5. Visualizing compositions

An important part of working with data is knowing how to visualize them in a meaningful and informative way. In the case of compositional data, we need to take their compositional nature into account or risk that we give erroneous impressions of the data.

5.1 Bar plots

One of the simplest ways of presenting a composition is to just plot the parts as bars in a bar chart. However, while there is nothing wrong in doing that, it is probably the least informative way of presenting compositions. Table 5.1 shows the nutritional content of a number of vegetables and these data are presented in a simple bar chart in Fig. 5.1. The part named "other" in the table is an amalgamated part containing fibers, minerals and vitamins, all of which are present in very small amounts. There is no natural ordering and it is difficult to compare the vegetables that are more similar and the ones that are more different. There is no natural sorting of the samples or the parts in a bar plot like this, so the analyst presenting the data can choose the sorting in a way that facilitates interpretation of the data. This should be done with caution however, since it is possible to fool the viewer into seeing a trend which may not be supported by the data itself. In a bar plot like Fig. 5.1, we do not take into account that the data are compositional and consequently, it has limited use. However, it is unfortunately the most common way to see compositional data presented.

By taking the compositional nature of the data into account, we can create a stacked bar chart which is a considerably better way of presenting the data. In a stacked bar chart, we can easily sort the vegetables on one of the parts, in this case protein, and we can immediately get and idea of how they compare (Fig. 5.2). For both cases, bars and stacked bars, it is absolutely necessary to apply closure with the same constant to all the samples shown in the plot, since they need to have the same unit.

In principle, pie charts are a viable alternative to (stacked) bar charts. In most cases,

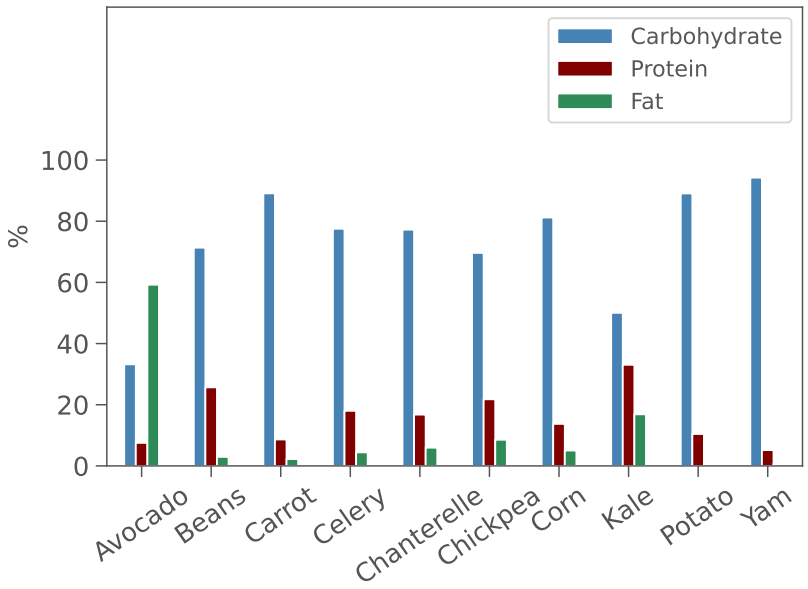

Figure 5.1: Bar chart showing the nutritional content of vegetables.

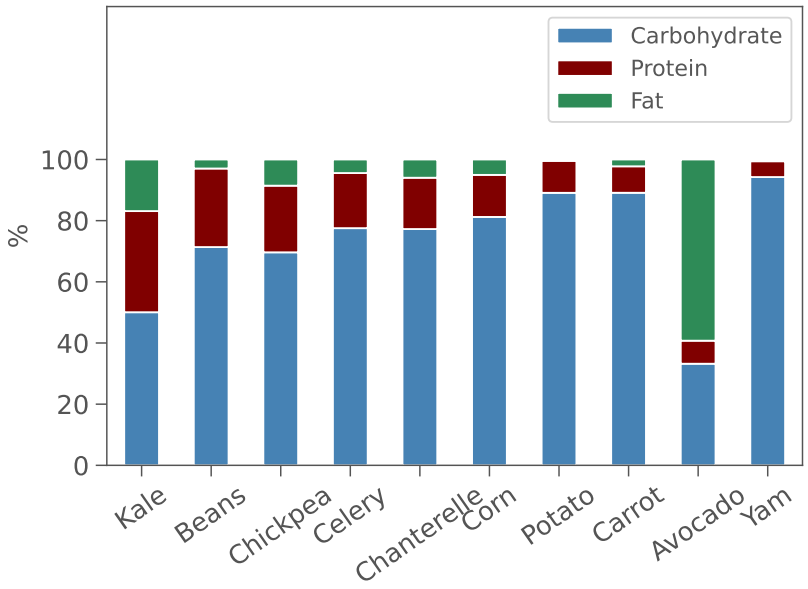

Figure 5.2: Stacked bar chart showing the nutritional content of vegetables.

Vegetable	Carbohydrate	Protein	Fat	Water	Other
Avocado	8.64	1.96	15.41	72.33	1.67
Beans	14.50	5.22	0.60	78.04	1.64
Carrot	9.58	0.93	0.24	88.29	0.96
Celery	2.97	0.69	0.17	95.43	0.74
Chanterelle	6.86	1.49	0.53	89.85	1.27
Chickpea	22.53	7.05	2.77	66.72	0.93
Corn	19.02	3.22	1.18	75.96	0.62
Kale	4.42	2.92	1.49	89.63	1.54
Potato	17.49	2.05	0.09	79.25	1.12
Yam	27.88	1.53	0.17	69.60	0.82

Table 5.1: Nutritional values, shown as percentages, of ten different vegetables (Data from US department of agriculture).

it is a matter of personal preference, but in general, pie charts are harder to read as there are no axis, they take up more space, and they are difficult to compare side by side. Many commercial software packages provide 3D versions of both bar and pie charts. While this may add a fancy look, it does not help to visually interpret the data and should be avoided in all serious applications. Never sacrifice data interpretability for fancy looking graphics. As Fig. 5.3 shows, it is misleading and oftentimes silly.

A final note on stacked bar plots and pie charts. Because they are, by construction, plotted on a linear scale, it is only possible to discern one, or at most two orders of magnitude. This means that very small proportions are just shown as a line and cannot be compared visually to other small proportions. This is why pie charts often have a 'other' category, where all the small parts are amalgamated. Likewise, if there is one dominating part in the composition, for instance, if one part makes up 99.5%, the entire bar (or pie) is just shown in the same color and very little information can be drawn from the plot. Figure 5.4 shows the composition of the atmosphere of three planets in the Solar System. The left panel is a stacked bar plot and by just looking at that plot, it is extremely difficult to compare the fraction of oxygen on Mars to the fraction of CO_2 on Earth. To overcome this problem of low dynamical range in stacked bar plots, we can choose to plot CLR values instead for a much clearer view. The downside of this approach is of course, that the reader does not necessarily know what CLR values are and additional explanation is therefore required when presenting data in that way. The atmospheric compositions are shown as CLR values in the right panel of Fig. 5.4.

5.2 Log-ratio scatter plots

In the case of bar plots and stacked bar plots, it is possible to show compositions with as many parts as we want, although with increasing number of parts it becomes harder to extract information. In the special case where the composition has three parts, we have a few other visualization options. Even if our data contains more than three parts, it can often be useful to reduce the dimensionality, either by extracting a sub-composition or by amalgamate parts, so that we only have three parts to visualize.

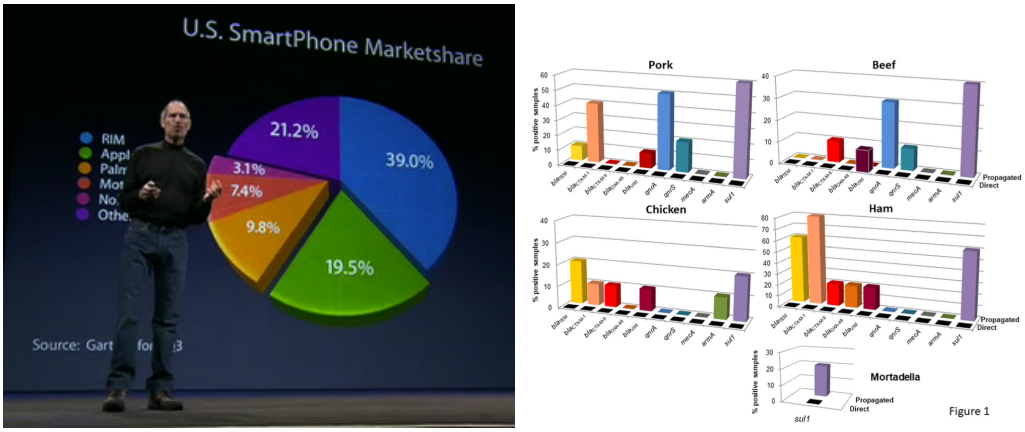

Figure 5.3: Left: Steve Jobs of Apple inc. presenting the iPhone market share. Due to the (unnecessary) 3D perspective, Apples market share of 19.5% appears bigger than the 21.2% share marked 'Other'. Right: Unnecessary use of 3D effects in a bar chart. Again, due to the depths effect, foreground features appear more prominent than background features despite having the same values (from Gómez-Gómez et al., Sci. Rep., 2019, 9, 13281).

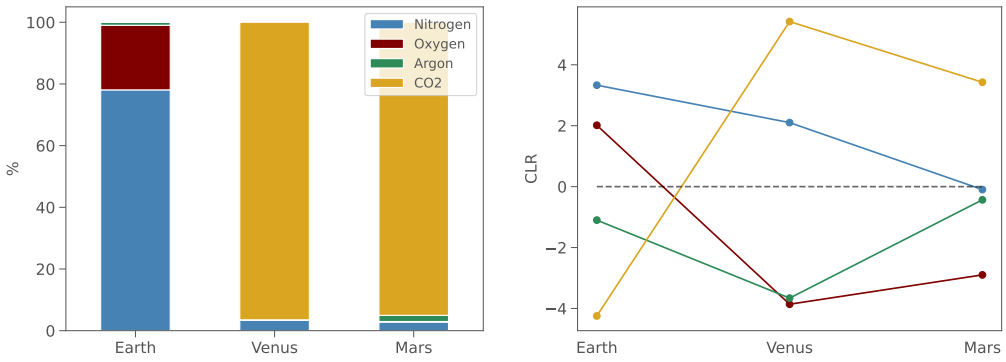

Figure 5.4: A stacked bar plot and categorical CLR values of the same data. It is much easier to compare small proportions in CLR space.

For a three-part composition we can make ratio scatter plots. Although entirely optional, these plots can be shown in logarithmic axis, making them log-ratio scatter plots. For a three-part composition we can form three ratios (six actually, but if we take the logarithm, only three are unique up to a sign), but of these three, only two are independent. The third ratio can be calculated from the two others

$$\text{Protein}/\text{Fat} = \frac{\text{Carbohydrate}/\text{Fat}}{\text{Carbohydrate}/\text{Protein}} \tag{5.1}$$

or, by taking the logarithm on both sides

$$\log(\text{Protein}/\text{Fat}) = \log(\text{Carbohydrate}/\text{Fat}) - \log(\text{Carbohydrate}/\text{Protein}) \tag{5.2}$$

If the two ratios are plotted against each other, like it is shown in Fig. 5.5, the third ratio is given by the orthogonal projection onto a line with a slope of -1.

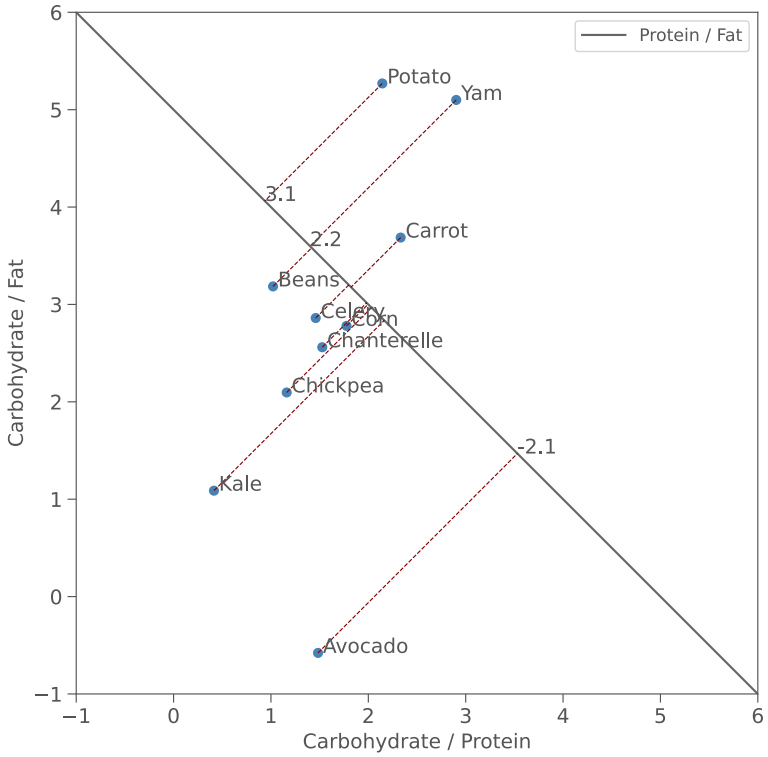

Figure 5.5: A log-ratio scatter plot, showing the vegetable data in two dimensions. In this type of plot we can easily identify similar vegetables.

5.3 Ternary diagrams

The best way of visualizing a 3-part composition is by plotting it in a ternary diagram. The ternary diagram is a direct visual representation of the simplex and is often referred to as a simplex plot. Unfortunately, ternary plots can only show three-part compositions, like in the case of the compositional scatter plot.

Figure 5.6, left panel, shows the vegetable nutrition data in a ternary diagram. It is easier to spot outlying points and in general identify samples which group together, which is very difficult to do by looking at the bar plots. The parts are read off the axis by following the grid lines that are parallel to the tick marks on the axis. The axis themselves and the triangle vertices are not part of the diagram since we do not allow parts to be zero.

In Chap. 3 we discussed the two arithmetic operations which make the simplex a vector space, perturbation and powering (Def. 3.2.1). Recall how these operations are analogous to addition and scalar multiplication in real vector spaces. By plotting the compositions in a ternary diagram, we can directly visualize the effects of perturbation and powering. An example is shown in Fig. 5.7. It is clear from this figure that perturbation shifts the points (translate) while powering scales them, and thus we can visually interpret these operation as adding a vector and multiplying by a constant respectively.

With perturbation and powering, we can define linear transformations

$$\mathbf{y} = (\alpha \odot \mathbf{x}) \oplus \mathbf{x}_0,$$

which is the compositional equivalent to the well known $\mathbf{y} = a\mathbf{x} + b$ for real vectors.

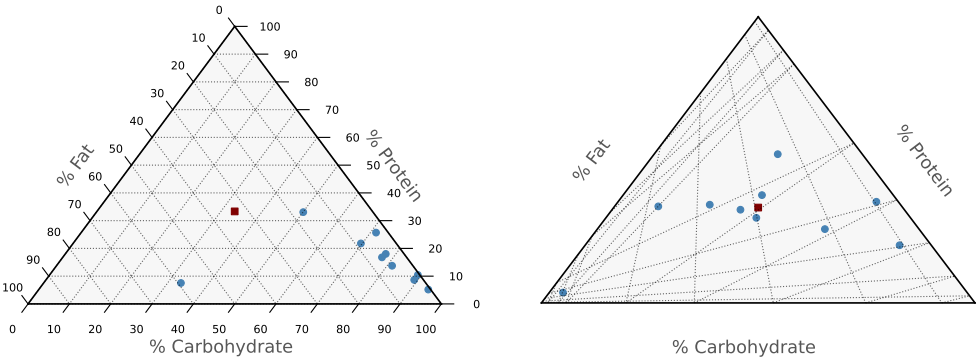

Figure 5.6: Left) Vegetable data plotted in a ternary diagram. Right) The same after centering of the data. The red square marks the barycenter of the triangle, which coincides with the neutral element of the simplex.

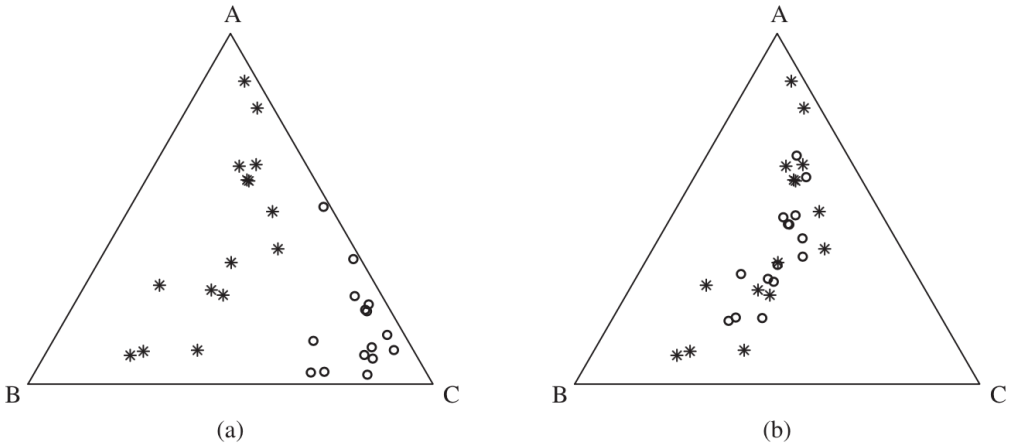

Figure 5.7: a) perturbation of a set of composition by another composition. b) powering of a set of composition by a constant.

Choosing a set of $\mathbf{x}_0$ and α which are equidistant in Aitchison-distance, we can plot compositional lines in a ternary diagram, as shown in Fig. 5.8. These lines are the compositional equivalent to a orthogonal grid in real space. It can be shown that the Aitchison inner product between the leading vectors of intersecting lines is 0, so in a general way, the dashed and full lines in Fig. 5.8 are orthogonal.

5.3.1 Centering

Sometimes, a set of samples plotted in a ternary diagram will fall very close to an edge or a vertex, making if difficult to see the actual distribution of points. There are two ways we can deal with this problem. We can either cut out the area containing the points, and magnify this part or we can center the data.

In order to center the data, we need to apply a perturbation to the composition, which is the equivalent of translation in real space. If we perturb a composition by its inverse

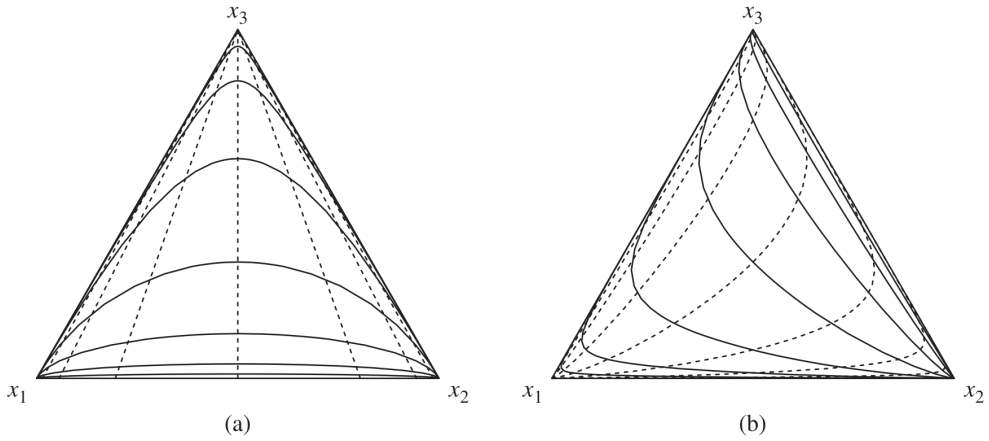

Figure 5.8: Orthogonal grids in the 3-simplex. The grid in b) is rotated by 45° with respect to the grid in a).

we obtain the neutral element according to the additive inverse law (Eq. 3.9). In the Aitchison geometry, the neutral element is the closure of a vector of 1's, equal to $\mathbf{0} = (1/D, 1/D, ..., 1/D)$. In the ternary plot, the neutral element coincides with the barycenter of the triangle, also known as its centroid. The centroid is the point of the intersections of the three lines that connect each vertex to the midpoint of the opposite edge. This means, that if we perturb the samples by their inverse, we move the points to the centroid of the triangle. If we instead perturb each sample by the inverse of the geometric mean of all the samples, we find that the set of samples, after perturbation, will gravitate around the centroid. Their new geometric mean will coincide with the triangle centroid. The data has been centered. This makes it a lot easier to visually explore the distribution of the data in the simplex, as long as we remember to take into account the the axis have been centered as well. The right hand side panel of Fig. 5.6 shows the vegetable data after centering.

5.3.2 Coordinate representation in ILR space

Finally, we can move out of the simplex, by making a coordinate transformation into real Euclidean space. In this case, we need to make use of the ILR transformation, because we need coordinates (recall that CLR values are coefficients, and not coordinates) that can be represented in a coordinate system. ALR does also provide coordinates, but the ALR transform is not isometric (meaning it does not conserve distances) and this is also a requirement for proper visualization. Only ILR fulfills both of these requirements.

Figure 5.9 shows the vegetable nutritional data in the ternary diagram and in ILR coordinate space. We have used the sequential binary partition talbe to make the transformation,

x_1	x_2	x_3
1	1	-1
1	-1	0

which then needs to be normalized. The axis in the plot are unit less and, as we discussed in Chap. 3, are defined by the balances of the contrast matrix. Euclidean distances are conserved, which means that two samples that are twice as far apart as two other samples

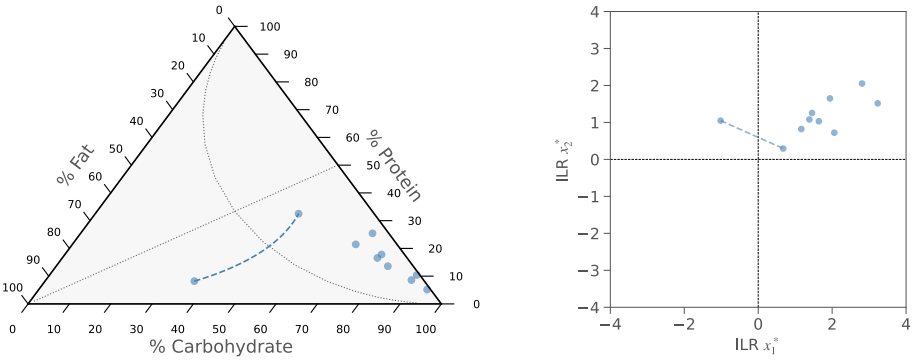

Figure 5.9: Left: The vegetable nutrition data in a ternary diagram. Right: The same data represented in ILR space. The dotted line segment represents the straight line connecting avocado and kale.

are really twice as different. This can be utilized in the following way: if we want to know which vegetable is nutritionally most similar to kale, we can identify kale in this plot and measure the Euclidean distance between kale and the nine other vegetables. The vegetable with the shortest distance to kale is the one with the closest matching nutritional parameters. It turns out that avocado is the closest match to kale, and we have for illustrative purposes connected the two with a straight line in Fig. 5.9. The dotted lines in the ILR coordinate plot, both the blue and the black lines, have been inverse transformed and plotted in the ternary diagram as well, to illustrate how line segments behave in the ternary diagram. In fact, if we plotted the centered and uncentered data points and connected the two sets with trajectories, they would show up as straight lines in the ILR coordinate plot.

5.3.3 Line segments in ternary diagrams

As we have just seen above, Aitchison straight lines appears as curves in a ternary diagram. Plotting curves in ternary diagrams can be useful when working with time-dependent compositions (see Chap. 9), but it can be a little tricky to get right. The best way is to discretize the curve into a series of points and then make a piecewise linear approximation by connecting the dots. Arbitrary smoothness can be achieved by increasing the number of points, but for all practical purposes, 100 points are sufficient.

Using this technique, we can plot various shapes and see how they appear in the simplex. Figure 5.10 shows three examples. The top row shows an example of perturbation of a line segment. In ILR space, the solid line has been translated by the vector corresponding to the dotted line, by adding the two vectors. In the simplex, this becomes perturbation. Notice that, while in ILR space it is clear that the dotted line was used, the same is not obvious in the simplex. Line segments lengths are not constant in the simplex! The middle row show an example of powering (scaling in ILR space). A vector $(1, 0.5)$ has been scalar multiplied by a range of integers to form a continuous line segment in ILR space. When translated into the simplex, the operation becomes powering. Notice again how the position of the power integers is not equidistant along the line segment. The last row shows a collection of circles and ellipses in real space and their corresponding shapes in the simplex. The closer we are to the origin of the real space coordinate system, the more

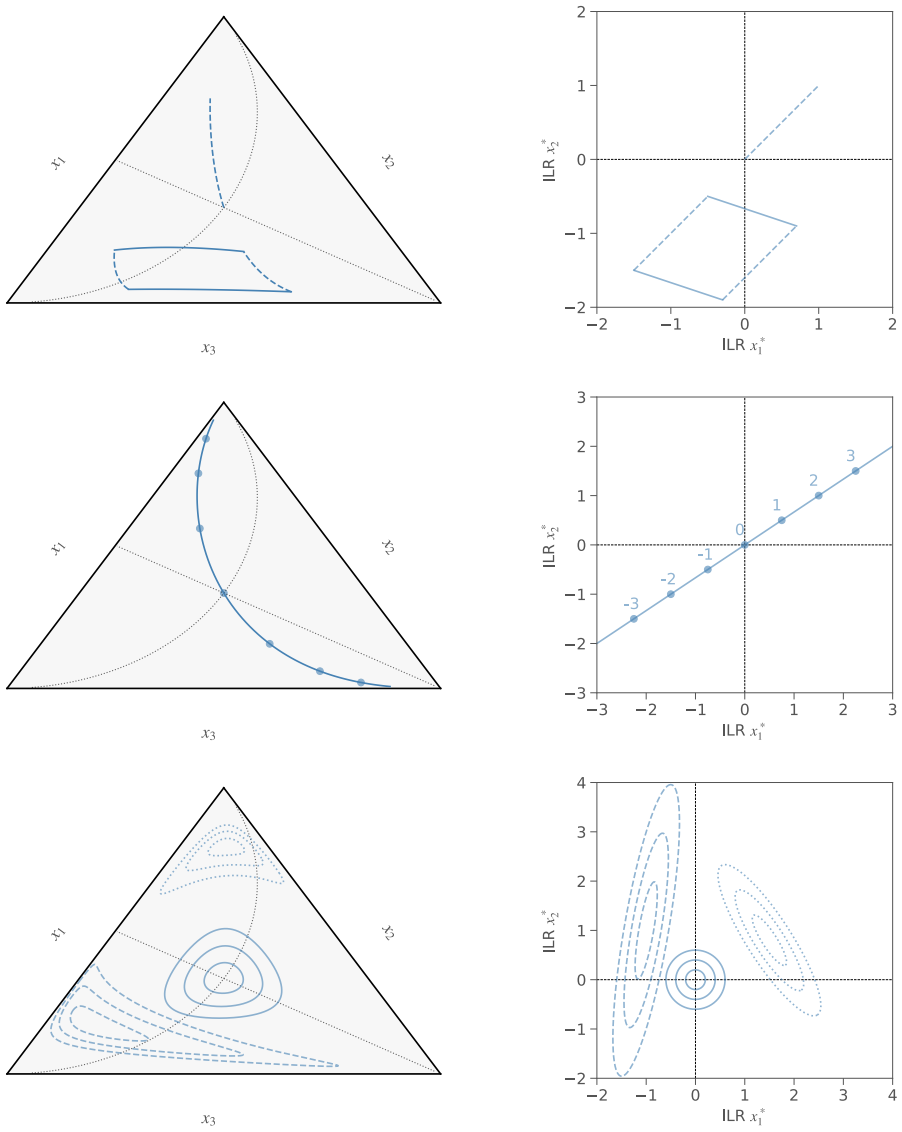

Figure 5.10: Top row) Line segments in the ternary plot and in ILR space. The solid line segment has been perturbed by the dotted line segment in the simplex, corresponding to a translation in real space. Middle row) A composition powered by a range of integers in the simplex and in ILR space. Bottom row) Circles and ellipses in ILR space and in the simplex.

similar the shape appears in the simplex, while the further we get from the origin, the more warped the transformed shape is, to the point where they are no longer recognizable as ellipses. This gives us a geometric understanding of why real vector algebra does not apply to compositions, and it also reveals that, sufficiently close to the origin (or centroid), real multivariate algebra gives approximately the correct result on the simplex.

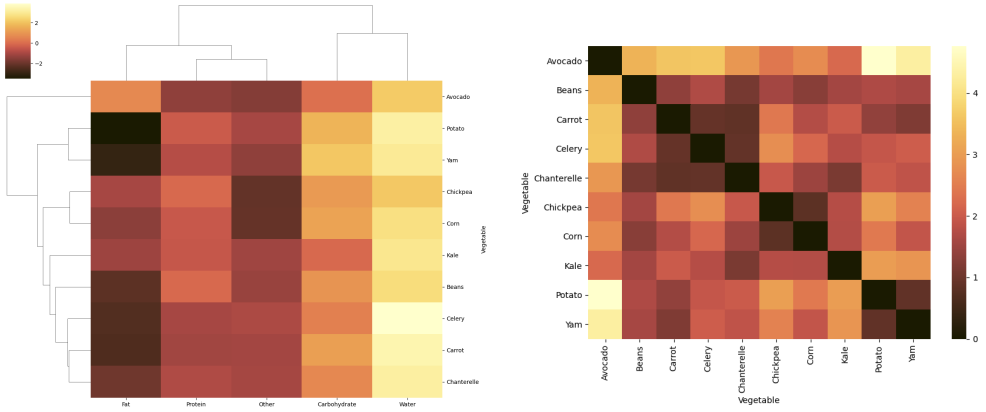

Figure 5.11: Left panel shows a cluster map of the vegetable data. Right panel shows a heatmap of the distance matrix. Darker colors mean closer together.

5.4 Heatmaps and cluster maps

The final visualization techniques that we will discuss in this chapter are the so-called heatmaps and the related cluster map. In all simplicity, heatmaps can be thought of as 2D versions of the ordinary bar chart and they are often more confusing than helpful. In a heatmap, one cannot see the heights of the bars and therefore we assign a color which represents the height of a bar. A color palette is therefore needed and choosing a proper palette is crucial for the information value of a heatmap. The details of choosing a good palette is beyond the present scope, but an interesting discussion on the topic can be found in Crameri, Shephard & Heron, Nature comm., 2020, 11, 5444. Heatmaps also suffer from the same dynamical range problem that we saw in the example with planetary atmospheres above. Again, this can be solved by transforming the data into CLR space, where a large dynamic range is shown on a linear scale.

A heatmap has no particular ordering of features or samples, but a cluster map does. The samples and/or the features can be linked and rearranged to form hierarchical clusters which results in a heatmap where samples (or features) that are more similar, i.e., has a smaller Aitchison distance from each other, are grouped together. Various methods exists for calculating the linkage, but since the data has already been CLR transformed, the Euclidean metric seems an obvious choice. Cluster maps are often plotted with corresponding dendrograms along the axis to show the hierarchical clustering.

Other data set specific properties can be shown in a heatmap. The distance matrix, for instance, which is a symmetric matrix with zeros in the diagonal. It contains the Aichison distance between each pair of samples (or if the data set is transposed, each pair of features) and this is sometimes plotted in a heatmap as well. Likewise, the variation matrix, which we will introduce in the following chapter, is ideally visualized in a heatmap. It is important to stress here, that derived data products, such as the distance or variation matrices, are not suitable for cluster maps, since calculating the distance between column or row vectors in these matrices is meaningless. Figure 5.11 shows a cluster map and a heatmap of the distance matrix of the full vegetable data set.

5.5 Exercises

In the following exercises, we will once again use the data from the Exercise 2.3. For convenience, the data is given here as well,

	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

Exercise 5.1 Make a bar chart and a stacked bar chart of the data. ∎

Exercise 5.2 Make a log-ratio scatter plot of the data. ∎

Exercise 5.3 Make a ternary diagram of the data. ∎

Exercise 5.4 Build a basis for the data, ILR transform the data and plot them in a Cartesian coordinate system. ∎

Exercise 5.5 Perturb the data by $\mathbf{s} = [0.1, 0.1, 0.8]$ and plot the initial and the perturbed data set in a ternary diagram and in ILR coordinates. In each case, join each pair of samples (unperturbed and perturbed) by a line segment. Observe the effect of perturbation. ∎

Exercise 5.6 Apply powering with α ranging from -8 to $+8$ in steps of 1 to the composition $\mathbf{t} = [0.7, 0.5, 0.8]$ and plot the resulting set of compositions in a ternary diagram and in ILR coordinates. Observe the effect of powering. ∎