Chapter 6

Measures of distance and correlation between variables

In Chapters 4 and 5 we concentrated on distances between samples of a data matrix, which are usually the rows. We now turn our attention to the variables, usually the columns. Two variables have a pair of values for each sample, and we can consider measures of distance and dissimilarity between these two column vectors. More often, however, we measure the similarity between variables: this can be in the form of correlation coefficients or other measures of association. In this chapter we shall look at the geometric properties of variables, and various measures of correlation between them. In particular we shall look at the geometric concept called a scalar product, which is highly related to the concept of Euclidean distance. The decision about which type of correlation function to use depends on the measurement scales of the variables, as we already saw briefly in Chapter 1. Finally, we also consider statistical tests of correlation, introducing the idea of permutation testing.

Contents

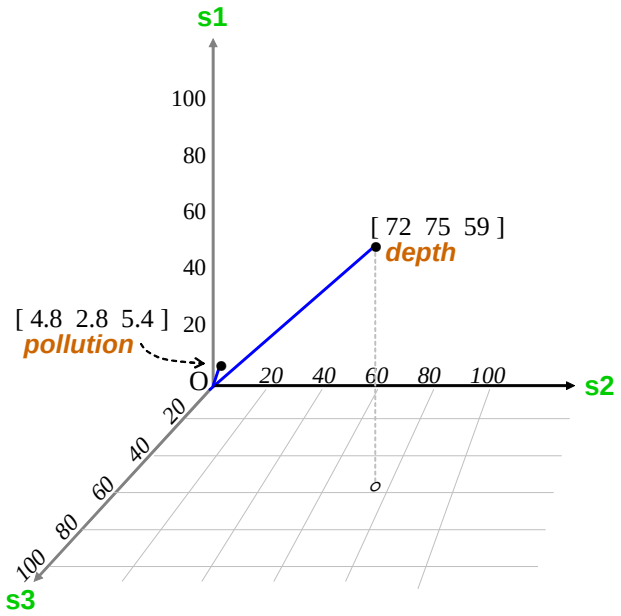
The geometry of variables
Correlation coefficient
Scalar product
Distances based on correlation coefficients
Distances between count variables
Distances between categorical variables and between categories
Testing correlations: permutation testing

The geometry of variables

In Exhibits 4.3 and 5.5 in the previous chapters we have been encouraging the notion of samples being points in a multidimensional space. Even though we cannot draw points in more than three dimensions, we can easily extend the mathematical definitions of distance to samples (usually the rows of the data matrix) for which we have J measurements for any J. We now consider the variables, which are usually the columns of the data matrix, and their sets of observed values across the I samples. In this case a two-dimensional, even a three-dimensional figure as a starting point for our thinking is rather trivial, because having only 2 or 3 samples in a study is ridiculous from a statistical point of view. Nevertheless, in Exhibit 6.1 we have attempted a picture of two variables in the three-dimensional space of a data set with sample size $I = 3$, because of the high pedagogical value of the diagram.

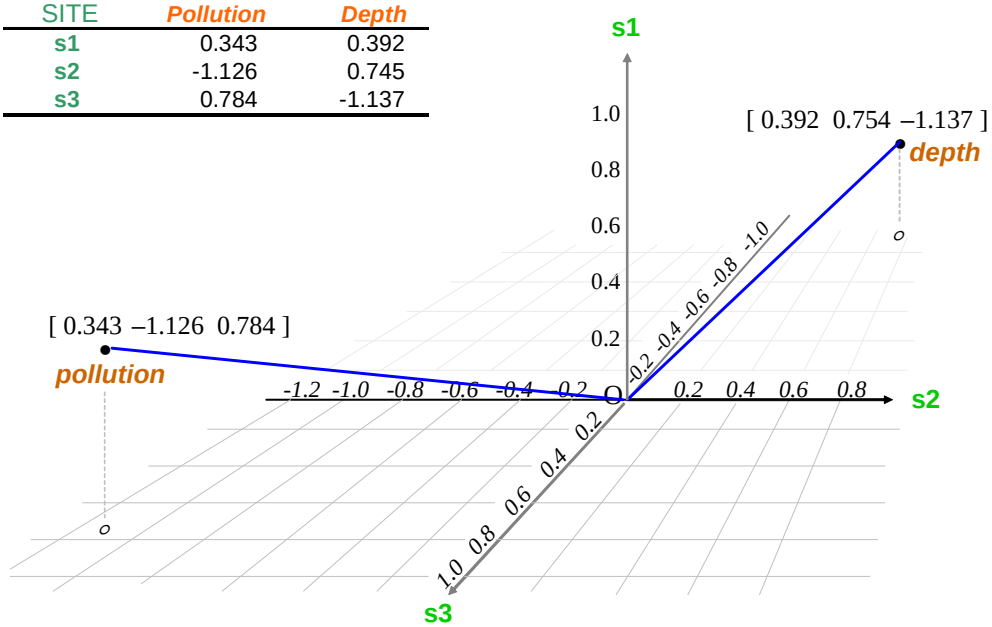
Exhibit 6.1 Two variables measured in three samples (sites in this case), viewed in three dimensions: (a) original scale; (b) standardized scale, where each set of three values has been centred with respect to its mean and divided by its standard deviation (standardized values as shown). Projections of some points onto the 'floor' of the s2–s3 plane are shown, to assist in understanding the three-dimensional positions of the points.

(a)



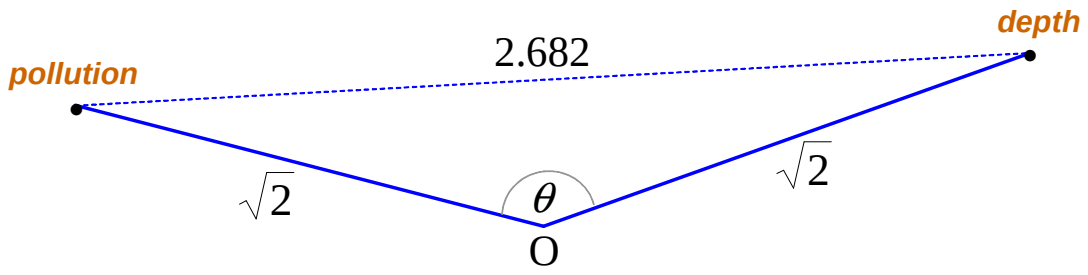
(b)

SITE	Pollution	Depth
s1	0.343	0.392
s2	-1.126	0.745
s3	0.784	-1.137



Note that the two variables pollution and depth have been standardized with respect to the mean and standard deviation of this sample of size 3, hence the values here do not coincide with the standardized values in the complete data set, given in Exhibit 4.4. Exhibit 6.2 now shows the triangle formed by the two vectors and the origin O, taken out of the three-dimensional space of Exhibit 6.1(b), and laid flat.

Exhibit 6.2 Triangle of pollution and depth vectors with respect to origin (O) taken out of Exhibit 6.1(b) and laid flat, with lengths of sides indicated.



The standardization has imposed equal lengths on the two vectors, because of the unit variance of each variable j=1 and 2:

$$\frac{1}{I-1}\sum_{i=1}^{I}(x_{ij}-\bar{x}_j)^2 = \frac{1}{I-1}\sum_{i=1}^{I}x_{ij}^2 = 1$$

where the means of the standardized variables are zero and, in this example, I = 3. So the length of each vector, i.e. the square root of the sum of squared coordinates (see Exhibit 4.3), is:

$$\sqrt{\sum_j x_j^2} = \sqrt{I-1} \qquad (6.1)$$

which is $\sqrt{2}$ in this example, as indicated in Exhibit 6.2. You can check that the (column) sums of squares of the standardized values in Exhibit 6.1(b) are indeed 2. The length of the third side of the triangle, between the pollution and depth points, is similarly calculated using the Euclidean distance formula (4.4) to be the square root of 7.190, i.e., 2.682 – check by summing the squared differences between the two columns and taking the square root. Hence we know the three sides of the triangle, so by using the cosine rule (which we all learnt at school) we can calculate the cosine of the angle θ between the vectors:

$$c^2 = a^2 + b^2 - 2ab\cos(\theta) \qquad (6.2)$$

i.e., $2.682^2 = 2 + 2 - 2\sqrt{2}\sqrt{2}\cos(\theta) = 4 - 4\cos(\theta)$

hence, $\cos(\theta) = 1 - \frac{1}{4} \times 7.190 = -0.7975$

and the angle is θ = 2.494 radians, or 142.9 degrees.

Correlation coefficient

After all that work, so what?! (we hear readers cry…) The punch line is that –0.7975 is actually the correlation coefficient between pollution and depth (in this sample of size 3) – so we have illustrated the result that the cosine of the angle between two standardized variables is the correlation. But in the above geometric explanation it is clear that the angle θ, and thus also the correlation which is $\cos(\theta)$, does not depend on the length of the two vectors which subtend the angle, only on the centering of these vectors. Since we can choose the lengths at will, there is (and will be later) a great advantage if these vectors have length 1. For example, in the cosine rule of (6.2), if $a = b = 1$, then there is the following simple relationship between the correlation $r = \cos(\theta)$ and the distance c between the two variable points, irrespective of the sample size:

$$r = 1 - \tfrac{1}{2} c^2 \qquad (6.3)$$

Standardized variables, whose lengths are equal to $\sqrt{I-1}$ (see (6.1)), can be converted to have length 1 simply by dividing them by $\sqrt{I-1}$, and then we call them *unit variables*. In our $I = 3$ example, the unit variables are:

SITE	*Pollution*	*Depth*
s1	0.242	0.277
s2	-0.796	0.527
s3	0.554	-0.804

and (6.3) can be easily verified:

$$-0.7975 = 1 - \tfrac{1}{2} [\, (0.242-0.277)^2 + (-0.796-0.527)^2 + (0.554-(-0.804))^2]$$

Scalar product

But there is yet another way of interpreting the correlation coefficient geometrically. Look at what you get when you take the sum of the products of the elements of the unit variables:

$$(0.242\times0.277) + (-0.796\times0.527) + (0.554\times(-0.804)) = -0.7975$$

Almost miraculously, the correlation coefficient can be calculated by what is called the *scalar product*:

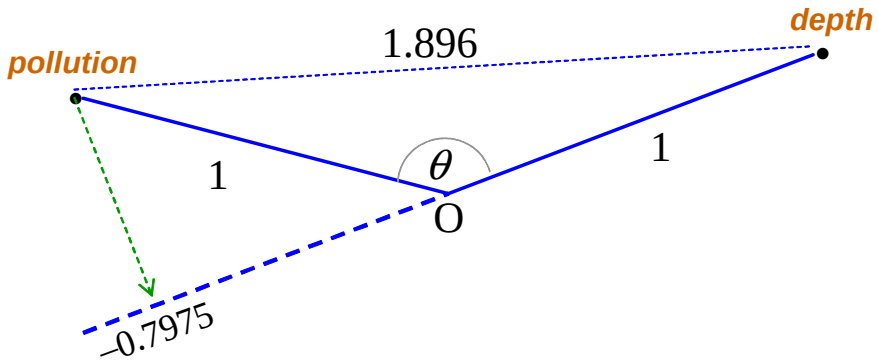
$$r_{jj'} = \sum_{i=1}^{I} x_{ij} x_{ij'} \qquad (6.4)$$

as long as the x_{ij} are the values of the unit variables. In fact, this result is easy to prove, since the definition of the correlation is a sum of products, but we leave details like that to the Theoretical Appendix A.

The concept of a scalar product underlies many multivariate techniques which we shall introduce later. It is very much related to the operation of *projection*, which is crucial later when we project points in high-dimensional spaces onto lower-dimensional ones. As an illustration of this, consider Exhibit 6.3, which depicts the unit variables (with length 1, so

they are shorter than those of length 1.414 in Exhibit 6.2).

Exhibit 6.3 Same triangle as in Exhibit 6.2, but with variables having unit length (i.e., unit variables, obtained by dividing standardized variables by the square root of $I-1$, where I is the sample size). The projection of either variable onto the direction defined by the other variable vector will give the value of the correlation, $\cos(\theta)$. (The origin O is the zero point and the scale is given by the unit length of the variables.)



Distances based on correlation coefficients

From a distance point of view, if the variables are expressed as unit variables, with sum of squares equal to 1, then from (6.3) the distance between variables j and j' (denoted by c in the previous discussion, denoted here by $d_{jj'}$) is directly related to the correlation coefficient $r_{jj'}$ as follows:

$$d_{jj'} = \sqrt{2-2r_{jj'}} = \sqrt{2}\sqrt{1-r_{jj'}} \qquad (6.4)$$

where d_{jj}' has a minimum of 0 when $r = 1$ (i.e., the two variables coincide), a maximum of 2 when $r = -1$ (i.e., the two variables go in exact opposite directions), and the value $\sqrt{2}$ when $r = 0$ (i.e., the two variables are uncorrelated and are at right-angles to each other).

An inter-variable distance can also be defined in the same way for other types of correlation coefficients and measures of association that lie between –1 and +1, for example the (*Spearman*) *rank correlation*. This so-called nonparametric measure of correlation is the regular correlation coefficient applied to the *ranks* of the data. In the sample of size 3 in Exhibit 6.1(a) pollution and depth have the following ranks:

SITE	*Pollution*	*Depth*
s1	2	2
s2	1	3
s3	3	1

where, for example in the pollution column, the value 2.8 for site 2 is the lowest value, hence rank 1, then 4.8 is the next lowest value, hence rank 2, and 5.4 is the highest value, hence rank 2. The correlation between these two vectors is –1, since the ranks are indeed direct opposites – therefore, the distance between them based on the rank correlation is equal to 2. Exhibit 6.4 shows the usual linear correlation coefficient, the Spearman rank correlation, and their associated distances, for the three variables based on their complete set of 30 sample values. This example confirms empirically that the results are more of less the same using ranks instead of the original values: most of the correlation is in the ordering of the values rather than their actual numerical amounts. The rank correlation is also more *robust*, which means that it is less affected by unusual or extreme values in the data.

Exhibit 6.4 Correlations and associated distances between the three continuous variables of Exhibit 1.1: first the regular correlation coefficient on the continuous data, and second the rank correlation.

Correlation	*Poll.*	*Depth*	*Temp.*
Pollution	1	-0.3955	-0.0921
Depth	-0.3955	1	-0.0034
Temperature	-0.0921	-0.0034	1

Distance	*Poll.*	*Depth*	*Temp.*
Pollution	0	1.6706	1.4779
Depth	1.6706	0	1.4166
Temperature	1.4779	1.4166	0

Rank correlation			
Pollution	1	-0.4233	-0.0525
Depth	-0.4233	1	-0.0051
Temperature	-0.0525	-0.0051	1

Distance			
Pollution	0	1.6872	1.4509
Depth	1.6872	0	1.4178
Temperature	1.4509	1.4178	0

Distances between count variables

When it comes to the count data of Exhibit 1.1, the various distance measures considered in Chapter 5 can be used, except that it makes little sense to apply the chi-square distance or the Bray-Curtis dissimilarity to the raw data – these should be expressed as proportions, (i.e., relativized) with respect to their column sums. The two measures then turn out as as in Exhibit 6.5, which in the accompanying scatterplot shows them to be very similar, apart from their respective scales, of course. The scatterplot is shown using the same horizontal and vertical scales as in Exhibit 5.4 in order to demonstrate that the spread of the distances between the columns are much less than the corresponding spread between the rows.

Distances between categorical variables and between categories

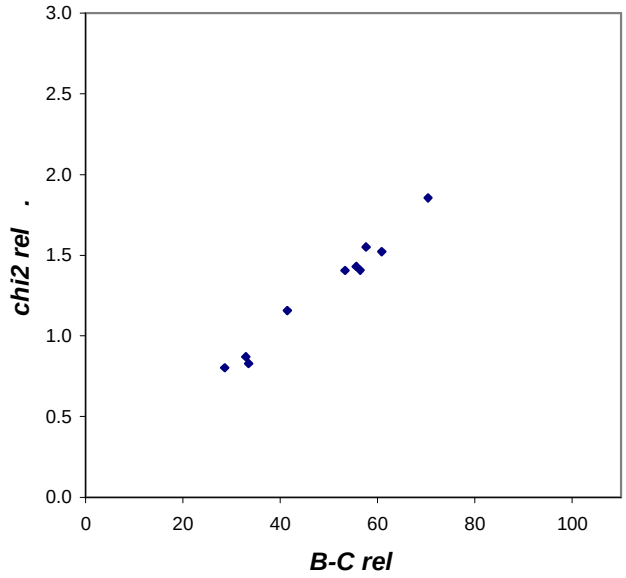
Measures of distance between samples based on a set of dichotomous variables were defined on the basis of a 2×2 table of counts of matches and mismatches, and this same idea can be applied to the dichotomous variables based on their values across the samples: in other words, the first types of match count would be the number of samples for which both variables were ‘present’, and so on. Then the various measures of dissimilarity (5.5), (5.6) and (5.7) apply, in particular the one based on the correlation coefficient. But after (5.7) we proposed that $1-r$ would make a reasonable measure of dissimilarity (or ½ $(1-r)$ to give it a

range of 0 to 1) – now, based on our study of the geometry of variables in this chapter, a better choice would be $\sqrt{2}\sqrt{1-r}$ (or $\sqrt{1-r}/\sqrt{2}$ if again one prefers a value between 0 and 1), because this is a Euclidean distance and is therefore a true metric, whereas the previous definition turns out to be a squared Euclidean distance.

Exhibit 6.5 Chi-square distances and Bray-Curtis dissimilarities between the five count variables, in both cases based on their proportions across the samples (i.e., removing the effect of different levels of abundances for each group of species).

chi2	a	b	c	d
b	0.802			
c	1.522	1.407		
d	0.87	0.828	1.157	
e	1.406	1.55	1.855	1.43

B-C	a	b	c	d
b	28.6			
c	60.9	56.4		
d	32.9	33.5	41.4	
e	53.3	57.6	70.4	55.6



For categorical variables with more than two categories, there are two types of distances in question: distances between variables, and distances between categories of variables, both not easy to deal with. At the level of the variable, we can define a measure of similarity, or association, and there are quite a few different ways to do this. The easiest way is to use a variation on the chi-square statistic computed on the cross-tabulation of the pair of variables. In our introductory data of Exhibit 1.1 there is only one categorical variable, but let us categorize depth into three categories: low, medium and high depth, by simply cutting up the range of depth into three parts, so there are 10 sites in each category; in other words, the crisp coding of a continuous variable described in Chapter 3. The cross-tabulation of depth and sediment is then as follows (notice that the counts of the depth categories are not exactly 10 each, because of some tied values in the depth data):

Exhibit 6.6 Cross-tabulation of depth, categorized into three categories, and sediment type, for the data of Exhibit 1.1.

		Sediment		
		C	S	G
Depth	low	6	5	0
	medium	3	5	1
	high	2	1	7

The chi-square statistic for this table equals 15.58 (we discuss significance testing below), but this depends on the sample size, so an alternative measure divides the chi-square statistic by the sample size, 30 in this case, to obtain the mean-square contingency coefficient, denoted by $\phi^2 = 15.58/30 = 0.519$. We will rediscover ϕ^2 in later chapters, since it is identical to what is called the inertia in correspondence analysis, which measures the total variance of a data matrix.

Now ϕ^2 measures how similar the variables are, but we need to invert this measure somehow to get a measure of dissimilarity. The maximum value of ϕ^2 turns out to be one less than the number of rows or columns of the cross-tabulation, whichever is the smaller: in this case there are 3 rows and 3 columns, so one less than the minimum is 2. You can verify that if a 3×3 cross-tabulation has only one nonzero count in each row (likewise in each column), that is there is perfect association between the two variables, then $\phi^2 = 2$. So a dissimilarity could be defined as $2-\phi^2$, equal to 1.481 in this example.

There are many alternatives, and we only mention one more. Since the maximum of ϕ^2 for an $I \times J$ cross-tabulation is $\min\{I-1, J-1\}$, we could divide ϕ^2 by this maximum. The so-called Cramer's V coefficient does this but also takes the square root of the result:

$$V = \sqrt{\frac{\phi^2}{\min\{I-1, J-1\}}} \qquad (6.5)$$

This coefficient has the properties of a correlation coefficient, but is never negative because the idea of negative correlation for categorical variables has no meaning: variables are either not associated or have some level of (positive) association. Once again, subtracting V from 1 would give an alternative measure of dissimilarity.

Distances between categories

For a categorical variable such as sediment in Exhibit 1.1, measuring the distance between the categories C, S and G makes no sense at all, because they never co-occur in this data set (in Chapter 5 we looked at other data where they could co-occur, but then we split the sediment types into separate dichotomous categorical variables). In this sense their correlations are always –1, and they are all at maximum distance apart. We can only measure their similarity in their relation to other variables. For example, in Exhibit 6.6 the sediment categories are cross-tabulated with depth, and this induces a measure of distance

between the sediment types. An appropriate measure of distance would be the chi-square distance between the column profiles of this table, which gives the following distances:

chi2	C	S
S	0.397	
G	1.525	1.664

This shows that G is the most dissimilar to the other two sediment types, in their respective relations with depth, which can be seen clearly in Exhibit 6.6.

Testing correlations: permutation testing

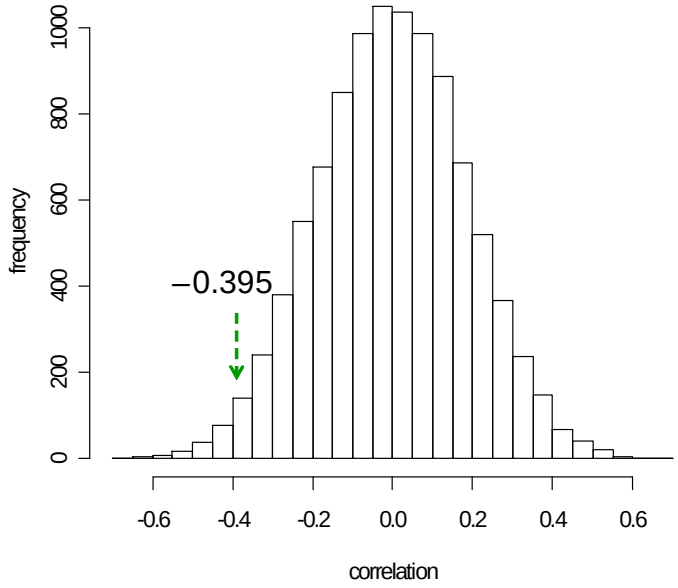
Researchers usually like to have some indication of statistical significance of the relationships between their variables, so the question arises how to test the correlation coefficients and dissimilarity measures that have been described in this chapter. Tests do exist of some of these statistical quantities, for example there are several ways to test for the correlation coefficient, assuming that data are normally distributed, or with some other known distribution that lends itself to working out the distribution of the correlation. An alternative way of obtaining those precious *p*-values is to perform permutation testing, which does not rely on knowledge of the underlying distribution of the variables. The idea is simple, all that one needs is a fast computer and the right software, and this presents no problem these days. Under the null hypothesis of no correlation between the two variables, any pair of observations in the same sample has no relation between them at all and could have been with any of the samples. So, we can generate as many values of the correlation coefficient under the null hypothesis by simply permuting the values across the samples. This process generates what is called the *permutation distribution*, and the exact permutation distribution can be determined if we consider all the possible permutations of the data set. But even with a sample of size 30, the 30! possible permutations are too many to compute, so we estimate the distribution by using a random sample of permutations.

This is exactly what we did in Chapter 1 to estimate the *p*-value for the correlation between pollution and depth. A total of 9999 permutations were made of the 30 observations of one of the variables (the other one can be kept fixed), and Exhibit 6.7 is the histogram of the resulting correlations, with the actually observed correlation of -0.395 indicated. The p-value is the probability of the observed results and any more extreme one, and since this is a two-sided testing problem, we have to count how many of the 10000 permutations (including the observed one, this is why we generate 9999) are equal or more extreme than –0.395. It turns out there are 159 values more extreme on the negative side (≤ –0.395) and 137 on the positive side (≥ 0.395), giving an estimated p-value of 296/10000 = 0.0296. This is very close to the *p*-value of 0.0305, which is calculated from the classical *t*-test for the correlation coefficient:

$$t = \frac{r}{\sqrt{(1-r^2)/(n-2)}}, \quad t\text{-distribution with } n\text{–}2 \text{ degrees of freedom } (n\text{=}30 \text{ here})$$

$$= -2.279 \text{ , corresponding to a 2-sided } p\text{-value of } 0.0305$$

Exhibit 6.7 Estimated permutation distribution for the correlation between pollution and depth (data from Exhibit 1.1). The observed value of -0.395



SUMMARY: Measures of distance and correlation between variables

1. Two variables that have been centred define two directions in the multidimensional space of the samples.
2. The cosine of the angle subtended by these two direction vectors is the classic linear correlation coefficient between the variables.
3. There are advantages in having the variable vectors unit length: this means that the standardized variables (which have been divided by their respective standard deviations) are further divided by $\sqrt{I-1}$, where I is the sample size. These are then called unit variables.
4. The distance d between the points defining the unit variables is $d = \sqrt{2}\sqrt{1-r}$, where r is the correlation coefficient. Conversely, the correlation is $r = 1-\frac{1}{2}d^2$.
5. Distances between count variables can be calculated in a similar way to distances between samples based on count data, with the restriction that the variables be expressed as profiles, that is as proportions relative to their total across the samples.
6. Distances between dichotomous categorical variables can be calculated as before for distances between samples based on dichotomous variables.
7. Distances between categories of a polychotomous variable can only be calculated in respect of the relation of this variable with another variable.
8. Permutation tests are convenient computer-based methods of arriving at p-values for quantifying the significance of relationships between variables.