set of ordered quadruples of real numbers. That is,

  $$x \in \mathbb{R}^4 \Leftrightarrow x = (x_1, x_2, x_3, x_4),$$

  where $x_1, x_2, x_3, x_4 \in \mathbb{R}$.

This is harder to visualize, but we will be able to define operations on it in a way that cleanly generalizes familiar operations on $\mathbb{R}^2$ and $\mathbb{R}^3$, and thus we will be able to talk about the properties of points in $\mathbb{R}^4$.

- **n-dimensional Euclidean space** $\mathbb{R}^n$: This is the set of ordered n-tuples of real numbers. That is,

  $$x \in \mathbb{R}^n \Leftrightarrow x = (x_1, x_2, ..., x_n),$$

  where $x_1, x_2, ..., x_n \in \mathbb{R}$.

  This is even harder to visualize, but here too we will be able to define operations on it in a way that cleanly generalizes familiar operations on $\mathbb{R}^2$ and $\mathbb{R}^3$.

  Importantly, as n gets larger, e.g., for 10 or 20, many properties of $\mathbb{R}^n$ are very different than the corresponding properties of $\mathbb{R}^2$ or $\mathbb{R}^3$, but a few properties do generalize cleanly. We can use these latter properties to analyze data that are modeled as points in $\mathbb{R}^n$ and thus understand better data modeled as matrices and graphs.

These examples of $\mathbb{R}^n$ will be key players for us in the linear algebra part of the class, and in the probability part of the class we will see that there are strong not-immediately-obvious connections between them and discrete probability.

Some very basic properties of $\mathbb{R}^n$

As a first step toward the goal of defining operations that cleanly generalize familiar operations from $\mathbb{R}^2$ and $\mathbb{R}^3$ to $\mathbb{R}^n$, let's ask what are some of the most basic intuitions that we have about points on a line or on a plane or in three dimensional space? Here are several:

- In $\mathbb{R}$:

  Size or absolute value. Given a real number x, some are larger, and some are smaller, and this is quantified by the absolute value, which is typically denoted $|x|$. Given this absolute value, we can define a distance between two points, i.e., given $x, y \in \mathbb{R}$, we can define

  $$\mathrm{dist}(x, y) = |x - y|.$$

  In many cases, one often has the intuition that points that have smaller distance between them are more alike.

- In $\mathbb{R}^2$:

  Size or norm. Given a point on the plane, let's reference it to an origin, in which case we will call it a vector. (We'll be more precise on this below, but this is analogous to choosing an origin for $\mathbb{R}$.) Then, we might want to measure the size of the vector. As opposed to the line, where any two numbers are comparable in terms of being larger or smaller, that is not true on the plane. But we can associate a real number to any point on the plane which measures it's size. There are actually many ways to do this, and we will discuss several in detail below, but the simplest version of this should be familiar as the Euclidean norm, which we will denote as $\|\cdot\|_2$. Here, given $x = (x_1, x_2) \in \mathbb{R}$, we define that

$$\|x\|_2 = \sqrt{x_1^2 + x_2^2}.$$

For this and other norms, we can define a distance between two vectors as the norm of the difference, i.e., given $x, y \in \mathbb{R}$, we can define

$$\mathrm{dist}(x, y) = \|x - y\|_2.$$

This generalizes the notion of size or absolute value to $\mathbb{R}$. Note that since there are other notions of size, and since we can associate a distance with a norm as the norm of the vector difference, there are other notions of distance, even for $\mathbb{R}^2$.

One reason this is of interest is that one often has the intuition that points that have smaller distance between them are more alike.

Angle. In addition, for $x, y \in \mathbb{R}^2$, by using this particular norm (i.e., the Euclidean norm), we can define an angle θ between x and y as

$$\cos(\theta) = \frac{x^T y}{\|x\|_2 \|y\|_2} = \frac{x_1 y_1 + x_2 y_2}{\|x\|_2 \|y\|_2}.$$

- In $\mathbb{R}^3$:

  Norms, distances, and angles. Here too, the notion of norms, distances, and angles arise. Other notions arise too (e.g., curls and divergences, if you have heard of those; if not, okay, since we won't be interested in them); and, while they are important in physical applications, they generalize less well to very high dimensional spaces. They are less important in data science, and so we won't go into them. We will see that norms and angles do generalize to $\mathbb{R}^2$, for $n > 3$. These are very useful in data science, and so we will spend a great deal of time on them.

2.2 Some very basic subsets of $\mathbb{R}^n$

Understanding the properties of high-dimensional spaces and data modeled by high-dimensional spaces can be difficult, i.e., it can be difficult to have an intuition of what $\mathbb{R}^n$ "looks like," and it is often best to do this by understanding the properties of very simple subsets of $\mathbb{R}^n$.

For the moment, let's restrict ourselves to $\mathbb{R}^2$. Here are three very simple subsets of the plane.

Unit ball. This is the set of points (x_1, x_2) defined by the equation: $x_1^2 + x_2^2 \leq 1$. In other words, this is the set

$$\left\{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2 : \; x_1^2 + x_2^2 \leq 1 \right\}.$$

The generalization to $\mathbb{R}^n$ is immediate:

$$\left\{ x \in \mathbb{R}^n : \; \sum_1^n x_i^2 \leq 1 \right\}.$$

Positive orthant. An orthant is an n-dimensional generalization of quadrant or octant. A positive orthant is defined as follows:

$$\left\{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2 : \; x_1 \geq 0, x_2 \geq 0 \right\}.$$

The generalization to $\mathbb{R}^n$ is immediate:

$$\{x \in \mathbb{R}^n : \; x_i \geq 0, i = 1, 2, ..., n\}.$$

Probability Simplex. A **probability simplex** is a mathematical space where each point represents a probability distribution between a finite number of mutually exclusive events. Each event is often called a category (trial) and we use the variable K to denote the number of categories.

A point on a probability simplex can be represented by K non-negative numbers that add up to 1. Here are some examples:

A point in a simplex where $K = 2 : (0.6, 0.4)$
A point in a simplex where $K = 3 : (0.1, 0.1, 0.8)$
A point in a simplex where $K = 6 : (0.05, 0.2, 0.15, 0.1, 0.3, 0.2)$

When $K = 2$, this space is a line, when $K = 3$ it is a triangle, and when $K = 4$ it is a tetrahedron. In each case, the simplex is a $K-1$ dimensional object. The requirement that the numbers sum to 1 reduces the dimensionality by 1. A probability simplex is the set of points (x_1, x_2) defined by

$$\left\{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2 : \; x_1 + x_2 = 1, x_1 \geq 0, x_2 \geq 0 \right\}.$$

The generalization to $\mathbb{R}^n$ is immediate:

$$\left\{ x \in \mathbb{R}^n : \sum_1^n x_i = 1, x_i \geq 0, \right\}.$$

3 Measuring the size of vectors in $\mathbb{R}^n$

3.1 Norms

A norm is a function that measures the size of things. We used the Euclidean norm above (since it is probably most familiar and since it provides us the notion of an angle which provides important connections to geometry), but there are many other norms that are of interest, and we will discuss a few of the most important.

Norms in $\mathbb{R}^2$.

Perhaps the most well-known norm is the so-called **Euclidean norm**, also known as the l_2 norm. For a vector $x \in \mathbb{R}^2$,

$$L_2 : \|x\|_2 = \left(\sum_1^2 x_i^2 \right) = (x_1^2 + x_2^2)^{1/2}.$$

Here are the key properties of the L_2 norm that justify its interpretation as the length or magnitude of a vector:

Positivity/non-negativity. The length of a vector is always greater than 0, unless it is the zero vector, in which case its length is equal to 0.

Positive scalability. The length of product of a vector and a scalar real number is the length of the vector multiplied by the absolute value of the scalar.

Triangle inequality. The length of one side of a triangle is not larger than sum of the lengths of the other two sides of that triangle.

Each of these three properties should be easily-understood in terms of the familiar geometric properties of the two-dimensional plane. Indeed, it is these three properties that motivated the definition of a norm, which is a generalization of the Euclidean norm in $\mathbb{R}^2$.

Definition (Normed space)**.** *A mapping $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ is called a* ***norm****, if it satisfies the following properties:*

N1. $\|x\| \geq 0$ *for all* $x \in X$*.*

N2. $\|x\| = 0$ *if and only if* $x = 0$*,*

N3. $\|\alpha x\| = |\alpha|\|x\|$ *for all* $x \in X$ *and for all* $\alpha \in \mathbb{C}$*,*

N4. $\|x + y\| \leq \|x\| + \|y\|$*.*

We have defined L_2 norm on $\mathbb{R}$. But we can define a norm on $\mathbb{R}$ in many ways. The most popular norms on $\mathbb{R}$ are L_1 and L_∞. They are defined as follows:

$$L_1 : \|x\|_1 = \sum_1^2 |x_i| = |x_1| + |x_2|.$$

$$L_\infty : \|x\|_\infty = \max_{i\in\{1,2\}} |x_i| = \max\{|x_1|, |x_2|\}.$$

These provide two different ways other than the L_2 norm to measure the size of a vector.

Example 1.

Let $x = (1, 2)$. Then

- $\|x\|_1 = 1 + 2 = 3$
- $\|x\|_2 = (1 + 4)^{1/2} \approx 2.24$
- $\|x\|_\infty = \max\{1, 2\} = 2.$

This example illustrates the more general property that

$$\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1.$$

This means that the numerical value of the size of a vector depends on the norm used to measure its size.

Fact. For all of these inequalities, there exists vectors $x \in \mathbb{R}^2$ such that the inequality holds as an equality.

Question. Can you give an example of a vector $x \in \mathbb{R}^2$ such that each of the above inequalities holds as an equality?

Example 2.

Let $x = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}$. Then

- $\|x\|_1 = |\cos\theta| + |\sin\theta|.$
- $\|x\|_2 = \cos^2\theta + \sin^2\theta.$
- $\|x\|_\infty = \max\{\cos\theta, \sin\theta\}.$

Similar statements hold for the vector $x = \begin{pmatrix} -\sin\theta \\ \cos\theta \end{pmatrix}$.

These properties are not peculiar to $\mathbb{R}^2$. Consider, e.g., a vector in $\mathbb{R}^3$.

Example 3.

If we take $x = (1, 2, 3)$, then we have

- $\|x\|_1 = 6$.
- $\|x\|_2 = (1 + 2 + 3)^{1/2} \approx 3.71$.
- $\|x\|_\infty = 3$.

Again, this illustrates that $\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1$.

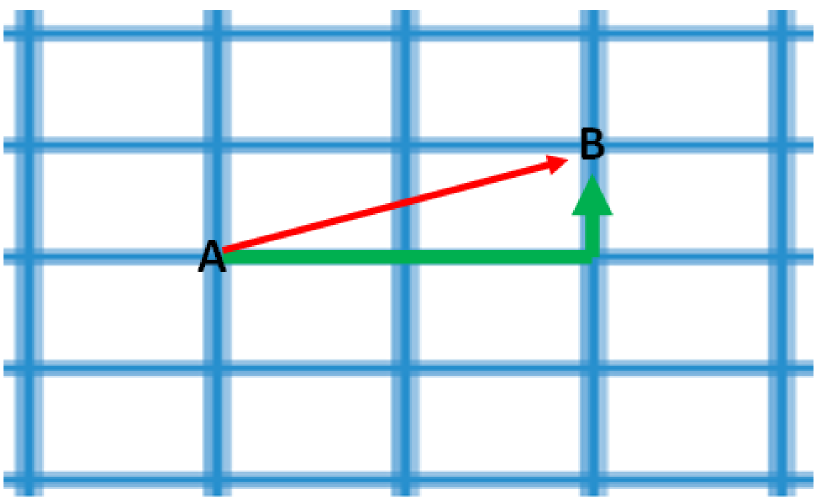


Figure 1: Roadmap of a city

Example 4.

While we have defined L_1, L_2 and L_3 norms, it might be less easy to see – initially at least – why these other notions of size or norm are interesting. To see one example of this, consider the above figure, which illustrates a grid-like street map of a city. Perhaps a bird can ignore the streets and fly directly from point A to point B, but to walk from point A to point B means that one must go along streets, horizontally and vertically, as opposed to along the hypotenuse. This is captured by the L_1 norm, and thus this notion of norm is a more meaningful notion of distance when walking around city blocks such as those shown in the above figure.

Remark. *The L_2 norm is important because it is associated with angles and dot products (which we will get later). This means that when we are measuring these quantities on the plane, there are two complementary perspectives: algebra; and geometry. Understanding these connections forms most of the basis for understanding high dimensional data. This connection will generalize to high dimensional Euclidean spaces and will be very important in what we will talk about in this course.*

3.2 Norms in $\mathbb{R}^n$.

We now generaliaze L_2 or L_3 norms to $\mathbb{R}^n$ for a positve integer $n \geq 1$. A point in $\mathbb{R}^n$ is an ordered list of n numbers, that is, n- tuple of real numbers and can be represented as

$$(x_1, x_2, ..., x_n) \text{ or } \begin{pmatrix} x_1 \\ x_2 \\ ... \\ x_n \end{pmatrix}.$$

This is often interpreted as a point representing position.

We now generaliaze L_1, L_2 and L_3 norms to $\mathbb{R}^n$ for a positve integer $n \geq 1$.

$$L_1 : \|x\|_1 = \sum_1^n |x_i| = |x_1| + |x_2| + ... + |x_n|.$$

$$L_2 : \|x\|_2 = \left(\sum_1^n x_i^2\right) = (x_1^2 + x_2^2 + ... + x_n^2)^{1/2}.$$

$$L_\infty : \|x\|_\infty = \max_{i \in \{1,2,...,n\}} |x_i| = \max\{|x_1|, |x_2|, ..., |x_n|\}.$$

These three norms (there are also many others we won't discuss) provide three different ways to measure the "size" of a vector in $\mathbb{R}^n$. Which norm is better is not the problem. Instead, you should ask what does a particular norm capture and when is it useful. All of these norms are useful in different situations, and we will see several examples of this later.

Observe that all three norms reduce to the same thing in $\mathbb{R}$. More precisely, if $x \in \mathbb{R}$, then

$$\|x\|_1 = \|x\|_2 = \|x\|_\infty = |x|.$$

Example 5.

Let $x = (1, 1, ..., 1) \in \mathbb{R}^n$. Then

$$\begin{aligned}\|x\|_1 &= n,\\ \|x\|_2 &= \sqrt{n},\\ \|x\|_\infty &= 1.\end{aligned}$$

Example 6.

Let $x = (1, 0, ..., 0) \in \mathbb{R}^n$. Then

$$\begin{aligned}\|x\|_1 &= 1,\\ \|x\|_2 &= 1,\\ \|x\|_\infty &= 1.\end{aligned}$$

These two examples are interesting since they represent two extreme cases—one in which the "mass" of the vector is spread out evenly over all the components of that vector, and the other in which the "mass" of the vector is spread out very unevenly.

Relationship between L_1, L_2 and L_3 norms

The L_1, L_2 and L_3 norms are related to each other as follows:

$$\|x\|_2 \le \|x\|_1 \le \sqrt{n}\|x\|_2, \tag{1}$$
$$\|x\|_\infty \le \|x\|_2 \le \sqrt{n}\|x\|_\infty, \tag{2}$$
$$\|x\|_\infty \le \|x\|_1 \le n\|x\|_\infty. \tag{3}$$

<u>Hints for the proof</u>: Let $M = \|x\|_\infty$. Then

$$\left(\sum_1^n |x_i|\right)^2 \ge \sum_1^n |x_i|^2 \ge M^2,$$

$$\sum_{1}^{n} |x_i| \leq nM.$$

Remark. *Whenever you see an inequality such as one of (1), (2) and (3), you should ask whether it is "tight," i.e., whether it can be "saturated." By that, we mean does there exist a vector $x \in \mathbb{R}^n$ such that the inequality holds as an equality. If the answer is yes, then it is a much more informative inequality than if the answer is no.*

Remark. *Note that the L_1, L_2 and L_3 norms are the special cases of the L_p norm defined as follows:*

$$\|x\|_p = \left(\sum_{1}^{n} |x_i|^p \right)^{1/p}$$

Balls

Having defined the size (norm) of a vector, we can now define a *ball* in $\mathbb{R}^n$ which is the generalization of a ball in two or three dimensions.

Definition. *Let $x_0 \in \mathbb{R}^n$ with a norm $\|\cdot\|$. Then*

- ***Unit sphere:***

$$S(x_0) = \{x \in \mathbb{R}^n : \ \|x - x_0\| = 1\}.$$

- ***Unit ball:***

$$B(x_0) = \{x \in \mathbb{R}^n : \ \|x - x_0\| < 1\}.$$

- ***Closed unit ball:***

$$\overline{B}(x_0) = \{x \in \mathbb{R}^n : \ \|x - x_0\| \leq 1\}.$$

Compare L_1, L_2 and L_∞ unit balls in the plane and space.

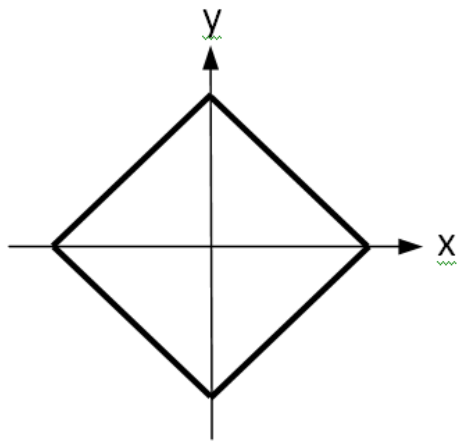


Figure 2: L_1 ball in the plane

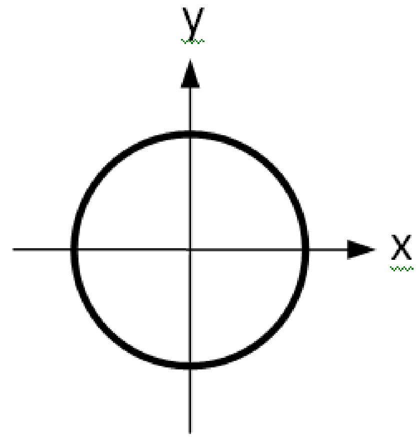


Figure 3: L_2 ball in the plane

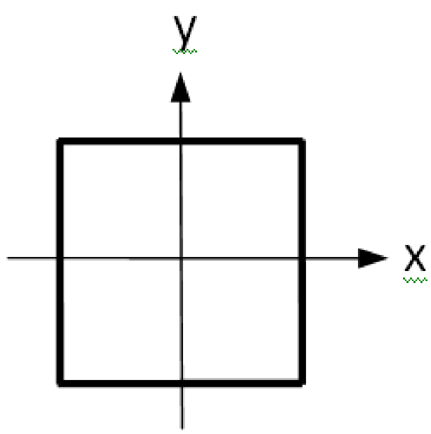


Figure 4: L_∞ ball in the plane

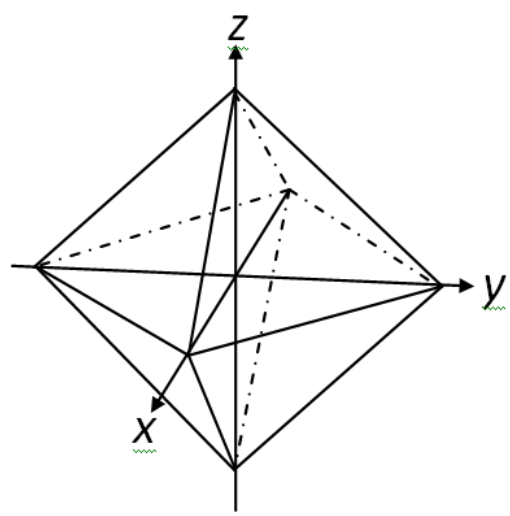
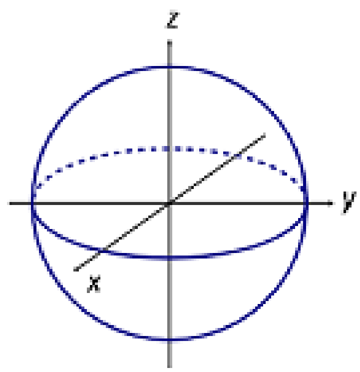


Figure 5: L_1 ball in the space



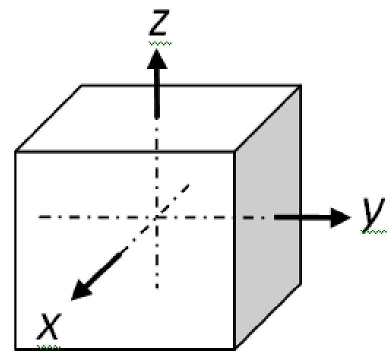


Figure 7: L_∞ ball in the space

Figure 6: L_2 ball in the space

	L_1 ball	L_2 ball	L_∞ ball
$n=1$	$2r$	$2r$	$2r$
$n=2$	$\frac{(2r)^2}{2!}$	πr^2	$(2r)^2$
$n=3$	$\frac{(2r)^3}{3!}$	$\frac{4}{3}\pi r^3$	$(2r)^3$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Table 1: Volumes of balls of radius r for different norms in different dimensions.

Three dimensional space looks different than two dimensional space (e.g., you can go up and down, in addition to left and right and back and forth), which in turn looks different than one-dimensional space (where you can only go left and right—or, of course, back and forth, or up and down, depending on what way you turn your head, but not two or three of those pairs at once). These differences are important since they inform your intuition about what is possible, and this intuition is important since it helps you to make decisions about what methods to use, when those methods might be returning a cray answer, etc.

High-dimensional vectors, i.e., vectors in $\mathbb{R}^n$, when n is 10^2 or 10^6 or even when it is only 10, have very different properties than vectors in $\mathbb{R}^n$, when $n = 1$ or 2 or 3. How can we begin to understand this?

To get an understanding of the geometry of $\mathbb{R}^n$, let's ask:

What is the difference between an L_2 ball in $\mathbb{R}$ versus $\mathbb{R}^2$ versus $\mathbb{R}^3$?

More specifically,

What about the ratio of the L_2 ball and L_∞ ball ?

Recall figures of balls, and the table for the volumes of balls of radius r for different norms in different dimensions.

- $\mathbb{R}$: $\quad \dfrac{L_2\text{-ball}}{L_\infty\text{-ball}} = \dfrac{2r}{2r} = 1$
- $\mathbb{R}^2$: $\quad \dfrac{L_2\text{-ball}}{L_\infty\text{-ball}} = \dfrac{\pi r^2}{(2r)^2} = \dfrac{\pi}{4} = 0.78$
- $\mathbb{R}^3$: $\quad \dfrac{L_2\text{-ball}}{L_\infty\text{-ball}} = \dfrac{4\pi r^3/3}{(2r)^3} = \dfrac{\pi}{6} = 0.55.$

We raise the same question for the ratio of the l_1 ball and l_∞ ball.

- $\mathbb{R}$: $\quad \dfrac{L_1\text{-ball}}{L_\infty\text{-ball}} = \dfrac{2r}{2r} = 1$
- $\mathbb{R}^2$: $\quad \dfrac{L_1\text{-ball}}{L_\infty\text{-ball}} = \dfrac{(2\pi)r^2/2!}{(2r)^2} = \dfrac{1}{2}$
- $\mathbb{R}^3$: $\quad \dfrac{L_1\text{-ball}}{L_\infty\text{-ball}} = \dfrac{(2\pi)r^3/3!}{(2r)^3} = \dfrac{1}{6}.$

Note that the ratio decreases as the dimension increases. We will see that this is true more generally, and it is this phenomenon that is responsible for many of the important and counterintuitive properties of $\mathbb{R}^n$, for $n > 1$.

For example,

- While it shouldn't be obvious, this is the phenomenon that is related to the fact that a fair coin flipped many times almost always comes up heads.
- It is also related to the fact that if you have a class with more than a few dozen people then it is very unlikely that there are no two people who share the same birthday.
- It is also related to many other related things in data science.

4 Lab Exercise

Example. Find the L_1, L_2 and L_∞ norms of the vector $(1, 2.\ 3)$..

```
x<-1:3 #creates a vector
x   #Print the vector x
y<-as.matrix(x) # Store the vector x as a matrix
y   #Print the vector y
norm(as.matrix(x), "1") #L_1 norm of x
norm(as.matrix(x), "2") #L_2 norm of x
norm(as.matrix(x), "M"). #L_\infty norm of x
```

Create the following vectors in R:

1. $(12, 7, 20), (1, 2, 3, 4)$
2. $\begin{pmatrix} 12 \\ 7 \\ 20 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}.$

5 Visualizing elements of $\mathbb{R}^n$

5.1 Ways to label points, elements, vectors, matrices

Consider three points on the plane. It may seem natural or more familiar to label them as $(x; y)$ pairs, with a number indicating which point it is, i.e., as in the following:

$$\begin{array}{ll} (x_1, y_1) & \text{is the first point} \\ (x_2, y_2) & \text{is the second point} \\ (x_3, y_3) & \text{is the third point.} \end{array}$$

That convention is fine, but we will adopt a numbering convention that will help us generalize to many dimensions, e.g., from $\mathbb{R}^2$ to $\mathbb{R}^{47}$ to $\mathbb{R}^n$, for arbitrary n. In general, given a vector x, i.e., x is a vector and not a number, we will label the different elements of x with subscripts, as in

$$\begin{array}{ll} x = (x_1, x_2) & \text{for } x \in \mathbb{R}^2 \\ x = (x_1, x_2, x_3) & \text{for } x \in \mathbb{R}^3 \\ \vdots & \\ x = (x_1, x_2, ..., x_n) & \text{for } x \in \mathbb{R}^n. \end{array}$$

This system works well if there are a small number of points. However, how to denote points if there are , say, 100 points? The above system is difficult to implement. So, we need a more *powerful system of notations* to denote many points in high dimension.

More powerful system of notations

Alternatively, we could label the elements in a way that seems more difficult, but that makes the interpretation of matrices as consisting of high-dimensional vectors more immediate.

In particular, if we label the three two-dimensional points as x_1, x_2, and x_3, where x_1, x_2, and x_3 are vectors and not numbers, and where the subscript denotes which point we are dealing with and not which component of a given vector, then we can let a second subscript denote the component of that vector and write it as follows:

$$\begin{aligned}
x_1 &= (x_{11}, x_{12}) \quad \text{is the first point} \\
x_2 &= (x_{21}, x_{22}) \quad \text{is the second point} \\
x_3 &= (x_{31}, x_{32}) \quad \text{is the third point.}
\end{aligned}$$

Given this notational convention, we could write this as a matrix:

$$A = \begin{pmatrix} x_{11} & x_{12} \\ x_{21}` & x_{22} \\ x_{31} & x_{32} \end{pmatrix}$$

Alternatively, we could change the order of the subscripts, letting the first subscript denote the component and the second subscript denote the point, to obtain

$$
\begin{aligned}
x_1 &= (x_{11}, x_{21}) \quad \text{is the first point} \\
x_2 &= (x_{12}, x_{22}) \quad \text{is the second point} \\
x_3 &= (x_{13}, x_{23}) \quad \text{is the third point.}
\end{aligned}
$$

In this case, we obtain the following matrix:

$$
A = \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \end{pmatrix}
$$

Here, we have still used the notation that element ij refers to the element that is in the ith horizontal row and the jth vertical column, but we have swapped what rows and columns mean. In this case, the columns of A are data points.

5.2 Visualizing elements

If, instead of working with a small number of points in $\mathbb{R}^2$ or $\mathbb{R}^3$, we have 10^9 points, each of which is described by a set of 10^6 numbers, then we have a much larger matrix. Not only would it be difficult for a human to write out that many points, but it is not possible to visualize points in that many dimensions.

Question: What do 10^9 points in 10^6 "look like"? For example, do they look similar to or different than on $\mathbb{R}^2$ or $\mathbb{R}^3$?

This is very common in data science, but it can be a very difficult question to answer.

Another question: What does it even mean to ask

> that previous question? That is, how can we quantify what "look like" means?

That ambiguity being noted, it can be difficult to tell what such a data set "looks like," and indeed one of the main challenges here is going to be that high dimensional spaces are very different that low dimensional spaces.

Data visualization is one of the main pillars supporting data analysis. It has been a powerful tool and has been widely adopted by organizations owing to its effectiveness in abstracting out the right information, understanding and interpreting results clearly and easily.

For $n = 1, 2, 3$, this is easy – it simply corresponds to the familiar one-dimensional line, two-dimensional plane, and three dimensional space, with which we are familiar, as well as familiar subsets of these spaces.

However, dealing with multi-dimensional datasets with typically more than two attributes start causing problems, since our medium of data analysis and communication is typically restricted to two dimensions.

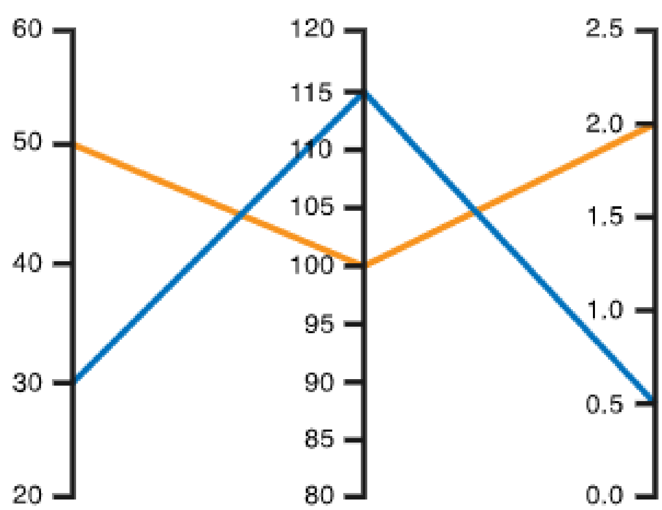
Here, we will describe one way to "visualize" higher dimensional vectors. Since people have such strong intuition about one and two and three dimensional spaces, this method tends to be less useful than the usual way; but since people have such weak intuition about higher dimensional spaces, for $n > 3$ it can be very helpful. We will call this the **parallel coordinates method** (or **augmented visualization method**).

Parallel coordinates method: To show a set of points in an n-dimensional space, n equally spaced vertical lines are drawn. A point in n-dimensional space is represented as a polyline (broken lines) with vertices on the parallel axes. The position of the vertex on the i-th axis corresponds to the i-th coordinate of the point.

Thus, the basic idea of the augmented visualization method is to view a vector $x \in \mathbb{R}^n$ as n separate numbers $x_i \in \mathbb{R}$, which can be represented on a two dimensional piece of paper by n points on n separate number lines, plotted vertically and parallel to each other. On the ith vertical line, we plot the point x_i.

Example 7.

Let $x = (30, 115, 0.5)$ and $y = (50, 100, 2)$ be two vectors in $(R)^3$. Using parsllel coordinates method, we plot them as in



This type of visualisation is used for plotting multivariate, numerical data. Parallel Coordinates Plots are ideal for comparing many variables together and seeing the relationships between them. For example, if you had to compare an array of products with the same attributes (comparing computer or cars specs across different models). In a Parallel Coordinates Plot, each variable is given its own axis and all the axes are placed in parallel to each other. Each axis can have a different scale, as each variable works off a different unit of measurement, or all the axes can be normalised to keep all the scales uniform. Values are plotted as a series of lines that connected across all the axes. This means that each line is a collection of points placed on each axis, that have all been connected together. The order the axes are arranged in can impact the way how the reader understands the data. One reason for this is that the relationships between adjacent variables are easier to perceive, then for non-adjacent variables. So re-ordering the axes can help in discovering patterns or correlations across variables. The downside to Parallel Coordinates Plots, is that they can become over-cluttered and therefore, illegible when theyre very data-dense. The best way to remedy this problem is through interactivity and a technique known as Brushing. Brushing highlights a selected line or collection of lines while fading out all the others. This allows you to isolate sections of the plot youre interested in while filtering out the noise.

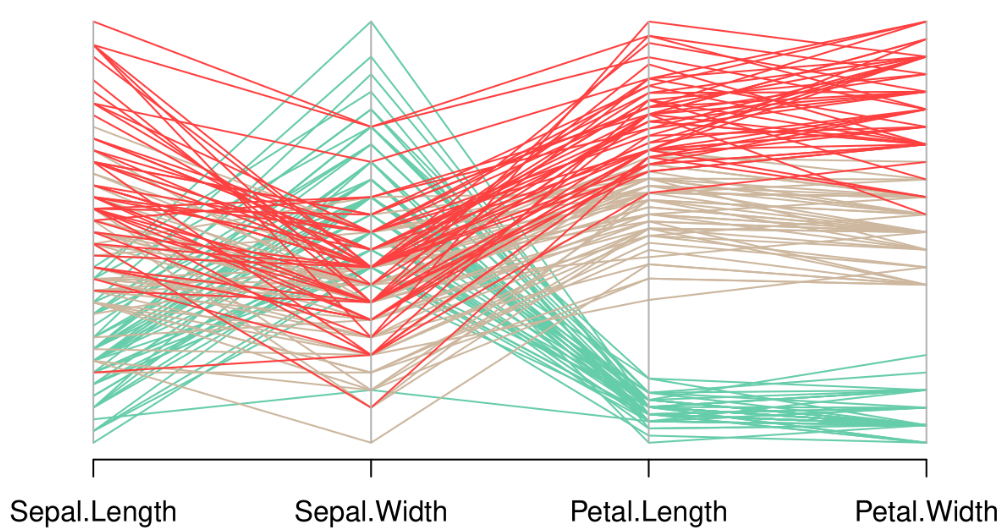
Parallel coordinates method is a common way of visualizing and analyzing high-dimensional datasets.

This visualization is closely related to time series visualization, except that it is applied to data where the axes do not correspond to points in time, and therefore do not have a natural order. Therefore, different axis arrangements may be of interest.

The MASS library provides the parcoord() function that automatically builds parallel coordinates chart.

The input dataset must be a data frame composed by numeric variables only. Each variable will be used to build one vertical axis of the chart.

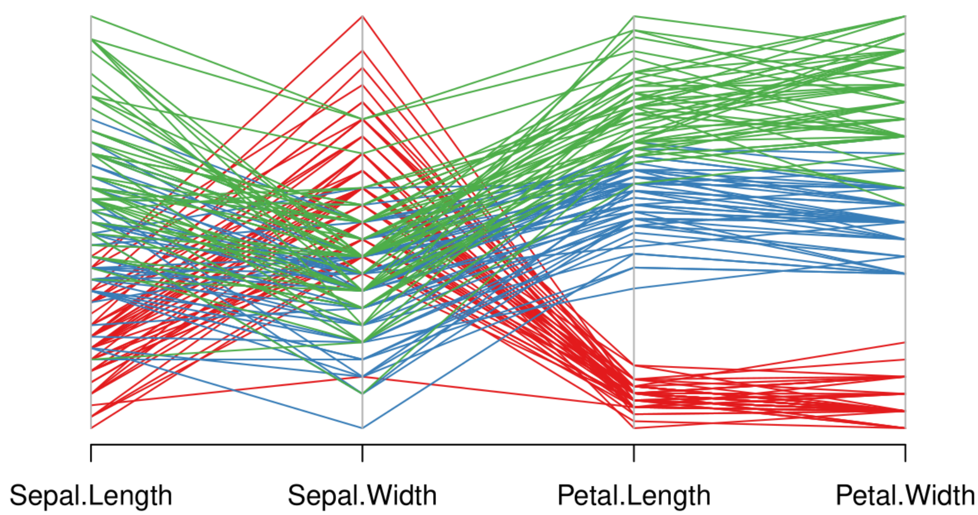
```
# You need the MASS library
library(MASS)
# Vector color
my_colors <- colors()[as.numeric(iris$Species)*11]
# Make the graph !
parcoord(iris[,c(1:4)] , col= my_colors  )
```



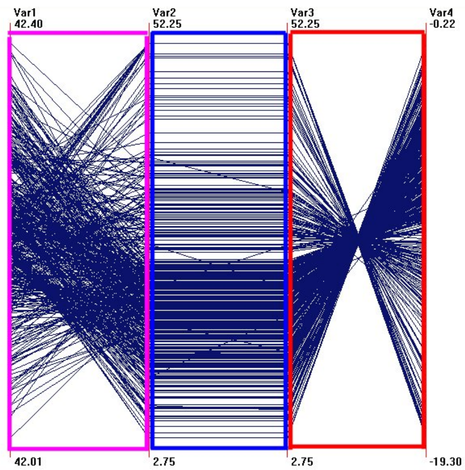
It is important to find the best variable order in your parallel coordinates chart. To change it, just change the order in the input dataset.

Note: the RColorBrewer package is used to generate a nice and reliable color palette.

```
# You need the MASS library
library(MASS)
# Vector color
library(RColorBrewer)
palette <- brewer.pal(3, "Set1")
my_colors <- palette[as.numeric(iris$Species)]
# Make the graph !
parcoord(iris[,c(1:4)] , col= my_colors  )
```



Parallel Coordinates and Scatter Plot Matrix: This example shows that Var1-Var2 has no correlation; Var2-Var3 has very strong positive correlation; Var3-Var4 has very strong negative (inverse) correlation.



5.3 Vector addition and scalar multiplication

Vector addition: If $x, y \in \mathbb{R}^n$, then

$$z = x + y = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{pmatrix} \in \mathbb{R}^n.$$

Scalar multiplication: If $x, y \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$, then

$$y = \alpha x = \alpha \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} \alpha x_1 \\ \alpha x_2 \\ \vdots \\ \alpha x_n \end{pmatrix} \in \mathbb{R}^n.$$

Both of the above operations give as output a vector in the same vector space, and thus we can apply them recursively.

Using norms to measure distances

Let $x, y \in \mathbb{R}^n$. The distance between x and y is given by $\|x - y\|$.

Example. If $x = (0, 0)$ and $y = (5, 5)$, then $x - y = (0 - 5, 0 - 5) = (5, 5)$.

1. $\|x - y\|_1 = \|(5, 5)\|_1 = 5 + 5 = 10.$
2. $\|x - y\|_2 = \|(5, 5)\|_2 = \sqrt{5^2 + 5^2} = 7.07.$
3. $\|x - y\|_\infty = \|(5, 5)\|_\infty = \max\{5, 5\} = 5.$

Here, we have considered two vectors in $\mathbb{R}^2$ and computed the norm of the difference vector which is also a vector in $\mathbb{R}^2$. Again,

$$\|x - y\|_\infty \leq \|x - y\|_2 \leq \|x - y\|_1,$$

$$\|x - y\|_1 \leq \sqrt{2}\|x - y\|_2 \leq 2\|x - y\|_\infty,$$

Two things should be noted.

- First, the distance between two vectors depends on the norm you use to measure distance.
- Second, the relationship between those distances is the same as the relationship between the corresponding norms (that we discussed eariler).

Clearly, these are true, since the distance is just a norm of the difference vector.

Using norms to normalize

Another way in which norms are useful is that they can be used to normalize vectors.

Normalize : Divide a vector by its norm (or size) to get a unit vector (or normalized vector) in the direction of the original vector.

For example, the unit vector (or normalized vector) $u \in \mathbb{R}^n$ in the direction of the vector $x \in \mathbb{R}^n$ is given by

$$u = \frac{x}{\|x\|}.$$

This is an example of multiplying a vector by a scalar to construct another vector. Any vector, except for a vector in which each entry equals 0 and which thus has norm equal to 0, can be normalized.

Example. Let $x = (1, 2)$. In this case:

1. Since $\|x\|_1 = 3$, the normalized vector is $x' = \frac{1}{3}(1, 2) = (1/3, 2/3)$.
2. Since $\|x\|_2 = \sqrt{5}$, the normalized vector is $x' = \frac{1}{\sqrt{5}}(1, 2) = (1/\sqrt{5}, 2/\sqrt{5})$.
3. Since $\|x\|_\infty = 2$, the normalized vector is $x' = \frac{1}{2}(1, 2) = (1/2, 1)$.

Example. Let $x = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}$. In this case:

1. Since $\|x\|_1 = |\cos\theta| + |\sin\theta|$, the normalized vector is

$$x' = \frac{1}{|\cos\theta| + |\sin\theta|}\begin{pmatrix}\cos\theta\\ \sin\theta\end{pmatrix}$$
$$= \begin{pmatrix}\cos\theta/(|\cos\theta| + |\sin\theta|)\\ \sin\theta/(|\cos\theta| + |\sin\theta|)\end{pmatrix}$$

2. Since $\|x\|_2 = 1$, the normalized vector is $x' = x$. That is, for every value of θ, the vector x is normalized, with respect to the Euclidean norm.

3. Since $\|x\|_\infty = \max\{|\cos\theta|, |\sin\theta|\}$, the normalized vector is
$x' = \frac{1}{2}(1,2) = \begin{pmatrix}\cos\theta/\|x\|_\infty\\ \sin\theta/\|x\|_\infty\end{pmatrix}$.

Example. Let $x = (1, 0, ..., 0) \in \mathbb{R}^n$. In this case:

1. Since $\|x\|_1 = 1$, the normalized vector is $x' = x$.

2. Since $\|x\|_2 = 1$, the normalized vector is $x' = x$.

3. Since $\|x\|_\infty = 1$, the normalized vector is $x' = x$.