1

Multivariate Data and Multivariate Analysis

1.1 Introduction

Multivariate data arise when researchers record the values of several random variables on a number of subjects or objects or perhaps one of a variety of other things (we will use the general term "units") in which they are interested, leading to a *vector-valued* or *multidimensional* observation for each. Such data are collected in a wide range of disciplines, and indeed it is probably reasonable to claim that the majority of data sets met in practise are multivariate. In some studies, the variables are chosen by design because they are known to be essential descriptors of the system under investigation. In other studies, particularly those that have been difficult or expensive to organise, many variables may be measured simply to collect as much information as possible as a matter of expediency or economy.

Multivariate data are ubiquitous as is illustrated by the following four examples:

- Psychologists and other behavioural scientists often record the values of several different cognitive variables on a number of subjects.
- Educational researchers may be interested in the examination marks obtained by students for a variety of different subjects.
- Archaeologists may make a set of measurements on artefacts of interest.
- Environmentalists might assess pollution levels of a set of cities along with noting other characteristics of the cities related to climate and human ecology.

Most multivariate data sets can be represented in the same way, namely in a rectangular format known from spreadsheets, in which the elements of each row correspond to the variable values of a particular unit in the data set and the elements of the columns correspond to the values taken by a particular variable. We can write data in such a rectangular format as

B. Everitt and T. Hothorn, *An Introduction to Applied Multivariate Analysis with R*: Use R!, DOI 10.1007/978-1-4419-9650-3_1, © Springer Science+Business Media, LLC 2011

Unit	Variable 1	...	Variable q
1	x_{11}	$\dots$	x_{1q}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_{n1}	$\dots$	x_{nq}

where n is the number of units, q is the number of variables recorded on each unit, and x_{ij} denotes the value of the jth variable for the ith unit. The observation part of the table above is generally represented by an $n \times q$ *data matrix*, $\mathbf{X}$. In contrast to the *observed* data, the theoretical entities describing the univariate distributions of each of the q variables and their joint distribution are denoted by so-called *random variables* $X_1, \dots, X_q$.

Although in some cases where multivariate data have been collected it may make sense to isolate each variable and study it separately, in the main it does not. Because the whole set of variables is measured on *each* unit, the variables will be related to a greater or lesser degree. Consequently, if each variable is analysed in isolation, the full structure of the data may not be revealed. *Multivariate statistical analysis* is the *simultaneous* statistical analysis of a collection of variables, which improves upon separate univariate analyses of each variable by using information about the *relationships* between the variables. Analysis of each variable separately is very likely to miss uncovering the key features of, and any interesting "patterns" in, the multivariate data.

The units in a set of multivariate data are sometimes sampled from a population of interest to the investigator, a population about which he or she wishes to make some inference or other. More often perhaps, the units cannot really be said to have been sampled from some population in any meaningful sense, and the questions asked about the data are then largely exploratory in nature. with the ubiquitous p-value of univariate statistics being notable by its absence. Consequently, there are methods of multivariate analysis that are essentially exploratory and others that can be used for statistical inference.

For the exploration of multivariate data, formal models designed to yield specific answers to rigidly defined questions are not required. Instead, methods are used that allow the detection of possibly unanticipated patterns in the data, opening up a wide range of competing explanations. Such methods are generally characterised both by an emphasis on the importance of graphical displays and visualisation of the data and the lack of any associated probabilistic model that would allow for formal inferences. Multivariate techniques that are largely exploratory are described in Chapters 2 to 6.

A more formal analysis becomes possible in situations when it is realistic to assume that the individuals in a multivariate data set have been sampled from some population and the investigator wishes to test a well-defined hypothesis about the parameters of that population's probability density function. Now the main focus will not be the sample data per se, but rather on using information gathered from the sample data to draw inferences about the population. And the probability density function almost universally assumed as the basis of inferences for multivariate data is the *multivariate normal*. (For a brief description of the multivariate normal density function and ways of assessing whether a set of multivariate data conform to the density, see Section 1.6). Multivariate techniques for which formal inference is of importance are described in Chapters 7 and 8. But in many cases when dealing with multivariate data, this implied distinction between the exploratory and the inferential may be a red herring because the general aim of *most* multivariate analyses, whether implicitly exploratory or inferential is to uncover, display, or extract any "signal" in the data in the presence of noise and to discover what the data have to tell us.

1.2 A brief history of the development of multivariate analysis

The genesis of multivariate analysis is probably the work carried out by Francis Galton and Karl Pearson in the late 19th century on quantifying the relationship between offspring and parental characteristics and the development of the correlation coefficient. And then, in the early years of the 20th century, Charles Spearman laid down the foundations of factor analysis (see Chapter 5) whilst investigating correlated intelligence quotient (IQ) tests. Over the next two decades, Spearman's work was extended by Hotelling and by Thurstone.

Multivariate methods were also motivated by problems in scientific areas other than psychology, and in the 1930s Fisher developed linear discriminant function analysis to solve a taxonomic problem using multiple botanical measurements. And Fisher's introduction of analysis of variance in the 1920s was soon followed by its multivariate generalisation, multivariate analysis of variance, based on work by Bartlett and Roy. (These techniques are not covered in this text for the reasons set out in the Preface.)

In these early days, computational aids to take the burden of the vast amounts of arithmetic involved in the application of the multivariate methods being proposed were very limited and, consequently, developments were primarily mathematical and multivariate research was, at the time, largely a branch of linear algebra. However, the arrival and rapid expansion of the use of electronic computers in the second half of the 20th century led to increased practical application of existing methods of multivariate analysis and renewed interest in the creation of new techniques.

In the early years of the 21st century, the wide availability of relatively cheap and extremely powerful personal computers and laptops allied with flexible statistical software has meant that all the methods of multivariate analysis can be applied routinely even to very large data sets such as those generated in, for example, genetics, imaging, and astronomy. And the application of multivariate techniques to such large data sets has now been given its own name, *data mining*, which has been defined as "the nontrivial extraction of implicit, previously unknown and potentially useful information from data." Useful books on data mining are those of Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy (1996) and Hand, Mannila, and Smyth (2001).

1.3 Types of variables and the possible problem of missing values

A hypothetical example of multivariate data is given in Table 1.1. The special symbol NA denotes missing values (being Not Available); the value of this variable for a subject is missing.

Table 1.1: `hypo` data. Hypothetical Set of Multivariate Data.

`individual`	`sex`	`age`	`IQ`	`depression`	`health`	`weight`
1	Male	21	120	Yes	Very good	150
2	Male	43	NA	No	Very good	160
3	Male	22	135	No	Average	135
4	Male	86	150	No	Very poor	140
5	Male	60	92	Yes	Good	110
6	Female	16	130	Yes	Good	110
7	Female	NA	150	Yes	Very good	120
8	Female	43	NA	Yes	Average	120
9	Female	22	84	No	Average	105
10	Female	80	70	No	Good	100

Here, the number of units (people in this case) is $n = 10$, with the number of variables being $q = 7$ and, for example, $x_{34} = 135$. In R, a "`data.frame`" is the appropriate data structure to represent such rectangular data. Subsets of units (rows) or variables (columns) can be extracted via the `[` subset operator; i.e.,

```
R> hypo[1:2, c("health", "weight")]

     health weight
1 Very good    150
2 Very good    160
```

extracts the values x_{15}, x_{16} and x_{25}, x_{26} from the hypothetical data presented in Table 1.1. These data illustrate that the variables that make up a set of multivariate data will not necessarily all be of the same type. Four levels of measurements are often distinguished:

Nominal: Unordered categorical variables. Examples include treatment allocation, the sex of the respondent, hair colour, presence or absence of depression, and so on.

Ordinal: Where there is an ordering but no implication of equal distance between the different points of the scale. Examples include social class, self-perception of health (each coded from I to V, say), and educational level (no schooling, primary, secondary, or tertiary education).

Interval: Where there are equal differences between successive points on the scale but the position of zero is arbitrary. The classic example is the measurement of temperature using the Celsius or Fahrenheit scales.

Ratio: The highest level of measurement, where one can investigate the *relative magnitudes* of scores as well as the differences between them. The position of zero is fixed. The classic example is the absolute measure of temperature (in Kelvin, for example), but other common ones includes age (or any other time from a fixed event), weight, and length.

In many statistical textbooks, discussion of different types of measurements is often followed by recommendations as to which statistical techniques are suitable for each type; for example, analyses on nominal data should be limited to summary statistics such as the number of cases, the mode, etc. And, for ordinal data, means and standard deviations are not suitable. But Velleman and Wilkinson (1993) make the important point that restricting the choice of statistical methods in this way may be a dangerous practise for data analysis–in essence the measurement taxonomy described is often too strict to apply to real-world data. This is not the place for a detailed discussion of measurement, but we take a fairly pragmatic approach to such problems. For example, we will not agonise over treating variables such as measures of depression, anxiety, or intelligence as if they are interval-scaled, although strictly they fit into the ordinal category described above.

1.3.1 Missing values

Table 1.1 also illustrates one of the problems often faced by statisticians undertaking statistical analysis in general and multivariate analysis in particular, namely the presence of *missing values* in the data; i.e., observations and measurements that should have been recorded but for one reason or another, were not. Missing values in multivariate data may arise for a number of reasons; for example, non-response in sample surveys, dropouts in *longitudinal data* (see Chapter 8), or refusal to answer particular questions in a questionnaire. The most important approach for dealing with missing data is to try to avoid them during the data-collection stage of a study. But despite all the efforts a researcher may make, he or she may still be faced with a data set that contains a number of missing values. So what can be done? One answer to this question is to take the *complete-case analysis* route because this is what most statistical software packages do automatically. Using complete-case analysis on multivariate data means omitting *any* case with a missing value on any of the variables. It is easy to see that if the number of variables is large, then even a sparse pattern of missing values can result in a substantial number of incomplete cases. One possibility to ease this problem is to simply drop any variables that have many missing values. But complete-case analysis is not recommended for two reasons:

- Omitting a possibly substantial number of individuals will cause a large amount of information to be discarded and lower the effective sample size of the data, making any analyses less effective than they would have been if all the original sample had been available.
- More worrisome is that dropping the cases with missing values on one or more variables can lead to serious biases in both estimation and inference unless the discarded cases are essentially a random subsample of the observed data (the term *missing completely at random* is often used; see Chapter 8 and Little and Rubin (1987) for more details).

So, at the very least, complete-case analysis leads to a loss, and perhaps a substantial loss, in power by discarding data, but worse, analyses based just on complete cases might lead to misleading conclusions and inferences.

A relatively simple alternative to complete-case analysis that is often used is *available-case analysis*. This is a straightforward attempt to exploit the incomplete information by using all the cases available to estimate quantities of interest. For example, if the researcher is interested in estimating the *correlation matrix* (see Subsection 1.5.2) of a set of multivariate data, then available-case analysis uses all the cases with variables X_i and X_j present to estimate the correlation between the two variables. This approach appears to make better use of the data than complete-case analysis, but unfortunately available-case analysis has its own problems. The sample of individuals used changes from correlation to correlation, creating potential difficulties when the missing data are not missing completely at random. There is no guarantee that the estimated correlation matrix is even positive-definite which can create problems for some of the methods, such as *factor analysis* (see Chapter 5) and *structural equation modelling* (see Chapter 7), that the researcher may wish to apply to the matrix.

Both complete-case and available-case analyses are unattractive unless the number of missing values in the data set is “small”. An alternative answer to the missing-data problem is to consider some form of *imputation*, the practise of “filling in” missing data with plausible values. Methods that impute the missing values have the advantage that, unlike in complete-case analysis, observed values in the incomplete cases are retained. On the surface, it looks like imputation will solve the missing-data problem and enable the investigator to progress normally. But, from a statistical viewpoint, careful consideration needs to be given to the *method* used for imputation or otherwise it may cause more problems than it solves; for example, imputing an observed variable mean for a variable’s missing values preserves the observed sample means but distorts the *covariance matrix* (see Subsection 1.5.1), biasing estimated variances and covariances towards zero. On the other hand, imputing predicted values from regression models tends to inflate observed correlations, biasing them away from zero (see Little 2005). And treating imputed data as if they were “real” in estimation and inference can lead to misleading standard errors and p-values since they fail to reflect the uncertainty due to the missing data.

The most appropriate way to deal with missing values is by a procedure suggested by Rubin (1987) known as *multiple imputation*. This is a Monte Carlo technique in which the missing values are replaced by $m > 1$ simulated versions, where m is typically small (say 3–10). Each of the simulated complete data sets is analysed using the method appropriate for the investigation at hand, and the results are later combined to produce, say, estimates and confidence intervals that incorporate missing-data uncertainty. Details are given in Rubin (1987) and more concisely in Schafer (1999). The great virtues of multiple imputation are its simplicity and its generality. The user may analyse the data using virtually any technique that would be appropriate if the data were complete. However, one should always bear in mind that the imputed values are not real measurements. We do not get something for nothing! And if there is a substantial proportion of individuals with large amounts of missing data, one should clearly question whether *any* form of statistical analysis is worth the bother.

1.4 Some multivariate data sets

This is a convenient point to look at some multivariate data sets and briefly ponder the type of question that might be of interest in each case. The first data set consists of chest, waist, and hip measurements on a sample of men and women and the measurements for 20 individuals are shown in Table 1.2. Two questions might be addressed by such data;

- Could body size and body shape be summarised in some way by combining the three measurements into a single number?
- Are there subtypes of body shapes amongst the men and amongst the women within which individuals are of similar shapes and between which body shapes differ?

The first question might be answered by *principal components analysis* (see Chapter 3), and the second question could be investigated using *cluster analysis* (see Chapter 6).

(In practise, it seems intuitively likely that we would have needed to record the three measurements on many more than 20 individuals to have any chance of being able to get convincing answers from these techniques to the questions of interest. The question of how many units are needed to achieve a sensible analysis when using the various techniques of multivariate analysis will be taken up in the respective chapters describing each technique.)

Table 1.2: measure data. Chest, waist, and hip measurements on 20 individuals (in inches).

chest	waist	hips	gender	chest	waist	hips	gender
34	30	32	male	36	24	35	female
37	32	37	male	36	25	37	female
38	30	36	male	34	24	37	female
36	33	39	male	33	22	34	female
38	29	33	male	36	26	38	female
43	32	38	male	37	26	37	female
40	33	42	male	34	25	38	female
38	30	40	male	36	26	37	female
40	30	37	male	38	28	40	female
41	32	39	male	35	23	35	female

Our second set of multivariate data consists of the results of chemical analysis on Romano-British pottery made in three different regions (region 1 contains kiln 1, region 2 contains kilns 2 and 3, and region 3 contains kilns 4 and 5). The complete data set, which we shall meet in Chapter 6, consists of the chemical analysis results on 45 pots, shown in Table 1.3. One question that might be posed about these data is whether the chemical profiles of each pot suggest different types of pots and if any such types are related to kiln or region. This question is addressed in Chapter 6.

Table 1.3: pottery data. Romano-British pottery data.

Al2O3	Fe2O3	MgO	CaO	Na2O	K2O	TiO2	MnO	BaO	kiln
18.8	9.52	2.00	0.79	0.40	3.20	1.01	0.077	0.015	1
16.9	7.33	1.65	0.84	0.40	3.05	0.99	0.067	0.018	1
18.2	7.64	1.82	0.77	0.40	3.07	0.98	0.087	0.014	1
16.9	7.29	1.56	0.76	0.40	3.05	1.00	0.063	0.019	1
17.8	7.24	1.83	0.92	0.43	3.12	0.93	0.061	0.019	1
18.8	7.45	2.06	0.87	0.25	3.26	0.98	0.072	0.017	1
16.5	7.05	1.81	1.73	0.33	3.20	0.95	0.066	0.019	1
18.0	7.42	2.06	1.00	0.28	3.37	0.96	0.072	0.017	1
15.8	7.15	1.62	0.71	0.38	3.25	0.93	0.062	0.017	1
14.6	6.87	1.67	0.76	0.33	3.06	0.91	0.055	0.012	1
13.7	5.83	1.50	0.66	0.13	2.25	0.75	0.034	0.012	1
14.6	6.76	1.63	1.48	0.20	3.02	0.87	0.055	0.016	1
14.8	7.07	1.62	1.44	0.24	3.03	0.86	0.080	0.016	1
17.1	7.79	1.99	0.83	0.46	3.13	0.93	0.090	0.020	1
16.8	7.86	1.86	0.84	0.46	2.93	0.94	0.094	0.020	1
15.8	7.65	1.94	0.81	0.83	3.33	0.96	0.112	0.019	1

Table 1.3: pottery data (continued).

Al2O3	Fe2O3	MgO	CaO	Na2O	K2O	TiO2	MnO	BaO	kiln
18.6	7.85	2.33	0.87	0.38	3.17	0.98	0.081	0.018	1
16.9	7.87	1.83	1.31	0.53	3.09	0.95	0.092	0.023	1
18.9	7.58	2.05	0.83	0.13	3.29	0.98	0.072	0.015	1
18.0	7.50	1.94	0.69	0.12	3.14	0.93	0.035	0.017	1
17.8	7.28	1.92	0.81	0.18	3.15	0.90	0.067	0.017	1
14.4	7.00	4.30	0.15	0.51	4.25	0.79	0.160	0.019	2
13.8	7.08	3.43	0.12	0.17	4.14	0.77	0.144	0.020	2
14.6	7.09	3.88	0.13	0.20	4.36	0.81	0.124	0.019	2
11.5	6.37	5.64	0.16	0.14	3.89	0.69	0.087	0.009	2
13.8	7.06	5.34	0.20	0.20	4.31	0.71	0.101	0.021	2
10.9	6.26	3.47	0.17	0.22	3.40	0.66	0.109	0.010	2
10.1	4.26	4.26	0.20	0.18	3.32	0.59	0.149	0.017	2
11.6	5.78	5.91	0.18	0.16	3.70	0.65	0.082	0.015	2
11.1	5.49	4.52	0.29	0.30	4.03	0.63	0.080	0.016	2
13.4	6.92	7.23	0.28	0.20	4.54	0.69	0.163	0.017	2
12.4	6.13	5.69	0.22	0.54	4.65	0.70	0.159	0.015	2
13.1	6.64	5.51	0.31	0.24	4.89	0.72	0.094	0.017	2
11.6	5.39	3.77	0.29	0.06	4.51	0.56	0.110	0.015	3
11.8	5.44	3.94	0.30	0.04	4.64	0.59	0.085	0.013	3
18.3	1.28	0.67	0.03	0.03	1.96	0.65	0.001	0.014	4
15.8	2.39	0.63	0.01	0.04	1.94	1.29	0.001	0.014	4
18.0	1.50	0.67	0.01	0.06	2.11	0.92	0.001	0.016	4
18.0	1.88	0.68	0.01	0.04	2.00	1.11	0.006	0.022	4
20.8	1.51	0.72	0.07	0.10	2.37	1.26	0.002	0.016	4
17.7	1.12	0.56	0.06	0.06	2.06	0.79	0.001	0.013	5
18.3	1.14	0.67	0.06	0.05	2.11	0.89	0.006	0.019	5
16.7	0.92	0.53	0.01	0.05	1.76	0.91	0.004	0.013	5
14.8	2.74	0.67	0.03	0.05	2.15	1.34	0.003	0.015	5
19.1	1.64	0.60	0.10	0.03	1.75	1.04	0.007	0.018	5

Source: Tubb, A., et al., *Archaeometry*, 22, 153–171, 1980. With permission.

Our third set of multivariate data involves the examination scores of a large number of college students in six subjects; the scores for five subjects are shown in Table 1.4. Here the main question of interest might be whether the exam scores reflect some underlying trait in a student that cannot be measured directly, perhaps "general intelligence"? The question could be investigated by using *exploratory factor analysis* (see Chapter 5).

Table 1.4: `exam` data. Exam scores for five psychology students.

subject	maths	english	history	geography	chemistry	physics
1	60	70	75	58	53	42
2	80	65	66	75	70	76
3	53	60	50	48	45	43
4	85	79	71	77	68	79
5	45	80	80	84	44	46

The final set of data we shall consider in this section was collected in a study of air pollution in cities in the USA. The following variables were obtained for 41 US cities:

`SO2`: SO_2 content of air in micrograms per cubic metre;
`temp`: average annual temperature in degrees Fahrenheit;
`manu`: number of manufacturing enterprises employing 20 or more workers;
`popul`: population size (1970 census) in thousands;
`wind`: average annual wind speed in miles per hour;
`precip`: average annual precipitation in inches;
`predays`: average number of days with precipitation per year.

The data are shown in Table 1.5.

Table 1.5: `USairpollution` data. Air pollution in 41 US cities.

	SO2	temp	manu	popul	wind	precip	predays
Albany	46	47.6	44	116	8.8	33.36	135
Albuquerque	11	56.8	46	244	8.9	7.77	58
Atlanta	24	61.5	368	497	9.1	48.34	115
Baltimore	47	55.0	625	905	9.6	41.31	111
Buffalo	11	47.1	391	463	12.4	36.11	166
Charleston	31	55.2	35	71	6.5	40.75	148
Chicago	110	50.6	3344	3369	10.4	34.44	122
Cincinnati	23	54.0	462	453	7.1	39.04	132
Cleveland	65	49.7	1007	751	10.9	34.99	155
Columbus	26	51.5	266	540	8.6	37.01	134
Dallas	9	66.2	641	844	10.9	35.94	78
Denver	17	51.9	454	515	9.0	12.95	86
Des Moines	17	49.0	104	201	11.2	30.85	103
Detroit	35	49.9	1064	1513	10.1	30.96	129
Hartford	56	49.1	412	158	9.0	43.37	127
Houston	10	68.9	721	1233	10.8	48.19	103
Indianapolis	28	52.3	361	746	9.7	38.74	121
Jacksonville	14	68.4	136	529	8.8	54.47	116

Table 1.5: `USairpollution` data (continued).

	`SO2`	`temp`	`manu`	`popul`	`wind`	`precip`	`predays`
Kansas City	14	54.5	381	507	10.0	37.00	99
Little Rock	13	61.0	91	132	8.2	48.52	100
Louisville	30	55.6	291	593	8.3	43.11	123
Memphis	10	61.6	337	624	9.2	49.10	105
Miami	10	75.5	207	335	9.0	59.80	128
Milwaukee	16	45.7	569	717	11.8	29.07	123
Minneapolis	29	43.5	699	744	10.6	25.94	137
Nashville	18	59.4	275	448	7.9	46.00	119
New Orleans	9	68.3	204	361	8.4	56.77	113
Norfolk	31	59.3	96	308	10.6	44.68	116
Omaha	14	51.5	181	347	10.9	30.18	98
Philadelphia	69	54.6	1692	1950	9.6	39.93	115
Phoenix	10	70.3	213	582	6.0	7.05	36
Pittsburgh	61	50.4	347	520	9.4	36.22	147
Providence	94	50.0	343	179	10.6	42.75	125
Richmond	26	57.8	197	299	7.6	42.59	115
Salt Lake City	28	51.0	137	176	8.7	15.17	89
San Francisco	12	56.7	453	716	8.7	20.66	67
Seattle	29	51.1	379	531	9.4	38.79	164
St. Louis	56	55.9	775	622	9.5	35.89	105
Washington	29	57.3	434	757	9.3	38.89	111
Wichita	8	56.6	125	277	12.7	30.58	82
Wilmington	36	54.0	80	80	9.0	40.25	114

Source: Sokal, R. R., Rohlf, F. J., *Biometry*, W. H. Freeman, San Francisco, 1981. With permission.

What might be the question of most interest about these data? Very probably it is “how is pollution level as measured by sulphur dioxide concentration related to the six other variables?” In the first instance at least, this question suggests the application of multiple linear regression, with sulphur dioxide concentration as the response variable and the remaining six variables being the independent or explanatory variables (the latter is a more acceptable label because the “independent” variables are rarely independent of one another). But in the model underlying multiple regression, only the response is considered to be a random variable; the explanatory variables are strictly assumed to be fixed, not random, variables. In practise, of course, this is rarely the case, and so the results from a multiple regression analysis need to be interpreted as being conditional on the observed values of the explanatory variables. So when answering the question of most interest about these data, they should not really be considered multivariate–there is only a single random variable involved–a more suitable label is *multivariable* (we know this sounds pedantic, but we are statisticians after all). In this book, we shall say only a little about the multiple linear model for multivariable data in Chapter 8. but essentially only to enable such regression models to be introduced for situations where there is a multivariate response; for example, in the case of *repeated-measures data* and *longitudinal data*.

The four data sets above have not exhausted either the questions that multivariate data may have been collected to answer or the methods of multivariate analysis that have been developed to answer them, as we shall see as we progress through the book.

1.5 Covariances, correlations, and distances

The main reason why we should analyse a multivariate data set using multivariate methods rather than looking at each variable separately using one or another familiar univariate method is that any structure or pattern in the data is as likely to be implied either by "relationships" between the variables or by the relative "closeness" of different units as by their different variable values; in some cases perhaps by both. In the first case, any structure or pattern uncovered will be such that it "links" together the columns of the data matrix, $\mathbf{X}$, in some way, and in the second case a possible structure that might be discovered is that involving interesting subsets of the units. The question now arises as to how we quantify the relationships between the variables and how we measure the distances between different units. This question is answered in the subsections that follow.

1.5.1 Covariances

The *covariance* of two random variables is a measure of their *linear* dependence. The population (theoretical) covariance of two random variables, X_i and X_j, is defined by

$$\mathsf{Cov}(X_i, X_j) = \mathsf{E}(X_i - \mu_i)(X_j - \mu_j),$$

where $\mu_i = \mathsf{E}(X_i)$ and $\mu_j = \mathsf{E}(X_j)$; E denotes expectation.

If $i = j$, we note that the covariance of the variable with itself is simply its variance, and therefore there is no need to define variances and covariances independently in the multivariate case. If X_i and X_j are independent of each other, their covariance is necessarily equal to zero, but the converse is not true. The covariance of X_i and X_j is usually denoted by σ_{ij}. The variance of variable X_i is $\sigma_i^2 = \mathsf{E}\left((X_i - \mu_i)^2\right)$. Larger values of the covariance imply a greater degree of linear dependence between two variables.

In a multivariate data set with q observed variables, there are q variances and $q(q-1)/2$ covariances. These quantities can be conveniently arranged in a $q \times q$ symmetric matrix, $\mathbf{\Sigma}$, where

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1q} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{q1} & \sigma_{q2} & \dots & \sigma_q^2 \end{pmatrix}.$$

Note that $\sigma_{ij} = \sigma_{ji}$. This matrix is generally known as the *variance-covariance matrix* or simply the *covariance matrix* of the data.

For a set of multivariate observations, perhaps sampled from some population, the matrix $\boldsymbol{\Sigma}$ is estimated by

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top,$$

where $\mathbf{x}_i^\top = (x_{i1}, x_{i2}, \dots, x_{iq})$ is the vector of (numeric) observations for the ith individual and $\bar{\mathbf{x}} = n^{-1} \sum_{i=1}^{n} \mathbf{x}_i$ is the mean vector of the observations. The diagonal of $\mathbf{S}$ contains the sample variances of each variable, which we shall denote as s_i^2.

The covariance matrix for the data in Table 1.2 can be obtained using the `var()` function in R; however, we have to “remove” the categorical variable `gender` from the `measure` data frame by subsetting on the numerical variables first:

```
R> cov(measure[, c("chest", "waist", "hips")])
```

```
      chest  waist  hips
chest 6.632  6.368 3.000
waist 6.368 12.526 3.579
hips  3.000  3.579 5.945
```

If we require the separate covariance matrices of men and women, we can use

```
R> cov(subset(measure, gender == "female")[,
+              c("chest", "waist", "hips")])
```

```
      chest waist  hips
chest 2.278 2.167 1.556
waist 2.167 2.989 2.756
hips  1.556 2.756 3.067
```

```
R> cov(subset(measure, gender == "male")[,
+              c("chest", "waist", "hips")])
```

```
       chest  waist  hips
chest 6.7222 0.9444 3.944
waist 0.9444 2.1000 3.078
hips  3.9444 3.0778 9.344
```

where the `subset()` returns all observations corresponding to females (first statement) or males (second statement).

1.5.2 Correlations

The covariance is often difficult to interpret because it depends on the scales on which the two variables are measured; consequently, it is often standardised by dividing by the product of the standard deviations of the two variables to give a quantity called the *correlation coefficient*, ρ_{ij}, where

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j},$$

where $\sigma_i = \sqrt{\sigma_i^2}$.

The advantage of the correlation is that it is independent of the scales of the two variables. The correlation coefficient lies between -1 and $+1$ and gives a measure of the *linear* relationship of the variables X_i and X_j. It is positive if high values of X_i are associated with high values of X_j and negative if high values of X_i are associated with low values of X_j. If the relationship between two variables is non-linear, their correlation coefficient can be misleading.

With q variables there are $q(q-1)/2$ distinct correlations, which may be arranged in a $q \times q$ correlation matrix the diagonal elements of which are unity. For observed data, the correlation matrix contains the usual estimates of the ρs, namely Pearson's correlation coefficient, and is generally denoted by $\mathbf{R}$. The matrix may be written in terms of the sample covariance matrix $\mathbf{S}$

$$\mathbf{R} = \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2},$$

where $\mathbf{D}^{-1/2} = \text{diag}(1/s_1, \dots, 1/s_q)$ and $s_i = \sqrt{s_i^2}$ is the sample standard deviation of variable i. (In most situations considered in this book, we will be dealing with covariance and correlation matrices of full rank, q, so that both matrices will be *non-singular*, that is, invertible, to give matrices $\mathbf{S}^{-1}$ or $\mathbf{R}^{-1}$.)

The sample correlation matrix for the three variables in Table 1.1 is obtained by using the function `cor()` in R:

```
R> cor(measure[, c("chest", "waist", "hips")])

       chest  waist   hips
chest 1.0000 0.6987 0.4778
waist 0.6987 1.0000 0.4147
hips  0.4778 0.4147 1.0000
```

1.5.3 Distances

For some multivariate techniques such as *multidimensional scaling* (see Chapter 4) and *cluster analysis* (see Chapter 6), the concept of *distance* between the units in the data is often of considerable interest and importance. So, given the variable values for two units, say unit i and unit j, what serves as a measure of distance between them? The most common measure used is *Euclidean distance*, which is defined as

$$d_{ij} = \sqrt{\sum_{k=1}^{q} (x_{ik} - x_{jk})^2},$$

where x_{ik} and $x_{jk}, k = 1, \dots, q$ are the variable values for units i and j, respectively. Euclidean distance can be calculated using the `dist()` function in R.

When the variables in a multivariate data set are on different scales, it makes more sense to calculate the distances *after* some form of standardisation. Here we shall illustrate this on the body measurement data and divide each variable by its standard deviation using the function `scale()` before applying the `dist()` function–the necessary R code and output are

```
R> dist(scale(measure[, c("chest", "waist", "hips")],
+        center = FALSE))
```

```
      1    2    3    4    5    6    7    8    9   10   11
2  0.17
3  0.15 0.08
4  0.22 0.07 0.14
5  0.11 0.15 0.09 0.22
6  0.29 0.16 0.16 0.19 0.21
7  0.32 0.16 0.20 0.13 0.28 0.14
8  0.23 0.11 0.11 0.12 0.19 0.16 0.13
9  0.21 0.10 0.06 0.16 0.12 0.11 0.17 0.09
10 0.27 0.12 0.13 0.14 0.20 0.06 0.09 0.11 0.09
11 0.23 0.28 0.22 0.33 0.19 0.34 0.38 0.25 0.24 0.32
12 0.22 0.24 0.18 0.28 0.18 0.30 0.32 0.20 0.20 0.28 0.06

...
```

(Note that only the distances for the first 12 observations are shown in the output.)

1.6 The multivariate normal density function

Just as the normal distribution dominates univariate techniques, the *multivariate normal distribution* plays an important role in *some* multivariate procedures, although as mentioned earlier many multivariate analyses are carried out in the spirit of data exploration where questions of statistical significance are of relatively minor importance or of no importance at all. Nevertheless, researchers dealing with the complexities of multivariate data may, on occasion, need to know a little about the multivariate density function and in particular how to assess whether or not a set of multivariate data can be assumed to have this density function. So we will define the multivariate normal density and describe some of its properties.

For a vector of q variables, $\mathbf{x}^\top = (x_1, x_2, \dots, x_q)$, the multivariate normal density function takes the form

$$f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-q/2} \det(\boldsymbol{\Sigma})^{-1/2} \exp\left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\},$$

where $\boldsymbol{\Sigma}$ is the population covariance matrix of the variables and $\boldsymbol{\mu}$ is the vector of population mean values of the variables. The simplest example of the *multivariate normal density function* is the bivariate normal density with $q = 2$; this can be written explicitly as

$$\begin{aligned} & f((x_1, x_2); (\mu_1, \mu_2), \sigma_1, \sigma_2, \rho) = \\ & \quad \left(2\pi\sigma_1\sigma_2(1-\rho^2)\right)^{-1/2} \exp\left\{ -\frac{1}{2(1-\rho^2)} \times \right. \\ & \quad \left. \left(\left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 - 2\rho \frac{x_1 - \mu_1}{\sigma_1} \frac{x_2 - \mu_2}{\sigma_2} + \left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2 \right) \right\}, \end{aligned}$$

where μ_1 and μ_2 are the population means of the two variables, σ_1^2 and σ_2^2 are the population variances, and ρ is the population correlation between the two variables X_1 and X_2. Figure 1.1 shows an example of a bivariate normal density function with both means equal to zero, both variances equal to one, and correlation equal to 0.5.

The population mean vector and the population covariance matrix of a multivariate density function are estimated from a sample of multivariate observations as described in the previous subsections.

One property of a multivariate normal density function that is worth mentioning here is that *linear combinations* of the variables (i.e., $y = a_1X_1 + a_2X_2 + \cdots + a_qX_q$, where $a_1, a_2, \dots, a_q$ is a set of scalars) are themselves normally distributed with mean $\mathbf{a}^\top\boldsymbol{\mu}$ and variance $\mathbf{a}^\top\boldsymbol{\Sigma}\mathbf{a}$, where $\mathbf{a}^\top = (a_1, a_2, \dots, a_q)$. Linear combinations of variables will be of importance in later chapters, particularly in Chapter 3.

For many multivariate methods to be described in later chapters, the assumption of multivariate normality is not critical to the results of the analysis, but there may be occasions when testing for multivariate normality may be of interest. A start can be made perhaps by assessing each variable separately for univariate normality using a *probability plot*. Such plots are commonly applied in univariate analysis and involve ordering the observations and then plotting them against the appropriate values of an assumed cumulative distribution function. There are two basic types of plots for comparing two probability distributions, the *probability-probability plot* and the *quantile-quantile plot*. The diagram in Figure 1.2 may be used for describing each type.

A plot of points whose coordinates are the cumulative probabilities $p_1(q)$ and $p_2(q)$ for different values of q with

$$\begin{aligned} p_1(q) &= \mathsf{P}(X_1 \leq q), \\ p_2(q) &= \mathsf{P}(X_2 \leq q), \end{aligned}$$

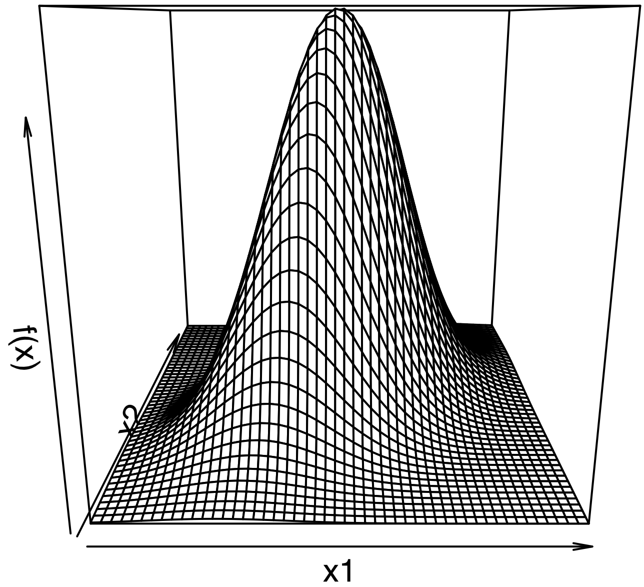

Fig. 1.1. Bivariate normal density function with correlation $\rho = 0.5$.

for random variables X_1 and X_2 is a probability-probability plot, while a plot of the points whose coordinates are the quantiles $(q_1(p), q_2(p))$ for different values of p with

$$\begin{aligned} q_1(p) &= p_1^{-1}(p), \\ q_2(p) &= p_2^{-1}(p), \end{aligned}$$

is a quantile-quantile plot. For example, a quantile-quantile plot for investigating the assumption that a set of data is from a normal distribution would involve plotting the ordered sample values of variable 1 (i.e.,$x_{(1)1}, x_{(2)1}, \dots, x_{(n)1}$) against the quantiles of a standard normal distribution, $\Phi^{-1}(p(i))$, where usually

$$p_i = \frac{i - \frac{1}{2}}{n} \quad \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du.$$

This is known as a *normal probability plot*.

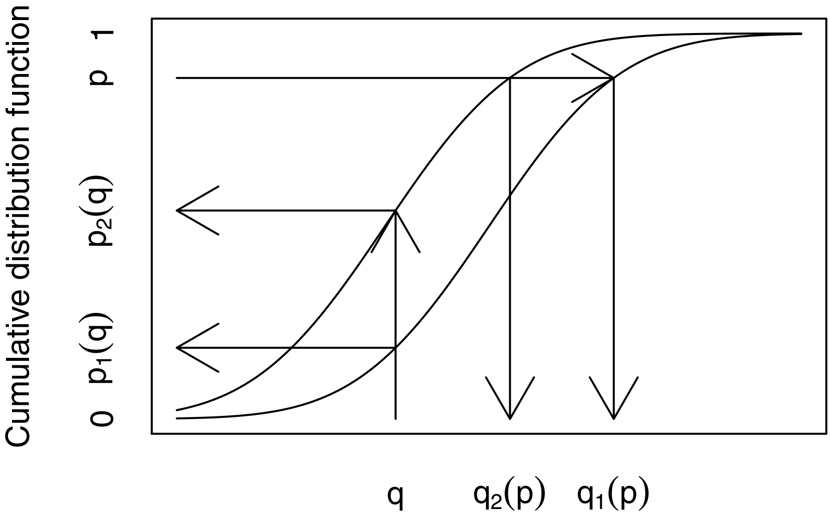

Fig. 1.2. Cumulative distribution functions and quantiles.

For multivariate data, normal probability plots may be used to examine each variable separately, although marginal normality does not necessarily imply that the variables follow a multivariate normal distribution. Alternatively (or additionally), each multivariate observation might be converted to a single number in some way before plotting. For example, in the specific case of assessing a data set for multivariate normality, each q-dimensional observation, $\mathbf{x}_i$, could be converted into a *generalised distance*, d_i^2, giving a measure of the distance of the particular observation from the mean vector of the complete sample, $\bar{\mathbf{x}}$; d_i^2 is calculated as

$$d_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})^\top \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}),$$

where $\mathbf{S}$ is the sample covariance matrix. This distance measure takes into account the different variances of the variables and the covariances of pairs of variables. If the observations do arise from a multivariate normal distribution, then these distances have approximately a *chi-squared distribution* with q degrees of freedom, also denoted by the symbol χ_q^2. So plotting the ordered distances against the corresponding quantiles of the appropriate chi-square distribution should lead to a straight line through the origin.

We will now assess the body measurements data in Table 1.2 for normality, although because there are only 20 observations in the sample there is really too little information to come to any convincing conclusion. Figure 1.3 shows separate probability plots for each measurement; there appears to be no evidence of any departures from linearity. The chi-square plot of the 20 generalised distances in Figure 1.4 does seem to deviate a little from linearity, but with so few observations it is hard to be certain. The plot is set up as follows. We first extract the relevant data

```
R> x <- measure[, c("chest", "waist", "hips")]
```

and estimate the means of all three variables (i.e., for each column of the data) and the covariance matrix

```
R> cm <- colMeans(x)
R> S <- cov(x)
```

The differences d_i have to be computed for all units in our data, so we iterate over the rows of `x` using the `apply()` function with argument `MARGIN = 1` and, for each row, compute the distance d_i:

```
R> d <- apply(x, MARGIN = 1, function(x)
+              t(x - cm) %*% solve(S) %*% (x - cm))
```

The sorted distances can now be plotted against the appropriate quantiles of the χ^2_3 distribution obtained from `qchisq()`; see Figure 1.4.

```
R> qqnorm(measure[,"chest"], main = "chest"); qqline(measure[,"chest"])
R> qqnorm(measure[,"waist"], main = "waist"); qqline(measure[,"waist"])
R> qqnorm(measure[,"hips"], main = "hips"); qqline(measure[,"hips"])
```

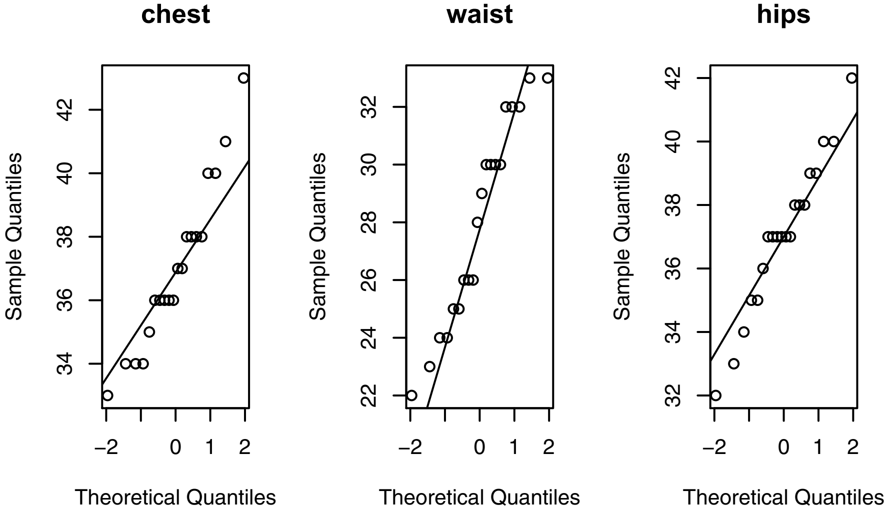

Fig. 1.3. Normal probability plots of chest, waist, and hip measurements.

```
R> plot(qchisq((1:nrow(x) - 1/2) / nrow(x), df = 3), sort(d),
+       xlab = expression(paste(chi[3]^2, " Quantile")),
+       ylab = "Ordered distances")
R> abline(a = 0, b = 1)
```

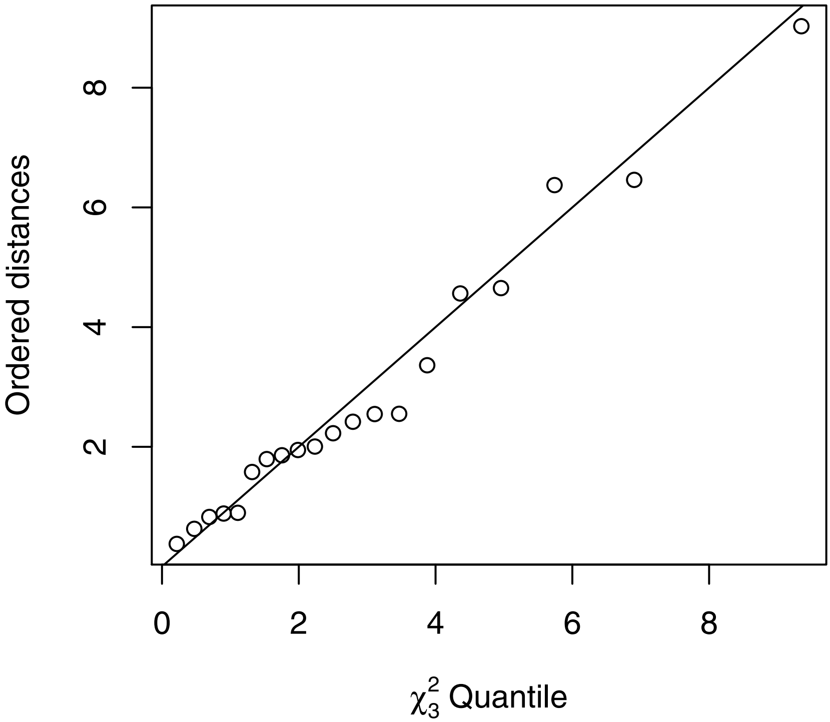

Fig. 1.4. Chi-square plot of generalised distances for body measurements data.

We will now look at using the chi-square plot on a set of data introduced early in the chapter, namely the air pollution in US cities (see Table 1.5). The probability plots for each separate variable are shown in Figure 1.5. Here, we also iterate over all variables, this time using a special function, `sapply()`, that loops over the variable names:

```
R> layout(matrix(1:8, nc = 2))
R> sapply(colnames(USairpollution), function(x) {
+      qqnorm(USairpollution[[x]], main = x)
+      qqline(USairpollution[[x]])
+  })
```

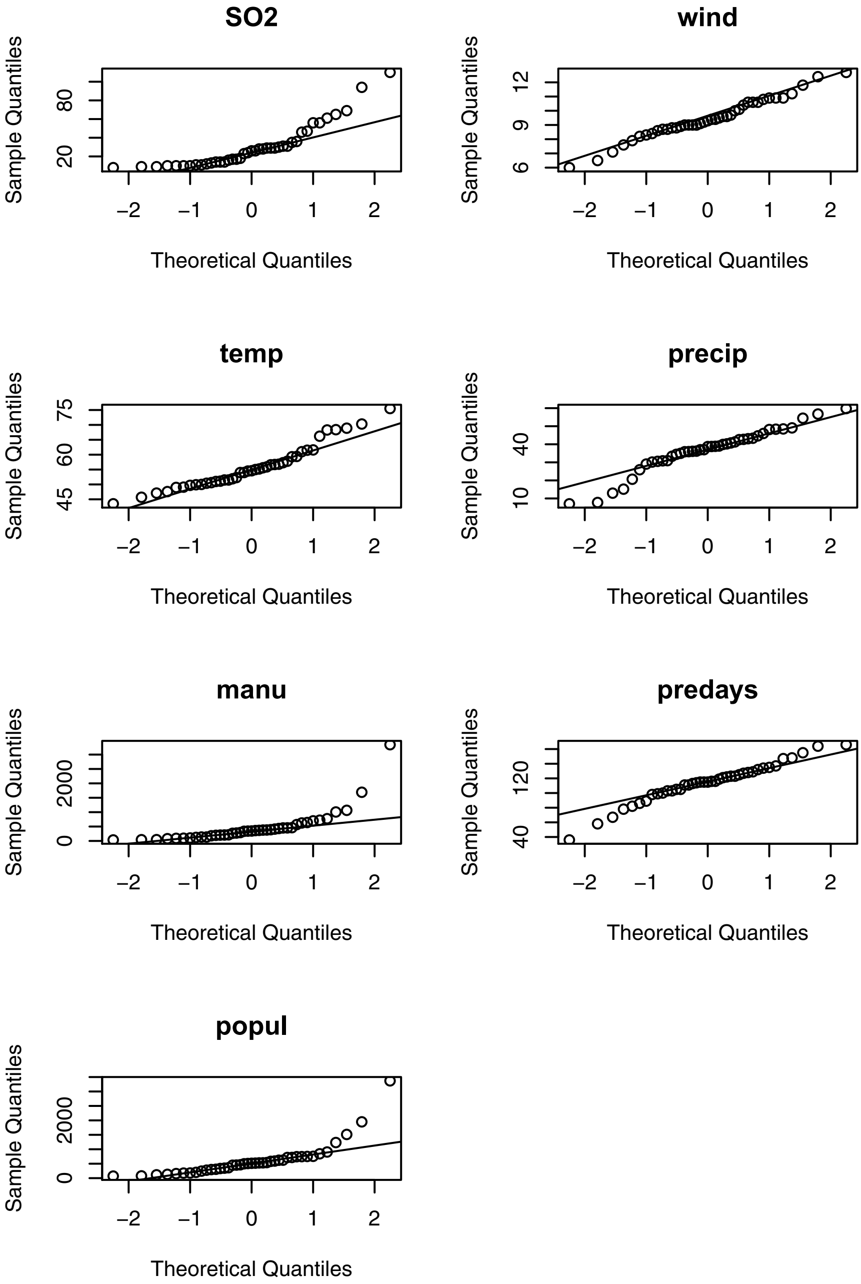

Fig. 1.5. Normal probability plots for `USairpollution` data.

The resulting seven plots are arranged on one page by a call to the `layout` matrix; see Figure 1.5. The plots for SO_2 concentration and precipitation both deviate considerably from linearity, and the plots for manufacturing and population show evidence of a number of outliers. But of more importance is the chi-square plot for the data, which is given in Figure 1.6; the R code is identical to the code used to produce the chi-square plot for the body measurement data. In addition, the two most extreme points in the plot have been labelled with the city names to which they correspond using `text()`.

```
R> x <- USairpollution
R> cm <- colMeans(x)
R> S <- cov(x)
R> d <- apply(x, 1, function(x) t(x - cm) %*% solve(S) %*% (x - cm))
R> plot(qc <- qchisq((1:nrow(x) - 1/2) / nrow(x), df = 6),
+      sd <- sort(d),
+      xlab = expression(paste(chi[6]^2, " Quantile")),
+      ylab = "Ordered distances", xlim = range(qc) * c(1, 1.1))
R> oups <- which(rank(abs(qc - sd), ties = "random") > nrow(x) - 3)
R> text(qc[oups], sd[oups] - 1.5, names(oups))
R> abline(a = 0, b = 1)
```

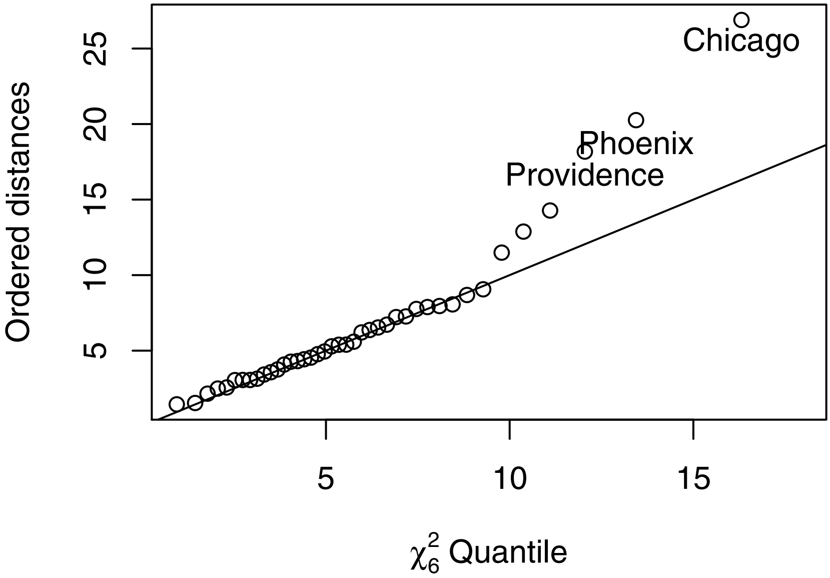

Fig. 1.6. χ^2 plot of generalised distances for `USairpollution` data.

This example illustrates that the chi-square plot might also be useful for detecting possible outliers in multivariate data, where informally outliers are "abnormal" in the sense of deviating from the natural data variability. Outlier identification is important in many applications of multivariate analysis either because there is some specific interest in finding anomalous observations or as a pre-processing task before the application of some multivariate method in order to preserve the results from possible misleading effects produced by these observations. A number of methods for identifying multivariate outliers have been suggested–see, for example, Rocke and Woodruff (1996) and Becker and Gather (2001)–and in Chapter 2 we will see how a number of the graphical methods described there can also be helpful for outlier detection.

1.7 Summary

The majority of data collected in all scientific disciplines are multivariate. To fully understand most such data sets, the variables need to be analysed simultaneously. The rest of this text is concerned with methods that have been developed to make this possible, some with the aim of discovering any patterns or structure in the data that may have important implications for future studies and some with the aim of drawing inferences about the data assuming they are sampled from a population with some particular probability density function, usually the multivariate normal.

1.8 Exercises

Ex. 1.1 Find the correlation matrix and covariance matrix of the data in Table 1.1.

Ex. 1.2 Fill in the missing values in Table 1.1 with appropriate mean values, and recalculate the correlation matrix of the data.

Ex. 1.3 Examine both the normal probability plots of each variable in the archaeology data in Table 1.3 and the chi-square plot of the data. Do the plots suggest anything unusual about the data?

Ex. 1.4 Convert the covariance matrix given below into the corresponding correlation matrix.

$$\begin{pmatrix} 3.8778 & 2.8110 & 3.1480 & 3.5062 \\ 2.8110 & 2.1210 & 2.2669 & 2.5690 \\ 3.1480 & 2.2669 & 2.6550 & 2.8341 \\ 3.5062 & 2.5690 & 2.8341 & 3.2352 \end{pmatrix}.$$

Ex. 1.5 For the small set of (10×5) multivariate data given below, find the (10×10) Euclidean distance matrix for the rows of the matrix. An alternative to Euclidean distance that might be used in some cases is what is known as *city block distance* (think New York). Write some R code to calculate the city block distance matrix for the data.

$$\begin{pmatrix} 3 & 6 & 4 & 0 & 7 \\ 4 & 2 & 7 & 4 & 6 \\ 4 & 0 & 3 & 1 & 5 \\ 6 & 2 & 6 & 1 & 1 \\ 1 & 6 & 2 & 1 & 4 \\ 5 & 1 & 2 & 0 & 2 \\ 1 & 1 & 2 & 6 & 1 \\ 1 & 1 & 5 & 4 & 4 \\ 7 & 0 & 1 & 3 & 3 \\ 3 & 3 & 0 & 5 & 1 \end{pmatrix}.$$