Multivariate Data Analysis

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?

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

The first question might be answered by **principal components analysis**
The second question could be investigated using **cluster analysis.**

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 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

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

: 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 reect some underlying trait in a student that cannot be measured directly, perhaps \general intelligence"?

The question could be investigated by

using exploratory factor analysis

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 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, 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.

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

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

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 $\mathbf{\Sigma}$ 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).

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 $\mathbf{D}^{-1/2} = \text{diag}(1/s_1, \ldots, 1/s_q)$ 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
```

Distances

For some multivariate techniques such as multidimensional scaling and cluster analysis 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

The multivariate normal density function

For a vector of q variables, $\mathbf{x}^\top = (x_1, x_2, \ldots, x_q)$, 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

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

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 $a_1, a_2, \ldots, a_q$ is a set of scalars) are themselves normally distributed with mean $\mathbf{a}^\top \boldsymbol{\mu}$ and variance $\mathbf{a}^\top \Sigma \mathbf{a}$, where $\mathbf{a}^\top = (a_1, a_2, \ldots, a_q)$. Linear combinations of variables will be of importance

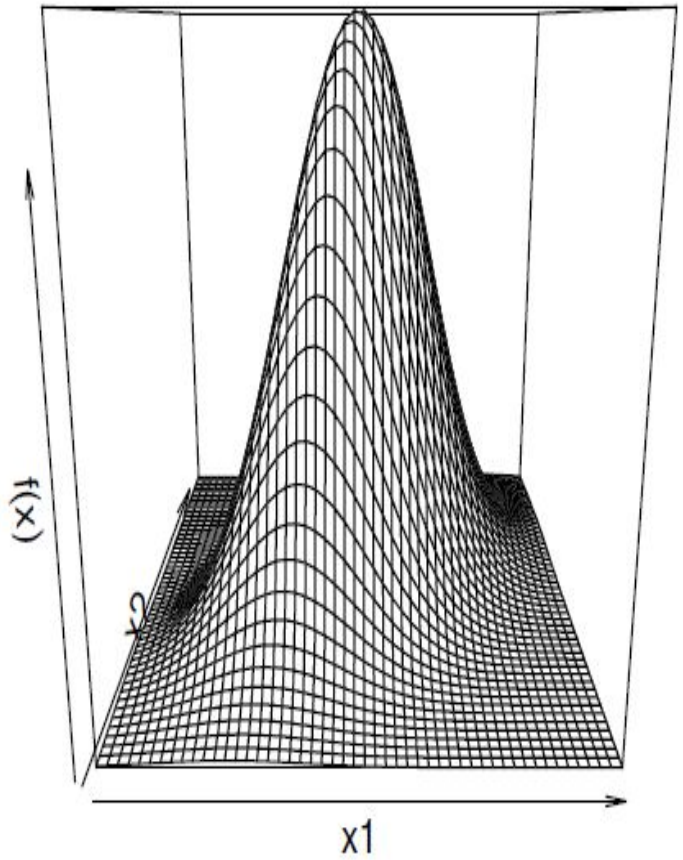


Fig. 1.1. Bivariate normal density function with correlation $\rho = 0.5$.

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.

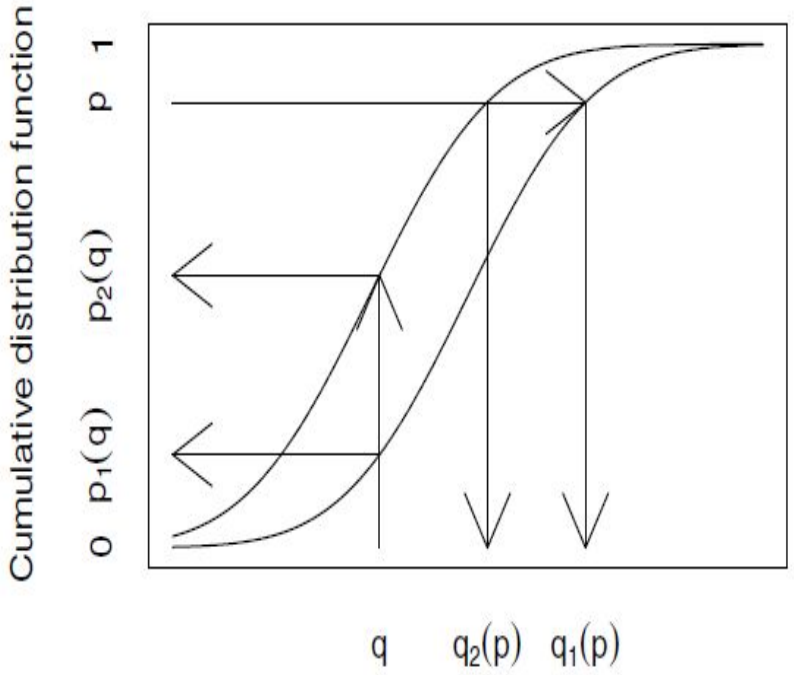


Fig. 1.2. Cumulative distribution functions and quantiles.

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

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

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

$$q_1(p) = p_1^{-1}(p),$$
$$q_2(p) = p_2^{-1}(p),$$

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.,$x_{(1)1}, x_{(2)1}, \ldots, x_{(n)1}$) 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*.

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"])
```

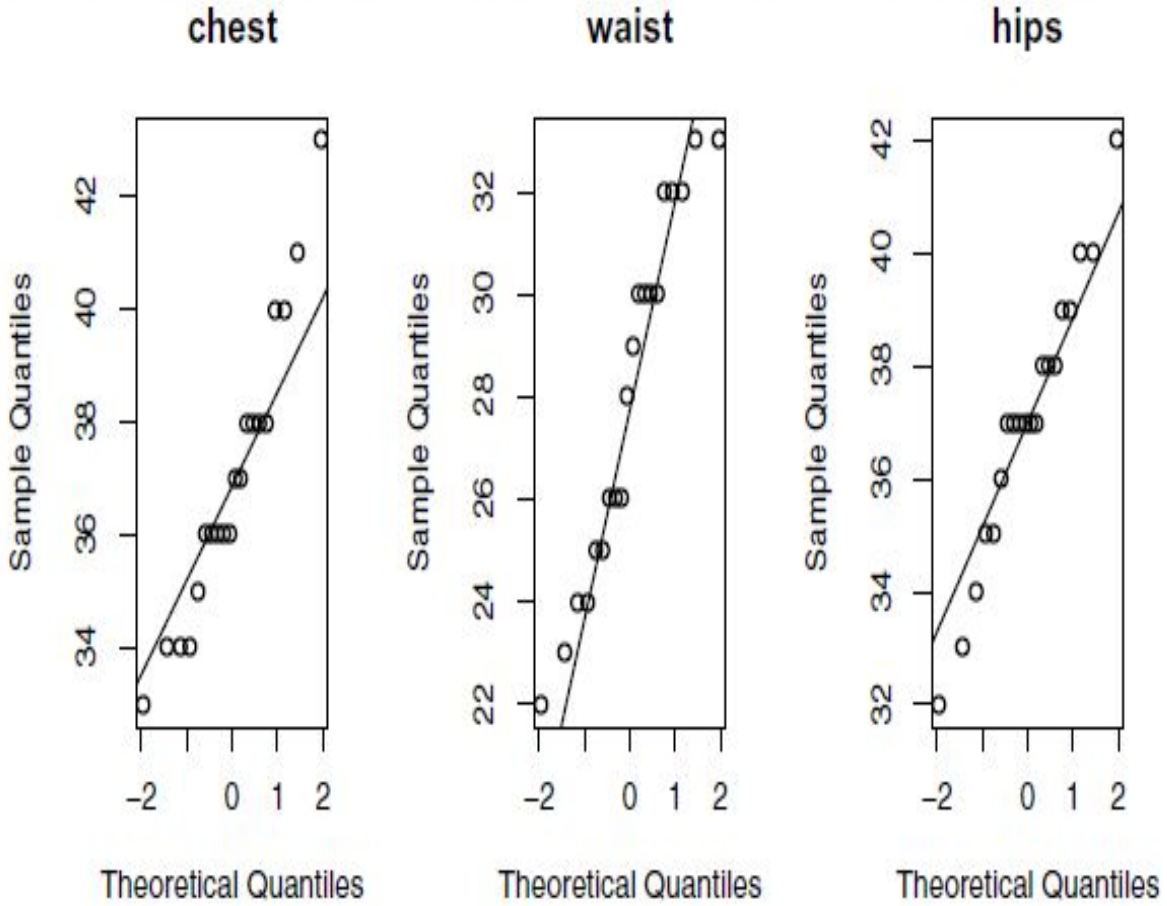


Fig. 1.3. Normal probability plots of chest, waist, and hip measurements.

```
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)
```

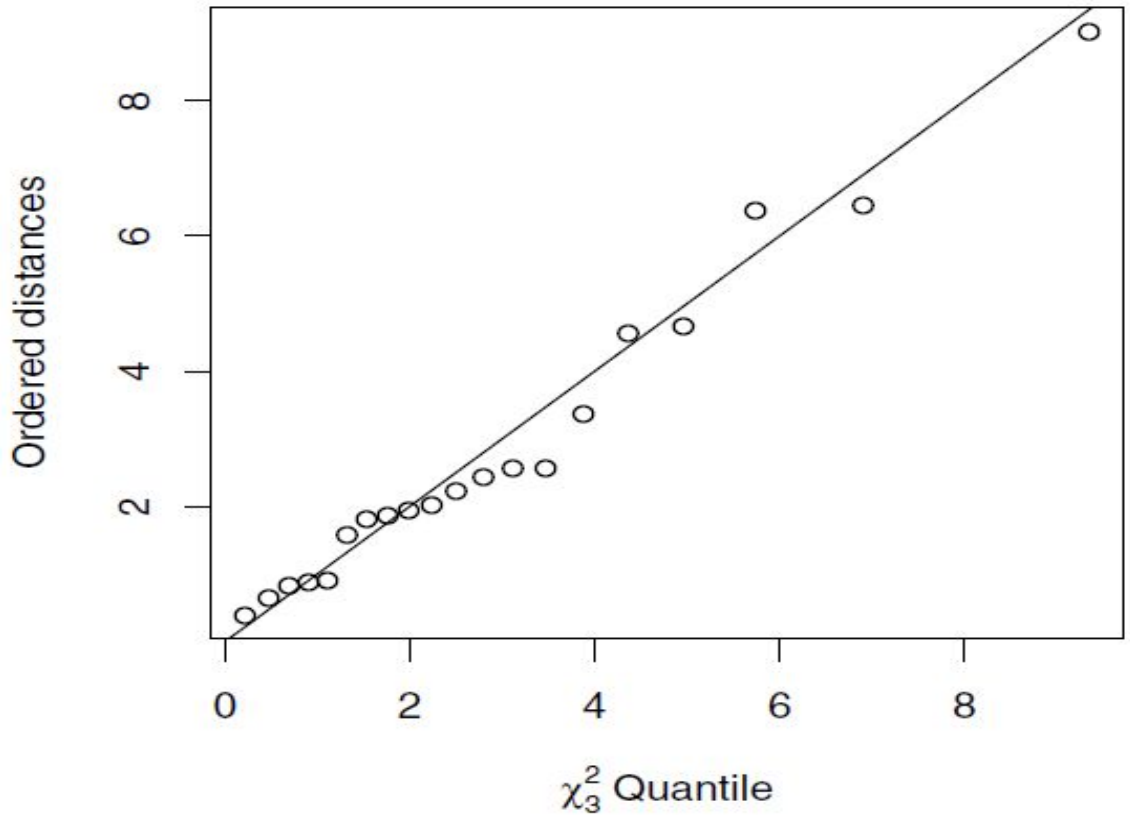


Fig. 1.4. Chi-square plot of generalised distances for body measurements data.

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.

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

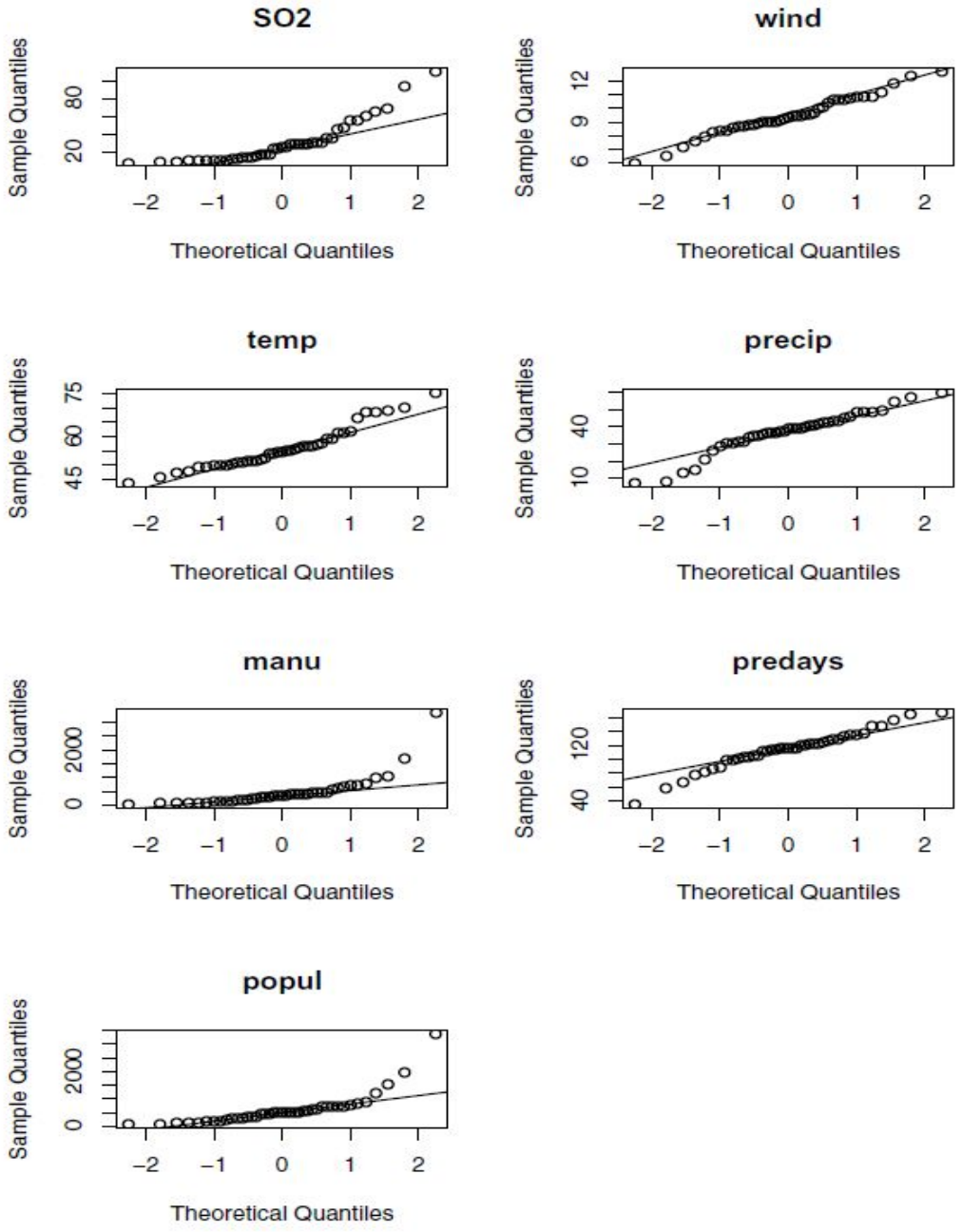


Fig. 1.5. Normal probability plots for USairpollution data.

The resulting seven plots are arranged on one page by a call to the layou matrix; see Figure 1.5. The plots for SO_2 concentration and precipitatio both deviate considerably from linearity, and the plots for manufacturing an population show evidence of a number of outliers. But of more importanc is the chi-square plot for the data, which is given in Figure 1.6; the R cod is identical to the code used to produce the chi-square plot for the bod measurement data. In addition, the two most extreme points in the plot hav 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)
```

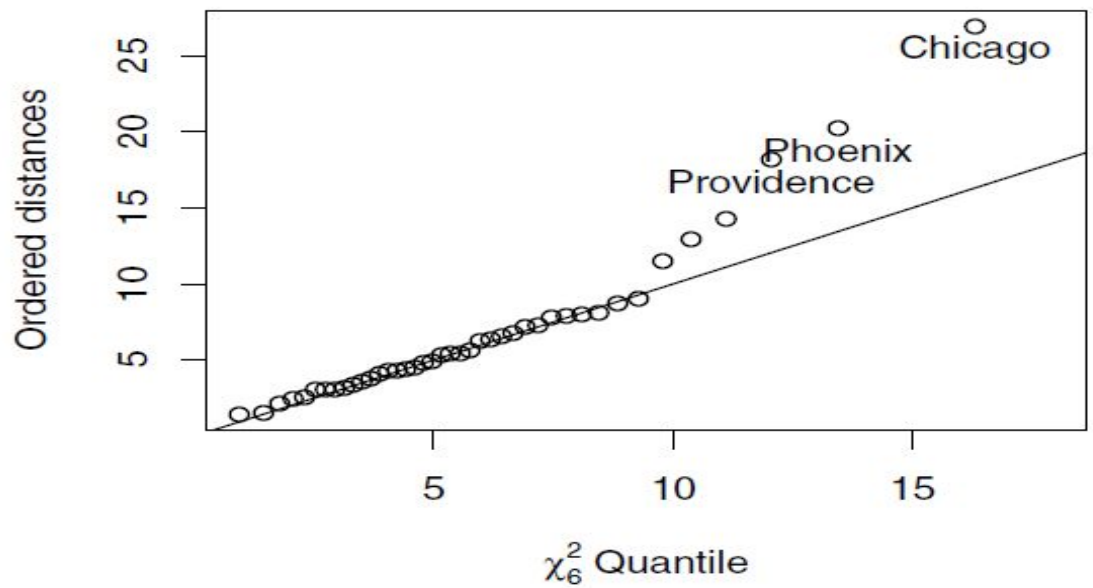


Fig. 1.6. χ^2 plot of generalised distances for USairpollution data.

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)

Data graphics visually display measured quantities by means of the combined use of points, lines, a coordinate system, numbers, symbols, words, shading and color.

Some of the advantages of graphical methods have been listed by Schmid (1954):

- In comparison with other types of presentation, well-designed charts are more effective in creating interest and in appealing to the attention of the reader.
- Visual relationships as portrayed by charts and graphs are more easily grasped and more easily remembered.
- The use of charts and graphs saves time since the essential meaning of large measures of statistical data can be visualised at a glance.
- Charts and graphs provide a comprehensive picture of a problem that makes for a more complete and better balanced understanding than could be derived from tabular or textual forms of presentation.
- Charts and graphs can bring out hidden facts and relationships and can stimulate, as well as aid, analytical thinking and investigation.

```
R> pairs(USairpollution, pch = ".", cex = 1.5)
```

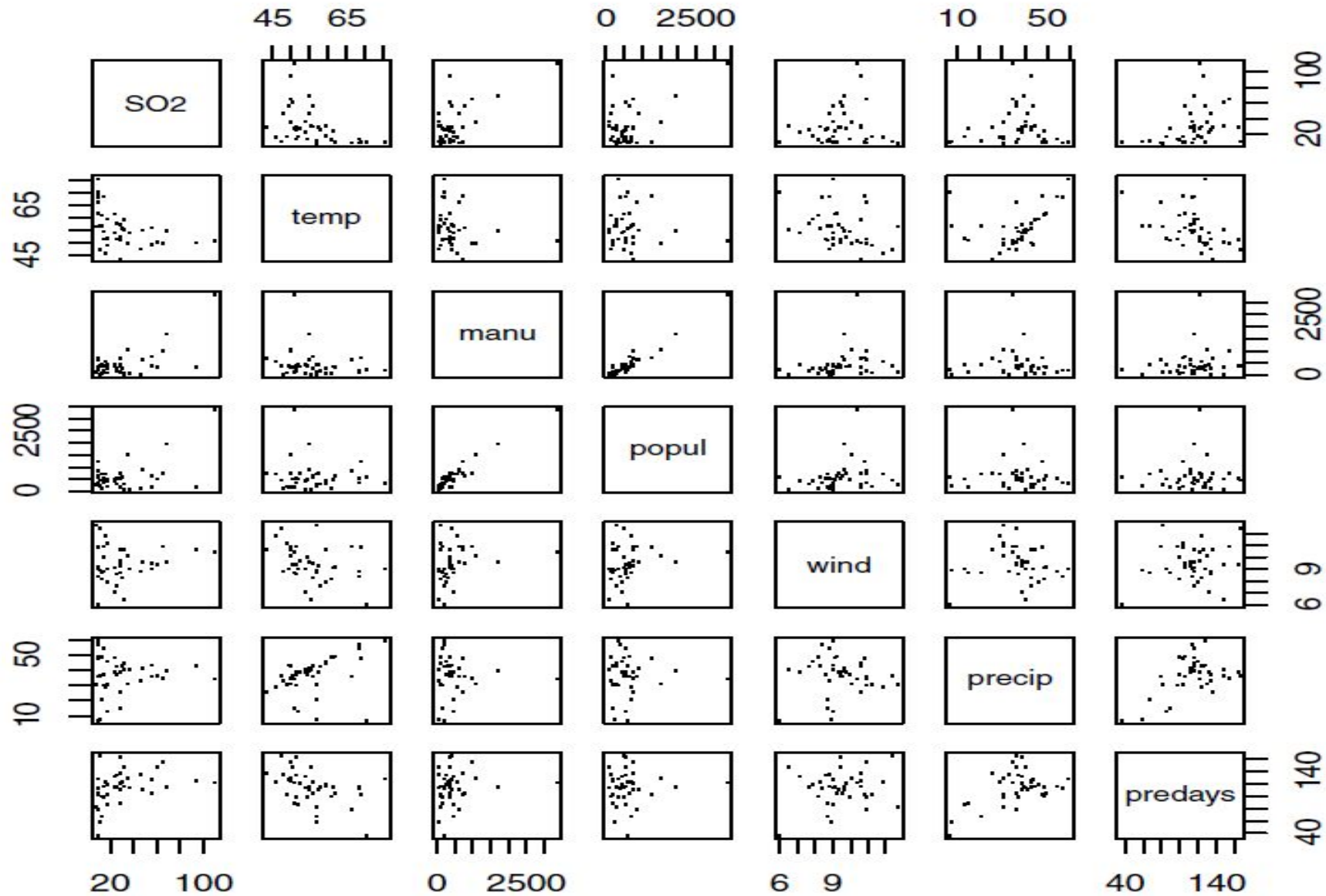


Fig. 2.10. Scatterplot matrix of the air pollution data.

```
R> round(cor(USairpollution), 4)
```

	SO2	temp	manu	popul	wind	precip	predays
SO2	1.0000	-0.4336	0.6448	0.4938	0.0947	0.0543	0.3696
temp	-0.4336	1.0000	-0.1900	-0.0627	-0.3497	0.3863	-0.4302
manu	0.6448	-0.1900	1.0000	0.9553	0.2379	-0.0324	0.1318
popul	0.4938	-0.0627	0.9553	1.0000	0.2126	-0.0261	0.0421
wind	0.0947	-0.3497	0.2379	0.2126	1.0000	-0.0130	0.1641
precip	0.0543	0.3863	-0.0324	-0.0261	-0.0130	1.0000	0.4961
predays	0.3696	-0.4302	0.1318	0.0421	0.1641	0.4961	1.0000

Focussing on the correlations between `SO2` and the six other variables, we see that the correlation for `SO2` and `precip` is very small and that for `SO2` and `predays` is moderate.

Positive Definite Matrices

The study of the variation and interrelationships in multivariate data is often based upon distances and the assumption that the data are multivariate normally distributed. Squared distances and the multivariate normal density can be expressed in terms of matrix products called *quadratic forms* . Consequently, it should not be surprising that quadratic forms play a central role in multivariate analysis. We consider quadratic forms that are always nonnegative and the associated *positive definite* matrices. Results involving quadratic forms and symmetric matrices are, in many cases, a direct consequence of an expansion for symmetric matrices known as the *spectral decomposition.* The spectral decomposition of *k x k* symmetric matrix A is given by

$$\underset{(k\times k)}{\mathbf{A}} = \lambda_1 \underset{(k\times 1)(1\times k)}{\mathbf{e}_1\mathbf{e}_1'} + \lambda_2 \underset{(k\times 1)(1\times k)}{\mathbf{e}_2\mathbf{e}_2'} + \cdots + \lambda_k \underset{(k\times 1)(1\times k)}{\mathbf{e}_k\mathbf{e}_k'} \qquad (2\text{-}16)$$

where $\lambda_1, \lambda_2, \ldots, \lambda_k$ are the eigenvalues of $\mathbf{A}$ and $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_k$ are the associated normalized eigenvectors. (See also Result 2A.14 in Supplement 2A). Thus, $\mathbf{e}_i'\mathbf{e}_i = 1$ for $i = 1, 2, \ldots, k$, and $\mathbf{e}_i'\mathbf{e}_j = 0$ for $i \neq j$.

Example 2.10 (The spectral decomposition of a matrix) Consider the symmetric matrix

$$\mathbf{A} = \begin{bmatrix} 13 & -4 & 2 \\ -4 & 13 & -2 \\ 2 & -2 & 10 \end{bmatrix}$$

The eigenvalues obtained from the characteristic equation $|\mathbf{A} - \lambda\mathbf{I}| = 0$ are $\lambda_1 = 9$, $\lambda_2 = 9$, and $\lambda_3 = 18$ (Definition 2A.30). The corresponding eigenvectors $\mathbf{e}_1, \mathbf{e}_2$, and $\mathbf{e}_3$ are the (normalized) solutions of the equations $\mathbf{A}\mathbf{e}_i = \lambda_i\mathbf{e}_i$ for $i = 1, 2, 3$. Thus, $\mathbf{A}\mathbf{e}_1 = \lambda\mathbf{e}_1$ gives

$$\begin{bmatrix} 13 & -4 & 2 \\ -4 & 13 & -2 \\ 2 & -2 & 10 \end{bmatrix} \begin{bmatrix} e_{11} \\ e_{21} \\ e_{31} \end{bmatrix} = 9 \begin{bmatrix} e_{11} \\ e_{21} \\ e_{31} \end{bmatrix}$$

or

$$\begin{aligned} 13e_{11} - 4e_{21} + 2e_{31} &= 9e_{11} \\ -4e_{11} + 13e_{21} - 2e_{31} &= 9e_{21} \\ 2e_{11} - 2e_{21} + 10e_{31} &= 9e_{31} \end{aligned}$$

Moving the terms on the right of the equals sign to the left yields three homogeneous equations in three unknowns, but two of the equations are redundant. Selecting one of the equations and arbitrarily setting $e_{11} = 1$ and $e_{21} = 1$, we find that $e_{31} = 0$. Consequently, the normalized eigenvector is $\mathbf{e}_1' = [1/\sqrt{1^2 + 1^2 + 0^2}, 1/\sqrt{1^2 + 1^2 + 0^2}, 0/\sqrt{1^2 + 1^2 + 0^2}] = [1/\sqrt{2}, 1/\sqrt{2}, 0]$, since the sum of the squares of its elements is unity. You may verify that $\mathbf{e}_2' = [1/\sqrt{18}, -1/\sqrt{18}, -4/\sqrt{18}]$ is also an eigenvector for $9 = \lambda_2$, and $\mathbf{e}_3' = [2/3, -2/3, 1/3]$ is the normalized eigenvector corresponding to the eigenvalue $\lambda_3 = 18$. Moreover, $\mathbf{e}_i'\mathbf{e}_j = 0$ for $i \neq j$.

The spectral decomposition of $\mathbf{A}$ is then

$$\mathbf{A} = \lambda_1 \mathbf{e}_1 \mathbf{e}_1' + \lambda_2 \mathbf{e}_2 \mathbf{e}_2' + \lambda_3 \mathbf{e}_3 \mathbf{e}_3'$$

or

$$\begin{bmatrix} 13 & -4 & 2 \\ -4 & 13 & -2 \\ 2 & -2 & 10 \end{bmatrix} = 9 \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \end{bmatrix}$$

$$+ 9 \begin{bmatrix} \frac{1}{\sqrt{18}} \\ \frac{-1}{\sqrt{18}} \\ \frac{-4}{\sqrt{18}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{18}} & \frac{-1}{\sqrt{18}} & \frac{-4}{\sqrt{18}} \end{bmatrix} + 18 \begin{bmatrix} \frac{2}{3} \\ -\frac{2}{3} \\ \frac{1}{3} \end{bmatrix} \begin{bmatrix} \frac{2}{3} & -\frac{2}{3} & \frac{1}{3} \end{bmatrix}$$

$$= 9 \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 \end{bmatrix} + 9 \begin{bmatrix} \frac{1}{18} & -\frac{1}{18} & -\frac{4}{18} \\ -\frac{1}{18} & \frac{1}{18} & \frac{4}{18} \\ -\frac{4}{18} & \frac{4}{18} & \frac{16}{18} \end{bmatrix}$$

$$+ 18 \begin{bmatrix} \frac{4}{9} & -\frac{4}{9} & \frac{2}{9} \\ -\frac{4}{9} & \frac{4}{9} & -\frac{2}{9} \\ \frac{2}{9} & -\frac{2}{9} & \frac{1}{9} \end{bmatrix}$$

as you may readily verify. ■

The spectral decomposition is an important analytical tool. With it, we are very easily able to demonstrate certain statistical results. The first of these is a matrix explanation of distance, which we now develop.

Because $\mathbf{x}'\mathbf{A}\mathbf{x}$ has only squared terms x_i^2 and product terms $x_i x_k$, it is called a *quadratic form*. When a $k \times k$ symmetric matrix $\mathbf{A}$ is such that

$$0 \leq \mathbf{x}'\mathbf{A}\mathbf{x} \tag{2-17}$$

for all $\mathbf{x}' = [x_1, x_2, \ldots, x_k]$, both the matrix $\mathbf{A}$ and the quadratic form are said to be *nonnegative definite*. If equality holds in (2-17) only for the vector $\mathbf{x}' = [0, 0, \ldots, 0]$, then $\mathbf{A}$ or the quadratic form is said to be *positive definite*. In other words, $\mathbf{A}$ is positive definite if

$$0 < \mathbf{x}'\mathbf{A}\mathbf{x} \tag{2-18}$$

for all vectors $\mathbf{x} \neq \mathbf{0}$.

A Square-Root Matrix

The spectral decomposition allows us to express the inverse of a square matrix in terms of its eigenvalues and eigenvectors, and this leads to a useful *square-root matrix*.

Let $\mathbf{A}$ be a $k \times k$ positive definite matrix with the spectral decomposition $\mathbf{A} = \sum_{i=1}^{k} \lambda_i \mathbf{e}_i \mathbf{e}_i'$. Let the normalized eigenvectors be the columns of another matrix $\mathbf{P} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_k]$. Then

$$\underset{(k\times k)}{\mathbf{A}} = \sum_{i=1}^{k} \lambda_i \underset{(k\times 1)}{\mathbf{e}_i} \underset{(1\times k)}{\mathbf{e}_i'} = \underset{(k\times k)}{\mathbf{P}} \underset{(k\times k)}{\mathbf{\Lambda}} \underset{(k\times k)}{\mathbf{P}'} \tag{2-20}$$

where $\mathbf{PP}' = \mathbf{P}'\mathbf{P} = \mathbf{I}$ and $\mathbf{\Lambda}$ is the diagonal matrix

$$\underset{(k\times k)}{\mathbf{\Lambda}} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k \end{bmatrix} \quad \text{with } \lambda_i > 0$$

Thus,

$$\mathbf{A}^{-1} = \mathbf{P}\boldsymbol{\Lambda}^{-1}\mathbf{P}' = \sum_{i=1}^{k} \frac{1}{\lambda_i}\mathbf{e}_i\mathbf{e}_i' \tag{2-21}$$

since $(\mathbf{P}\boldsymbol{\Lambda}^{-1}\mathbf{P}')\mathbf{P}\boldsymbol{\Lambda}\mathbf{P}' = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}'(\mathbf{P}\boldsymbol{\Lambda}^{-1}\mathbf{P}') = \mathbf{P}\mathbf{P}' = \mathbf{I}$.

Next, let $\boldsymbol{\Lambda}^{1/2}$ denote the diagonal matrix with $\sqrt{\lambda_i}$ as the ith diagonal element. The matrix $\sum_{i=1}^{k} \sqrt{\lambda_i}\,\mathbf{e}_i\mathbf{e}_i' = \mathbf{P}\boldsymbol{\Lambda}^{1/2}\mathbf{P}'$ is called the *square root* of $\mathbf{A}$ and is denoted by $\mathbf{A}^{1/2}$.

The square-root matrix, of a positive definite matrix $\mathbf{A}$,

$$\mathbf{A}^{1/2} = \sum_{i=1}^{k} \sqrt{\lambda_i}\,\mathbf{e}_i\mathbf{e}_i' = \mathbf{P}\boldsymbol{\Lambda}^{1/2}\mathbf{P}' \tag{2-22}$$

has the following properties:

1. $(\mathbf{A}^{1/2})' = \mathbf{A}^{1/2}$ (that is, $\mathbf{A}^{1/2}$ is symmetric).
2. $\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{A}$.
3. $(\mathbf{A}^{1/2})^{-1} = \sum_{i=1}^{k} \frac{1}{\sqrt{\lambda_i}}\mathbf{e}_i\mathbf{e}_i' = \mathbf{P}\boldsymbol{\Lambda}^{-1/2}\mathbf{P}'$, where $\boldsymbol{\Lambda}^{-1/2}$ is a diagonal matrix with $1/\sqrt{\lambda_i}$ as the ith diagonal element.
4. $\mathbf{A}^{1/2}\mathbf{A}^{-1/2} = \mathbf{A}^{-1/2}\mathbf{A}^{1/2} = \mathbf{I}$, and $\mathbf{A}^{-1/2}\mathbf{A}^{-1/2} = \mathbf{A}^{-1}$, where $\mathbf{A}^{-1/2} = (\mathbf{A}^{1/2})^{-1}$.

Random Vectors and Matrices

A *random vector* is a vector whose elements are random variables. Similarly, a *random matrix* is a matrix whose elements are random variables. The expected value of a random matrix (or vector) is the matrix (vector) consisting of the expected values of each of its elements. Specifically, let $\mathbf{X} = \{X_{ij}\}$ be an $n \times p$ random matrix. Then the expected value of $\mathbf{X}$, denoted by $E(\mathbf{X})$, is the $n \times p$ matrix of numbers (if they exist)

$$E(\mathbf{X}) = \begin{bmatrix} E(X_{11}) & E(X_{12}) & \cdots & E(X_{1p}) \\ E(X_{21}) & E(X_{22}) & \cdots & E(X_{2p}) \\ \vdots & \vdots & \ddots & \vdots \\ E(X_{n1}) & E(X_{n2}) & \cdots & E(X_{np}) \end{bmatrix} \tag{2-23}$$

where, for each element of the matrix,[2]

$$E(X_{ij}) = \begin{cases} \displaystyle\int_{-\infty}^{\infty} x_{ij} f_{ij}(x_{ij})\, dx_{ij} & \text{if } X_{ij} \text{ is a continuous random variable with probability density function } f_{ij}(x_{ij}) \\ \displaystyle\sum_{\text{all } x_{ij}} x_{ij} p_{ij}(x_{ij}) & \text{if } X_{ij} \text{ is a discrete random variable with probability function } p_{ij}(x_{ij}) \end{cases}$$

Example 2.12 (Computing expected values for discrete random variables) Suppose $p = 2$ and $n = 1$, and consider the random vector $\mathbf{X}' = [X_1, X_2]$. Let the discrete random variable X_1 have the following probability function:

x_1	-1	0	1
$p_1(x_1)$	.3	.3	.4

Then $E(X_1) = \sum_{\text{all } x_1} x_1 p_1(x_1) = (-1)(.3) + (0)(.3) + (1)(.4) = .1.$

Similarly, let the discrete random variable X_2 have the probability function

x_2	0	1
$p_2(x_2)$	.8	.2

Then $E(X_2) = \sum_{\text{all } x_2} x_2 p_2(x_2) = (0)(.8) + (1)(.2) = .2.$

Thus,

$$E(\mathbf{X}) = \begin{bmatrix} E(X_1) \\ E(X_2) \end{bmatrix} = \begin{bmatrix} .1 \\ .2 \end{bmatrix}$$

■

Two results involving the expectation of sums and products of matrices follow directly from the definition of the expected value of a random matrix and the univariate properties of expectation, $E(X_1 + Y_1) = E(X_1) + E(Y_1)$ and $E(cX_1) = cE(X_1)$. Let $\mathbf{X}$ and $\mathbf{Y}$ be random matrices of the same dimension, and let $\mathbf{A}$ and $\mathbf{B}$ be conformable matrices of constants. Then (see Exercise 2.40)

$$E(\mathbf{X} + \mathbf{Y}) = E(\mathbf{X}) + E(\mathbf{Y}) \tag{2-24}$$
$$E(\mathbf{AXB}) = \mathbf{A}E(\mathbf{X})\mathbf{B}$$

Mean Vectors and Covariance Matrices

Suppose $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ is a $p \times 1$ random vector. Then each element of $\mathbf{X}$ is a random variable with its own marginal probability distribution. (See Example 2.12.) The marginal means μ_i and variances σ_i^2 are defined as $\mu_i = E(X_i)$ and $\sigma_i^2 = E(X_i - \mu_i)^2$, $i = 1, 2, \ldots, p$, respectively. Specifically,

$$\mu_i = \begin{cases} \displaystyle\int_{-\infty}^{\infty} x_i f_i(x_i)\, dx_i & \text{if } X_i \text{ is a continuous random variable with probability density function } f_i(x_i) \\ \displaystyle\sum_{\text{all } x_i} x_i p_i(x_i) & \text{if } X_i \text{ is a discrete random variable with probability function } p_i(x_i) \end{cases}$$

$$\sigma_i^2 = \begin{cases} \displaystyle\int_{-\infty}^{\infty} (x_i - \mu_i)^2 f_i(x_i)\, dx_i & \text{if } X_i \text{ is a continuous random variable with probability density function } f_i(x_i) \\ \displaystyle\sum_{\text{all } x_i} (x_i - \mu_i)^2 p_i(x_i) & \text{if } X_i \text{ is a discrete random variable with probability function } p_i(x_i) \end{cases} \tag{2-25}$$

It will be convenient in later sections to denote the marginal variances by σ_{ii} rather than the more traditional σ_i^2, and consequently, we shall adopt this notation.

The behavior of any pair of random variables, such as X_i and X_k, is described by their joint probability function, and a measure of the linear association between them is provided by the covariance

$$\sigma_{ik} = E(X_i - \mu_i)(X_k - \mu_k)$$

$$= \begin{cases} \displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x_i - \mu_i)(x_k - \mu_k) f_{ik}(x_i, x_k)\, dx_i\, dx_k & \text{if } X_i, X_k \text{ are continuous random variables with the joint density function } f_{ik}(x_i, x_k) \\ \displaystyle\sum_{\text{all } x_i}\sum_{\text{all } x_k} (x_i - \mu_i)(x_k - \mu_k) p_{ik}(x_i, x_k) & \text{if } X_i, X_k \text{ are discrete random variables with joint probability function } p_{ik}(x_i, x_k) \end{cases} \tag{2-26}$$

and μ_i and μ_k, $i, k = 1, 2, \ldots, p$, are the marginal means. When $i = k$, the covariance becomes the marginal variance.

More generally, the collective behavior of the p random variables $X_1, X_2, \ldots, X_p$ or, equivalently, the random vector $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$, is described by a joint probability density function $f(x_1, x_2, \ldots, x_p) = f(\mathbf{x})$. As we have already noted in

f(x) will often be the multivariate normal density function.

If the joint probability $P[X_i \leq x_i \text{ and } X_k \leq x_k]$ can be written as the product of the corresponding marginal probabilities, so that

$$P[X_i \leq x_i \text{ and } X_k \leq x_k] = P[X_i \leq x_i]P[X_k \leq x_k] \tag{2-27}$$

for all pairs of values x_i, x_k, then X_i and X_k are said to be *statistically independent.* When X_i and X_k are continuous random variables with joint density $f_{ik}(x_i, x_k)$ and marginal densities $f_i(x_i)$ and $f_k(x_k)$, the independence condition becomes

$$f_{ik}(x_i, x_k) = f_i(x_i)f_k(x_k)$$

for all pairs (x_i, x_k).

The p continuous random variables $X_1, X_2, \ldots, X_p$ are *mutually statistically independent* if their joint density can be factored as

$$f_{12\cdots p}(x_1, x_2, \ldots, x_p) = f_1(x_1)f_2(x_2)\cdots f_p(x_p) \tag{2-28}$$

for all p-tuples $(x_1, x_2, \ldots, x_p)$.

Statistical independence has an important implication for covariance. The factorization in (2-28) implies that $\text{Cov}(X_i, X_k) = 0$. Thus,

$$\text{Cov}(X_i, X_k) = 0 \qquad \text{if } X_i \text{ and } X_k \text{ are independent} \tag{2-29}$$

The converse of (2-29) is not true in general; there are situations where $\text{Cov}(X_i, X_k) = 0$, but X_i and X_k are not independent.

The means and covariances of the $p \times 1$ random vector $\mathbf{X}$ can be set out as matrices. The expected value of each element is contained in the vector of means $\boldsymbol{\mu} = E(\mathbf{X})$, and the p variances σ_{ii} and the $p(p-1)/2$ distinct covariances $\sigma_{ik}(i < k)$ are contained in the symmetric variance-covariance matrix $\boldsymbol{\Sigma} = E(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})'$. Specifically,

$$E(\mathbf{X}) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} = \boldsymbol{\mu} \tag{2-30}$$

and

$$= E(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})'$$

$$= E\left(\begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \\ \vdots \\ X_p - \mu_p \end{bmatrix} [X_1 - \mu_1, X_2 - \mu_2, \ldots, X_p - \mu_p] \right)$$

$$= E\begin{bmatrix} (X_1 - \mu_1)^2 & (X_1 - \mu_1)(X_2 - \mu_2) & \cdots & (X_1 - \mu_1)(X_p - \mu_p) \\ (X_2 - \mu_2)(X_1 - \mu_1) & (X_2 - \mu_2)^2 & \cdots & (X_2 - \mu_2)(X_p - \mu_p) \\ \vdots & \vdots & \ddots & \vdots \\ (X_p - \mu_p)(X_1 - \mu_1) & (X_p - \mu_p)(X_2 - \mu_2) & \cdots & (X_p - \mu_p)^2 \end{bmatrix}$$

$$= \begin{bmatrix} E(X_1-\mu_1)^2 & E(X_1-\mu_1)(X_2-\mu_2) & \cdots & E(X_1-\mu_1)(X_p-\mu_p) \\ E(X_2-\mu_2)(X_1-\mu_1) & E(X_2-\mu_2)^2 & \cdots & E(X_2-\mu_2)(X_p-\mu_p) \\ \vdots & \vdots & \ddots & \vdots \\ E(X_p-\mu_p)(X_1-\mu_1) & E(X_p-\mu_p)(X_2-\mu_2) & \cdots & E(X_p-\mu_p)^2 \end{bmatrix}$$

or

$$\boldsymbol{\Sigma} = \text{Cov}(\mathbf{X}) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix} \tag{2-31}$$

Example 2.13 (Computing the covariance matrix) Find the covariance matrix for the two random variables X_1 and X_2 introduced in Example 2.12 when their joint probability function $p_{12}(x_1, x_2)$ is represented by the entries in the body of the following table:

x_1 \ x_2	0	1	$p_1(x_1)$
−1	.24	.06	.3
0	.16	.14	.3
1	.40	.00	.4
$p_2(x_2)$	.8	.2	1

We have already shown that $\mu_1 = E(X_1) = .1$ and $\mu_2 = E(X_2) = .2$. (See Example 2.12.) In addition,

$$\sigma_{11} = E(X_1 - \mu_1)^2 = \sum_{\text{all } x_1} (x_1 - .1)^2 p_1(x_1)$$

$$= (-1 - .1)^2(.3) + (0 - .1)^2(.3) + (1 - .1)^2(.4) = .69$$

$$\sigma_{22} = E(X_2 - \mu_2)^2 = \sum_{\text{all } x_2} (x_2 - .2)^2 p_2(x_2)$$

$$= (0 - .2)^2(.8) + (1 - .2)^2(.2)$$

$$= .16$$

$$\sigma_{12} = E(X_1 - \mu_1)(X_2 - \mu_2) = \sum_{\text{all pairs } (x_1, x_2)} (x_1 - .1)(x_2 - .2) p_{12}(x_1, x_2)$$

$$= (-1 - .1)(0 - .2)(.24) + (-1 - .1)(1 - .2)(.06)$$

$$+ \cdots + (1 - .1)(1 - .2)(.00) = -.08$$

$$\sigma_{21} = E(X_2 - \mu_2)(X_1 - \mu_1) = E(X_1 - \mu_1)(X_2 - \mu_2) = \sigma_{12} = -.08$$

Consequently, with $\mathbf{X}' = [X_1, X_2]$,

$$\boldsymbol{\mu} = E(\mathbf{X}) = \begin{bmatrix} E(X_1) \\ E(X_2) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} .1 \\ .2 \end{bmatrix}$$

and

$$\boldsymbol{\Sigma} = E(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})'$$

$$= E\begin{bmatrix} (X_1 - \mu_1)^2 & (X_1 - \mu_1)(X_2 - \mu_2) \\ (X_2 - \mu_2)(X_1 - \mu_1) & (X_2 - \mu_2)^2 \end{bmatrix}$$

$$= \begin{bmatrix} E(X_1 - \mu_1)^2 & E(X_1 - \mu_1)(X_2 - \mu_2) \\ E(X_2 - \mu_2)(X_1 - \mu_1) & E(X_2 - \mu_2)^2 \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} = \begin{bmatrix} .69 & -.08 \\ -.08 & .16 \end{bmatrix}$$

■

We note that the computation of means, variances, and covariances for *discrete* random variables involves summation (as in Examples 2.12 and 2.13), while analogous computations for *continuous* random variables involve integration.

Because $\sigma_{ik} = E(X_i - \mu_i)(X_k - \mu_k) = \sigma_{ki}$, it is convenient to write the matrix appearing in (2-31) as

$$\boldsymbol{\Sigma} = E(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})' = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \cdots & \sigma_{pp} \end{bmatrix} \tag{2-32}$$

We shall refer to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as the *population mean* (vector) and *population variance–covariance* (matrix), respectively.

Properties of the Covariance Matrix $\Sigma = \mathsf{Var}(X)$

$$\Sigma = (\sigma_{X_i X_j}), \quad \sigma_{X_i X_j} = \mathsf{Cov}(X_i, X_j), \quad \sigma_{X_i X_i} = \mathsf{Var}(X_i)$$

$$\Sigma = \mathsf{E}(XX^\top) - \mu\mu^\top$$

$$\Sigma \geq 0$$

$$\mathsf{Var}(a^\top X) = a^\top \mathsf{Var}(X) a = \sum_{i,j} a_i a_j \sigma_{X_i X_j}$$

$$\mathsf{Var}(AX + b) = A\,\mathsf{Var}(X) A^\top$$

$$\mathsf{Cov}(X + Y, Z) = \mathsf{Cov}(X, Z) + \mathsf{Cov}(Y, Z)$$

$$\mathsf{Var}(X + Y) = \mathsf{Var}(X) + \mathsf{Cov}(X, Y) + \mathsf{Cov}(Y, X) + \mathsf{Var}(Y)$$

$$\mathsf{Cov}(AX, BY) = A\,\mathsf{Cov}(X, Y) B^\top.$$

Theorem 4.6 *Let* $X \sim N_p(\mu, \Sigma)$ *and* $A(p \times p)$, $c \in \mathbb{R}^p$, *where* A *is nonsingular. Then* $Y = AX + c$ *is again a p-variate Normal, i.e.*

$$Y \sim N_p(A\mu + c, A\Sigma A^\top). \tag{4.50}$$

The multinormal distribution with mean μ and covariance $\Sigma > 0$ has the density

$$f(x) = |2\pi\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right\}. \tag{4.47}$$

We write $X \sim N_p(\mu, \Sigma)$.

How is this multinormal distribution with mean μ and covariance Σ related to the multivariate standard normal $N_p(0, \mathcal{I}_p)$? Through a linear transformation using

Theorem 4.5 *Let $X \sim N_p(\mu, \Sigma)$ and $Y = \Sigma^{-1/2}(X - \mu)$ (Mahalanobis transformation). Then*

$$Y \sim N_p(0, \mathcal{I}_p),$$

i.e. the elements $Y_j \in \mathbb{R}$ are independent, one-dimensional $N(0, 1)$ variables.

How can we create $N_p(\mu, \Sigma)$ variables on the basis of $N_p(0, \mathcal{I}_p)$ variables? We use the inverse linear transformation

$$X = \Sigma^{1/2} Y + \mu. \tag{4.49}$$

we can also check that $\mathsf{E}(X) = \mu$ and $\mathsf{Var}(X) = \Sigma$.

Theorem 4.7 *If $X \sim N_p(\mu, \Sigma)$, then the variable $U = (X-\mu)^\top \Sigma^{-1}(X-\mu)$ has a χ^2_p distribution.*

Theorem 4.8 *The characteristic function (cf) of a multinormal $N_p(\mu, \Sigma)$ is given by*

$$\varphi_X(t) = \exp\left(i\, t^\top \mu - \frac{1}{2} t^\top \Sigma t\right). \tag{4.52}$$

Geometry of the $N_p(\mu, \Sigma)$ Distribution

From (4.47) we see that the density of the $N_p(\mu, \Sigma)$ distribution is constant on ellipsoids of the form

$$(x-\mu)^\top \Sigma^{-1}(x-\mu) = d^2. \tag{4.51}$$

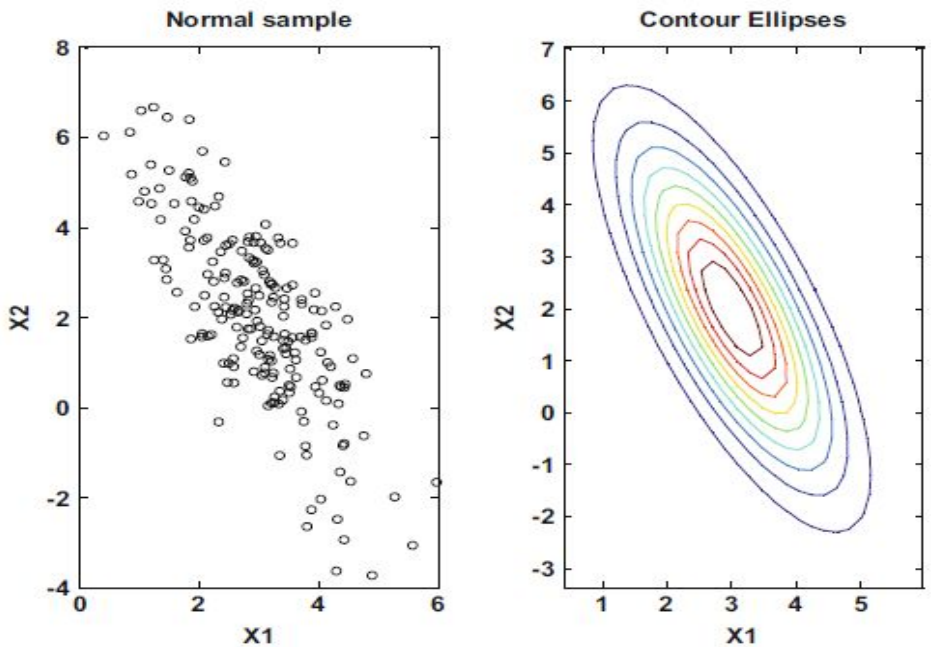


Fig. 4.3 Scatterplot of a normal sample and contour ellipses for $\mu = \binom{3}{2}$ and $\Sigma = \left(\begin{smallmatrix} 1 & -1.5 \\ -1.5 & 4 \end{smallmatrix}\right)$ MVAcontnorm

Singular Normal Distribution

Suppose that we have $\text{rank}(\Sigma) = k < p$, where p is the dimension of X. We define the (singular) density of X with the aid of the G-Inverse Σ^- of Σ,

$$f(x) = \frac{(2\pi)^{-k/2}}{(\lambda_1 \cdots \lambda_k)^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^\top \Sigma^- (x-\mu)\right\} \tag{4.53}$$

where

1. x lies on the hyperplane $\mathcal{N}^\top(x-\mu) = 0$ with $\mathcal{N}(p \times (p-k)) : \mathcal{N}^\top \Sigma = 0$ and $\mathcal{N}^\top \mathcal{N} = \mathcal{I}_k$.
2. Σ^- is the G-Inverse of Σ, and $\lambda_1, \ldots, \lambda_k$ are the nonzero eigenvalues of Σ.

What is the connection to a multinormal with k-dimensions? If

$$Y \sim N_k(0, \Lambda_1) \text{ and } \Lambda_1 = \text{diag}(\lambda_1, \ldots, \lambda_k), \tag{4.54}$$

then an orthogonal matrix $\mathcal{B}(p \times k)$ with $\mathcal{B}^\top \mathcal{B} = \mathcal{I}_k$ exists that means $X = \mathcal{B}Y + \mu$ where X has a singular pdf of the form (4.53).

It is frequently informative to separate the information contained in variances σ_{ii} from that contained in measures of association and, in particular, the measure of association known as the *population correlation coefficient* ρ_{ik}. The correlation coefficient ρ_{ik} is defined in terms of the covariance σ_{ik} and variances σ_{ii} and σ_{kk} as

$$\rho_{ik} = \frac{\sigma_{ik}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{kk}}} \tag{2-33}$$

The correlation coefficient measures the amount of *linear* association between the random variables X_i and X_k.

Let the population correlation matrix be the $p \times p$ symmetric matrix

$$\boldsymbol{\rho} = \begin{bmatrix} \frac{\sigma_{11}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{11}}} & \frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} & \cdots & \frac{\sigma_{1p}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{pp}}} \\ \frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} & \frac{\sigma_{22}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{22}}} & \cdots & \frac{\sigma_{2p}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{pp}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\sigma_{1p}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{pp}}} & \frac{\sigma_{2p}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{pp}}} & \cdots & \frac{\sigma_{pp}}{\sqrt{\sigma_{pp}}\sqrt{\sigma_{pp}}} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{12} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1p} & \rho_{2p} & \cdots & 1 \end{bmatrix} \tag{2-34}$$

and let the $p \times p$ *standard deviation* matrix be

$$\mathbf{V}^{1/2} = \begin{bmatrix} \sqrt{\sigma_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{\sigma_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\sigma_{pp}} \end{bmatrix} \tag{2-35}$$

Then it is easily verified (see Exercise 2.23) that

$$\mathbf{V}^{1/2}\boldsymbol{\rho}\mathbf{V}^{1/2} = \boldsymbol{\Sigma} \tag{2-36}$$

and

$$\boldsymbol{\rho} = (\mathbf{V}^{1/2})^{-1}\boldsymbol{\Sigma}(\mathbf{V}^{1/2})^{-1} \tag{2-37}$$

That is, $\boldsymbol{\Sigma}$ can be obtained from $\mathbf{V}^{1/2}$ and $\boldsymbol{\rho}$, whereas $\boldsymbol{\rho}$ can be obtained from $\boldsymbol{\Sigma}$. Moreover, the expression of these relationships in terms of matrix operations allows the calculations to be conveniently implemented on a computer.

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 $\mathbf{D}^{-1/2} = \text{diag}(1/s_1, \ldots, 1/s_q)$ 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
```

Example 2.14 (Computing the correlation matrix from the covariance matrix) Suppose

$$\Sigma = \begin{bmatrix} 4 & 1 & 2 \\ 1 & 9 & -3 \\ 2 & -3 & 25 \end{bmatrix} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{bmatrix}$$

Obtain $\mathbf{V}^{1/2}$ and $\boldsymbol{\rho}$.

Here

$$\mathbf{V}^{1/2} = \begin{bmatrix} \sqrt{\sigma_{11}} & 0 & 0 \\ 0 & \sqrt{\sigma_{22}} & 0 \\ 0 & 0 & \sqrt{\sigma_{33}} \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 5 \end{bmatrix}$$

and

$$(\mathbf{V}^{1/2})^{-1} = \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{5} \end{bmatrix}$$

Consequently, from (2-37), the correlation matrix $\boldsymbol{\rho}$ is given by

$$(\mathbf{V}^{1/2})^{-1}\boldsymbol{\Sigma}(\mathbf{V}^{1/2})^{-1} = \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{5} \end{bmatrix} \begin{bmatrix} 4 & 1 & 2 \\ 1 & 9 & -3 \\ 2 & -3 & 25 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{5} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & \frac{1}{6} & \frac{1}{5} \\ \frac{1}{6} & 1 & -\frac{1}{5} \\ \frac{1}{5} & -\frac{1}{5} & 1 \end{bmatrix}$$

The zero-order Pearson correlation between two random variables Y_i and Y_j is given by

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j} = \frac{\operatorname{cov}\left(Y_i, Y_j\right)}{\sqrt{\operatorname{var}\left(Y_i\right) \operatorname{var}\left(Y_j\right)}} \quad \text{where} \quad -1 \leq \rho_{ij} \leq 1$$

The correlation matrix for the random p-vector $\mathbf{Y}$ is

$$\mathbf{P} = \left[\rho_{ij}\right] \tag{3.2.4}$$

Letting $(\operatorname{diag} \Sigma)^{-1/2}$ represent the diagonal matrix with diagonal elements equal to the square root of the diagonal elements of Σ, the relationship between $\mathbf{P}$ and Σ is established

$$\mathbf{P} = (\operatorname{diag} \Sigma)^{-1/2} \Sigma (\operatorname{diag} \Sigma)^{-1/2}$$
$$\Sigma = (\operatorname{diag} \Sigma)^{1/2} \mathbf{P} (\operatorname{diag} \Sigma)^{1/2}$$

Because the correlation matrix does not depend on the scale of the random variables, it is used to express relationships among random variables measured on different scales. Furthermore, since the $|\Sigma| = |(\operatorname{diag} \Sigma)^{1/2} \mathbf{P} (\operatorname{diag} \Sigma)^{1/2}|$ we have that $0 \leq |\mathbf{P}|^2 \leq 1$. Takeuchi, et al. (1982, p. 246) call the $|\mathbf{P}|$ the generalized alienation coefficient. If the elements of $\mathbf{Y}$ are independent its value is one and if elements are dependent it value is zero. Thus, the determinant of the correlation matrix may be interpreted as an overall measure of association or nonassociation.

Partitioning the Covariance Matrix

In general, we can partition the p characteristics contained in the $p \times 1$ random vector $\mathbf{X}$ into, for instance, two groups of size q and $p - q$, respectively. For example, we can write

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_q \\ \hline X_{q+1} \\ \vdots \\ X_p \end{bmatrix} \begin{matrix} \} q \\ \\ \\ \\ \} p-q \end{matrix} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \hline \mathbf{X}^{(2)} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\mu} = E(\mathbf{X}) = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_q \\ \hline \mu_{q+1} \\ \vdots \\ \mu_p \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}^{(1)} \\ \hline \boldsymbol{\mu}^{(2)} \end{bmatrix} \tag{2-38}$$

From the definitions of the transpose and matrix multiplication,

$$(\mathbf{X}^{(1)} - \boldsymbol{\mu}^{(1)})(\mathbf{X}^{(2)} - \boldsymbol{\mu}^{(2)})'$$

$$= \begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \\ \vdots \\ X_q - \mu_q \end{bmatrix} [X_{q+1} - \mu_{q+1}, X_{q+2} - \mu_{q+2}, \ldots, X_p - \mu_p]$$

Upon taking the expectation of the matrix $(\mathbf{X}^{(1)} - \boldsymbol{\mu}^{(1)})(\mathbf{X}^{(2)} - \boldsymbol{\mu}^{(2)})'$, we get

$$E(\mathbf{X}^{(1)} - \boldsymbol{\mu}^{(1)})(\mathbf{X}^{(2)} - \boldsymbol{\mu}^{(2)})' = \begin{bmatrix} \sigma_{1,q+1} & \sigma_{1,q+2} & \cdots & \sigma_{1p} \\ \sigma_{2,q+1} & \sigma_{2,q+2} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{q,q+1} & \sigma_{q,q+2} & \cdots & \sigma_{qp} \end{bmatrix} = \boldsymbol{\Sigma}_{12} \quad (2\text{-}39)$$

which gives all the covariances, σ_{ij}, $i = 1, 2, \ldots, q$, $j = q + 1, q + 2, \ldots, p$, between a component of $\mathbf{X}^{(1)}$ and a component of $\mathbf{X}^{(2)}$. Note that the matrix $\boldsymbol{\Sigma}_{12}$ is not necessarily symmetric or even square.

Making use of the partitioning in Equation (2–38), we can easily demonstrate that

$$(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})' = \begin{bmatrix} (\underset{(q\times 1)}{\mathbf{X}^{(1)}} - \underset{(q\times 1)}{\boldsymbol{\mu}^{(1)}})(\underset{(1\times q)}{\mathbf{X}^{(1)}} - \underset{(1\times q)}{\boldsymbol{\mu}^{(1)}})' & (\underset{(q\times 1)}{\mathbf{X}^{(1)}} - \underset{(q\times 1)}{\boldsymbol{\mu}^{(1)}})(\underset{(1\times (p-q))}{\mathbf{X}^{(2)}} - \underset{(1\times (p-q))}{\boldsymbol{\mu}^{(2)}})' \\ (\underset{((p-q)\times 1)}{\mathbf{X}^{(2)}} - \underset{((p-q)\times 1)}{\boldsymbol{\mu}^{(2)}})(\underset{(1\times q)}{\mathbf{X}^{(1)}} - \underset{(1\times q)}{\boldsymbol{\mu}^{(1)}})' & (\underset{((p-q)\times 1)}{\mathbf{X}^{(2)}} - \underset{((p-q)\times 1)}{\boldsymbol{\mu}^{(2)}})(\underset{(1\times (p-q))}{\mathbf{X}^{(2)}} - \underset{(1\times (p-q))}{\boldsymbol{\mu}^{(2)}})' \end{bmatrix}$$

and consequently,

$$\underset{(p\times p)}{\boldsymbol{\Sigma}} = E(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})' = \begin{array}{c} \\ q \\ p-q \end{array} \begin{array}{c} \begin{array}{cc} q & p-q \end{array} \\ \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array}\right] \\ (p\times p) \end{array}$$

$$= \left[\begin{array}{ccc|ccc} \sigma_{11} & \cdots & \sigma_{1q} & \sigma_{1,q+1} & \cdots & \sigma_{1p} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{q1} & \cdots & \sigma_{qq} & \sigma_{q,q+1} & \cdots & \sigma_{qp} \\ \hline \sigma_{q+1,1} & \cdots & \sigma_{q+1,q} & \sigma_{q+1,q+1} & \cdots & \sigma_{q+1,p} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \cdots & \sigma_{pq} & \sigma_{p,q+1} & \cdots & \sigma_{pp} \end{array}\right] \quad (2\text{-}40)$$

Note that $\Sigma_{12} = \Sigma_{21}'$. The covariance matrix of $\mathbf{X}^{(1)}$ is Σ_{11}, that of $\mathbf{X}^{(2)}$ is Σ_{22}, and that of elements from $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ is Σ_{12} (or Σ_{21}).

It is sometimes convenient to use the Cov $(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$ notation where

$$\text{Cov}\,(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}) = \Sigma_{12}$$

is a matrix containing all of the covariances between a component of $\mathbf{X}^{(1)}$ and a component of $\mathbf{X}^{(2)}$.

Partitioning a random p-vector into two subvectors: $\mathbf{Y} = [\mathbf{Y}_1', \mathbf{Y}_2']'$, the covariance matrix of the partitioned vector is

$$\operatorname{cov}(\mathbf{Y}) = \begin{bmatrix} \operatorname{cov}(\mathbf{Y}_1, \mathbf{Y}_1) & \operatorname{cov}(\mathbf{Y}_1, \mathbf{Y}_2) \\ \operatorname{cov}(\mathbf{Y}_2, \mathbf{Y}_1) & \operatorname{cov}(\mathbf{Y}_2, \mathbf{Y}_2) \end{bmatrix} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

where $\Sigma_{ij} = \operatorname{cov}(\mathbf{Y}_i, \mathbf{Y}_j)$. To evaluate whether $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are uncorrelated, the following theorem is used.

Theorem 3.2.3 *The random vectors $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are uncorrelated if and only if $\Sigma_{12} = \mathbf{0}$.*

The individual components of $\mathbf{Y}_i$ are uncorrelated if and only if Σ_{ii} is a diagonal matrix. If $\mathbf{Y}_i$ has cumulative distribution function $(c.d.f)$, $F_{\mathbf{Y}_i}(\mathbf{y}_i)$, with mean $\boldsymbol{\mu}_i$ and covariance matrix Σ_{ii}, we write $\mathbf{Y}_i \sim (\boldsymbol{\mu}_i, \Sigma_{ii})$.

Definition 3.2.1 *Two (absolutely) continuous random vectors $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are (statistically) independent if the probability density function of $\mathbf{Y} = [\mathbf{Y}_1', \mathbf{Y}_2']'$ is obtained from the product of the marginal densities of $\mathbf{Y}_1$ and $\mathbf{Y}_2$:*

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{Y}_1}(\mathbf{y}_1)\, f_{\mathbf{Y}_2}(\mathbf{y}_2)$$

Lengths, Distances,

Knowledge of vector lengths, distances and angles between vectors helps one to understand relationships among multivariate vector observations. However, prior to discussing these concepts, the inner (scalar or dot) product of two vectors needs to be defined.

Definition 2.3.2 *The inner product of two vectors* $\mathbf{x}$ *and* $\mathbf{y}$, *each with n elements, is the scalar quantity*

$$\mathbf{x}'\mathbf{y} = \sum_{i=1}^{n} x_i y_i$$

In textbooks on linear algebra, the inner product may be represented as $(\mathbf{x}, \mathbf{y})$ or $\mathbf{x} \cdot \mathbf{y}$. Given Definition 2.3.2, inner products have several properties

If $\mathbf{x} = \mathbf{y}$ in Definition 2.3.2, then $\mathbf{x}'\mathbf{x} = \sum_{i=1}^{n} x_i^2$. The quantity $(\mathbf{x}'\mathbf{x})^{1/2}$ is called the Euclidean vector norm or length of $\mathbf{x}$ and is represented as $\|\mathbf{x}\|$. Thus, the norm of $\mathbf{x}$ is the positive square root of the inner product of a vector with itself. The norm squared of $\mathbf{x}$ is represented as $||\mathbf{x}||^2$. The Euclidean distance or length between two vectors $\mathbf{x}$ and $\mathbf{y}$ in V_n is $\|\mathbf{x} - \mathbf{y}\| = [(\mathbf{x} - \mathbf{y})'(\mathbf{x} - \mathbf{y})]^{1/2}$.

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

where x_{ik} and $x_{jk}, k = 1, \ldots, q$ 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.)

Trace and the Euclidean Matrix Norm

An important operation for square matrices is the trace operator. For a square matrix $\mathbf{A}_{n\times n} = [a_{ij}]$, the trace of $\mathbf{A}$, represented as $\text{tr}(\mathbf{A})$, is the sum of the diagonal elements of $\mathbf{A}$. Hence,

$$\text{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii} \tag{2.4.6}$$

Theorem 2.4.3 *For square matrices* $\mathbf{A}$ *and* $\mathbf{B}$ *and scalars* α *and* β*, the following properties hold for the trace of a matrix.*

1. $\text{tr}(\alpha\mathbf{A}+\beta\mathbf{B}) = \alpha\,\text{tr}(\mathbf{A}) + \beta\,\text{tr}(\mathbf{B})$
2. $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$
3. $\text{tr}(\mathbf{A}') = \text{tr}(\mathbf{A})$
4. $\text{tr}(\mathbf{A}'\mathbf{A}) = \text{tr}(\mathbf{A}\mathbf{A}') = \sum_{i,j} a_{ij}^2$ *and equals* 0*, if and only if* $\mathbf{A} = \mathbf{0}$.

Property (4) is an important property for matrices since it generalizes the Euclidean vector norm squared to matrices. The Euclidean norm squared of $\mathbf{A}$ is defined as

$$\|\mathbf{A}\|^2 = \sum_i \sum_j a_{ij}^2 = \text{tr}(\mathbf{A}'\mathbf{A}) = \text{tr}(\mathbf{A}\mathbf{A}')$$

The Euclidean matrix norm is defined as

$$\|\mathbf{A}\| = \left\{\text{tr}\left(\mathbf{A}'\mathbf{A}\right)\right\}^{1/2} = \left\{\text{tr}\left(\mathbf{A}\mathbf{A}'\right)\right\}^{1/2} \tag{2.4.7}$$

and is zero only if $\mathbf{A} = \mathbf{0}$.

Solving the characteristic equation $\left|\mathbf{A}'\mathbf{A} - \lambda\mathbf{I}\right| = 0$, the Euclidean norm becomes $\|\mathbf{A}\|_2 = \left\{\sum_i \lambda_i\right\}^{1/2}$ where λ_i is a root of $\mathbf{A}'\mathbf{A}$. The spectral norm is the square root of the maximum root of $\mathbf{A}'\mathbf{A}$. Thus, $\|\mathbf{A}\|_s = \max \sqrt{\lambda_i}$. Extending the Minkowski vector norm to a matrix, a general matrix (L_p norm) norm is $\|\mathbf{A}\|_p = \left\{\sum_i \lambda_i^{p/2}\right\}^{1/p}$ where λ_i are the roots of $\mathbf{A}'\mathbf{A}$, also called the von Neumann norm. For $p = 2$. it reduces to the Euclidean norm.

While Euclidean distances and norms are useful concepts in statistics since they help to visualize statistical sums of squares, non-Euclidean distance and non-Euclidean norms are often useful in multivariate analysis. We have seen that the Euclidean norm generalizes to a more general function that maps a vector to a scalar. In a similar manner, we may generalize the concept of distance. A non-Euclidean distance important in multivariate analysis is the statistical or Mahalanobis distance.

To motivate the definition, consider a normal random variable X with mean zero and variance one, $X \sim N(0, 1)$. An observation $\mathbf{x}_o$ that is two standard deviations from the mean lies a distance of two units from the origin since the $\|\mathbf{x}_o\| = (0^2 + 2^2)^{1/2} = 2$ and the probability that $0 \leq x \leq 2$ is 0.4772. Alternatively, suppose $Y \sim N(0, 4)$ where the distance from the origin for $\mathbf{y}_o = \mathbf{x}_o$ is still 2. However, the probability that $0 \leq y \leq 2$ becomes 0.3413 so that y is closer to the origin than x. To compare the distances, we must take into account the variance of the random variables. Thus, the squared distance between x_i and x_j is defined as

$$D_{ij}^2 = (x_i - x_j)^2/\sigma^2 = (x_i - x_j)(\sigma^2)^{-1}(x_i - x_j) \tag{2.3.5}$$

where σ^2 is the population variance. For our example, the point $\mathbf{x}_o$ has a squared statistical distance $D_{ij}^2 = 4$ while the point $\mathbf{y}_o = 2$ has a value of $D_{ij}^2 = 1$ which maintains the inequality in probabilities in that Y is "closer" to zero statistically than X. D_{ij} is the distance between x_i and x_j, in the metric of σ^2 called the Mahalanobis distance between x_i and x_j. When $\sigma^2 = 1$, Mahalanobis' distance reduces to the Euclidean distance.

we defined the Mahalanobis distance for a random variable. It was an "adjusted" Euclidean distance which represented statistical closeness in the metric of $1/\sigma^2$ or $\left(\sigma^2\right)^{-1}$. With the first two moments of a random vector defined, suppose we want to calculate the distance between $\mathbf{Y}$ and $\boldsymbol{\mu}$. Generalizing (2.3.5), the Mahalanobis distance between $\mathbf{Y}$ and $\boldsymbol{\mu}$ in the metric of Σ is

$$D_{\Sigma}(\mathbf{Y}, \boldsymbol{\mu}) = [(\mathbf{Y} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{Y} - \boldsymbol{\mu})]^{1/2} \tag{3.2.5}$$

If $\mathbf{Y} \sim (\boldsymbol{\mu}_1, \Sigma)$ and $\mathbf{X} \sim (\boldsymbol{\mu}_2, \Sigma)$, then the Mahalanobis distance between $\mathbf{Y}$ and $\mathbf{X}$, in the metric of Σ, is the square root of

$$D_{\Sigma}^2(\mathbf{X}, \mathbf{Y}) = (\mathbf{X} - \mathbf{Y})' \Sigma^{-1} (\mathbf{X} - \mathbf{Y})$$

which is invariant under linear transformations $\mathbf{z}_{\mathbf{X}} = \mathbf{A}\mathbf{X} + \mathbf{a}$ and $\mathbf{z}_{\mathbf{Y}} = \mathbf{A}\mathbf{Y} + \mathbf{b}$. The covariance matrix Σ of $\mathbf{X}$ and $\mathbf{Y}$ becomes $\Omega = \mathbf{A}\Sigma\mathbf{A}'$ under the transformations so that $D_{\Sigma}^2(\mathbf{X}, \mathbf{Y}) = \mathbf{z}_{\mathbf{X}}' \Omega \mathbf{z}_{\mathbf{Y}} = D_{\Omega}^2(\mathbf{z}_{\mathbf{X}}, \mathbf{z}_{\mathbf{Y}})$.

The Mahalanobis distances, D, arise in a natural manner when investigating the separation between two or more multivariate populations, the topic of discriminant analysis

Having defined the mean and covariance matrix for a random vector $\mathbf{Y}_{p\times 1}$ and the first two moments of a random vector, we extend the classical measures of skewness and kurtosis, $E[(Y-\mu)^3]/\sigma^3 = \mu_3'/\sigma^3$ and $E[(Y-\mu)^4]/\sigma^4 = \mu_4'/\sigma^4$ of a univariate variable Y, respectively, to the multivariate case. Following Mardia (1970), multivariate skewness and kurtosis measures for a random p-variate vector $\mathbf{Y}_p \sim (\boldsymbol{\mu}, \Sigma)$ are, respectively, defined as

$$\beta_{1,\,p} = E\left\{(\mathbf{Y}-\boldsymbol{\mu})'\,\Sigma^{-1}\,(\mathbf{X}-\boldsymbol{\mu})\right\}^3 \tag{3.2.6}$$

$$\beta_{2,\,p} = E\left\{(\mathbf{Y}-\boldsymbol{\mu})'\,\Sigma^{-1}\,(\mathbf{Y}-\boldsymbol{\mu})\right\}^2 \tag{3.2.7}$$

where $\mathbf{Y}_p$ and $\mathbf{X}_p$ are identically and independent identically distributed (*i.i.d.*). Because $\beta_{1,p}$ and $\beta_{2,p}$ have the same form as Mahalanobis' distance, they are also seen to be invariant under linear transformations.

The multivariate measures of skewness and kurtosis are natural generalizations of the univariate measures

$$\sqrt{\beta_1} = \sqrt{\beta_{1,1}} = \mu_3'/\sigma^3 \tag{3.2.8}$$

and

$$\beta_2 = \beta_{2,\,1} = \mu_4'/\sigma^4 \tag{3.2.9}$$

For a univariate normal random variable, $\gamma_1 = \sqrt{\beta_1} = 0$ and $\gamma_2 = \beta_2 - 3 = 0$.

The Multivariate Normal (MVN) Distribution

This is the joint density function of an independent multivariate normal distribution, written as $\mathbf{Y} \sim N_p(\boldsymbol{\mu}, \sigma^2\mathbf{I})$, where the mean vector and covariance matrix are

$$E(\mathbf{Y}) = \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}, \quad \text{and} \quad \text{cov}(\mathbf{Y}) = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2\mathbf{I}_p,$$

respectively.

More generally, replacing $\sigma^2\mathbf{I}_p$ with a positive definite covariance matrix Σ, a generalization of the independent multivariate normal density to the multivariate normal (MVN) distribution is established

$$f(\mathbf{y}) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\left\{-(\mathbf{y}-\boldsymbol{\mu})' \Sigma^{-1} (\mathbf{y}-\boldsymbol{\mu})/2\right\} \qquad -\infty < y_i < \infty \tag{3.3.2}$$

Theorem 3.3.1 *A random p-vector* $\mathbf{Y}$ *is said to have a p-variate normal or multivariate normal (MVN) distribution with mean* $\boldsymbol{\mu}$ *and p.d. covariance matrix* Σ *written* $\mathbf{Y} \sim N_p(\boldsymbol{\mu}, \Sigma)$, *if it has the joint density function given in (3.3.2). If* Σ *is not p.d., the density function of* $\mathbf{Y}$ *does not exist and* $\mathbf{Y}$ *is said to have a singular multivariate normal distribution.*

Observe that the joint density of the MVN distribution is constant whenever the quadratic form in the exponent is constant. The constant density ellipsoid $(\mathbf{Y}-\boldsymbol{\mu})' \Sigma^{-1} (\mathbf{Y}-\boldsymbol{\mu}) = c$ has center at $\boldsymbol{\mu}$ while Σ determines its shape and orientation. In the bivariate case,

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}, \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

For the MVN to be nonsingular, we need $\sigma_1^2 > 0, \sigma_2^2 > 0$ and the $|\Sigma| = \sigma_1^2\sigma_2^2(1-\rho^2) > 0$ so that $-1 < \rho < 1$. Then

$$\Sigma^{-1} = \frac{1}{1-\rho^2} \begin{bmatrix} \frac{1}{\sigma_1^2} & \frac{-\rho}{\sigma_1\sigma_2} \\ \frac{-\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2} \end{bmatrix}$$

and the joint probability density of $\mathbf{Y}$ yields the bivariate normal density

$$f(\mathbf{y}) = \frac{\exp\left\{\frac{-1}{2(1-\rho^2)}\left[\left(\frac{y_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{y_1-\mu_1}{\sigma_1}\right)\left(\frac{y_2-\mu_2}{\sigma_2}\right) + \left(\frac{y_2-\mu_2}{\sigma_2}\right)^2\right]\right\}}{2\pi\sigma_1\sigma_2\left(1-\rho^2\right)^{1/2}}$$

Letting $Z_i = \left(Y_i - \mu_i\right)/\sigma_i (i = 1, 2)$, the joint bivariate normal becomes the standard bivariate normal

$$f(\mathbf{z}) = \frac{\exp\frac{-1}{2(1-\rho^2)}\left(z_1^2 - 2\rho z_1 z_2 + z_2^2\right)}{2\pi\left(1-\rho^2\right)^{1/2}} \qquad -\infty < z_i < \infty$$

The exponent in the standard bivariate normal distribution is a quadratic form

$$Q = [z_1, z_2]' \, \Sigma^{-1} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \frac{z_1^2 - 2\rho z_1 z_2 + z_2^2}{1-\rho^2} > 0$$

where

$$\Sigma^{-1} = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

which generates concentric ellipses about the origin. Setting $\rho = 1/2$, the ellipses have the form $Q = z_1^2 - z_1 z_2 + z_2^2$ for $Q > 0$. Graphing this function in the plane with axes z_1 and z_2 for $Q = 1$ yields the constant density ellipse with semi-major axis a and semi-minor axis b, Figure 3.3.1.

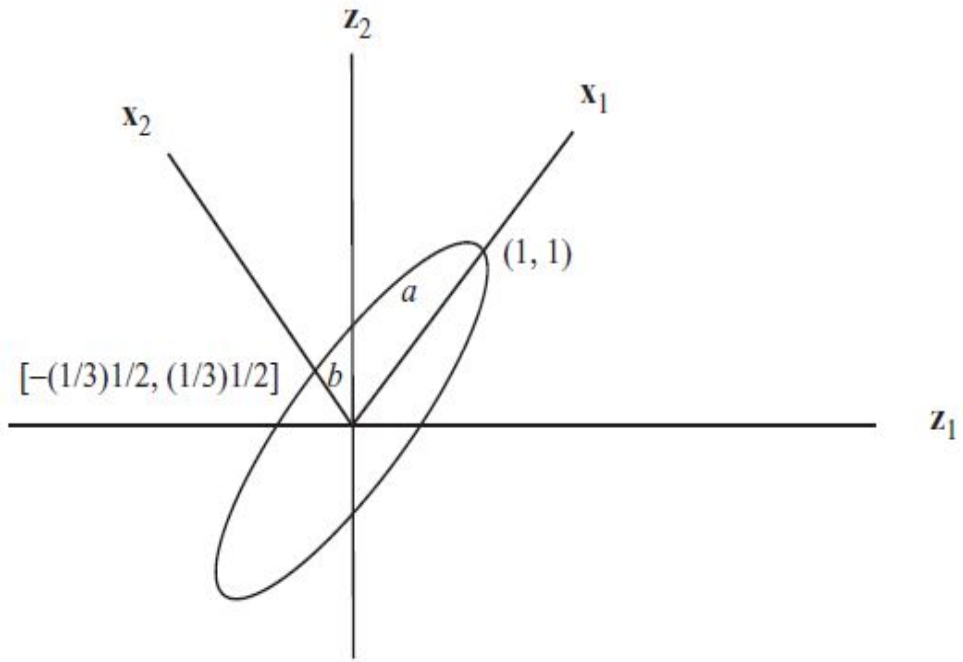


FIGURE 3.3.1. $\mathbf{z}'\Sigma^{-1}\mathbf{z} = z_1^2 - z_1 z_2 + z_2^2 = 1$

Performing an orthogonal rotation of $\mathbf{x} = \mathbf{P}'\mathbf{z}$, the quadratic form for the exponent of the standard MVN becomes

$$\mathbf{z}'\Sigma^{-1}\mathbf{z} = \lambda_1^* x_1^2 + \lambda_2^* x_2^2 = \frac{1}{\lambda_2} x_1^2 + \frac{1}{\lambda_1} x_2^2$$

where $\lambda_1 = 1 + \rho = 3/2$ and $\lambda_2 = 1 - \rho = 1/2$ are the roots of $|\Sigma - \lambda\mathbf{I}| = 0$ and $\lambda_2^* = 1/\lambda_1$ and $\lambda_1^* = 1/\lambda_2$ are the roots of $|\Sigma^{-1} - \lambda^*\mathbf{I} = 0$. From analytic geometry, the equation of an ellipse, for $Q = 1$, is given by

$$\left(\frac{1}{b}\right)^2 x_1^2 + \left(\frac{1}{a}\right)^2 x_2^2 = 1$$

Hence $a^2 = \lambda_1$ and $b^2 = \lambda_2$ so that each half-axis is proportional to the inverse of the squared lengths of the eigenvalues of Σ. As Q varies, concentric ellipsoids are generated so that $a = \sqrt{Q\lambda_1}$ and $b = \sqrt{Q\lambda_2}$.

Theorem 3.3.2 *Properties of normally distributed random variables.*

1. *Linear combinations of the elements of* $\mathbf{Y} \sim \mathbf{N}[\boldsymbol{\mu}, \boldsymbol{\Sigma}]$ *are normally distributed. For a constant vector* $\mathbf{a} \neq \mathbf{0}$ *and* $\mathbf{X} = \mathbf{a}'\mathbf{Y}$, *then* $\mathbf{X} \sim N_1(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a})$.

2. *The normal distribution of* $\mathbf{Y}_p \sim N_p[\boldsymbol{\mu}, \boldsymbol{\Sigma}]$ *is invariant to linear transformations. For a constant matrix* $\mathbf{A}_{q\times p}$ *and vector* $\mathbf{b}_{q\times 1}$, $\mathbf{X} = \mathbf{A}\mathbf{Y}_p + \mathbf{b} \sim N_q(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$.

3. *Partitioning* $\mathbf{Y} = [\mathbf{Y}_1', \mathbf{Y}_2']'$, $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}$ and $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$, *the subvectors of* $\mathbf{Y}$ *are normally distributed.* $\mathbf{Y}_1 \sim N_{p_1}[\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}]$ *and* $\mathbf{Y}_2 \sim N_{p_2}[\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}]$ *where* $p_1 + p_2 = p$. *More generally, all marginal distributions for any subset of random variables are normally distributed. However, the converse is not true, marginal normality does not imply multivariate normality.*

4. *The random subvectors* $\mathbf{Y}_1$ *and* $\mathbf{Y}_2$ *of* $\mathbf{Y} = [\mathbf{Y}_1', \mathbf{Y}_2']'$ *are independent if and only if* $\Sigma = \text{diag}[\Sigma_{11}, \Sigma_{22}]$. *Thus, uncorrelated normal subvectors are independent under multivariate normality.*

5. *The conditional distribution of* $\mathbf{Y}_1 \mid \mathbf{Y}_2$ *is normally distributed,*

$$\mathbf{Y}_1 \mid \mathbf{Y}_2 \sim N_{p_1}\left[\boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_1), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right]$$

Writing the mean of the conditional normal distribution as

$$\begin{aligned}\boldsymbol{\mu} &= (\boldsymbol{\mu}_1 - \Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{\mu}_2) + \Sigma_{12}\Sigma_{22}^{-1}\mathbf{y}_2 \\ &= \boldsymbol{\mu}_0 + \mathbf{B}_1'\mathbf{y}_2\end{aligned}$$

$\boldsymbol{\mu}$ *is called the regression function of* $\mathbf{Y}_1$ *on* $\mathbf{Y}_2 = \mathbf{y}_2$ *with regression coefficients* $\mathbf{B}_1'$. *The matrix* $\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ *is called the partial covariance matrix with elements* $\sigma_{ij.\,p_1+1,\ldots,\,p_1+p_2}$. *A similar result holds for* $\mathbf{Y}_2 \mid \mathbf{Y}_1$.

6. *Letting* $\mathbf{Y}_1 = Y$, *a single random variable and letting the random vector* $\mathbf{Y}_2 = \mathbf{X}$, *a random vector of independent variables, the population coefficient of determination or population squared multiple correlation coefficient is defined as the maximum correlation between* $\mathbf{Y}$ *and linear functions* $\boldsymbol{\beta}'\mathbf{X}$. *The population coefficient of determination or the squared population multiple correlation coefficient is*

$$\rho^2_{Y\mathbf{X}} = \boldsymbol{\sigma}'_{Y\mathbf{X}}\Sigma^{-1}_{\mathbf{XX}}\boldsymbol{\sigma}_{\mathbf{X}Y}/\sigma_{YY}$$

If the random vector $\mathbf{Z} = (Y, \mathbf{X}')'$ *follows a multivariate normal distribution, the population coefficient of determination is the square of the zero-order correlation between the random variable* Y *and the population predicted value of* Y *which we see from (5) has the form* $\widehat{Y} = \mu_Y + \boldsymbol{\sigma}'_{Y\mathbf{X}}\Sigma^{-1}_{\mathbf{XX}}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}})$.

7. *For* $\mathbf{X} = \Sigma^{-1/2}(\mathbf{Y} - \boldsymbol{\mu})$ *where* $\Sigma^{-1/2}$*is the symmetric positive definite square root of* Σ^{-1} *then* $\mathbf{X} \sim N_p(\mathbf{0}, \mathbf{I})$ or $X_i \sim IN(0, 1)$.

8. *If* $\mathbf{Y}_1$ and $\mathbf{Y}_2$ *are independent multivariate normal random vectors, then the sum* $\mathbf{Y}_1 + \mathbf{Y}_2 \sim N(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2, \Sigma_{11} + \Sigma_{22})$. *More generally, if* $\mathbf{Y}_i \sim IN_p(\boldsymbol{\mu}_i, \Sigma_i)$ and $a_1, a_2, \ldots, a_n$ *are fixed constants, then the sum of n p-variate vectors*

$$\sum_{i=1}^{n} a_i \mathbf{Y}_i \sim N_p\left[\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \Sigma_i\right]$$

Theorem 3.3.3 *If* $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$ *are independent MVN random vectors with common mean* $\boldsymbol{\mu}$ *and covariance matrix* Σ, *then* $\overline{\mathbf{Y}} = \sum_{i=1}^{n} \mathbf{Y}_i / n$ *is MVN with mean* $\boldsymbol{\mu}$ *and covariance matrix* Σ/n, $\overline{\mathbf{Y}} \sim N_p(\boldsymbol{\mu}, \Sigma/n)$.

Estimating $\boldsymbol{\mu}$ *and* Σ

From Theorem 3.3.3, observe that for a random sample from a normal population that $\overline{\mathbf{Y}}$ is an unbiased and consistent estimator of $\boldsymbol{\mu}$, written as $\widehat{\boldsymbol{\mu}} = \overline{\mathbf{Y}}$. Having estimated $\boldsymbol{\mu}$, the $p \times p$ sample covariance matrix is

$$\begin{aligned}
\mathbf{S} &= \sum_{i=1}^{n} (\mathbf{y}_i - \overline{\mathbf{y}})(\mathbf{y}_i - \overline{\mathbf{y}})' / (n-1) \\
&= \sum_{i=1}^{n} [(\mathbf{y}_i - \boldsymbol{\mu}) - (\overline{\mathbf{y}} - \boldsymbol{\mu})][(\mathbf{y}_i - \boldsymbol{\mu}) - (\overline{\mathbf{y}} - \boldsymbol{\mu})]' / (n-1) \\
&= \left[\sum_{i=1}^{n} (\mathbf{y}_i - \boldsymbol{\mu})(\mathbf{y}_i - \boldsymbol{\mu})' + n(\overline{\mathbf{y}}_i - \boldsymbol{\mu})(\overline{\mathbf{y}}_i - \boldsymbol{\mu})' \right] / (n-1) \qquad (3.3.3)
\end{aligned}$$

where $E(\mathbf{S}) = \Sigma$ so that $\mathbf{S}$ is an unbiased estimator of Σ. Representing the sample as a matrix $\mathbf{Y}_{n \times p}$ so that

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1' \\ \mathbf{y}_2' \\ \vdots \\ \mathbf{y}_n' \end{bmatrix}$$

$\mathbf{S}$ may be written as

$$(n-1)\,\mathbf{S} = \mathbf{Y}'\left[\mathbf{I}_n - \mathbf{1}_n\left(\mathbf{1}_n'\mathbf{1}_n\right)^{-1}\mathbf{1}_n'\right]\mathbf{Y}$$
$$= \mathbf{Y}'\mathbf{Y} - n\overline{\mathbf{y}}\overline{\mathbf{y}}' \tag{3.3.4}$$

where $\mathbf{I}_n$ is the identity matrix and $\mathbf{1}_n$ is a vector of n 1s. While the matrix $\mathbf{S}$ is an unbiased estimator, a biased estimator, called the maximum likelihood estimator under normality is $\widehat{\Sigma} = \frac{(n-1)}{n}\mathbf{S} = \sum_{i=1}^{n}(\mathbf{y}_i - \overline{\mathbf{y}})(\mathbf{y}_i - \overline{\mathbf{y}})'/n = \mathbf{E}/n$. The matrix $\mathbf{E}$ is called the sum of squares and cross-products matrix, SSCP and the $|\mathbf{S}|$ is the sample estimate of the generalized variance.

In Theorem 3.3.3, we assumed that the observations $\mathbf{Y}_i$ represent a sample from a normal distribution. More generally, suppose $\mathbf{Y}_i \sim (\boldsymbol{\mu}, \Sigma)$ is an independent sample from any distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ. Theorem 3.3.4 is a multivariate version of the Central Limit Theorem (CLT).

Theorem 3.3.4 *Let* $\{\mathbf{Y}_i\}_{i=1}^{\infty}$ *be a sequence of random p-vectors with finite mean* $\boldsymbol{\mu}$ *and covariance matrix* Σ. *Then*

$$n^{1/2}(\overline{\mathbf{Y}} - \boldsymbol{\mu}) = n^{-1/2}\sum_{i-1}^{n}(\mathbf{Y}_i - \boldsymbol{\mu}) \xrightarrow{d} N_p(\mathbf{0}, \Sigma)$$

Theorem 3.3.4 is used to show that $\mathbf{S}$ is a consistent estimator of Σ. To obtain the distribution of a random matrix $\mathbf{Y}_{n\times p}$, the vec $(\cdot)$ operator is used. Assuming a random sample of n p-vectors $\mathbf{Y}_i \sim (\boldsymbol{\mu}, \Sigma)$, consider the random matrix $\mathbf{X}_i = (\mathbf{Y}_i - \boldsymbol{\mu})(\mathbf{Y}_i - \boldsymbol{\mu})'$. By Theorem 3.3.4,

$$n^{-1/2} \sum_{i=1}^{n} [\text{vec}(\mathbf{X}_i) - \text{vec}\,(\Sigma)] \xrightarrow{d} N_{p^2}\,(\mathbf{0}, \Omega)$$

where

$$\Omega = \text{cov}[\text{vec}(\mathbf{X}_i)]$$

and

$$n^{-1/2}\,(\bar{\mathbf{y}} - \boldsymbol{\mu}) \xrightarrow{d} N_p\,(\mathbf{0}, \Sigma)$$

so that

$$n^{-1/2}[\text{vec}(\mathbf{E}) - n\ \text{vec}(\Sigma)] \xrightarrow{d} N_{p^2}(\mathbf{0}, \Omega).$$

Because $\mathbf{S} = (n-1)^{-1}\mathbf{E}$ and the replacement of n by $n-1$ does not effect the limiting distribution, we have the following theorem.

Theorem 3.3.5 *Let $\{\mathbf{Y}_i\}_{i=1}^{\infty}$ be a sequence of independent and identically distributed $p \times 1$ vectors with finite fourth moments and mean $\boldsymbol{\mu}$ and covariance matrix Σ. Then*

$$(n-1)^{-1/2}\,\text{vec}(\mathbf{S} - \Sigma) \xrightarrow{d} N_{p^2}(\mathbf{0}, \Omega)$$

a. Chi-Square Distribution

Recall that if $Y_1, Y_2, \ldots, Y_n$ are independent normal random variables with mean $\mu_i = 0$ and variance $\sigma^2 = 1$, $Y_i \sim IN(0, 1)$, or, employing vector notation $\mathbf{Y} \sim N_n(\mathbf{0}, \mathbf{I})$, then

$$Q = \mathbf{Y}'\mathbf{Y} = \sum_{i=1}^{n} Y_i^2 \sim \chi^2(n) \qquad 0 < Q < \infty$$

$Q = \mathbf{Y}'\mathbf{Y}$ has a central χ^2 distribution with n degrees of freedom. Letting $Y_i \sim IN(\mu_i, \sigma^2)$, results in the noncentral chi-square distribution.

Definition 3.4.1 *If the random n-vector* $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \sigma^2\mathbf{I})$, *then* $\mathbf{Y}'\mathbf{Y}/\sigma^2$ *has a noncentral* χ^2 *distribution with n degrees of freedom and noncentrality parameter* $\gamma = \boldsymbol{\mu}'\boldsymbol{\mu}/\sigma^2$.

For $\boldsymbol{\mu} = \mathbf{0}$, the noncentral chi-square distribution reduces to a central chi-square distribution. For $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \mathbf{I})$, then $\mathbf{Y}'\mathbf{Y} \sim \chi^2(n, \gamma)$ with $\gamma = \boldsymbol{\mu}'\boldsymbol{\mu}$ so that $\gamma = \|\boldsymbol{\mu}\|^2$ is a norm squared. The further $\boldsymbol{\mu}$ is from zero, the larger the noncentrality parameter γ or the norm squared of $\boldsymbol{\mu}$. Because $\mathbf{Y}'\mathbf{Y}$ in Definition 3.4.1 is a special case of the quadratic form $\mathbf{Y}'\mathbf{A}\mathbf{Y}$, with $\mathbf{A} = \mathbf{I}$ and since $\mathbf{I}^2 = \mathbf{I}$, we have the following more general result.

Theorem 3.4.1 *Let* $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \sigma^2\mathbf{I})$ *and* $\mathbf{A}$ *be a symmetric matrix of rank* r*. Then we have* $\mathbf{Y}'\mathbf{A}\mathbf{Y}/\sigma^2 \sim \chi^2(r, \gamma)$*, where* $\gamma = \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}/\sigma^2$*, if and only if* $\mathbf{A} = \mathbf{A}^2$*.*

Example 3.4.1 *As an example of Theorem 3.4.1, suppose* $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \sigma^2\mathbf{I}_n)$*. Then*

$$\frac{(n-1)s^2}{\sigma^2} = \frac{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}{\sigma^2} = \frac{\mathbf{Y}'[\mathbf{I} - \mathbf{1}\left(\mathbf{1}'\mathbf{1}\right)^{-1}\mathbf{1}]\mathbf{Y}}{\sigma^2} = \frac{\mathbf{Y}'\mathbf{A}\mathbf{Y}}{\sigma^2}$$

However, $\mathbf{A}' = \mathbf{A}$ *and* $\mathbf{A}^2 = \mathbf{A}$ *since* $\mathbf{A}$ *is a projection matrix and the* $r(\mathbf{A}) = \mathrm{tr}(\mathbf{A}) = n-1$*. Hence*

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1,\ \gamma = 0)$$

since $\gamma = E\left(\mathbf{Y}'\right)\mathbf{A}E(\mathbf{Y})/\sigma^2 = 0$*. Thus,* $(n-1)s^2 \sim \sigma^2\chi^2(n-1)$*.*

Theorem 3.4.2 generalizes Theorem 3.4.1 to a vector of dependent variables in a natural manner by setting $\mathbf{Y} = \mathbf{F}\mathbf{X}$ and $\mathbf{F}\mathbf{F}' = \Sigma$.

Theorem 3.4.2 *If* $\mathbf{Y} \sim N_p(\boldsymbol{\mu}, \Sigma)$*. Then the quadratic form* $\mathbf{Y}'\mathbf{A}\mathbf{Y} \sim \chi^2(r, \gamma)$*, where* $\gamma = \left(\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}\right)$ *and the* $r(\mathbf{A}) = r$*, if and only if* $\mathbf{A}\Sigma\mathbf{A} = \mathbf{A}$ *or* $\mathbf{A}\Sigma$ *is idempotent.*

Example 3.4.2 *An important application of Theorem 3.4.2 follows:*

Let $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$ *be* n *independent* p*-vectors from any distribution with mean* $\boldsymbol{\mu}$ *and nonsingular covariance matrix* Σ*. Then by the CLT,* $\sqrt{n}(\overline{\mathbf{Y}} - \boldsymbol{\mu}) \xrightarrow{d} N_p(\mathbf{0}, \Sigma)$*. By Theorem 3.4.2,* $T^2 = n(\overline{\mathbf{Y}} - \boldsymbol{\mu})' \Sigma^{-1} (\overline{\mathbf{Y}} - \boldsymbol{\mu}) = nD^2 \xrightarrow{d} \chi^2(p)$ *for* $n - p$ *large since* $\Sigma^{-1}\Sigma\Sigma^{-1} = \Sigma$*. The distribution is exactly* $\chi^2(p)$ *if the sample is from a multivariate normal distribution.*

Thus, comparing nD^2 with a χ^2 critical value may be used to evaluate multivariate normality. Furthermore, nD^2 for known Σ may be used to test $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ vs. $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$. The critical value of the test with significance level α is represented as

$$\Pr[nD^2 \geq \chi^2_{1-\alpha}(p) \mid H_0] = \alpha$$

where $\chi^2_{1-\alpha}$ is the upper $1 - \alpha$ chi-square critical value. For $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$, the noncentrality parameter is

$$\gamma = n(\boldsymbol{\mu} - \boldsymbol{\mu}_0)' \Sigma^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)$$

The above result is for a single quadratic form. More generally we have Cochran's Theorem.

Theorem 3.4.3 *If* $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n)$ *and* $\mathbf{Y}'\mathbf{Y}/\sigma^2 = \sum_{i=1}^n \mathbf{Y}'\mathbf{A}_i\mathbf{Y}$ *where* $r(\mathbf{A}_i) = r$ *and* $\sum_{i=1}^n \mathbf{A}_i = \mathbf{I}_n$*, then the quadratic forms* $\mathbf{Y}'\mathbf{A}_i\mathbf{Y}/\sigma^2 \sim \chi^2(r_i, \gamma_i)$*, where* $\gamma_i = \boldsymbol{\mu}'\mathbf{A}_i\boldsymbol{\mu}/\sigma^2$ *are statistically independent for all* i *if and only if* $\sum_{i=1}^n r_i = n$ and $\sum_i r(\mathbf{A}_i) = r\left(\sum_i \mathbf{A}_i\right)$.

Example 2.6.3 *Consider a model that relates one dependent variable* y *to* $x_1, x_2, \ldots, x_k$ *linearly independent variables by the linear relationship*

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$$

where e *is a random error. This model is the multiple linear regression model, which, using matrix notation, may be written as*

$$\underset{n\times 1}{\mathbf{y}} = \underset{n\times(k+1)}{\mathbf{X}} \quad \underset{(k+1)\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\mathbf{e}}$$

Letting $\mathbf{X}$ *represent the space spanned by the columns of* $\mathbf{X}$*, the projection of* $\mathbf{y}$ *onto* $\mathbf{X}$ *is*

$$\widehat{\mathbf{y}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$$

Assuming $\mathbf{e} = \mathbf{0}$*, the system of equations is solved to obtain the best estimate of* $\boldsymbol{\beta}$*. Then, the best estimate of* $\mathbf{y}$ *using the linear model is* $\mathbf{X}\widehat{\boldsymbol{\beta}}$ *where* $\widehat{\boldsymbol{\beta}}$ *is the solution to the system* $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$ *for unknown* $\boldsymbol{\beta}$*. The least squares estimate* $\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$ *minimizes the sum of squared errors for the fitted model* $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$*. Furthermore,*

$$\begin{aligned}\|(\mathbf{y} - \widehat{\mathbf{y}})\|^2 &= (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \\ &= \mathbf{y}'(\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}')\mathbf{y} \\ &= \left\|\mathbf{P}_{V^\perp}\mathbf{y}\right\|^2\end{aligned}$$

is the squared distance of the projection of $\mathbf{y}$ *onto the orthocomplement of* $V_r \subseteq V_n$*.*

Example 3.4.3 *Let* $Y \sim N_4(\mu, \sigma^2 I)$

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} = [\mathbf{A}_1\ \mathbf{A}_2] \quad \textit{and} \quad \mathbf{y} = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \end{bmatrix}$$

where

$$\mathbf{A}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad \textit{and} \quad \mathbf{A}_2 \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

In Example 2.6.2, projection matrices of the form

$$\mathbf{P}_1 = \mathbf{A}_1 \left(\mathbf{A}_1'\mathbf{A}_1\right)^- \mathbf{A}_1'$$
$$\mathbf{P}_2 = \mathbf{A} \left(\mathbf{A}'\mathbf{A}\right)^- \mathbf{A}' - \mathbf{A}_1 \left(\mathbf{A}_1'\mathbf{A}_1\right)^- \mathbf{A}_1'$$
$$\mathbf{P}_3 = \mathbf{I} - \mathbf{A} \left(\mathbf{A}'\mathbf{A}\right)^- \mathbf{A}'$$

were constructed to project the observation vector $\mathbf{y}$ *onto orthogonal subspaces. The projection matrices were constructed such that* $\mathbf{I} = \mathbf{P}_1 + \mathbf{P}_2 + \mathbf{P}_3$ *where* $\mathbf{P}_i\mathbf{P}_j = \mathbf{0}$ *for* $i \neq j$ *and each* $\mathbf{P}_i$ *is symmetric and idempotent so that the* $r(\mathbf{I}) = \sum_i r(\mathbf{P}_i)$*. Forming an equation of quadratic forms, we have that*

$$\mathbf{y}'\mathbf{y} = \sum_{i=1}^{3} \mathbf{y}'\mathbf{P}_i\mathbf{y}$$

or

$$\|\mathbf{y}\|^2 = \sum_{i=1}^{3} \|\mathbf{P}_i\mathbf{y}\|^2$$

For $\mathbf{P}_1$, $\mathbf{P}_2$, *and* $\mathbf{P}_3$ *given in Example 2.6.3, it is easily verified that*

$$\|\mathbf{P}_1\mathbf{y}\|^2 = \mathbf{y}'\mathbf{P}_1\mathbf{y} = 4^2 y_{\cdot\cdot}$$

$$\|\mathbf{P}_2\mathbf{y}\|^2 = \mathbf{y}'\mathbf{P}_2\mathbf{y} = \sum_i 2\,(y_{i\cdot} - y_{\cdot\cdot})^2$$

$$\|\mathbf{P}_3\mathbf{y}\|^2 = \mathbf{y}'\mathbf{P}_3\mathbf{y} = \sum_i \sum_j 2\,(y_{ij} - y_{i\cdot})^2$$

Hence, the total sum of squares has the form

$$\mathbf{y}'\mathbf{I}\mathbf{y} = \sum_i \sum_j y_{ij}^2 = 4y_{\cdot\cdot}^2 + \sum_i 2(y_{i\cdot} - y_{\cdot\cdot})^2 + \sum_i \sum_j (y_{ij} - y_{i\cdot})^2$$

or

$$\sum_i \sum_j (y_{ij} - y_{..})^2 = \sum_i 2(y_{i.} - y_{..})^2 + \sum_i \sum_j (y_{ij} - y_{i.})^2$$

"Total about the Mean" SS = Between SS + Within SS

where the degrees of freedom are the ranks of $r(\mathbf{I} - \mathbf{P}_1) = n - 1$, $r(\mathbf{P}_2) = I - 1$, *and* $r(\mathbf{P}_3) = n - I$ *for* $n = 4$ *and* $I = 2$. *By Theorem 3.4.3, the sum of squares (SS) are independent and may be used to test hypotheses in analysis of variance, by forming ratios of independent chi-square statistics.*

Definition 3.4.2 *If* $\mathbf{Q} = \mathbf{Y}'\mathbf{Y}$ *and the matrix* $\mathbf{Y} \sim N_{n,\,p}(\mathbf{0},\, \Sigma \otimes \mathbf{I}_n)$. *Then* $\mathbf{Q}$ *has a central p-dimensional Wishart distribution with n degrees of freedom and covariance matrix* Σ, *written as* $\mathbf{Q} \sim W_p(n, \Sigma)$.

For $E(\mathbf{Y}) = \mathbf{M}$ and $\mathbf{M} \neq \mathbf{0}$, $\mathbf{Q}$ has a noncentral Wishart distribution with noncentrality parameter $\Gamma = \mathbf{M}'\mathbf{M}\Sigma^{-1}$, written as $\mathbf{Q} \sim W_p(n, \Sigma, \Gamma)$. More formally, $\mathbf{Q} \sim W_p(n, \Sigma, \Gamma = \mathbf{M}'\mathbf{M}\Sigma^{-1})$ if and only if $\mathbf{a}'\mathbf{Q}\,\mathbf{a}/\mathbf{a}'\Sigma\mathbf{a} \sim \chi^2(n, \mathbf{a}'\mathbf{M}'\mathbf{M}\mathbf{a}\,/\mathbf{a}'\mathbf{M}\mathbf{a})$ for all nonnull vectors $\mathbf{a}$. In addition $E(\mathbf{Q}) = n\Sigma + \Sigma\Gamma = n\Sigma + \mathbf{M}'\mathbf{M}$ and $E(\mathbf{Y}'\mathbf{A}\mathbf{Y}) = \mathrm{tr}(\mathbf{A})\Sigma + \mathbf{M}'\mathbf{A}\mathbf{M}$ for a symmetric matrix $\mathbf{A}_{n\times n}$. For a comprehensive treatment of the noncentral Wishart distribution, see Muirhead (1982).

If $\mathbf{Q} \sim W_p(n, \Sigma)$, then the distribution of $\mathbf{Q}^{-1}$ is called an inverted Wishart distribution. That is $\mathbf{Q}^{-1} \sim W_p^{-1}(n + p + 1,\ \Sigma^{-1})$ and

$$E(\mathbf{Q}^{-1}) = \Sigma^{-1}/(n - p - 1)$$

for $n - p - 1 > 0$. Or, if $\mathbf{P} \sim W_p^{-1}(n^*, \mathbf{V}^{-1})$ then $E(\mathbf{P}) = \mathbf{V}^{-1}/(n^* - 2p - 2)$.

The Wishart distribution is a multivariate extension of the chi-square distribution and arises in the derivation of the distribution of the sample covariance matrix $\mathbf{S}$. For a random sample of n p-vectors, $\mathbf{Y}_i \sim N_p(\boldsymbol{\mu}, \Sigma)$ for $i = 1, \ldots, n$ and $n \geq p$,

$$(n - 1)\mathbf{S} = \sum_{i=1}^{n} (\mathbf{Y}_i' - \overline{\mathbf{Y}})(\mathbf{Y}_i - \overline{\mathbf{Y}})' \sim W_p(n - 1,\ \Sigma) \tag{3.4.1}$$

or

$$\mathbf{S} \sim W_p[n - 1,\ \Sigma/(n - 1)]$$

In multivariate analysis, the sum of independent Wishart distributions follows the same rules as in the univariate case. Matrix quadratic forms are often used in multivariate mixed models. Also important in multivariate analysis are the ratios of independently distributed Wishart matrices or, more specifically, the determinant and trace of matrix products or ratios which are functions of the eigenvalues of matrices. To construct distributions of roots of Wishart matrices, independence needs to be established. The multivariate extension for Cochran's Theorem follows.

Theorem 3.4.5 *If* $\mathbf{Y}_i \sim IN_p(\mu, \Sigma)$ *for* $i = 1, \ldots, n$ *and* $\mathbf{Y}'\mathbf{Y} = \sum_{i=1}^{k} \mathbf{Y}'\mathbf{P}_i\mathbf{Y}$ *where* $\sum_{i=1}^{k} \mathbf{P}_i = \mathbf{I}_n$, *the forms* $\mathbf{Y}'\mathbf{P}_i\mathbf{Y} \sim W_p(r_i, \Sigma, \Gamma_i)$ *are statistically independent for all* i *if and only if* $\sum_{i=1}^{k} r_i = n$. *If* $r_i < p$, *the Wishart density does not exist.*

b. Hotelling's T^2 Distribution

A multivariate extension of Student's t distribution is Hotelling's T^2 distribution.

Definition 3.5.3 *Let* $\mathbf{Y}$ *and* $\mathbf{Q}$ *be independent random variables where* $\mathbf{Y} \sim N_p(\boldsymbol{\mu}, \Sigma)$ *and* $\mathbf{Q} \sim W_p(n, \Sigma)$, *and* $n > p$. *Then Hotelling's* T^2 *(1931) statistic*

$$T^2 = n\mathbf{Y}'\mathbf{Q}^{-1}\mathbf{Y}$$

has a distribution proportional to a noncentral F distribution

$$\frac{n-p+1}{p}\frac{T^2}{n} \sim F\,(p, n-p+1, \gamma)$$

where $\gamma = \boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu}$.

Example 3.5.1 *Let* $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$ *be a random sample from a MVN population,* $\mathbf{Y}_i \sim IN_p(\boldsymbol{\mu}, \Sigma)$. *Then* $\overline{\mathbf{Y}} \sim N_p(\boldsymbol{\mu}, \Sigma/n)$ *and* $(n-1)\mathbf{S} \sim W_p(n-1, \Sigma)$, *and* $\overline{\mathbf{Y}}$ *and* $\mathbf{S}$ *are independent. Hence, for testing* $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ *vs.* $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$, $T^2 = n(\overline{\mathbf{Y}} - \mu_0)'\mathbf{S}^{-1}\left(\overline{\mathbf{Y}} - \mu_0\right)$ *or*

$$\frac{n-p}{p}\frac{T^2}{n-1} = \frac{n\,(n-p)}{p\,(n-1)}\left(\overline{\mathbf{Y}} - \boldsymbol{\mu}_0\right)'\mathbf{S}^{-1}\left(\overline{\mathbf{Y}} - \boldsymbol{\mu}_0\right) \sim F\,(p, n-p, \gamma)$$

where

$$\gamma = n(\boldsymbol{\mu} - \boldsymbol{\mu}_0)'\Sigma^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)$$

is the noncentrality parameter. When H_0 *is true, the noncentrality parameter is zero and* T^2 *has a central F distribution.*

Example 3.5.3 *Replacing* **Q** *by* **S** *in Definition 3.5.3, Hotelling's* T^2 *statistic follows an F distribution*

$$(n-p)\,T^2/(n-1)\,p \sim F(p, n-p, \gamma)$$

For $\gamma = 0$,

$$E(T^2) = (n-1)\,p/(n-p-2)$$

$$\text{var}(T^2) = \frac{2p\,(n-1)^2\,(n-2)}{(n-p-2)^2\,(n-p-4)}$$

By Theorem 3.4.2, $T^2 \xrightarrow{d} \chi^2(p)$ *as* $n \longrightarrow \infty$. *However, for small values of n, the distribution of* T^2 *is far from chi-square. If* $X^2 \sim \chi^2(p)$, *then* $E\left(X^2\right) = p$ *and the* $\text{var}\left(X^2\right) = 2p$. *Thus, if one has a statistic* $T^2 \xrightarrow{d} \chi^2(p)$, *a better approximation for small to moderate sample sizes is the statistic*

$$\frac{(n-p)\,T^2}{(n-1)\,p} \sim F(p, n-p, \gamma)$$

c. *The Beta Distribution*

A distribution closely associated with the F distribution is the beta distribution.

Definition 3.5.4 *Let H and E be independent random variables such that* $H \sim \chi^2(v_h, \gamma)$ *and* $E \sim \chi^2(v_e, \gamma = 0)$. *Then*

$$B = \frac{H}{H+E} \sim \text{beta}\,(v_h/2, v_e/2, \gamma)$$

has a noncentral (Type I) beta distribution and

$$V = H/E \sim \textit{Inverted beta}\,(v_h/2, v_e/2, \gamma)$$

has a (Type II) beta or inverted noncentral beta distribution.

To generalize (3.6.3) to the multivariate (linear) regression model, a model is formulated for each of p correlated dependent, response variables

$$\begin{array}{rcccccccc} \mathbf{y}_1 & = & \beta_{01}\mathbf{1}_n & + & \beta_{11}\mathbf{x}_1 & +\cdots+ & \beta_{k1}\mathbf{x}_k & + & \mathbf{e}_1 \\ \mathbf{y}_2 & = & \beta_{02}\mathbf{1}_n & + & \beta_{12}\mathbf{x}_2 & +\cdots+ & \beta_{k2}\mathbf{x}_k & + & \mathbf{e}_2 \\ \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\ \mathbf{y}_p & = & \beta_{0p}\mathbf{1}_n & + & \beta_{1p}\mathbf{x}_p & +\cdots+ & \beta_{kp}\mathbf{x}_k & + & \mathbf{e}_p \end{array} \tag{3.6.15}$$

Each of the vectors $\mathbf{y}_j$, $\mathbf{x}_j$ and $\mathbf{e}_j$, for $j = 1, 2, \ldots, p$ are $n \times 1$ vectors. Hence, we have n observations for each of p variables. To represent (3.6.15) in matrix form, we construct matrices using each variable as a column vector. That is,

$$\begin{aligned} \mathbf{Y}_{n\times p} &= [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_p] \\ \mathbf{X}_{n\times q} &= [\mathbf{1}_n, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k] \\ \mathbf{B}_{q\times p} &= [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_p] \\ &= \begin{bmatrix} \beta_{01} & \beta_{01} & \cdots & \beta_{0p} \\ \beta_{11} & \beta_{11} & \cdots & \beta_{1p} \\ \vdots & \vdots & & \vdots \\ \beta_{k1} & \beta_{k2} & \cdots & \beta_{kp} \end{bmatrix} \\ \mathbf{E}_{n\times p} &= [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p] \end{aligned} \tag{3.6.16}$$

Then for $q = k + 1$, the matrix linear model for (3.6.15) becomes

$$\begin{aligned}\mathbf{Y}_{n\times p} &= \mathbf{X}_{n\times q}\mathbf{B}_{q\times p} + \mathbf{E}_{n\times p} \\ &= \left[\mathbf{X}\boldsymbol{\beta}_1, \mathbf{X}\boldsymbol{\beta}_2, \ldots, \mathbf{X}\boldsymbol{\beta}_p\right] + \left[\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p\right]\end{aligned} \tag{3.6.17}$$

Model (3.6.17) is called the multivariate (linear) regression (MR) model , or MLMR model. If the $r(\mathbf{X}) < q = k+1$, so that the design matrix is not of full rank, the model is called the multivariate analysis of variance (MANOVA) model. Partitioning $\mathbf{X}$ into two matrices as in the univariate regression model, $\mathbf{X} = [\mathbf{A}, \mathbf{Z}]$ and $\mathbf{B}' = [\Theta', \Gamma']$, model (3.6.17) becomes the multivariate analysis of covariance (MANCOVA) model.

To represent the MR model as a GLM, the vec $(\cdot)$ operator is employed. Let $\mathbf{y} = \text{vec}(\mathbf{Y})$, $\boldsymbol{\beta} = \text{vec}(\mathbf{B})$ and $\mathbf{e} = \text{vec}(\mathbf{Y})$. Since the design matrix $\mathbf{X}_{n\times q}$ is the same for each of the p dependent variables, the GLM for the MR model is as follows

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_p \end{bmatrix} = \begin{bmatrix} \mathbf{X} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{X} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_p \end{bmatrix} + \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \vdots \\ \mathbf{e}_p \end{bmatrix}$$

Or, for $N = np$ and $K = pq = p(k+1)$, we have the vector form of the MLMR model

$$\begin{aligned}\mathbf{y}_{N\times 1} &= \left(\mathbf{I}_p \otimes \mathbf{X}\right)_{N\times K} \boldsymbol{\beta}_{K\times 1} + \mathbf{e}_{K\times 1} \\ \text{cov}\,(\mathbf{y}) &= \Sigma \otimes \mathbf{I}_n\end{aligned} \tag{3.6.18}$$

To test hypotheses, we assume that $\mathbf{E}$ in (3.6.17) has a matrix normal distribution, $\mathbf{E} \sim N_{n,\,p}(\mathbf{0}, \Sigma \otimes \mathbf{I}_n)$ or using the row representation that $\mathbf{E}' \sim N_{n,\,p}(\mathbf{0}, \mathbf{I}_n \otimes \Sigma)$. Alternatively, by (3.6.18), $\mathbf{e} \sim N_{np}(\mathbf{0}, \Sigma \otimes \mathbf{I}_n)$. To obtain the ML estimate of $\boldsymbol{\beta}$ given (3.6.18), we associate the covariance structure $\Sigma$with $\Sigma \otimes \mathbf{I}_n$ and apply (3.6.8), even though Σ is unknown. The unknown matrix drops out of the product. To see this, we have by substitution that

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{ML} &= \left[(\mathbf{I}_p \otimes \mathbf{X})'(\Sigma \otimes \mathbf{I}_n)^{-1}(\mathbf{I}_p \otimes \mathbf{X})\right]^{-1}(\mathbf{I}_p \otimes \mathbf{X})'(\Sigma \otimes \mathbf{I}_n)^{-1}\mathbf{y} \\ &= \left(\Sigma^{-1} \otimes \mathbf{X}'\mathbf{X}\right)^{-1}\left(\Sigma^{-1} \otimes \mathbf{X}'\right)\mathbf{y} \qquad (3.6.19)\end{aligned}$$

However, by property (5) in Theorem 2.4.7, we have that

$$\widehat{\boldsymbol{\beta}}_{ML} = \text{vec}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\right]$$

by letting $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\mathbf{C}' = \mathbf{I}_p$. Thus,

$$\widehat{\mathbf{B}}_{ML} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \qquad (3.6.20)$$

using the matrix form of the model. This is also the OLS estimate of the parameter matrix.

Similarly using (3.6.19), the

$$\begin{aligned}\operatorname{cov}\left(\widehat{\boldsymbol{\beta}}_{ML}\right) &= \left[\left(\mathbf{I}_p \otimes \mathbf{X}\right)' \left(\Sigma \otimes \mathbf{I}_n\right)^{-1} \left(\mathbf{I}_p \otimes \mathbf{X}\right)\right]^{-1} \\ &= \Sigma \otimes \left(\mathbf{X}'\mathbf{X}\right)^{-1}\end{aligned}$$

Finally, the ML estimate of Σ is

$$\widehat{\Sigma} = \mathbf{Y}' \left[\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]\mathbf{Y}/n \tag{3.6.21}$$

or the restricted maximum likelihood (REML) unbiased estimate is $\mathbf{S} = \mathbf{E}/(n-q)$ where $q = r(\mathbf{X})$. Furthermore $\widehat{\boldsymbol{\beta}}_{ML}$ and $\widehat{\Sigma}$ are independent, and $n\,\widehat{\Sigma} \sim W_p(n-q,\ \Sigma)$. Again, the Wishart density only exists if $n \geq p+q$.

In the above discussion, we have assumed that $\mathbf{X}$ has full column rank q. If the $r(\mathbf{X}) = r < q$, then $\widehat{\mathbf{B}}$ is not unique since $(\mathbf{X}'\mathbf{X})^{-1}$ is replaced with a g-inverse. However, $\widehat{\Sigma}$ is still unique since $\left(\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\right)$ is a unique projection matrix by property (4), Theorem

The null hypothesis tested for the matrix form of the MR model takes the general form

$$H : \mathbf{CBM} = \mathbf{0} \tag{3.6.22}$$

where $\mathbf{C}_{g\times q}$ is a known matrix of full row rank g, $g \leq q$ and $\mathbf{M}_{p\times u}$ is a matrix of full column rank $u \leq p$. Hypothesis (3.6.22) is called the standard multivariate hypothesis. To test (3.6.22) using the vector form of the model, observe that $\text{vec}(\mathbf{C}\widehat{\mathbf{B}}\mathbf{M}) = (\mathbf{M}' \otimes \mathbf{C})\,\text{vec}\,\widehat{\mathbf{B}}$ so that (3.6.22) is equivalent to testing $H : \mathbf{L}\boldsymbol{\beta} = \mathbf{0}$ when $\mathbf{L}$ is a matrix of order $gu \times pq$ of rank $v = gu$. Assuming $\Omega = \Sigma \otimes \mathbf{I}_n$ is known,

$$\widehat{\boldsymbol{\beta}}_{ML} \sim N_{gu}(\boldsymbol{\beta}, \mathbf{L}[(\mathbf{I}_n \otimes \mathbf{X})'\Omega^{-1}(\mathbf{I}_p \otimes \mathbf{X})]^{-1}\mathbf{L}') \tag{3.6.23}$$

Simplifying the structure of the covariance matrix,

$$\text{cov}\left(\widehat{\boldsymbol{\beta}}_{ML}\right) = \left(\mathbf{M}'\Sigma\mathbf{M}\right) \otimes \left(\mathbf{C}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{C}'\right) \tag{3.6.24}$$

For known Σ, the likelihood ratio test of H is to reject H if $X^2 > c_\alpha$ where c_α is chosen such that the $P(X^2 > c_\alpha \mid H) = \alpha$ and $X^2 = \widehat{\boldsymbol{\beta}}'_{ML}[(\mathbf{M}'\Sigma\mathbf{M}) \otimes (\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}')]^{-1}\widehat{\boldsymbol{\beta}}_{ML}$.

However, we can simplify X^2 since

$$\begin{aligned} X^2 &= [\text{vec}(\mathbf{C}\hat{\mathbf{B}}\mathbf{M})]'\left[\left(\mathbf{M}'\Sigma\mathbf{M}\right)^{-1} \otimes \left(\mathbf{C}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)\right]\text{vec}(\mathbf{C}\hat{\mathbf{B}}\mathbf{M}) \\ &= [\text{vec}(\mathbf{C}\hat{\mathbf{B}}\mathbf{M})]'\,\text{vec}[\left(\mathbf{C}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{C}'\right]^{-1}(\mathbf{C}\hat{\mathbf{B}}\mathbf{M})\left(\mathbf{M}'\Sigma\mathbf{M}\right)^{-1} \\ &= \text{tr}[(\mathbf{C}\hat{\mathbf{B}}\mathbf{M})'[\mathbf{C}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{C}']^{-1}(\mathbf{C}\hat{\mathbf{B}}\mathbf{A})\left(\mathbf{M}'\Sigma\mathbf{M}\right)^{-1}] \end{aligned} \tag{3.6.25}$$

Thus to test $H : \mathbf{L}\boldsymbol{\beta} = \mathbf{0}$, the hypothesis is rejected if X^2 in (3.6.25) is larger than a chi-square critical value with $v = gu$ degrees of freedom. Again, by finding a consistent estimate of Σ, $X^2 \xrightarrow{d} \chi^2(v = gu)$. Thus an approximate test of H is available if Σ is estimated by $\hat{\Sigma} \xrightarrow{p} \Sigma$.

Evaluating Normality

The study of robust estimation for location and dispersion of model parameters, the identification of outliers, the analysis of multivariate residuals, and the assessment of the effects of model assumptions on tests of significance and power are as important in multivariate analysis as they are in univariate analysis. However, the problems are much more complex. In multivariate data analysis there is no natural one-dimensional order to the observations, hence we can no longer just investigate the extremes of the distribution to locate outliers or identify data clusters in only one dimension. Clusters can occur in some subspace and outliers may not be extreme in any one dimension. Outliers in multivariate samples effect not only the location and variance of a variable, but also its orientation in the sample as measured by the covariance or correlation with other variables. Residuals formed from fitting a multivariate model to a data set in the presence of extreme outliers may lead to the identification of spurious outliers. Upon replotting the data, they are often removed. Finally, because non-normality can occur in so many ways robustness studies of Type I errors and power are difficult to design and evaluate.

The two most important problems in multivariate data analysis are the detection of outliers and the evaluation of multivariate normality. The process is complex and first begins with the assessment of marginal normality, a variable at a time; see Looney (1995). The evaluation process usually proceeds as follows.

1. Evaluate univariate normality by performing the Shapiro and Wilk (1965) W test a variable at a time when sample sizes are less than or equal to 50. The test is known to show a reasonable sensitivity to nonnormality; see Shapiro et al. (1968). For $50 < n \leq 2000$, Royston's (1982, 1992) approximation is recommended and is implemented

2. Construct normal probability quantile-vs-quantile (Q-Q) plots a variable at a time which compare the cumulative empirical distribution with the expected order values of a normal density to informally assess the lack of linearity and the presence of extreme values; see Wilk and Gnanadesikan (1968) and Looney and Gulledge (1985).

3. If variables are found to be non-normal, transform them to normality using perhaps a Box and Cox (1964) power transformation or some other transformation such as a logit.

4. Locate and correct outliers using graphical techniques or tests of significance as outlined by Barnett and Lewis (1994).

Example 3.7.1 (Generating MVN Distributions) *To illustrate the analysis of multivariate data, several multivariate normal distributions are generated. The data generated are used to demonstrate several of the procedures for evaluating multivariate normality and testing hypotheses about means and covariance matrices.*

By using the properties of the MVN distribution, recall that if $\mathbf{z} \sim N_p(\mathbf{0}, \mathbf{I}_p)$, *then* $\mathbf{y} = \mathbf{zA} + \boldsymbol{\mu} \sim N_p(\boldsymbol{\mu}, \Sigma = \mathbf{A}'\mathbf{A})$. *Hence, to generate a MVN distribution with mean* $\boldsymbol{\mu}$ *and covariance matrix* Σ, *one proceeds as follows.*

1. Specify $\boldsymbol{\mu}$ and Σ.
2. Obtain a Cholesky decomposition for Σ; call it $\mathbf{A}$.
3. Generate a $n \times p$ matrix of $N(0, 1)$ random variables named $\mathbf{Z}$.
4. Transform $\mathbf{Z}$ to $\mathbf{Y}$ using the expression $\mathbf{Y} = \mathbf{ZA} + \mathbf{U}$ where $\mathbf{U}$ is created by repeating the row vector $\mathbf{u}'$ n times producing an $n \times p$ matrix.

Multivariate Analysis Data into R

Data sets are available from the UCI Machine LearningRepository http://archive.ics.uci.edu/ml

For example, the file http://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data contains data on concentrations of 13 different chemicals in wines grown in the same region in Italy that are derived from three different cultivars.

The data set looks like this:

```
1,14.23,1.71,2.43,15.6,127,2.8,3.06,.28,2.29,5.64,1.04,3.92,1065
1,13.2,1.78,2.14,11.2,100,2.65,2.76,.26,1.28,4.38,1.05,3.4,1050
1,13.16,2.36,2.67,18.6,101,2.8,3.24,.3,2.81,5.68,1.03,3.17,1185
1,14.37,1.95,2.5,16.8,113,3.85,3.49,.24,2.18,7.8,.86,3.45,1480
1,13.24,2.59,2.87,21,118,2.8,2.69,.39,1.82,4.32,1.04,2.93,735
...
```

There is one row per wine sample. The first column contains the cultivar of a wine sample (labelled 1, 2 or 3), and the following thirteen columns contain the concentrations of the 13 different chemicals in that sample. The columns are separated by commas.

When we read the file into R using the read.table() function, we need to use the "sep=" argument in read.table() to tell it that the columns are separated by commas. That is, we can read in the file using the read.table() function as follows:

```
> wine <- read.table("http://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data",
          sep=",")
> wine
     V1    V2   V3   V4   V5  V6   V7   V8   V9  V10       V11   V12  V13  V14
1     1 14.23 1.71 2.43 15.6 127 2.80 3.06 0.28 2.29  5.640000 1.040 3.92 1065
2     1 13.20 1.78 2.14 11.2 100 2.65 2.76 0.26 1.28  4.380000 1.050 3.40 1050
3     1 13.16 2.36 2.67 18.6 101 2.80 3.24 0.30 2.81  5.680000 1.030 3.17 1185
4     1 14.37 1.95 2.50 16.8 113 3.85 3.49 0.24 2.18  7.800000 0.860 3.45 1480
5     1 13.24 2.59 2.87 21.0 118 2.80 2.69 0.39 1.82  4.320000 1.040 2.93  735
...
176   3 13.27 4.28 2.26 20.0 120 1.59 0.69 0.43 1.35 10.200000 0.590 1.56  835
177   3 13.17 2.59 2.37 20.0 120 1.65 0.68 0.53 1.46  9.300000 0.600 1.62  840
178   3 14.13 4.10 2.74 24.5  96 2.05 0.76 0.56 1.35  9.200000 0.610 1.60  560
```

In this case the data on 178 samples of wine has been read into the variable 'wine'.

A Matrix Scatterplot

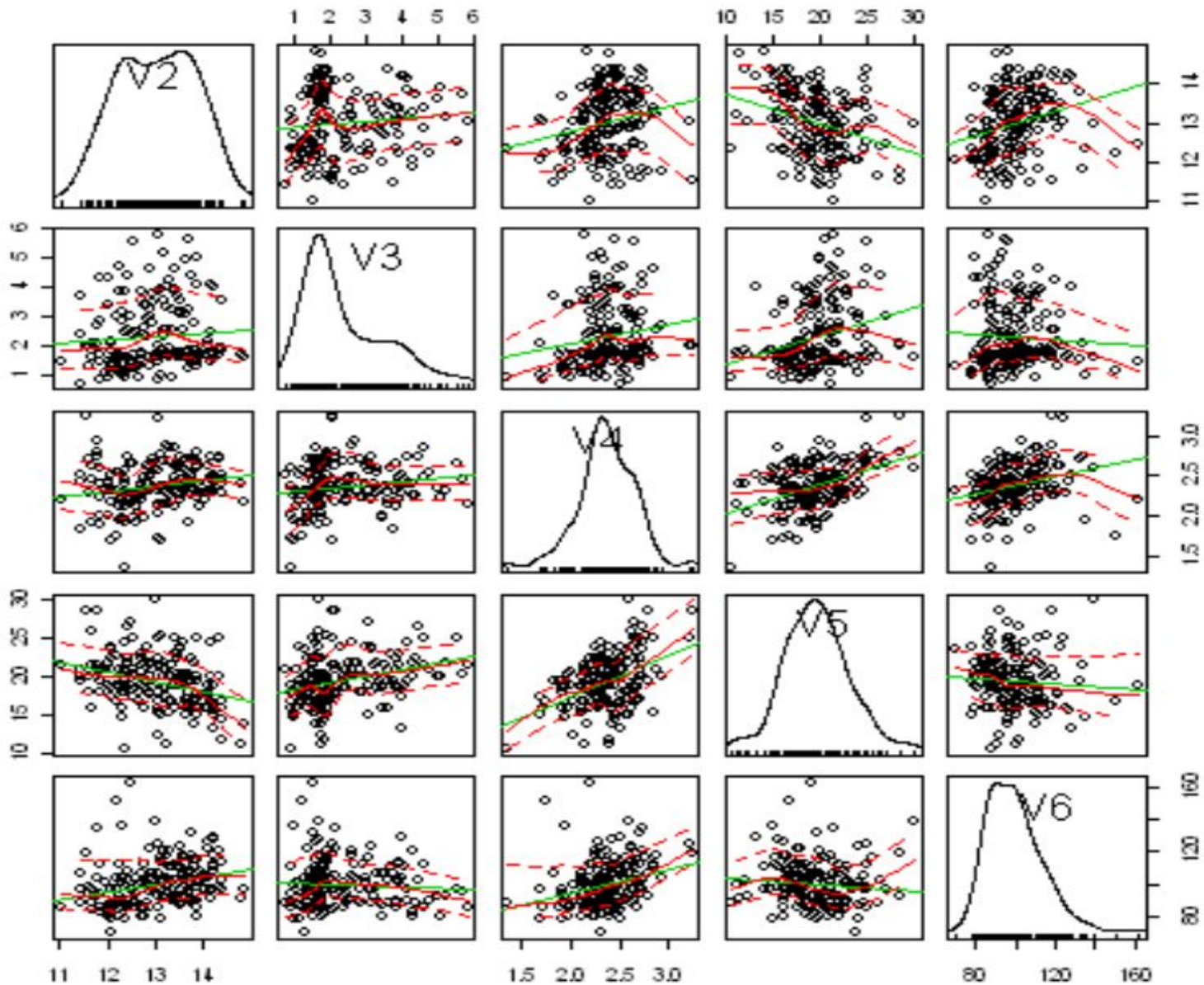
One common way of plotting multivariate data is to make a "matrix scatterplot", showing each pair of variables plotted against each other. We can use the "scatterplotMatrix()" function from the "car" R package to do this. To use this function, we first need to install the "car" R package (for instructions on how to install an R package

In this matrix scatterplot, the diagonal cells show histograms of each of the variables, in this case the concentrations of the first five chemicals (variables V2, V3, V4, V5, V6).
Each of the off-diagonal cells is a scatterplot of two of the five chemicals, for example, the second cell in the first row is a scatterplot of V2 (y-axis) against V3 (x-axis).

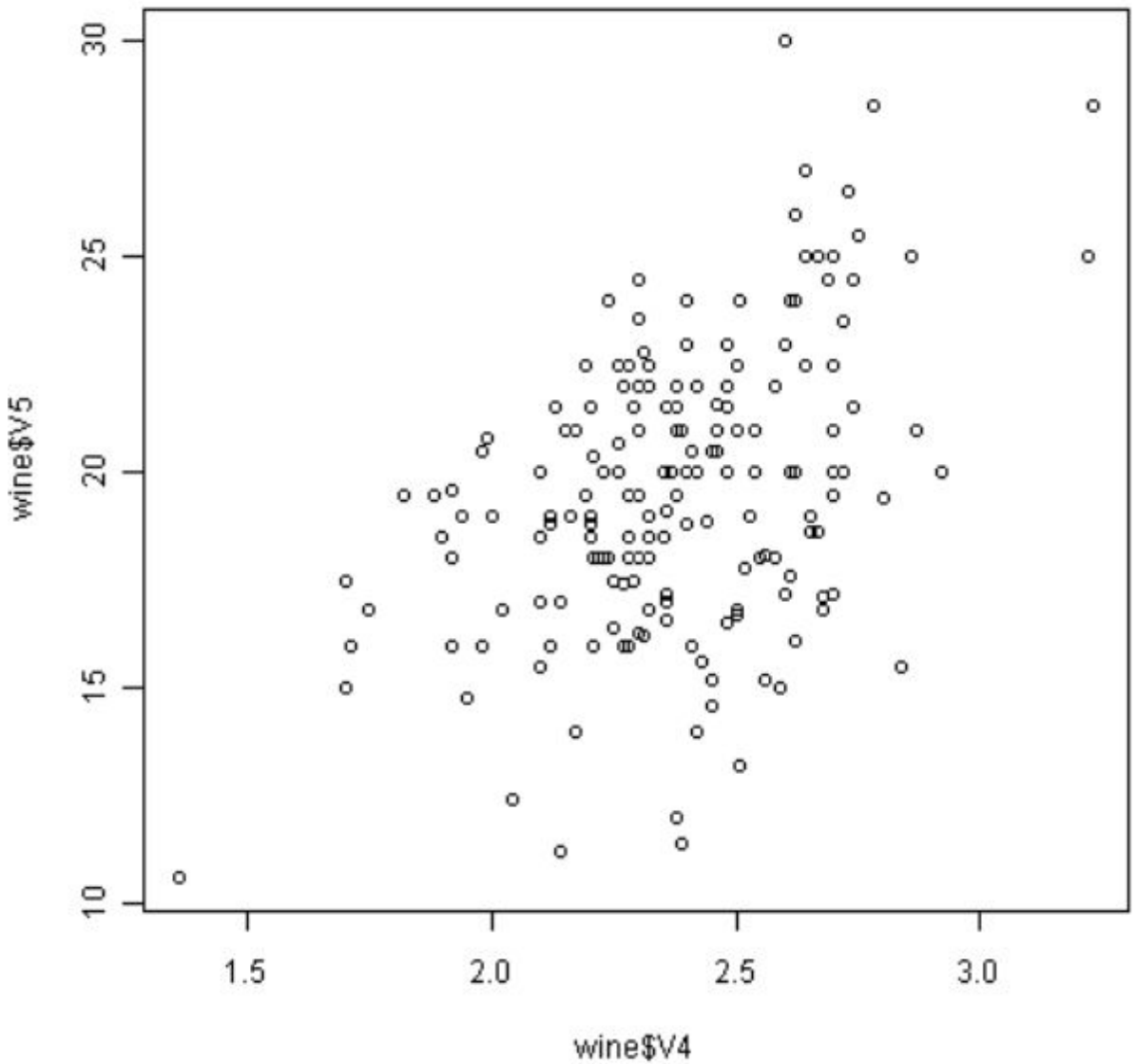
If you see an interesting scatterplot for two variables in the matrix scatterplot, you may want to plot that scatterplot in more detail, with the data points labelled by their group (their cultivar in this case).
For example, in the matrix scatterplot above, the cell in the third column of the fourth row down is a scatterplot of V5 (x-axis) against V4 (y-axis). If you look at this scatterplot, it appears that there may be a positive relationship between V5 and V4.

We may therefore decide to examine the relationship between V5 and V4 more closely, by plotting a scatterplot of these two variable, with the data points labelled by their group (their cultivar). To plot a scatterplot of two variables, we can use the "plot" R function. The V4 and V5 variables are stored in the columns V4 and V5 of the variable "wine", so can be accessed by typing wine$V4 or wine$V5. Therefore, to plot the scatterplot, we type:

```
> scatterplotMatrix(wine[2:6])
```

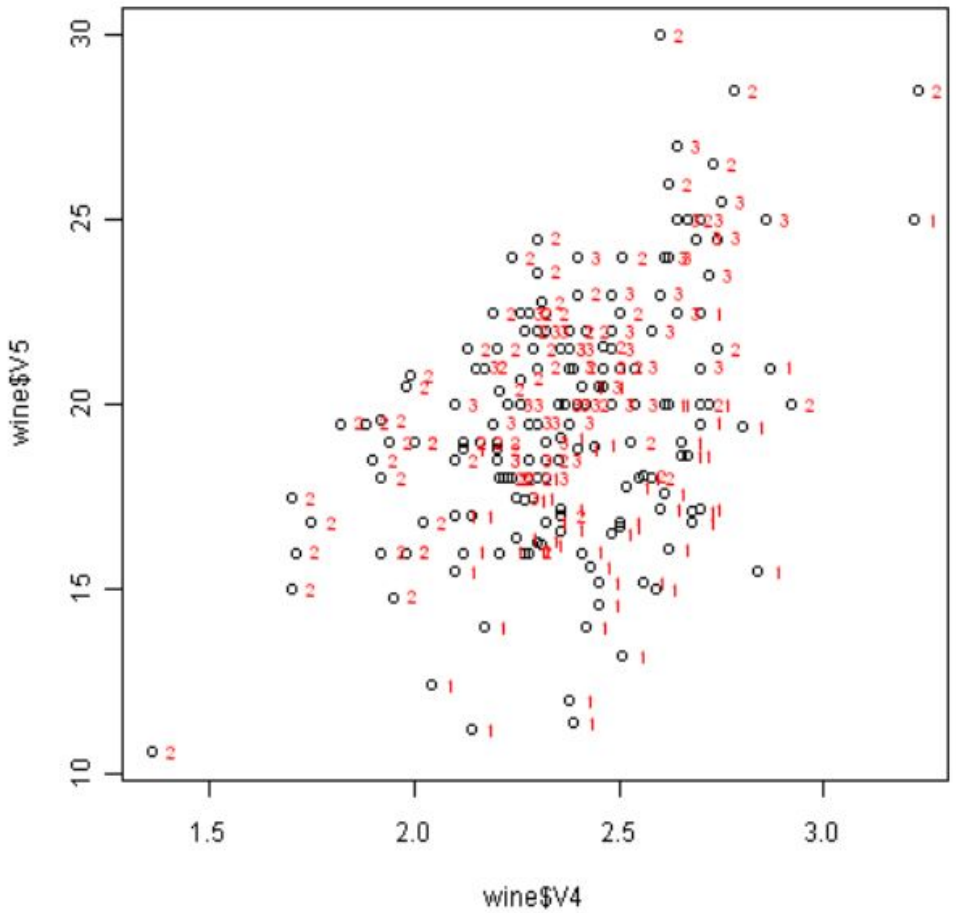


```
> plot(wine$V4, wine$V5)
```



```
> text(wine$V4, wine$V5, wine$V1, cex=0.7, pos=4, col="red")
```

If you look at the help page for the "text" function, you will see that "pos=4" will plot the text just to the right of the symbol for a data point. The "cex=0.5" option will plot the text at half the default size, and the "col=red" option will plot the text in red. This gives us the following plot:



We can see from the scatterplot of V4 versus V5 that the wines from cultivar 2 seem to have lower values of V4 compared to the wines of cultivar 1.

A Profile Plot

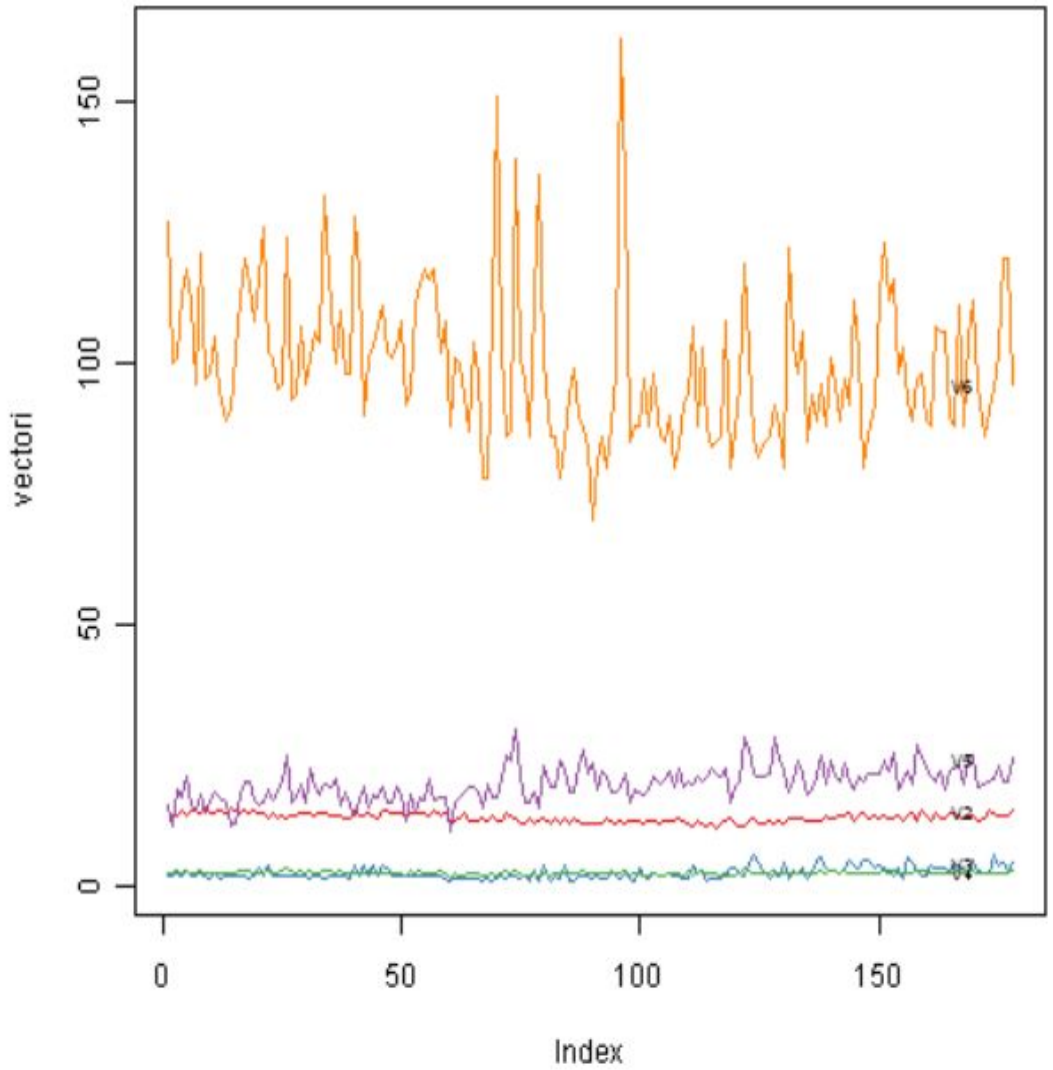
Another type of plot that is useful is a "profile plot", which shows the variation in each of the variables, by plotting the value of each of the variables for each of the samples.

The function "makeProfilePlot()" below can be used to make a profile plot. This function requires the "RColorBrewer"

To use this function, you first need to copy and paste it into R. The arguments to the function are a vector containing the names of the varibles that you want to plot, and a list variable containing the variables themselves.

For example, to make a profile plot of the concentrations of the first five chemicals in the wine samples (stored in columns V2, V3, V4, V5, V6 of variable "wine"), we type:

```
> library(RColorBrewer)
> names <- c("V2","V3","V4","V5","V6")
> mylist <- list(wine$V2,wine$V3,wine$V4,wine$V5,wine$V6)
> makeProfilePlot(mylist,names)
```



It is clear from the profile plot that the mean and standard deviation for V6 is quite a lot higher than that for the other variables.

Calculating Summary Statistics for Multivariate Data

Another thing that you are likely to want to do is to calculate summary statistics such as the mean and standard deviation for each of the variables in your multivariate data set.

This is easy to do, using the "mean()" and "sd()" functions in R. For example, say we want to calculate the mean and standard deviations of each of the 13 chemical concentrations in the wine samples. These are stored in columns 2-14 of the variable "wine". So we type

```
> sapply(wine[2:14],mean)
        V2         V3         V4         V5         V6         V7
13.0006180  2.3363483  2.3665169 19.4949438 99.7415730  2.2951124
        V8         V9        V10        V11        V12        V13
 2.0292697  0.3618539  1.5908989  5.0580899  0.9574494  2.6116854
       V14
746.8932584
```

Similarly, to get the standard deviations of the 13 chemical concentrations, we type:

```
> sapply(wine[2:14],sd)
        V2         V3         V4         V5         V6         V7
 0.8118265  1.1171461  0.2743440  3.3395638 14.2824835  0.6258510
        V8         V9        V10        V11        V12        V13
 0.9988587  0.1244533  0.5723589  2.3182859  0.2285716  0.7099904
       V14
314.9074743
```

We can see here that it would make sense to standardise in order to compare the variables because the variables have very different standard deviations – the standard deviation of V14 is 314.9074743, while the standard deviation of V9 is just 0.1244533. Thus, in order to compare the variables, we need to standardise each variable so that it has a sample variance of 1 and sample mean of 0. We will explain below how to standardise the variables.

Means and Variances Per Group ¶

It is often interesting to calculate the means and standard deviations for just the samples from a particular group, for example, for the wine samples from each cultivar. The cultivar is stored in the column "V1" of the variable "wine".

To extract out the data for just cultivar 2, we can type:

```
> cultivar2wine <- wine[wine$V1=="2",]
```

We can then calculate the mean and standard deviations of the 13 chemicals' concentrations, for just the cultivar 2 samples:

```
> sapply(cultivar2wine[2:14],mean)
      V2       V3       V4       V5       V6       V7       V8
12.278732  1.932676  2.244789 20.238028 94.549296  2.258873  2.080845
      V9      V10      V11      V12      V13      V14
 0.363662  1.630282  3.086620  1.056282  2.785352 519.507042
> sapply(cultivar2wine[2:14])
       V2        V3        V4        V5        V6        V7        V8
0.5379642 1.0155687 0.3154673 3.3497704 16.7534975 0.5453611 0.7057008
       V9       V10       V11       V12       V13       V14
0.1239613 0.6020678 0.9249293 0.2029368 0.4965735 157.2112204
```

To use the function "printMeanAndSdByGroup()", you first need to copy and paste it into R. The arguments of the function are the variables that you want to calculate means and standard deviations for, and the variable containing the group of each sample. For example, to calculate the mean and standard deviation for each of the 13 chemical concentrations, for each of the three different wine cultivars, we type:

```
> printMeanAndSdByGroup(wine[2:14],wine[1])
[1] "Means:"
  V1       V2       V3       V4       V5       V6       V7        V8       V9      V10      V11      
1  1 13.74475 2.010678 2.455593 17.03729 106.3390 2.840169 2.9823729 0.290000 1.899322 5.528305 1.06
2  2 12.27873 1.932676 2.244789 20.23803  94.5493 2.258873 2.0808451 0.363662 1.630282 3.086620 1.05
3  3 13.15375 3.333750 2.437083 21.41667  99.3125 1.678750 0.7814583 0.447500 1.153542 7.396250 0.68
[1] "Standard deviations:"
  V1        V2        V3        V4       V5       V6        V7        V8         V9       V10       
1  1 0.4621254 0.6885489 0.2271660 2.546322 10.49895 0.3389614 0.3974936 0.07004924 0.4121092 1.2385
2  2 0.5379642 1.0155687 0.3154673 3.349770 16.75350 0.5453611 0.7057008 0.12396128 0.6020678 0.9249
3  3 0.5302413 1.0879057 0.1846902 2.258161 10.89047 0.3569709 0.2935041 0.12413959 0.4088359 2.3109
[1] "Sample sizes:"
  V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14
1  1 59 59 59 59 59 59 59 59  59  59  59  59  59
2  2 71 71 71 71 71 71 71 71  71  71  71  71  71
3  3 48 48 48 48 48 48 48 48  48  48  48  48  48
```

The function "printMeanAndSdByGroup()" also prints out the number of samples in each group. In this case, we see that there are 59 samples of cultivar 1, 71 of cultivar 2, and 48 of cultivar 3.

Between-groups Variance and Within-groups Variance for a Variable

If we want to calculate the within-groups variance for a particular variable (for example, for a particular chemical's concentration), we can use the function "calcWithinGroupsVariance()" below:

You will need to copy and paste this function into R before you can use it. For example, to calculate the within-groups variance of the variable V2 (the concentration of the first chemical), we type:

```
> calcWithinGroupsVariance(wine[2],wine[1])
  [1] 0.2620525
```

Thus, the within-groups variance for V2 is 0.2620525.

We can calculate the between-groups variance for a particular variable (eg. V2) using the function "calcBetweenGroupsVariance()" below:

Once you have copied and pasted this function into R, you can use it to calculate the between-groups variance for a variable such as V2:

```
> calcBetweenGroupsVariance (wine[2],wine[1])
  [1] 35.39742
```

Thus, the between-groups variance of V2 is 35.39742.

We can calculate the "separation" achieved by a variable as its between-groups variance devided by its within-groups variance. Thus, the separation achieved by V2 is calculated as:

```
> 35.39742/0.2620525
  [1] 135.0776
```

If you want to calculate the separations achieved by all of the variables in a multivariate data set, you can use the function "calcSeparations()" below:

For example, to calculate the separations for each of the 13 chemical concentrations, we type:

```
> calcSeparations(wine[2:14],wine[1])
  [1] "variable V2 Vw= 0.262052469153907 Vb= 35.3974249602692 separation= 135.0776242428"
  [1] "variable V3 Vw= 0.887546796746581 Vb= 32.7890184869213 separation= 36.9434249631837"
  [1] "variable V4 Vw= 0.0660721013425184 Vb= 0.879611357248741 separation= 13.312901199991"
  [1] "variable V5 Vw= 8.00681118121156 Vb= 286.41674636309 separation= 35.7716374073093"
  [1] "variable V6 Vw= 180.65777316441 Vb= 2245.50102788939 separation= 12.4295843381499"
  [1] "variable V7 Vw= 0.191270475224227 Vb= 17.9283572942847 separation= 93.7330096203673"
  [1] "variable V8 Vw= 0.274707514337437 Vb= 64.2611950235641 separation= 233.925872681549"
  [1] "variable V9 Vw= 0.0119117022132797 Vb= 0.328470157461624 separation= 27.5754171469659"
  [1] "variable V10 Vw= 0.246172943795542 Vb= 7.45199550777775 separation= 30.2713831702276"
  [1] "variable V11 Vw= 2.28492308133354 Vb= 275.708000822304 separation= 120.664018441003"
  [1] "variable V12 Vw= 0.0244876469432414 Vb= 2.48100991493829 separation= 101.3167953903"
  [1] "variable V13 Vw= 0.160778729560982 Vb= 30.5435083544253 separation= 189.972320578889"
  [1] "variable V14 Vw= 29707.6818705169 Vb= 6176832.32228483 separation= 207.920373902178"
```

Thus, the individual variable which gives the greatest separations between the groups (the wine cultivars) is V8 (separation 233.9). As we will discuss below, the purpose of linear discriminant analysis (LDA) is to find the linear combination of the individual variables that will give the greatest separation between the groups (cultivars here). This hopefully will give a better separation than the best separation achievable by any individual variable (233.9 for V8 here).

Between-groups Covariance and Within-groups Covariance for Two Variables

If you have a multivariate data set with several variables describing sampling units from different groups, such as the wine samples from different cultivars, it is often of interest to calculate the within-groups covariance and between-groups variance for pairs of the variables.

For example, to calculate the within-groups covariance for variables V8 and V11, we type:

```
> calcWithinGroupsCovariance(wine[8],wine[11],wine[1])
  [1] 0.2866783
```

For example, to calculate the between-groups covariance for variables V8 and V11, we type:

```
> calcBetweenGroupsCovariance(wine[8],wine[11],wine[1])
  [1] -60.41077
```

Thus, for V8 and V11, the between-groups covariance is -60.41 and the within-groups covariance is 0.29. Since the within-groups covariance is positive (0.29), it means V8 and V11 are positively related within groups: for individuals from the same group, individuals with a high value of V8 tend to have a high value of V11, and vice versa. Since the between-groups covariance is negative (-60.41), V8 and V11 are negatively related between groups: groups with a high mean value of V8 tend to have a low mean value of V11, and vice versa.

Calculating Correlations for Multivariate Data

It is often of interest to investigate whether any of the variables in a multivariate data set are significantly correlated.

To calculate the linear (Pearson) correlation coefficient for a pair of variables, you can use the "cor.test()" function in R. For example, to calculate the correlation coefficient for the first two chemicals' concentrations, V2 and V3, we type:

```
> cor.test(wine$V2, wine$V3)
  Pearson's product-moment correlation
  data:  wine$V2 and wine$V3
  t = 1.2579, df = 176, p-value = 0.2101
  alternative hypothesis: true correlation is not equal to 0
  95 percent confidence interval:
  -0.05342959  0.23817474
  sample estimates:
   cor
  0.09439694
```

This tells us that the correlation coefficient is about 0.094, which is a very weak correlation. Furthermore, the P-value for the statistical test of whether the correlation coefficient is significantly different from zero is 0.21. This is much greater than 0.05 (which we can use here as a cutoff for statistical significance), so there is very weak evidence that that the correlation is non-zero.

If you have a lot of variables, you can use "cor.test()" to calculate the correlation coefficient for each pair of variables, but you might be just interested in finding out what are the most highly correlated pairs of variables. For this you can use the function "mosthighlycorrelated()" below.

The function "mosthighlycorrelated()" will print out the linear correlation coefficients for each pair of variables in your data set, in order of the correlation coefficient. This lets you see very easily which pair of variables are most highly correlated.

For example, to calculate correlation coefficients between the concentrations of the 13 chemicals in the wine samples, and to print out the top 10 pairwise correlation coefficients, you can type:

```
> mosthighlycorrelated(wine[2:14], 10)
     First.Variable Second.Variable Correlation
84               V7              V8   0.8645635
150              V8             V13   0.7871939
149              V7             V13   0.6999494
111              V8             V10   0.6526918
157              V2             V14   0.6437200
110              V7             V10   0.6124131
154             V12             V13   0.5654683
132              V3             V12  -0.5612957
118              V2             V11   0.5463642
137              V8             V12   0.5434786
```

This tells us that the pair of variables with the highest linear correlation coefficient are V7 and V8 (correlation = 0.86 approximately).

Calculating Correlations for Multivariate Data

It is often of interest to investigate whether any of the variables in a multivariate data set are significantly correlated.

To calculate the linear (Pearson) correlation coefficient for a pair of variables, you can use the "cor.test()" function in R. For example, to calculate the correlation coefficient for the first two chemicals' concentrations, V2 and V3, we type:

```
> cor.test(wine$V2, wine$V3)
  Pearson's product-moment correlation
  data:  wine$V2 and wine$V3
  t = 1.2579, df = 176, p-value = 0.2101
  alternative hypothesis: true correlation is not equal to 0
  95 percent confidence interval:
  -0.05342959  0.23817474
  sample estimates:
   cor
  0.09439694
```

This tells us that the correlation coefficient is about 0.094, which is a very weak correlation. Furthermore, the P-value for the statistical test of whether the correlation coefficient is significantly different from zero is 0.21. This is much greater than 0.05 (which we can use here as a cutoff for statistical significance), so there is very weak evidence that that the correlation is non-zero.

If you have a lot of variables, you can use "cor.test()" to calculate the correlation coefficient for each pair of variables, but you might be just interested in finding out what are the most highly correlated pairs of variables. For this you can use the function "mosthighlycorrelated()" below.

For example, to calculate correlation coefficients between the concentrations of the 13 chemicals in the wine samples, and to print out the top 10 pairwise correlation coefficients, you can type:

```
> mosthighlycorrelated(wine[2:14], 10)
    First.Variable Second.Variable Correlation
84              V7              V8   0.8645635
150             V8             V13   0.7871939
149             V7             V13   0.6999494
111             V8             V10   0.6526918
157             V2             V14   0.6437200
110             V7             V10   0.6124131
154            V12             V13   0.5654683
132             V3             V12  -0.5612957
118             V2             V11   0.5463642
137             V8             V12   0.5434786
```

This tells us that the pair of variables with the highest linear correlation coefficient are V7 and V8 (correlation = 0.86 approximately).

Standardising Variables

If you want to compare different variables that have different units, are very different variances, it is a good idea to first standardise the variables.

For example, we found above that the concentrations of the 13 chemicals in the wine samples show a wide range of standard deviations, from 0.1244533 for V9 (variance 0.01548862) to 314.9074743 for V14 (variance 99166.72). This is a range of approximately 6,402,554-fold in the variances.

As a result, it is not a good idea to use the unstandardised chemical concentrations as the input for a principal component analysis (PCA, see below) of the wine samples, as if you did that, the first principal component would be dominated by the variables which show the largest variances, such as V14.

Thus, it would be a better idea to first standardise the variables so that they all have variance 1 and mean 0, and to then carry out the principal component analysis on the standardised data. This would allow us to find the principal components that provide the best low-dimensional representation of the variation in the original data, without being overly biased by those variables that show the most variance in the original data.

For example, to standardise the concentrations of the 13 chemicals in the wine samples, we type:

```
> standardisedconcentrations <- as.data.frame(scale(wine[2:14]))
```

We can check that each of the standardised variables stored in "standardisedconcentrations" has a mean of 0 and a standard deviation of 1 by typing:

```
> sapply(standardisedconcentrations,mean)
      V2            V3            V4            V5            V6            V7
-8.591766e-16 -6.776446e-17  8.045176e-16 -7.720494e-17 -4.073935e-17 -1.395560e-17
      V8            V9           V10           V11           V12           V13
 6.958263e-17 -1.042186e-16 -1.221369e-16  3.649376e-17  2.093741e-16  3.003459e-16
     V14
-1.034429e-16
> sapply(standardisedconcentrations,sd)
 V2  V3  V4  V5  V6  V7  V8  V9 V10 V11 V12 V13 V14
  1   1   1   1   1   1   1   1   1   1   1   1   1
```

We see that the means of the standardised variables are all very tiny numbers and so are essentially equal to 0, and the standard deviations of the standardised variables are all equal to 1.

Principal Component Analysis

The purpose of principal component analysis is to find the best low-dimensional representation of the variation in a multivariate data set. For example, in the case of the wine data set, we have 13 chemical concentrations describing wine samples from three different cultivars. We can carry out a principal component analysis to investigate whether we can capture most of the variation between samples using a smaller number of new variables (principal components), where each of these new variables is a linear combination of all or some of the 13 chemical concentrations.

To carry out a principal component analysis (PCA) on a multivariate data set, the first step is often to standardise the variables under study using the "scale()" function (see above). This is necessary if the input variables have very different variances, which is true in this case as the concentrations of the 13 chemicals have very different variances (see above).

Once you have standardised your variables, you can carry out a principal component analysis using the "prcomp()" function in R.

You can get a summary of the principal component analysis results using the "summary()" function on the output of "prcomp()":

```
> summary(wine.pca)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10
Standard deviation     2.169 1.580 1.203 0.9586 0.9237 0.8010 0.7423 0.5903 0.5375 0.5009
Proportion of Variance 0.362 0.192 0.111 0.0707 0.0656 0.0494 0.0424 0.0268 0.0222 0.0193
Cumulative Proportion  0.362 0.554 0.665 0.7360 0.8016 0.8510 0.8934 0.9202 0.9424 0.9617
                          PC11   PC12    PC13
Standard deviation     0.4752 0.4108 0.32152
Proportion of Variance 0.0174 0.0130 0.00795
Cumulative Proportion  0.9791 0.9920 1.00000
```

This gives us the standard deviation of each component, and the proportion of variance explained by each component. The standard deviation of the components is stored in a named element called "sdev" of the output variable made by "prcomp":

```
> wine.pca$sdev
 [1] 2.1692972 1.5801816 1.2025273 0.9586313 0.9237035 0.8010350 0.7423128 0.5903367
 [9] 0.5374755 0.5009017 0.4751722 0.4108165 0.3215244
```

The total variance explained by the components is the sum of the variances of the components:

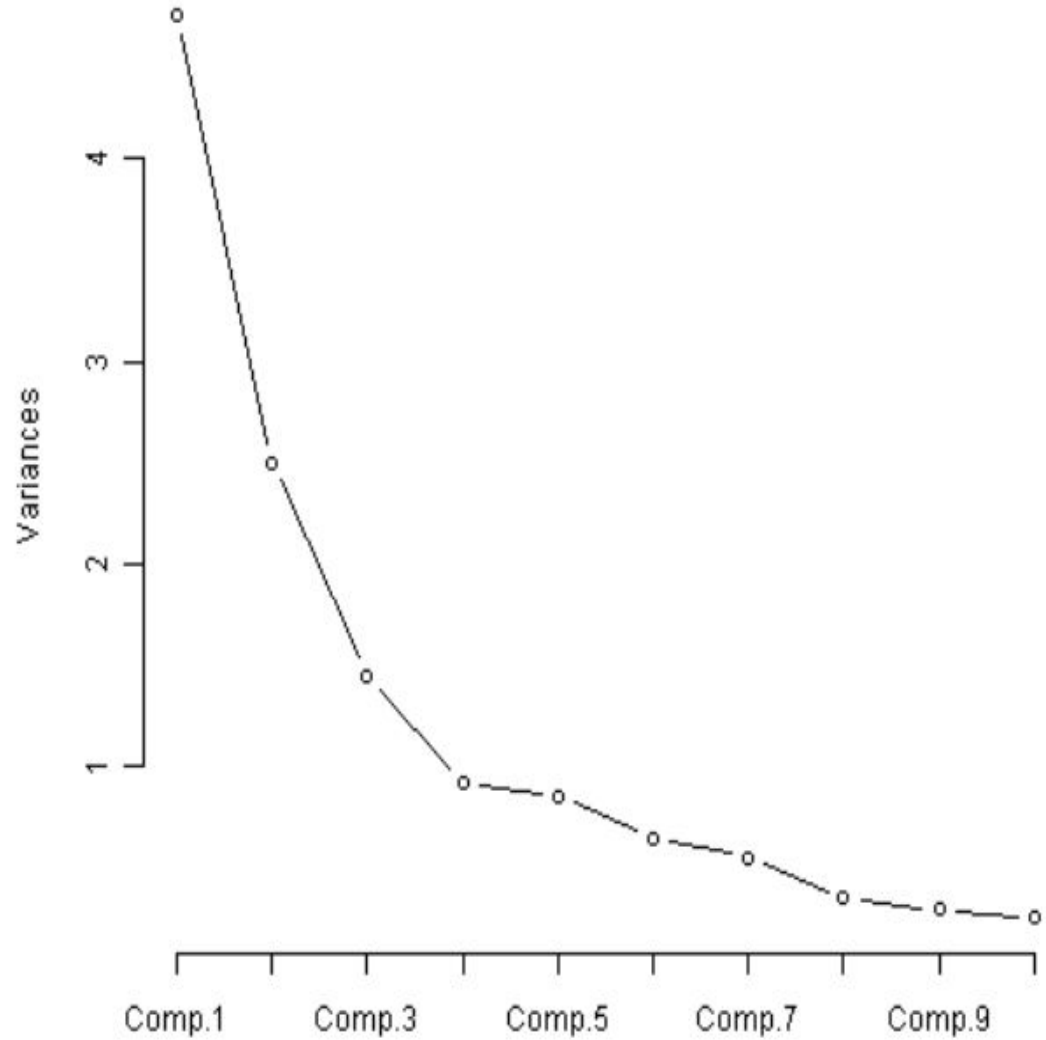
```
> sum((wine.pca$sdev)^2)
[1] 13
```

In this case, we see that the total variance is 13, which is equal to the number of standardised variables (13 variables). This is because for standardised data, the variance of each standardised variable is 1. The total variance is equal to the sum of the variances of the individual variables, and since the variance of each standardised variable is 1, the total variance should be equal to the number of variables (13 here).

Deciding How Many Principal Components to Retain

In order to decide how many principal components should be retained, it is common to summarise the results of a principal components analysis by making a scree plot, which we can do in R using the "screeplot()" function:

```
> screeplot(wine.pca, type="lines")
```



The most obvious change in slope in the scree plot occurs at component 4, which is the “elbow” of the scree plot. Therefore, it cound be argued based on the basis of the scree plot that the first three components should be retained.

Another way of deciding how many components to retain is to use Kaiser's criterion: that we should only retain principal components for which the variance is above 1 (when principal component analysis was applied to standardised data). We can check this by finding the variance of each of the principal components:

```
> (wine.pca$sdev)^2
 [1] 4.7058503 2.4969737 1.4460720 0.9189739 0.8532282 0.6416570 0.5510283 0.3484974
 [9] 0.2888799 0.2509025 0.2257886 0.1687702 0.1033779
```

We see that the variance is above 1 for principal components 1, 2, and 3 (which have variances 4.71, 2.50, and 1.45, respectively). Therefore, using Kaiser's criterion, we would retain the first three principal components.

A third way to decide how many principal components to retain is to decide to keep the number of components required to explain at least some minimum amount of the total variance. For example, if it is important to explain at least 80% of the variance, we would retain the first five principal components, as we can see from the output of "summary(wine.pca)" that the first five principal components explain 80.2% of the variance (while the first four components explain just 73.6%, so are not sufficient).

Loadings in PCA

Principal Component Analysis (PCA) is a widely used technique for dimensionality reduction and data simplification. One of the core outputs of PCA is the derivation of principal components. These components are linear combinations of the original features or variables in your dataset. The coefficients, or **weights, assigned to** these original variables within these **linear combinations** are termed **loadings**.

Definition: Loadings offer valuable insights into the nature of the principal components. Specifically, they tell us how much each original variable contributes to the respective principal component.

Interpretation: The magnitude of a loading points to the importance or relevance of its corresponding original variable to that component. A larger absolute value suggests that the variable significantly influences the component. Furthermore, the sign of the loading — whether positive or negative — denotes the directionality of this influence. A positive sign implies that the variable and the principal component are positively related, moving in the same direction. Conversely, a negative sign indicates an inverse relationship, where one increases as the other decreases. your dataset's original features or variables

Matrix Representation: In the mathematical underpinnings of PCA, loadings are connected with the eigenvectors of the covariance or correlation matrix of the data. These eigenvectors represent the loadings for the resultant principal components. When visualized in matrix form, each column corresponds to a specific principal component, and each row aligns with an original variable from the dataset.

Orthogonality: An intriguing aspect of PCA is the orthogonality of the principal components, meaning they remain uncorrelated with each other. Consequently, the loadings associated with different principal components are also orthogonal.

Scaling: The choice of scaling before applying PCA is another critical aspect. If PCA performed on a correlation matrix, it means that each variable was standardized to have a mean of 0 and variance of 1 before the analysis. In this scenario, the relationships between variables are evaluated based on their correlations. On the other hand, if PCA is performed on the covariance matrix, the relationships are interpreted based on covariances. In the latter case, variables with larger variances will naturally have more substantial loadings.

Contribution to Variance: A valuable insight from loadings is their contribution to variance. By squaring the value of a loading, one can ascertain the proportion of the variance of a variable captured by a principal component. For instance, if a loading value for a specific variable on the first principal component is 0.8, this indicates that 64% (which is 0.8 squared) of the variance of that variable is encapsulated by the first principal component.

In summary, loadings in PCA provide insights into how the original variables are combined to create each principal component, helping to interpret the nature and meaning of the principal components in the context of the original data.

Loadings for the Principal Components

The loadings for the principal components are stored in a named element "rotation" of the variable returned by "prcomp()". This contains a matrix with the loadings of each principal component, where the first column in the matrix contains the loadings for the first principal component, the second column contains the loadings for the second principal component, and so on.

Therefore, to obtain the loadings for the first principal component in our analysis of the 13 chemical concentrations in wine samples, we type:

```
> wine.pca$rotation[,1]
     V2           V3           V4           V5           V6           V7
-0.144329395  0.245187580  0.002051061  0.239320405 -0.141992042 -0.394660845
     V8           V9          V10          V11          V12          V13
-0.422934297  0.298533103 -0.313429488  0.088616705 -0.296714564 -0.376167411
    V14
-0.286752227
```

This means that the first principal component is a linear combination of the variables: -0.144*Z2 + 0.245*Z3 + 0.002*Z4 + 0.239*Z5 - 0.142*Z6 - 0.395*Z7 - 0.423*Z8 + 0.299*Z9 -0.313*Z10 + 0.089*Z11 - 0.297*Z12 - 0.376*Z13 - 0.287*Z14, where Z2, Z3, Z4...Z14 are the standardised versions of the variables V2, V3, V4...V14 (that each have mean of 0 and variance of 1).

Note that the square of the loadings sum to 1, as this is a constraint used in calculating the loadings:

```
> sum((wine.pca$rotation[,1])^2)
[1] 1
```

To calculate the values of the first principal component, we can define our own function to calculate a principal component given the loadings and the input variables' values:

We can then use the function to calculate the values of the first principal component for each sample in our wine data:

```
> calcpc(standardisedconcentrations, wine.pca$rotation[,1])
  [1] -3.30742097 -2.20324981 -2.50966069 -3.74649719 -1.00607049 -3.04167373 -2.44220051
  [8] -2.05364379 -2.50381135 -2.74588238 -3.46994837 -1.74981688 -2.10751729 -3.44842921
  [15] -4.30065228 -2.29870383 -2.16584568 -1.89362947 -3.53202167 -2.07865856 -3.11561376
  [22] -1.08351361 -2.52809263 -1.64036108 -1.75662066 -0.98729406 -1.77028387 -1.23194878
  [29] -2.18225047 -2.24976267 -2.49318704 -2.66987964 -1.62399801 -1.89733870 -1.40642118
  [36] -1.89847087 -1.38096669 -1.11905070 -1.49796891 -2.52268490 -2.58081526 -0.66660159
  ...
```

In fact, the values of the first principal component are stored in the variable wine.pca$x[,1] that was returned by the "prcomp()" function, so we can compare those values to the ones that we calculated, and they should agree:

```
> wine.pca$x[,1]
  [1] -3.30742097 -2.20324981 -2.50966069 -3.74649719 -1.00607049 -3.04167373 -2.44220051
  [8] -2.05364379 -2.50381135 -2.74588238 -3.46994837 -1.74981688 -2.10751729 -3.44842921
  [15] -4.30065228 -2.29870383 -2.16584568 -1.89362947 -3.53202167 -2.07865856 -3.11561376
  [22] -1.08351361 -2.52809263 -1.64036108 -1.75662066 -0.98729406 -1.77028387 -1.23194878
  [29] -2.18225047 -2.24976267 -2.49318704 -2.66987964 -1.62399801 -1.89733870 -1.40642118
  [36] -1.89847087 -1.38096669 -1.11905070 -1.49796891 -2.52268490 -2.58081526 -0.66660159
  ...
```

We see that they do agree

The first principal component has highest (in absolute value) loadings for V8 (−0.423), V7 (−0.395), V13 (−0.376), V10 (−0.313), V12 (−0.297), V14 (−0.287), V9 (0.299), V3 (0.245), and V5 (0.239). The loadings for V8, V7, V13, V10, V12 and V14 are negative, while those for V9, V3, and V5 are positive. Therefore, an interpretation of the first principal component is that it represents a contrast between the concentrations of V8, V7, V13, V10, V12, and V14, and the concentrations of V9, V3 and V5.

Similarly, we can obtain the loadings for the second principal component by typing:

```
> wine.pca$rotation[,2]
     V2           V3           V4           V5           V6           V7
0.483651548  0.224930935  0.316068814 -0.010590502  0.299634003  0.065039512
     V8           V9          V10          V11          V12          V13
-0.003359812  0.028779488  0.039301722  0.529995672 -0.279235148 -0.164496193
    V14
0.364902832
```

This means that the second principal component is a linear combination of the variables: 0.484*Z2 + 0.225*Z3 + 0.316*Z4 − 0.011*Z5 + 0.300*Z6 + 0.065*Z7 − 0.003*Z8 + 0.029*Z9 + 0.039*Z10 + 0.530*Z11 − 0.279*Z12 − 0.164*Z13 + 0.365*Z14, where Z1, Z2, Z3...Z14 are the standardised versions of variables V2, V3, ... V14 that each have mean 0 and variance 1.

Note that the square of the loadings sum to 1, as above:

```
> sum((wine.pca$rotation[,2])^2)
[1] 1
```

The second principal component has highest loadings for V11 (0.530), V2 (0.484), V14 (0.365), V4 (0.316), V6 (0.300), V12 (-0.279), and V3 (0.225). The loadings for V11, V2, V14, V4, V6 and V3 are positive, while the loading for V12 is negative. Therefore, an interpretation of the second principal component is that it represents a contrast between the concentrations of V11, V2, V14, V4, V6 and V3, and the concentration of V12. Note that the loadings for V11 (0.530) and V2 (0.484) are the largest, so the contrast is mainly between the concentrations of V11 and V2, and the concentration of V12.

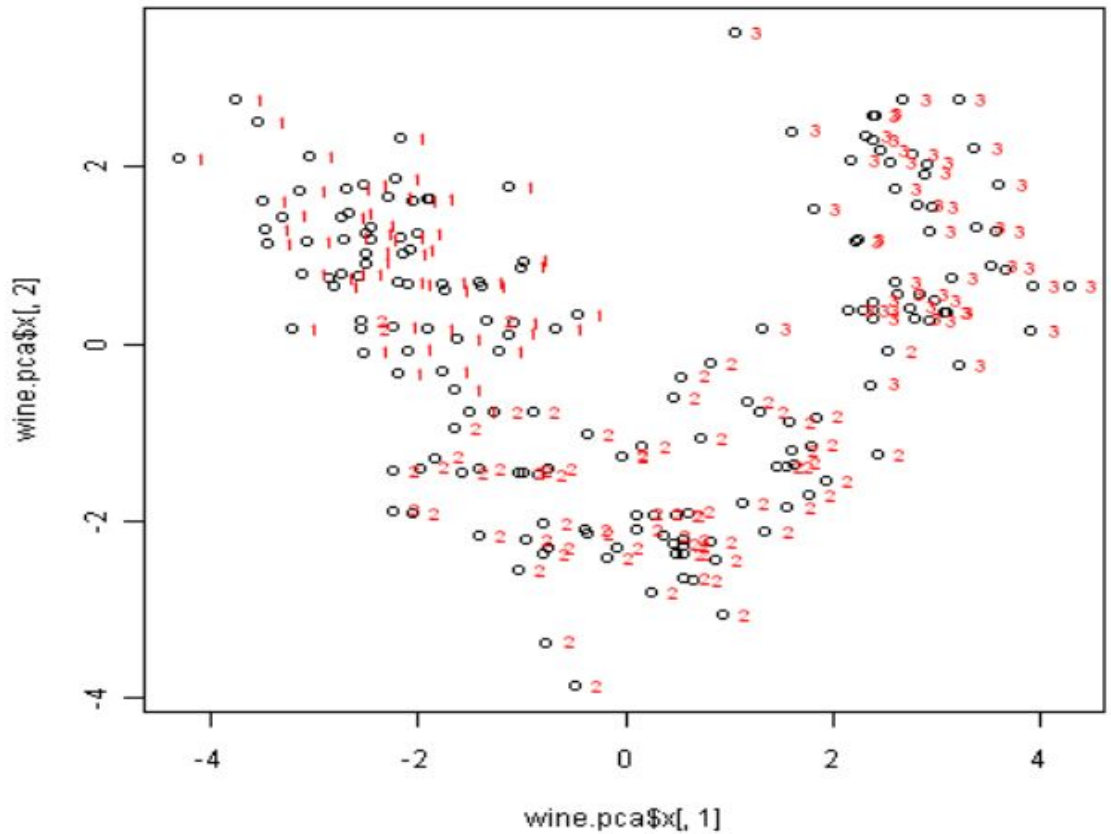
Scatterplots of the Principal Components

The values of the principal components are stored in a named element "x" of the variable returned by "prcomp()". This contains a matrix with the principal components, where the first column in the matrix contains the first principal component, the second column the second component, and so on.

Thus, in our example, "wine.pca$x[,1]" contains the first principal component, and "wine.pca$x[,2]" contains the second principal component.

We can make a scatterplot of the first two principal components, and label the data points with the cultivar that the wine samples come from, by typing:

```
> plot(wine.pca$x[,1],wine.pca$x[,2]) # make a scatterplot
> text(wine.pca$x[,1],wine.pca$x[,2], wine$V1, cex=0.7, pos=4, col="red") # add labels
```



The scatterplot shows the first principal component on the x-axis, and the second principal component on the y-axis. We can see from the scatterplot that wine samples of cultivar 1 have much lower values of the first principal component than wine samples of cultivar 3. Therefore, the first principal component separates wine samples of cultivars 1 from those of cultivar 3.

We can also see that wine samples of cultivar 2 have much higher values of the second principal component than wine samples of cultivars 1 and 3. Therefore, the second principal component separates samples of cultivar 2 from samples of cultivars 1 and 3.

Therefore, the first two principal components are reasonably useful for distinguishing wine samples of the three different cultivars.

Above, we interpreted the first principal component as a contrast between the concentrations of V8, V7, V13, V10, V12, and V14, and the concentrations of V9, V3 and V5. We can check whether this makes sense in terms of the concentrations of these chemicals in the different cultivars, by printing out the means of the standardised concentration variables in each cultivar, using the "printMeanAndSdByGroup()" function (see above):

```
> printMeanAndSdByGroup(standardisedconcentrations,wine[1])
[1] "Means:"
  V1         V2         V3         V4         V5          V6         V7         V8          V9         V10         V11         V12         V13         V14
1  1  0.9166093 -0.2915199  0.3246886 -0.7359212  0.46192317  0.87090552  0.95419225 -0.57735640  0.5388633  0.2028288  0.4575567  0.7691811  1.1711967
2  2 -0.8892116 -0.3613424 -0.4437061  0.2225094 -0.36354162 -0.05790375  0.05163434  0.01452785  0.0688079 -0.8503999  0.4323908  0.2446043 -0.7220731
3  3  0.1886265  0.8928122  0.2572190  0.5754413 -0.03004191 -0.98483874 -1.24923710  0.68817813 -0.7641311  1.0085728 -1.2019916 -1.3072623 -0.3715295
```

Does it make sense that the first principal component can separate cultivar 1 from cultivar 3? In cultivar 1, the mean values of V8 (0.954), V7 (0.871), V13 (0.769), V10 (0.539), V12 (0.458) and V14 (1.171) are very high compared to the mean values of V9 (-0.577), V3 (-0.292) and V5 (-0.736). In cultivar 3, the mean values of V8 (-1.249), V7 (-0.985), V13 (-1.307), V10 (-0.764), V12 (-1.202) and V14 (-0.372) are very low compared to the mean values of V9 (0.688), V3 (0.893) and V5 (0.575). Therefore, it does make sense that principal component 1 is a contrast between the concentrations of V8, V7, V13, V10, V12, and V14, and the concentrations of V9, V3 and V5; and that principal component 1 can separate cultivar 1 from cultivar 3.

Above, we intepreted the second principal component as a contrast between the concentrations of V11, V2, V14, V4, V6 and V3, and the concentration of V12. In the light of the mean values of these variables in the different cultivars, does it make sense that the second principal component can separate cultivar 2 from cultivars 1 and 3? In cultivar 1, the mean values of V11 (0.203), V2 (0.917), V14 (1.171), V4 (0.325), V6 (0.462) and V3 (-0.292) are not very different from the mean value of V12 (0.458). In cultivar 3, the mean values of V11 (1.009), V2 (0.189), V14 (-0.372), V4 (0.257), V6 (-0.030) and V3 (0.893) are also not very different from the mean value of V12 (-1.202). In contrast, in cultivar 2, the mean values of V11 (-0.850), V2 (-0.889), V14 (-0.722), V4 (-0.444), V6 (-0.364) and V3 (-0.361) are much less than the mean value of V12 (0.432). Therefore, it makes sense that principal component is a contrast between the concentrations of V11, V2, V14, V4, V6 and V3, and the concentration of V12; and that principal component 2 can separate cultivar 2 from cultivars 1 and 3.