Multivariate Analysis of Ecological Data

MICHAEL GREENACRE
Professor of Statistics at the Pompeu Fabra University in Barcelona, Spain

RAUL PRIMICERIO
Associate Professor of Ecology, Evolutionary Biology and Epidemiology
at the University of Tromsø, Norway

Chapter 6 Offprint

Measures of Distance and Correlation between Variables

First published: December 2013
ISBN: 978-84-92937-50-9

Supporting websites:
www.fbbva.es
www.multivariatestatistics.org

Fundación BBVA

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, and we can consider measures of distance and dissimilarity between these 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.

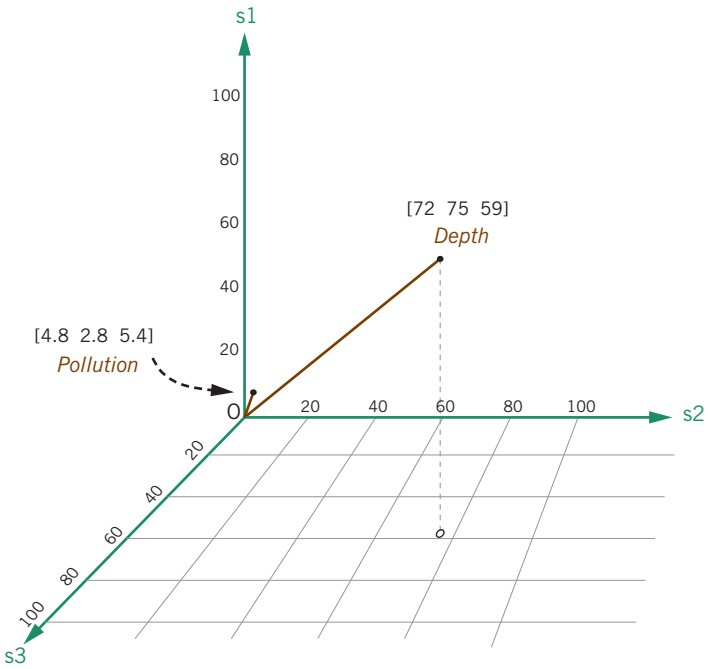
Contents

The geometry of variables

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 for which we have J measurements, for any J. Now, rather than considering the samples, the rows of the data matrix,

Exhibit 6.1: *(a) Two variables measured in three samples (sites in this case), viewed in three dimensions, using original scales; (b) Standardized values; (c) Same variables plotted in three dimensions using standardized values. 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*

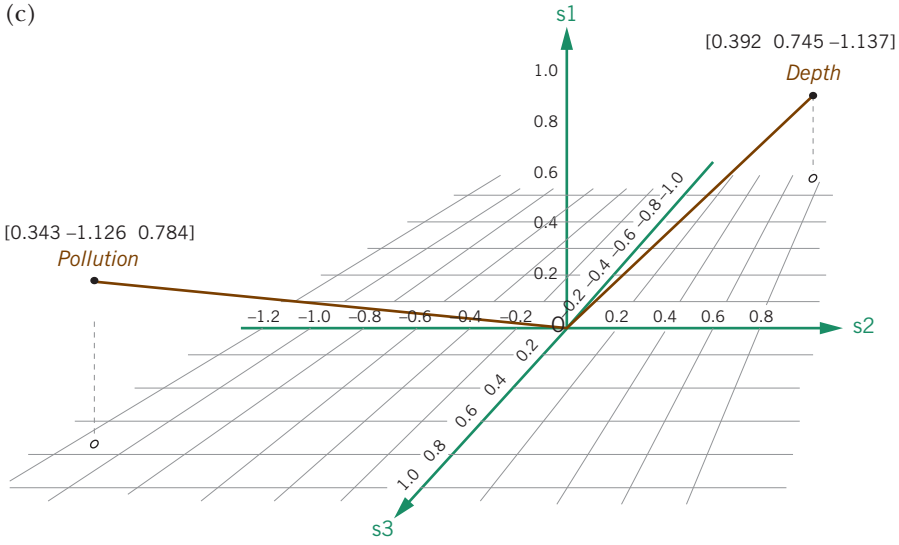
(a)



(b)

Site	*Depth*	*Pollution*
s1	0.392	0.343
s2	0.745	–1.126
s3	–1.137	0.784

(c)



we turn our attention to the variables (the columns of the data matrix) and their sets of observed values across the I samples. To be able to visualize two variables in I-dimensional space, we choose $I = 3$, since more than 3 is impossible to display or imagine. Exhibit 6.1(a) shows the variables depth and pollution according to the first three samples in Exhibit 1.1, with depth having values (i.e. coordinates) [72 75 59] and pollution [4.8 2.8 5.4]. Notice that it is the samples that now form the axes of the space. The much lower values for pollution compared to those for depth causes the distance between the variables to be dominated by this scale effect. Standardizing overcomes this effect – Exhibit 6.1(b) shows standardized values 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 6.1(c) shows the two variables plotted according to these standardized values.

Correlation coefficient as an angle cosine

Exhibit 6.2 now shows the triangle formed by the two vectors in Exhibit 6.1(c) and the origin O, taken out of the three-dimensional space, and laid flat. From the coordinates of the points we can easily calculate the lengths of the three sides a, b and c of the triangle (where the sides a and b subtend the angle θ shown), so by using the cosine rule ($c^2 = a^2 + b^2 - 2ab\cos(\theta)$, which we all learned at school) we can calculate the cosine of the angle θ between the vectors, which turns out to be –0.798, exactly the correlation between pollution and depth (the angle is 142.9°). Notice that this is the correlation calculated in this illustrative sample of size 3, not in the original sample of size 30, where the estimated correlation is –0.396.

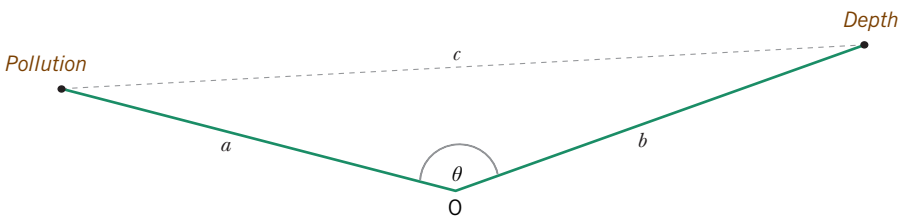


Exhibit 6.2:
Triangle of pollution and depth vectors with respect to origin (O) *taken out of Exhibit 6.1(c) and laid flat*

Hence we have illustrated the result that the cosine of the angle between two standardized variables, plotted as vectors in the space of dimensionality I, the number of samples, is their correlation coefficient.

Correlation coefficient as a scalar product

But there is yet another way of interpreting the correlation coefficient geometrically. First we have to convert the standardized pollution and depth values to so-called *unit variables*. At present they are standardized to have variance 1, but a unit variable has sum of squares equal to 1 – in other words, its length is 1. Since the variance of I centred values is defined as $1/(I-1)$ times their sum

of squares, it follows that the sum of squares equals $(I-1)$ times the variance. By dividing the standardized values of pollution and depth in Exhibit 6.1(b) by $\sqrt{I-1}$, equal to $\sqrt{2}$ in this example, the standardized variables are converted to unit variables:

Site	*Depth*	*Pollution*
s1	0.277	0.242
s2	0.527	–0.796
s3	–0.804	0.554

[it can be checked that $0.277^2 + 0.527^2 + (-0.804)^2 = 0.242^2 + (-0.796)^2 + 0.554^2 = 1$]. The correlation coefficient then has the alternative definition as the sum of the products of the elements of the unit variables:

$$(0.242 \times 0.277) + (-0.796 \times 0.527) + (0.554 \times (-0.804)) = -0.798$$

i.e., the *scalar product:*

$$r_{jj'} = \sum_{i=1}^{I} x_{ij} x_{ij'} \tag{6.1}$$

where x_{ij} are the values of the unit variables.

The concept of a scalar product underlies many multivariate techniques which we shall introduce later. It is closely 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 is the same as Exhibit 6.2 except that the sides *a* and *b* of the triangle are now shortened

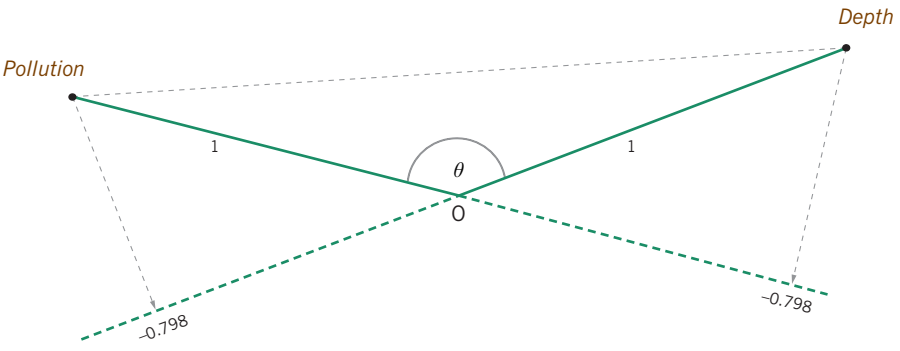


Exhibit 6.3: *Same triangle as in Exhibit 6.2, but with variables having unit length (i.e., unit variables. The projection of either variable onto the direction defined by the other variable vector will give the value of the correlation, cos(θ). (The origin O is the zero point – see Exhibit 6.1(c) – and the scale is given by the unit length of the variables.)*

to length 1 as unit variables (the subtended angle is still the same). The projection of either variable onto the axis defined by the other one gives the exact value of the correlation coefficient.

Distances based on correlation coefficients

When variables are plotted in their unit form as in Exhibit 6.3, the squared distance between the variable points is computed (again using the cosine rule) as $1 + 1 - 2\cos(\theta) = 2 - 2r$, where r is the correlation. In general, therefore, a distance $d_{jj'}$ between variables j and j' can be defined in terms of their correlation coefficient $r_{jj'}$ as follows:

$$d_{jj'} = \sqrt{2 - 2r_{jj'}} = \sqrt{2}\sqrt{1 - r_{jj'}} \tag{6.2}$$

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 variables go in exact opposite directions), and $d_{jj'} = \sqrt{2}$ when $r = 0$ (i.e., the two variables are uncorrelated and are at right-angles to each other). For example, the distance between pollution and depth in Exhibit 6.3 is $\sqrt{2}\sqrt{1 - (-0.798)} = 1.896$.

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

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 3. 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, the maximum distance possible. 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 or less the same using ranks instead of the original values: that is, most of the correlation is in the ordering of the

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	*Depth*	*Pollution*	*Temperature*
Depth	1	−0.3955	−0.0034
Pollution	−0.3955	1	−0.0921
Temperature	−0.0034	−0.0921	1

Distance	*Depth*	*Pollution*	*Temperature*
Depth	0	1.6706	1.4166
Pollution	1.6706	0	1.4779
Temperature	1.4166	1.4779	0

Rank correlation	*Depth*	*Pollution*	*Temperature*
Depth	1	−0.4233	−0.0051
Pollution	−0.4233	1	−0.0525
Temperature	−0.0051	−0.0525	1

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

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.

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 to measure distances between species. It makes little sense, however, 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 in Exhibit 6.5, where the scatterplot shows them to be very similar, apart from their different scales, of course. The scatterplot is shown using the same horizontal and vertical scales as in Exhibit 5.4(b) in order to demonstrate that the spread of the distances between the columns is 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: for example, 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 r. But after (5.7) we proposed that $1 - r$ would make a reasonable measure of dissimilarity (or $\frac{1}{2}(1 - r)$ 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.

chi2	a	b	c	d
b	0.802			
c	1.522	1.407		
d	0.870	0.828	1.157	
e	1.406	1.550	1.855	1.430

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

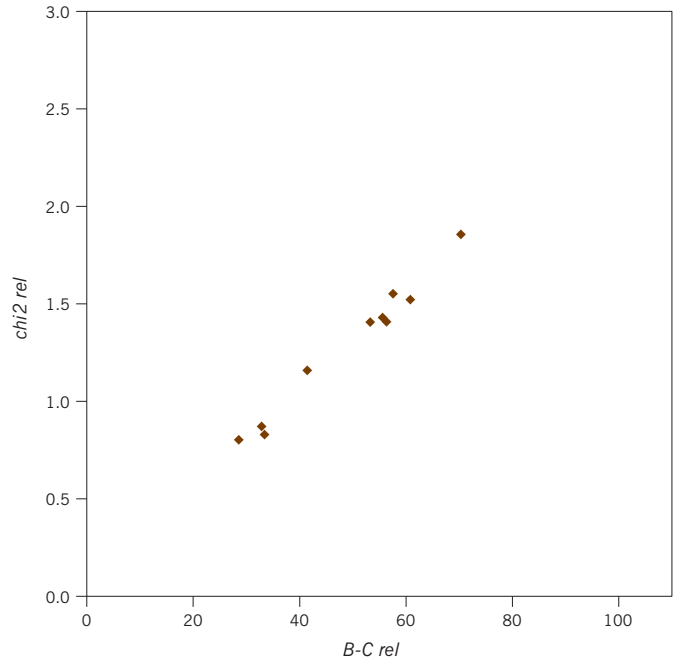


Exhibit 6.5: *Chi-square distances and Bray-Curtis dissimilarities between the five species variables, in both cases based on their proportions across the samples (i.e., removing the effect of different levels of abundances for each species). The two sets of values are compared in the scatterplot*

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 – this is the crisp coding of a continuous variable described in Chapter 3. The cross-tabulation of depth and sediment is then given in Exhibit 6.6 (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, 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 so-called *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:

$$V = \sqrt{\frac{\phi^2}{\min\{I-1, J-1\}}} \tag{6.3}$$

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 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 the table in Exhibit 6.6, which gives the following distances:

chi2	C	S
S	0.397	
G	1.525	1.664

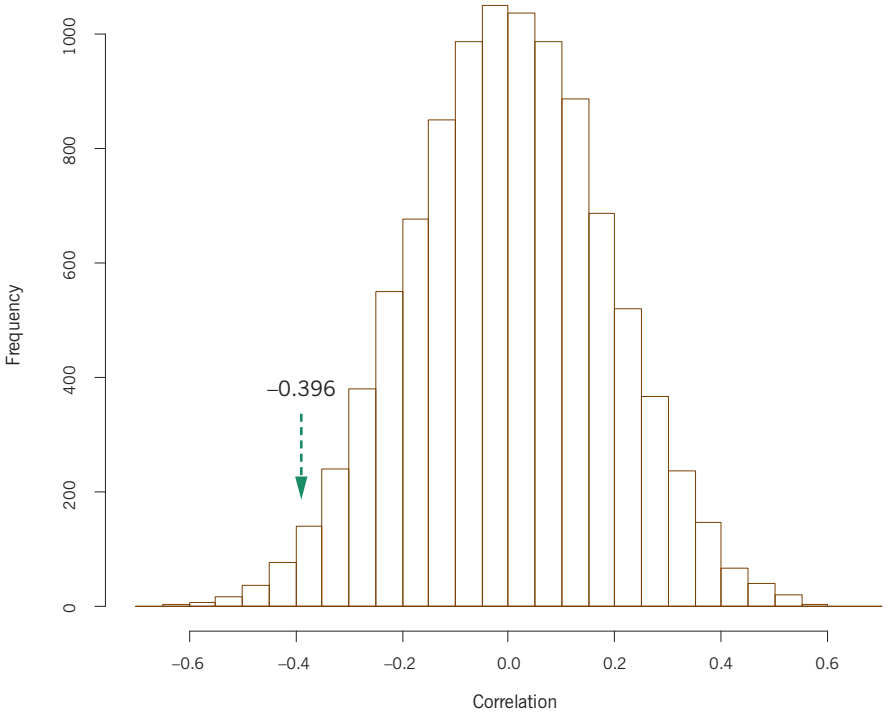
This shows that G is the most dissimilar to the other two sediment types, in terms of their respective relations with depth, which can be seen clearly in Exhibit 6.6.

Testing correlations: an introduction to 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 a *p*-value 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, say between depth and pollution, the pairing of observations in the same sample is irrelevant, so we can associate any value of depth, say, with any value of pollution. Thus we generate many values of the correlation coefficient under the null hypothesis by 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 9,999 random permutations were made of the 30 observations of one of the variables, say depth (with the order of pollution kept fixed), and Exhibit 6.7 is the histogram of the resulting correlations, with the actually observed correlation of –0.396 indicated. The *p*-value is the probability of the observed result and any more extreme ones, and since this is a two-sided

Exhibit 6.7: *Estimated permutation distribution for the correlation between pollution and depth (data from Exhibit 1.1), for testing the null hypothesis that the correlation is zero. The observed value of –0.396 is shown, and the permutation test consists in counting how many of the simulated correlations have an absolute value greater than or equal to 0.396*



testing problem, we have to count how many of the 10,000 permutations (including the observed one, this is why we generate 9,999) are equal or more extreme than 0.396 in absolute value. It turns out there are 159 values more extreme on the negative side (≤ -0.396) and 137 on the positive side (≥ 0.396), giving an estimated p-value of 296/10,000 = 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:

$$t = \frac{r}{\sqrt{(1-r^2)/(n-2)}} = -2.279,$$

corresponding to a two-sided p-value of 0.0305, using the t-distribution with $n-2$ degrees of freedom ($n = 30$ here).

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 set of observations for each variable of unit length. This is obtained by dividing the standardized variables by $\sqrt{I-1}$, where I is the sample size, so that the sum of squares of their values is equal to 1. These are then called *unit variables*.

4. The distance d between the points defined by 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 for 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.

List of Exhibits