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 4 Offprint

Measures of Distance between Samples: Euclidean

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

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

Fundación BBVA

Chapter 4

Measures of Distance between Samples: Euclidean

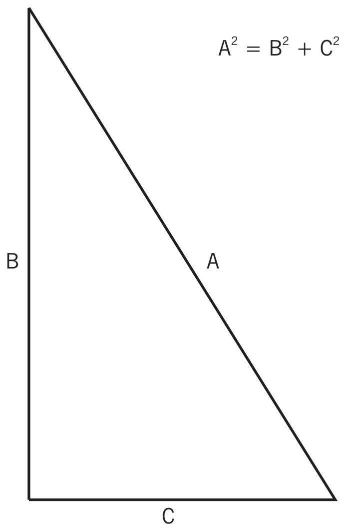
We will be talking a lot about distances in this book. The concept of distance between two samples or between two variables is fundamental in multivariate analysis – almost everything we do has a relation with this measure. If we talk about a single variable we take this concept for granted. If one sample has a pH of 6.1 and another a pH of 7.5, the absolute difference between them is 1.4. But on the pH line, the values 6.1 and 7.5 are at a distance apart of 1.4 units, and this is how we want to start thinking about data: points on a line, points in a plane, … even points in a 10-dimensional space! So, given two samples with not one measurement on them but several, how do we measure the difference between them? There are many possible answers to this question, and we devote three chapters to this topic. In the present chapter we consider what are called *Euclidean* distances, which coincide with our basic physical idea of distance, but generalized to multidimensional space.

Contents

Pythagoras' theorem

Pythagoras' theorem is at the heart of most of the multivariate analysis presented in this book, and particularly the graphical approach to data analysis that we are strongly promoting. When you see the word "square" mentioned in a statistical text (for example, chi-square or least squares), you can be almost sure that the corresponding theory has some relation to this theorem. We first

Exhibit 4.1: *Pythagoras' theorem in the familiar right-angled triangle, and the monument to this triangle in the port of Pythagorion, Samos island, Greece, with Pythagoras himself forming one of the sides (Photo: Michael Greenacre)*



show the theorem in its simplest and most familiar two-dimensional form, before showing how easy it is to generalize it to multidimensional space. In a right-angled triangle, the square on the hypotenuse (the side denoted by A in Exhibit 4.1) is equal to the sum of the squares on the other two sides (B and C); that is, $A^2 = B^2 + C^2$.

Euclidean distance

The immediate consequence of this is that the squared length of a vector $\mathbf{x} = [\,x_1 \;\; x_2\,]$ is the sum of the squares of its coordinates (see triangle OPA in Exhibit 4.2, or triangle OPB – $|OP|^2$ denotes the squared length of $\mathbf{x}$, that is the distance between point O, with both co-ordinates zero, and P); and the squared distance between two vectors $\mathbf{x} = [\,x_1 \;\; x_2\,]$ and $\mathbf{y} = [\,y_1 \;\; y_2\,]$ is the sum of squared differences in their coordinates (see triangle PQD in Exhibit 4.2; $|PQ|^2$ denotes the squared distance between points P and Q). To denote the distance between vectors $\mathbf{x}$ and $\mathbf{y}$ we can use the notation $d_{\mathbf{x},\mathbf{y}}$ so that this last result can be written as:

$$d_{x,y}^2 = (x_1 - y_1)^2 + (x_2 - y_2)^2 \qquad (4.1)$$

that is, the distance itself is the square root

$$d_{x,y} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \qquad (4.2)$$

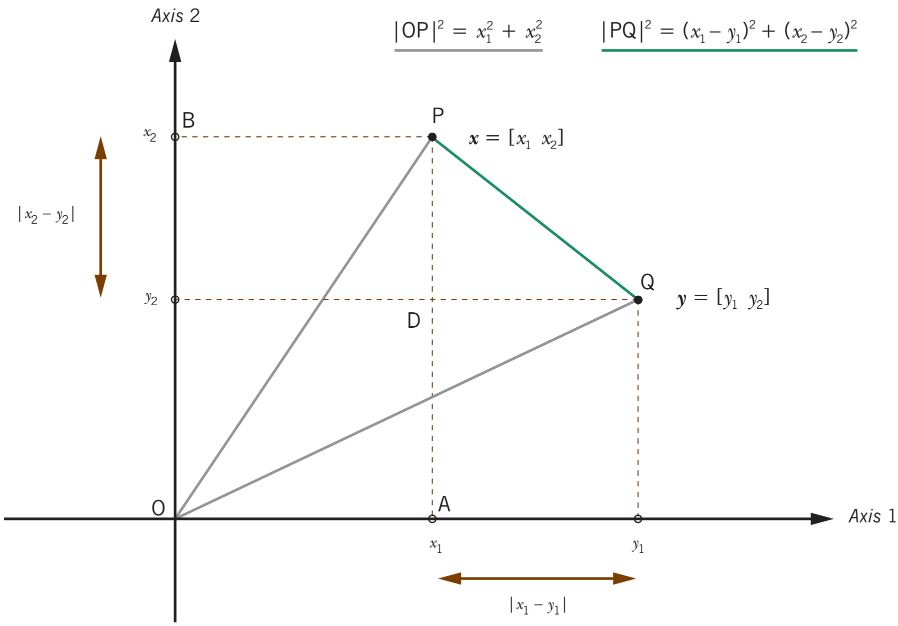


Exhibit 4.2:
Pythagoras' theorem applied to distances in two-dimensional space

What we called the *squared length* of **x**, the distance between points P and O in Exhibit 4.2, is the distance between the vector $\mathbf{x} = [x_1 \;\; x_2]$ and the zero vector $\mathbf{0} = [0 \;\; 0]$:

$$d_{x,0} = \sqrt{x_1^2 + x_2^2} \tag{4.3}$$

which we could just denote by $d_{\mathbf{x}}$. The zero vector is called the *origin* of the space.

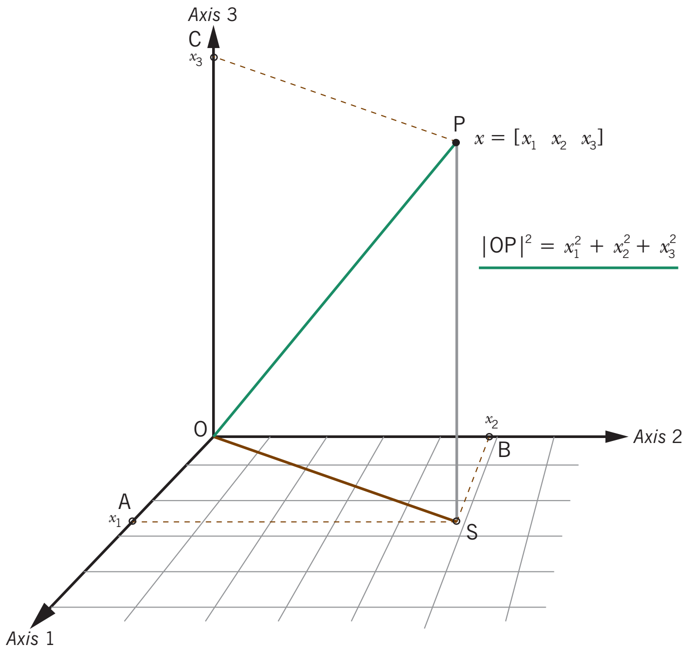
We move immediately to a three-dimensional point $\mathbf{x} = [x_1 \;\; x_2 \;\; x_3]$, shown in Exhibit 4.3. This figure has to be imagined in a room where the origin O is at the corner – to reinforce this idea "floor tiles" have been drawn on the plane of axes 1 and 2, which is the "floor" of the room. The three coordinates are at points A, B and C along the axes, and the angles AOB, AOC and COB are all 90° as well as the angle OSP at S, where the point P (depicting vector **x**) is projected onto the "floor". Using Pythagoras' theorem twice we have:

$$|OP|^2 = |OS|^2 + |PS|^2 \qquad \text{(because of right-angle at S)}$$
$$|OS|^2 = |OA|^2 + |AS|^2 \qquad \text{(because of right-angle at A)}$$

and so

$$|OP|^2 = |OA|^2 + |AS|^2 + |PS|^2$$

Exhibit 4.3:
Pythagoras' theorem extended into three dimensional space



that is, the squared length of **x** is the sum of its three squared coordinates, hence the length is

$$d_x = \sqrt{x_1^2 + x_2^2 + x_3^2}$$

It is also clear that placing a point Q in Exhibit 4.3 to depict another vector **y** and going through the motions to calculate the distance between **x** and **y** will lead to

$$d_{x,y} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2} \tag{4.4}$$

Furthermore, we can carry on like this into four or more dimensions, in general J dimensions, where J is the number of variables. Although we cannot draw the geometry any more, we can express the distance between two J-dimensional vectors **x** and **y** as:

$$d_{x,y} = \sqrt{\sum_{j=1}^{J} (x_j - y_j)^2} \tag{4.5}$$

This well-known distance measure, which generalizes our notion of physical distance in two- or three-dimensional space to multidimensional space, is called the *Euclidean distance*.

Standardized Euclidean distance

Let us consider measuring the distances between our 30 samples in Exhibit 1.1, using the three continuous variables depth, pollution and temperature. What would happen if we applied formula (4.5) to measure distance between the last two samples, s29 and s30, for example? Here is the calculation:

$$d_{\text{s29,s30}} = \sqrt{(51-99)^2 + (6.0-1.9)^2 + (3.0-2.9)^2}$$

$$= \sqrt{2304 + 16.81 + 0.01} = \sqrt{2320.82} = 48.17$$

The contribution of the first variable depth to this calculation is huge – one could say that the distance is practically just the absolute difference in the depth values (equal to |51-99| = 48) with only tiny additional contributions from pollution and temperature. This is the problem of standardization discussed in Chapter 3 – the three variables have completely different units of measurement and the larger depth values have larger inter-sample differences, so they will dominate in the calculation of Euclidean distances.

Some form of transformation of the data is necessary to balance out the contributions, and the conventional way to do this is to make all variables have the same variance of 1. At the same time we centre the variables at their means – this centring is not necessary for calculating distance, but it makes the variables all have mean zero and thus easier to compare. This transformation, commonly called *standardization*, is thus as follows:

$$\text{standardized value} = (\text{original value} - \text{mean}) \;/\; \text{standard deviation} \qquad (4.6)$$

The means and standard deviations (sd) of the three variables are:

	Depth	*Pollution*	*Temperature*
mean	74.433	4.517	3.057
sd	15.615	2.141	0.281

leading to the table of standardized values given in Exhibit 4.4. These values are now on comparable standardized scales, in units of standard deviation with respect to the mean. For example, the standardized pollution value 0.693 for row s29 would signify 0.693 standard deviations above the mean, while −1.222 for row s30 would signify 1.222 standard deviations below the mean. The distance calculation thus aggregates squared differences in standard deviation units of each variable. As an example, the distance between the last two sites of the table in Exhibit 4.4 is:

$$d_{s29,s30} = \sqrt{[-1.501 - 1.573]^2 + [0.693 - (-1.222)]^2 + [-0.201 - (-.557)]^2}$$

$$= \sqrt{9.449 + 3.667 + 0.127} = \sqrt{13.243} = 3.639$$

For this particular pair of sites the difference in temperatures is still small but pollution now has a higher contribution than before. Depth still plays the largest role in this particular example, even after standardization, but this contribution is

Exhibit 4.4: *Standardized values of the three continuous variables of Exhibit 1.1*

Site No.	Environmental variables		
	Depth	*Pollution*	*Temperature*
s1	–0.156	0.132	1.576
s2	0.036	–0.802	–1.979
s3	–0.988	0.413	–1.268
s4	–0.668	1.720	–0.557
s5	–0.860	–0.288	0.154
s6	1.253	–0.895	1.576
s7	–1.373	0.039	–0.557
s8	–0.860	0.272	0.865
s9	–0.412	–0.288	1.221
s10	–0.348	2.561	–0.201
s11	–1.116	0.926	0.865
s12	0.613	–0.335	0.154
s13	–1.373	2.281	–0.201
s14	0.549	0.086	–1.979
s15	1.637	1.020	–0.913
s16	0.613	–0.802	–0.201
s17	1.381	0.880	0.154
s18	–0.028	–0.054	–0.913
s19	0.292	–0.662	1.932
s20	–0.092	0.506	–0.201
s21	–0.988	–0.101	1.221
s22	–1.309	–1.222	–0.913
s23	1.317	–0.989	–0.557
s24	–0.668	–0.101	–0.201
s25	1.445	–1.175	–0.201
s26	0.228	–0.942	1.221
s27	0.677	–1.129	–0.201
s28	1.125	–0.522	0.865
s29	–1.501	0.693	–0.201
s30	1.573	–1.222	–0.557

	s1	s2	s3	s4	s5	s6	· · ·	s24	s25	s26	s27	s28	s29
s2	3.681												
s3	2.977	1.741											
s4	2.708	2.980	1.523										
s5	1.642	2.371	1.591	2.139									
s6	1.744	3.759	3.850	3.884	2.619								
s7	2.458	2.171	0.890	1.823	0.935	3.510							
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱						
s25	2.727	2.299	3.095	3.602	2.496	1.810	· · ·	2.371					
s26	1.195	3.209	3.084	3.324	1.658	1.086	· · ·	1.880	1.886				
s27	2.333	1.918	2.507	3.170	1.788	1.884	· · ·	1.692	0.770	1.503			
s28	1.604	3.059	3.145	3.204	2.122	0.813	· · ·	2.128	1.291	1.052	1.307		
s29	2.299	2.785	1.216	1.369	1.224	3.642	· · ·	1.150	3.488	2.772	2.839	3.083	
s30	3.062	2.136	3.121	3.699	2.702	2.182	· · ·	2.531	0.381	2.247	0.969	1.648	3.639

Exhibit 4.5:
Standardized Euclidean distances between the 30 samples, based on the three continuous environmental variables, showing part of the triangular distance matrix

justified now, since depth does show the biggest standardized difference between the samples. We call this the *standardized Euclidean distance*, meaning that it is the Euclidean distance calculated on standardized data. It will be assumed that standardization refers to the form defined by (4.6), unless specified otherwise.

We can repeat this calculation for all pairs of samples. Since the distance between sample A and sample B will be the same as between sample B and sample A, we can report these distances in a triangular matrix – Exhibit 4.5 shows part of this distance matrix, which contains a total of ½ × 30 × 29 = 435 distances.

Readers might ask how all this has helped them – why convert a data table with 90 numbers into one that has 435, almost five times more? Were the histograms and scatterplots in Exhibits 1.2 and 1.4 not enough to understand these three variables? This is a good question, but we shall have to leave the answer to Part 3 of the book, from Chapter 7 onwards, when we describe actual analyses of these distance matrices. At this early stage in the book, we can only ask readers to accept that the computation of interpoint distances is an intermediate step in a process that will lead to an eventual simplification in interpreting the data structure – having a measure of distance (i.e., difference) between samples based on several variables is the key to this process.

Weighted Euclidean distance

The standardized Euclidean distance between two J-dimensional vectors can be written as:

$$d_{x,y} = \sqrt{\sum_{j=1}^{J}\left(\frac{x_j}{s_j} - \frac{y_j}{s_j}\right)^2} \tag{4.7}$$

where s_j is the sample standard deviation of the j-th variable. Notice that we need not subtract the j-th mean from x_j and y_j because the means will just cancel out in the differencing. Now (4.7) can be rewritten in the following equivalent way:

$$d_{x,y} = \sqrt{\sum_{j=1}^{J}\frac{1}{s_j^2}(x_j - y_j)^2} = \sqrt{\sum_{j=1}^{J} w_j (x_j - y_j)^2} \tag{4.8}$$

where $w_j = 1/s_j^2$ is the inverse of the j-th variance. We can think of w_j as a *weight* attached to the j-th variable: in other words, we compute the usual squared differences between the variables on their original scales, as we did in the (unstandardized) Euclidean distance, but then multiply these squared differences by their corresponding weights. Notice in this case how the weight of a variable with high variance is low, while the weight of a variable with low variance is high, which is another way of thinking about the compensatory effect produced by standardization. The weights of the three variables in our example are (to 4 significant figures) 0.004101, 0.2181 and 12.64 respectively, showing how much the depth variable is downweighted and the temperature variable upweighted: depth has over 3000 times the variance of temperature, so each squared difference in (4.8) is downweighted relatively by that much. We call (4.8) *weighted Euclidean distance*.

Distances for count data

So far we have looked at the distances between samples based on continuous data, now we consider distances on count data, for example the abundance data for the five species labelled *a*, *b*, *c*, *d* and *e* in Exhibit 1.1. First, notice that these five variables apparently do not have the problem of different measurement units that we had for the continuous environmental variables – all variables are counts. There are, however, different average frequencies of counts, and as we mentioned in Chapter 3, variances of count variables can be positively related to their means. The means and variances of these five variables are as follows:

	a	*b*	*c*	*d*	*e*
mean	13.47	8.73	8.40	10.90	2.97
variance	157.67	83.44	73.62	44.44	15.69

Variable *a* with the highest mean also has the highest variance, while *e* with the lowest mean has the lowest variance. Only *d* is out of line with the others, having smaller variance than *b* and *c* but a higher mean. Because this variance–mean relationship is a natural phenomenon for count variables, not one that is just particular to any given example, some form of compensation of the variances needs to be performed, as before. It is not common for count data to be standardized as Z-scores (i.e., with mean 0, variance 1), as was the case for continuous variables in (4.6). The most common ways of balancing the contributions of count variables to the distance measure are:

- *a power transformation:* usually square root $n^{1/2}$, where n is the count value, but also double square root (i.e., fourth root $n^{1/4}$) when the variance increases faster than the mean (this situation is called *overdispersion* in the literature);
- *a "shifted log" transformation:* because of the many zeros in ecological count data, a positive number, usually 1, has to be added to the data before log-transforming; that is, $\log(1 + n)$;
- *chi-square distance:* this is a weighted Euclidean distance of the form (4.8), which we shall discuss now.

Chi-square distance

The chi-square distance is special because it is at the heart of correspondence analysis, used extensively in ecological research. The first premise of this distance function is that it is calculated on relative counts,[1] and not on the original ones, and the second is that it standardizes by the mean and not by the variance.

In our example, the count data are first converted into relative counts by dividing the rows by their row totals so that each row contains relative proportions across the species, which add up to 1. These sets of proportions are called *profiles*, site profiles in this example – see Exhibit 4.6.

The extra row at the end of Exhibit 4.6 gives the set of proportions called the *average profile*. These are the proportions calculated on the set of column totals, which are equal to 404, 262, 252, 327 and 89 respectively, with grand total 1334. Hence, 404/1334 = 0.303, 262/1334 = 0.196, etc. Chi-square distances are then calculated between the profiles, in a weighted Euclidean fashion, using the inverse of the average proportions as weights. Suppose c_j denotes the j-th element of the average profile, that is the abundance proportion of the j-th species in the whole data set. Then the *chi-square*[2] *distance*, denoted by χ, between two sites with profiles $\mathbf{x} = [x_1 \ x_2 \ \cdots \ x_J]$ and $\mathbf{y} = [y_1 \ y_2 \ \cdots \ y_J]$ is defined as:

[1] A definition of *chi-square distance* on raw counts is referred to in the bibliographical appendix.
[2] From the definition of this distance function it would have been better to call it the *chi distance function*, because it is not squared, as in the *chi-square statistic!* But the "chi-square" epithet persists in the literature, so when we talk of its square we say the "squared chi-square distance".

$$\chi_{x,y} = \sqrt{\sum_{j=1}^{J} \frac{1}{c_j}(x_j - y_j)^2} \tag{4.8}$$

Exhibit 4.6: *Profiles of the sites, obtained by dividing the rows of counts in Exhibit 1.1 by their respective row totals. The last row is the average profile, computed in the same way, as proportions of the column totals of the original table of counts*

Site No.	Species proportions				
	a	b	c	d	e
s1	0.000	0.074	0.333	0.519	0.074
s2	0.481	0.074	0.241	0.204	0.000
s3	0.000	0.370	0.333	0.296	0.000
s4	0.000	0.000	0.833	0.167	0.000
s5	0.342	0.132	0.079	0.263	0.184
s6	0.360	0.244	0.151	0.186	0.058
s7	0.321	0.214	0.000	0.393	0.071
s8	0.667	0.000	0.000	0.000	0.333
s9	0.315	0.130	0.185	0.259	0.111
s10	0.000	0.125	0.650	0.225	0.000
s11	0.000	0.276	0.276	0.207	0.241
s12	0.264	0.208	0.245	0.283	0.000
s13	0.000	0.000	0.760	0.000	0.240
s14	0.591	0.000	0.000	0.409	0.000
s15	0.154	0.000	0.385	0.462	0.000
s16	0.592	0.282	0.000	0.042	0.085
s17	1.000	0.000	0.000	0.000	0.000
s18	0.236	0.169	0.371	0.225	0.000
s19	0.053	0.132	0.316	0.421	0.079
s20	0.000	0.303	0.424	0.273	0.000
s21	0.444	0.000	0.000	0.222	0.333
s22	0.493	0.141	0.000	0.127	0.239
s23	0.146	0.171	0.024	0.415	0.244
s24	0.316	0.211	0.351	0.123	0.000
s25	0.395	0.321	0.000	0.284	0.000
s26	0.492	0.323	0.000	0.154	0.031
s27	0.333	0.236	0.000	0.347	0.083
s28	0.302	0.057	0.226	0.377	0.038
s29	0.423	0.000	0.269	0.308	0.000
s30	0.282	0.435	0.059	0.212	0.012
ave.	0.303	0.196	0.189	0.245	0.067

Exhibit 4.7 shows part of the 30 × 30 triangular matrix of chi-square distances. Once again, this is a large matrix with more numbers (435) than the original table of counts (150), and we shall see the benefit of calculating these distances

	s1	s2	s3	s4	s5	s6	· · ·	s24	s25	s26	s27	s28	s29
s2	1.139												
s3	0.855	1.137											
s4	1.392	1.630	1.446										
s5	1.093	0.862	1.238	2.008									
s6	1.099	0.539	0.887	1.802	0.597								
s7	1.046	0.845	1.081	2.130	0.573	0.555							
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮						
s25	1.312	0.817	1.057	2.185	0.858	0.495	· · ·	0.917					
s26	1.508	0.805	1.224	2.241	0.834	0.475	· · ·	0.915	0.338				
s27	1.100	0.837	1.078	2.136	0.520	0.489	· · ·	0.983	0.412	0.562			
s28	0.681	0.504	0.954	1.572	0.724	0.613	· · ·	0.699	0.844	0.978	0.688		
s29	0.951	0.296	1.145	1.535	0.905	0.708	· · ·	0.662	0.956	1.021	0.897	0.340	
s30	1.330	0.986	0.846	2.101	0.970	0.535	· · ·	0.864	0.388	0.497	0.617	1.001	1.142

Exhibit 4.7:
Chi-square distances between the 30 samples, based on the biological count data, showing part of the triangular distance matrix

from Part 3 onwards. For the moment, think of Exhibit 4.5 as a way of measuring similarities and differences between the 30 samples based on the (continuous) environmental data, while Exhibit 4.7 is the similar idea but based on the count data. Notice that the scale of distances in Exhibit 4.5 is not comparable to that of Exhibit 4.7, but the ordering of the values does have some meaning: for example, in Exhibit 4.5 the smallest standardized Euclidean distance (amongst those that we report there) is 0.381, between sites s30 and s25. In Exhibit 4.7 these two sites have one of the smallest chi-square distances as well. This means that these two sites are relatively similar in their environmental variables and also in their biological compositions. This might be an interesting finding, but we will need to study all the pairwise distances, and not just this isolated one, in order to see if there is any connection between the biological abundances and the environmental variables (this will come later).

Distances for categorical data

In our introductory example we have only one categorical variable (sediment), so the question of computing distance is fairly trivial: if two samples have the same sediment then their distance is 0, and if different then it is 1. But what if there were several categorical variables, say *K* of them? There are several possibilities, one of the simplest being to count how many matches and mismatches there are between samples, with optional averaging over variables. For example, suppose that there are five categorical variables, *C1* to *C5*, each with three categories, which we denote by a/b/c and that there are two samples with the following characteristics:

	C1	C2	C3	C4	C5
Sample 1	a	c	c	b	a
Sample 2	b	c	b	a	a

Then the number of matches is 2 and the number of mismatches is 3, hence the distance between the two samples is 3 divided by 5 (the number of variables), that is 0.6. This is called the *simple matching coefficient*. Sometimes this coefficient is expressed in terms of similarity, not dissimilarity, in which case the similarity would be equal to 0.4, the relative number of matches – so one should check which way it is being defined. Here we stick to distances, in other words dissimilarities or mismatches. Note that this coefficient is directly proportional to the squared Euclidean distance calculated between these data in dummy variable form, where each category defines a zero-one variable:

	C1a	C1b	C1c	C2a	C2b	C2c	C3a	C3b	C3c	C4a	C4b	C4c	C5a	C5b	C5c
Sample 1	1	0	0	0	0	1	0	0	1	0	1	0	1	0	0
Sample 2	0	1	0	0	0	1	0	1	0	1	0	0	1	0	0

The squared Euclidean distance sums the squared differences between these two vectors: if there is an agreement (there are two matches in this example) there is zero sum of squared differences, but if there is a discrepancy there are two differences, +1 and –1, which give a sum of squares of 2. So the sum of squared differences here is 6, and if this is expressed relative to the maximum discrepancy that can be achieved, namely 10 when there are no matches in the 5 variables, then this gives exactly the same value 0.6 as before.

There are several variations on the theme of the matching coefficient, and one of them is the chi-square distance for multivariate categorical data, which introduces a weighting of each category inverse to its mean value, as for profile data based on counts. Suppose that there are J categories in total (in the above example $J = 15$) and that the total occurrences of each category are denoted by $n_1,\ldots,n_J$, with total $n = \Sigma_j\, n_j$ (since the totals for each variable equal the sample size, n will be the sample size times the number of variables). Then define c_j as follows: $c_j = n_j/n$ and use $1/c_j$ as weights in a weighted Euclidean distance between the samples coded in dummy variable form. The idea here is, as before, that mismatches on a rare category should have a higher weight in the distance calculation than that of a frequent category. Just like the chi-square distance function is at the heart of correspondence analysis of abundance data, so this form of the chi-square for multivariate categorical data is at the heart of *multiple correspondence analysis*. We do not treat multiple correspondence analysis specifically in this book, as it is more common in the social sciences where almost all the data are categorical, for example in survey research.

SUMMARY:
Measures of distance between samples: Euclidean

1. Pythagoras' theorem extends to sets of observations (called *vectors*) in multidimensional space, for example sets of observations corresponding to a series of samples: the squared length of a vector is the sum of squares of its coordinates.
2. As a consequence, squared distances between two vectors (e.g., between two samples) in multidimensional space are the sum of squared differences in their coordinates. This multidimensional distance is called the *Euclidean distance*, and is the natural generalization of our three-dimensional notion of physical distance to more dimensions.
3. When variables are on different measurement scales, standardization is necessary to balance the contributions of the variables in the computation of distance. The Euclidean distance computed on standardized variables is called the *standardized Euclidean distance*.
4. Standardization in the calculation of distances is equivalently thought of as *weighting* the variables – this leads to the notion of Euclidean distances with any choice of weights, called *weighted Euclidean distance*.
5. A particular weighted Euclidean distance applicable to count data is the *chi-square distance*, which is calculated between the relative counts for each sample, called *profiles*, and weights each variable by the inverse of the variable's overall mean count.

List of Exhibits