# 2. Basic mathematical concepts

## 2.1 Defining compositions

In this chapter we will define compositions as mathematical objects. We will start with some formal definitions.

**Definition 2.1.1 — Composition.** A composition is defined by a vector of positive, non-zero values, which only carry relative information. The composition is said to contain D parts if the length of the vector equals D:

$$\mathbf{x} = (x_1, x_2, ..., x_D), \qquad x_i \in \mathbb{R}_+ \tag{2.1}$$

Relative information means that each individual part in the composition carries no information on its own. If we did a poll among students on campus, whether they would like to have lectures on Saturdays and you would only be told a single part of the result, lets us say, 25 students agree, this would not provide any information on the outcome of the poll. You would have to know the number of students who answered 'no', in order to extract any information. If the number of students who says 'no' is 17, only then would you know that a majority of students would like to have lectures on Saturdays. The number of yaysayers only carry information relative to the number of naysayers.

In the above example we asked 42 students about their opinion. If we had asked twice as many, 84, and 50 had said 'yes' and 34 had said 'no', the result would have been exactly the same. Multiplying two numbers by a constant, does not change the ratio between them. Thus we can define compositions as equivalence classes.

**Definition 2.1.2 — Compositional equivalence.** Two compositions $\mathbf{x}, \mathbf{y}$ are compositionally equivalent if a positive, real constant $\lambda$ exists, so that $\mathbf{x} = \lambda \cdot \mathbf{y}$.

The total number of students that were asked is irrelevant, only the ratio of 'yes' to

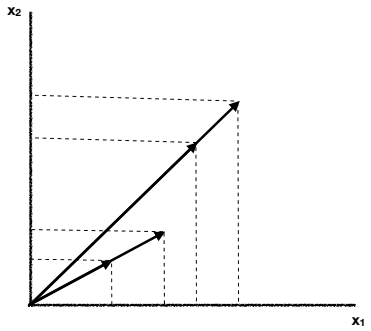


Figure 2.1: Compositional equivalence means that vectors that point in the same direction are equivalent. The length of compositional vectors is irrelevant.

'no' answers matter. However if we wish to compare several compositions, it is sometimes useful to rescale the compositions to the same total sum, for instance 1 (for proportions) or 100 (for percentages). Rescaling corresponds to changing the unit of the parts. When we rescale a composition to a certain constant $\kappa$, we say that we close the composition to $\kappa$.

**Definition 2.1.3 — Closure.** The closure of a composition $\mathbf{x}$ to a positive, real number $\kappa$ is defined as

$$\mathscr{C}(\mathbf{x}) = \frac{\kappa}{\sum_{i=1}^{D} x_i} \cdot \mathbf{x} \tag{2.2}$$

As an example of closure, let us consider the situation where, on two consecutive days, we ask a number of students whether they would favor lectures on Saturdays. The first day we ask 42 students and their replies are $\mathbf{x} = (25, 17)$. The second day we ask 33 student and their answer is $\mathbf{y} = (19, 14)$. Does the answer differ on day two? It is difficult to judge by eye, because the total number of respondents differ. However, we can close both compositions to 100 using definition 2.1.3, so that $\mathbf{x} = (60, 40)$ and $\mathbf{y} = (58, 42)$. Now the compositions are given as percentages and it is immediately clear that the answers have shifted by two percent points on day 2. Notice that even if it seems that the parts carry absolute information after closure, this is not the case. If you were told that 40% answered 'no', this does not mean that 60% said 'yes'. It could be that 30% said 'yes' and 30% said 'don't know', in which case 'no' would be in majority. The parts still only carry relative information.

It should be clear already, that compositions are different from ordinary real vectors. They can only be positive and their lengths are determined by an arbitrary closure constant. We can therefore now understand the mathematical reason why the politician talking about library usage in the last chapter was wrong. She was dealing with two compositions with parts describing non-users and users of libraries, $\mathbf{m} = (39, 61)$ and $\mathbf{f} = (30, 70)$. Adding them would give the composition $(69, 131)$, which is not closed to 100 and therefore does not have the unit 'percent'. If we close the sum back to 100, we get $(34.5, 65.5)$, which

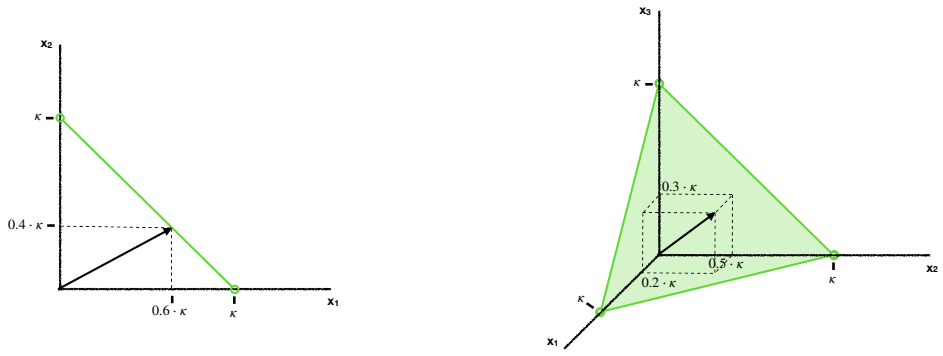


Figure 2.2: 1- and 2-simplices (green) as the sampling space for a 2- and a 3-part composition.

really is the average and not the sum. Furthermore, in order for 'addition' to be a valid operation, we need a 'neutral element' and 'inverse elements', defined such that the sum of an element and the neutral element gives the element itself and the sum of an element and its inverse element gives the neutral element. For real numbers, the neutral element is 0 and the inverse elements are the negative numbers, i.e., $a+0=a$ and $a+(-a)=0$. From the definition of compositions we can see that we have neither the inverse elements nor the neutral element (only positive numbers are allowed), hence adding compositions together is ill-defined.

With these definitions, we can define the space in which all compositions exists, the so-called sample space.

**Definition 2.1.4 — Sample space.** The sample space, i.e., the space containing all possible compositions, is the *simplex*, defined as

$$\mathscr{S}^D = \left\{ \mathbf{x} = (x_1, x_2, ...x_D) \,\middle|\, x_i \in \mathbb{R}_+, i = 1,2,...,D; \sum_{i=1}^{D} x_i = \kappa \right\} \tag{2.3}$$

The simplex is a generalization of a triangle to arbitrary dimensions. A triangle is a 2-simplex defined by its 3 vertices. The simplex that defines the sample space of a particular composition has a dimensionality which is always 1 lower than the number of parts. This is because the last part is constrained by the closure relation, so that only D-1 parts are independent. Figure 2.2 show the 1- and 2-simplices. The 3-simplex is a tetrahedron and the 4-simplex is known as a 5-cell (and so on for higher dimensions), which can only be viewed in projection.

In many circumstances, we are only interested in some parts of a composition, either because the remaining parts are irrelevant or impossible to obtain. In the example from above with the student poll, we might get a certain number of 'yes' answers, another number of 'no' answers and in addition, a few who answers 'don't know'. We are not interested in the 'don't know's, so we chose to consider only the 'yes' and the 'no' answers. This is called a subcomposition.

**Definition 2.1.5 — Subcomposition.** Given a composition $\mathbf{x}$ and a set of indices $S = i_1, i_2, ...i_s$, a subcomposition is obtained by applying the closure operation to the subvector $\mathbf{x}_s$.

If our poll shows $\mathbf{x} = (60, 30, 30)$ for 'yes'/'no'/'don't know', we can form the subcomposition $\mathbf{x}_{yes/no} = (66.67, 33.33)$ by closing the subvector to 100. Most compositions are already subcompositions. For instance, the 'yes'/'no'/'don't know'-composition is already a subcomposition of the composition 'yes'/'no'/'don't know'/'don't care' which again can be seen as a subcomposition of a composition that contains even more possible answers.

Another way to reduce the dimensionality of a composition is to amalgamate parts, by summing them into a new part.

**Definition 2.1.6 — Amalgamation.** Given a composition $\mathbf{x}$ of D parts and a set of indices $A = i_1, i_2, ..., i_a$ and another set $\bar{A} = D_i \setminus A$, the composition

$$\mathbf{x}' = (\mathbf{x}_{\bar{A}}, x_A), \qquad x_A = \sum_{i \in A} x_i \tag{2.4}$$

is called the amalgamated composition in $\mathcal{S}^{D-a+1}$.

An example of an amalgamated composition is when we take the poll from above and include 5 'don't care'-responses, so that we get $\mathbf{x} = (60, 30, 30, 5)$ for 'yes'/'no'/'don't know'/'don't care'. We can then amalgamate the last two parts into one new part, which we could call 'Other', and we would get $\mathbf{x}_{yes/no/other} = (48, 24, 28)$, by summing the parts we want to amalgamate and apply closure. The original 4-part compositions is defined on $\mathcal{S}^4$, whereas after amalgamating two parts, the new composition is defined on $\mathcal{S}^{4-2+1} = \mathcal{S}^3$

Subcompositions and amalgamations are often encountered in metagenomic data analysis. When the read sequences are mapped or aligned against reference databases, it is expected that a certain number of reads do not map, either because the databases are incomplete or because the reads originate from DNA that belongs to an organism which we are not interested in, e.g., mammal DNA if we only map against microorganisms. These unassigned reads will form a part of their own in the composition, but they are, in most cases, irrelevant for the down stream analysis. We could then chose to look at the subcomposition that is everything but unassigned reads.

An example of amalgamation in metagenomics is when we look at reads that map to antimicrobial resistance gene references. Many different genes give bacteria resistance to the same class of antimicrobial agent and it is sometimes more useful to look at the composition where these genes have been amalgamated, so that, rather than having a composition where the parts represents individual genes, the parts represents phenotypical resistance classes.

## 2.2 Principles of compositional data analysis

Three principles should be respected by any mathematical method applied to compositional data. The first two principles, scale and permutation invariance are simple and easily accepted as necessary truths, whereas the third principle, subcompositional coherence, requires a bit more insight to understand.

### 2.2.1 Scale invariance

Scale invariance follows from the fact that only ratios between parts in a composition matters. It shouldn't matter if we conduct our student poll among 50, 117, or 500 students (if we disregard counting noise and uncertainties), the result should be the same. A function $f$ that operates on a composition $\mathbf{x}$ is called scale invariant if $f(\lambda\mathbf{x}) = f(\mathbf{x})$ for any positive real value $\lambda \in \mathbb{R}_+$. The function should give the same result for all compositionally equivalent vectors. Many mathematical functions obey this criterium, but some functions are a more practical choice than others. One example of such a function is $f(\mathbf{x}) = x_1/x_2$ since $x_1/x_2 = (\lambda x_1)/(\lambda x_2)$. It is clear that the constant $\lambda$ cancels and the function is scale invariant. This function corresponds to changing the unit for example from percent to ppm and is essentially the function we use for closure. The downside of ratios is that they are strictly positive and that they depend on the ordering of the parts, since $x_1/x_2 \neq x_2/x_1$. One can get around this conveniently by taking the logarithm of the ratio, $f(\mathbf{x}) = \ln(x_z/x_2)$. This transformation is symmetric with respect to the (arbitrary) ordering of the parts and maps to the entire set of real numbers.

One can define more complex logratios, for instance the so called logcontrast,

$$f(\mathbf{x}) = \sum_{i=1}^{D} \alpha_i \ln(x_i), \qquad \sum_{i=1}^{D} \alpha_i = 0. \tag{2.5}$$

This is also a scale invariant function and can be used to determine equilibrium conditions in, for instance, chemical reactions and thermodynamics. Combustion of methane, for instance, can be described by the balanced equation,

$$1\mathrm{CH_4} + 2\mathrm{O_2} \rightarrow 1\mathrm{CO_2} + 2\mathrm{H_2O},$$

where the stoichiometric coefficients are $1+2$ on the left hand side and $1+2$ on the right hand side, so that their total sum is 0. If $\mathbf{x}$ is a 4-part compositions, describing the concentration of methane, oxygen, carbon dioxide and water, we can write the reaction as a logcontrast,

$$1\ln(x_1) + 2\ln(x_2) - 1\ln(x_3) - 2\ln(x_4) = \ln\left(\frac{x_1 \cdot x_2^2}{x_3 \cdot x_4^2}\right)$$

When the reaction is in equilibrium, the logcontrast will stay constant, no matter the concentration of the parts.

### 2.2.2 Permutation invariance

Permutation invariance means that the order of parts in a composition does not influence the result of the analysis. Obviously, if two compositions are to be compared, the ordering of the parts needs to be the same in the two compositions, but if the order is changed in both, the result will be the same. The function above, $f(\mathbf{x}) = x_1/x_2$, which provide scale invariance does not provide permutation invariance, since $x_1/x_2 \neq x_2/x_1$. By taking the logarithm of the ratio, inverting the ratio (rearranging the parts) only produces a sign change and thus gives symmetry to $f$ with respect to permutation. We can square the log-ratios to obtain perfect permutational invariance.

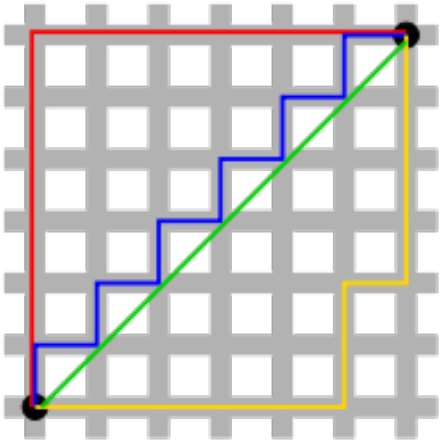

Figure 2.3: Comparison between the Euclidean metric and the taxicab metric on the plane. In the taxicab metric, the red, the blue, and the yellow routes all give the shortest distance, whereas in the Euclidean metric, only a single distance exists (green).

### 2.2.3 Subcompositional coherence

The equivalent of a subcomposition in real analysis is an orthogonal projection. We know that the length of the projection of a real vector onto a subspace is always shorter or, at most equal, to the length of the full vector. In compositional analysis, the length of vectors is meaningless, but a similar principle should apply. This is known as *subcompositional coherence*. Subcompositional coherence (sometimes also referred to as subcompositional dominance) means that the distance between to arbitrary subcompositions should always be smaller or, at most, equal to the distance between the full compositions. Also, the principle of scale invariance must be preserved within an arbitrary subcomposition. The distance between two compositions depends on the choice of metric, also know as the distance function. While different metrics exist, the best known is probably the Euclidean metric which gives the "straight-line" distance between two points in Euclidean space. The general definition of the Euclidean metric gives the distance between two points $\mathbf{p}$ and $\mathbf{q}$ as $d(\mathbf{p},\mathbf{q}) = \sqrt{(\sum_{i=1}^{n}(q_i - p_i)^2}$, which reduces to the famous Pythagorean theorem, $c^2 = a^2 + b^2$ for the two points $\mathbf{p} = (a,0)$ and $\mathbf{q} = (0,b)$ in $\mathbb{R}^2$.

The Euclidean metric, when applied to compositional data, does not obey the principle of subcompositional coherence (see exercise 2.5) and we can therefore not directly base any analysis of compositional data on this metric. In the field of ecology, the Euclidean and Manhattan (taxicab) metrics (shown in figure 2.3) are often used when comparing samples, but neither provides compositional coherence. Even worse, the Bray-Curtis dissimilarity,

$$BC_{ij} = 1 - \frac{2C_{ij}}{S_i S_j}$$

is often used to provide a distance measure between samples. While it is certainly valid to apply Bray-Curtis to compositional data, it is not a metric, since it does not satisfy the triangle inequality. It is a dissimilarity and not a distance and should not be used as such, for instance in machine learning applications.

The proper metric, which provides compositional coherence, is called the Aitchison metric and we will introduce this in the next chapter.

## 2.3 Exercises

**Exercise 2.1** During the COVID-19 pandemic, SSI (Statens Serum Institut) would publish the result of the national test effort on a daily basis. They provided, among other things, the number of positive tests. They also provided the number of tests conducted, from which one could calculate the positive percentage. There has been some discussion, particularly in the media, of which of these two numbers is the appropriate number to report. When is it appropriate to use the number of positive test and when is it appropriate to use the positive percentage? Why? (Discuss in class) ∎

**Exercise 2.2** When applying closure to a composition, the choice of $\kappa$ depends on the unit of the data. What are the values of $\kappa$ when data is measured as percentages? As parts-per-million (ppm)? What unit is metagenomic data measured in and what is the corresponding $\kappa$? ∎

**Exercise 2.3** Consider this table of faux data:

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $x_1$ | 79.07 | 31.74 | 18.61 | 49.51 | 29.22 |
| $x_2$ | 12.83 | 56.69 | 72.05 | 15.11 | 52.36 |
| $x_3$ | 8.10 | 11.57 | 9.34 | 35.38 | 18.42 |

Verify that the data could be treated as compositional. ∎

**Exercise 2.4** Form a two-part amalgamated composition, $(x_1, x_2 + x_3)$ of the data from Exercise 2.3. Does amalgamation preserve closure? ∎

**Exercise 2.5** Compute the Euclidean distance between the first two vectors in the data table from Exercise 2.3. Imagine originally a fourth variable $x_4$ was measured, constant for all samples and equal to 5%. Take the first three vectors, close them to 95%, add the fourth variable (so that they sum up to 100%), and compute the Euclidean distance between the vectors. Is the distance greater or smaller than the distance between the three-part compositions? What about the Manhattan distance ($d_{pq} = \sum_i |p_i - q_i|$)? ∎