Chapter 3

Sample Geometry and Random Sampling

3.1 Introduction

With the vector concepts introduced in the previous chapter, we can now delve deeper into the geometrical interpretations of the descriptive statistics $\bar{\mathbf{x}}$, $\mathbf{S}_n$, and $\mathbf{R}$; we do so in Section 3.2. Many of our explanations use the representation of the columns of $\mathbf{X}$ as p vectors in n dimensions. In Section 3.3 we introduce the assumption that the observations constitute a random sample. Simply stated, random sampling implies that (1) measurements taken on different items (or trials) are unrelated to one another and (2) the joint distribution of all p variables remains the same for all items. Ultimately, it is this structure of the random sample that justifies a particular choice of distance and dictates the geometry for the n-dimensional representation of the data. Furthermore, when data can be treated as a random sample, statistical inferences are based on a solid foundation.

Returning to geometric interpretations in Section 3.4, we introduce a single number, called *generalized variance*, to describe variability. This generalization of variance is an integral part of the comparison of multivariate means. In later sections we use matrix algebra to provide concise expressions for the matrix products and sums that allow us to calculate $\bar{\mathbf{x}}$ and $\mathbf{S}_n$ directly from the data matrix $\mathbf{X}$. The connection between $\bar{\mathbf{x}}$, $\mathbf{S}_n$, and the means and covariances for linear combinations of variables is also clearly delineated, using the notion of matrix products.

3.2 The Geometry of the Sample

A single multivariate observation is the collection of measurements on p different variables taken on the same item or trial. As in Chapter 1, if n observations have been obtained, the entire data set can be placed in an $n \times p$ array (matrix):

$$\underset{(n\times p)}{\mathbf{X}} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

Each row of $\mathbf{X}$ represents a multivariate observation. Since the entire set of measurements is often one particular realization of what might have been observed, we say that the data are a *sample* of size n from a p-variate "population." The sample then consists of n measurements, each of which has p components.

As we have seen, the data can be plotted in two different ways. For the p-dimensional scatter plot, the *rows* of $\mathbf{X}$ represent n points in p-dimensional space. We can write

$$\underset{(n\times p)}{\mathbf{X}} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix} \begin{matrix} \leftarrow \text{1st (multivariate) observation} \\ \\ \\ \leftarrow n\text{th (multivariate) observation} \end{matrix} \tag{3-1}$$

The row vector $\mathbf{x}_j'$, representing the jth observation, contains the coordinates of a point.

The scatter plot of n points in p-dimensional space provides information on the locations and variability of the points. If the points are regarded as solid spheres, the sample mean vector $\bar{\mathbf{x}}$, given by (1-8), is the center of balance. Variability occurs in more than one direction, and it is quantified by the sample variance–covariance matrix $\mathbf{S}_n$. A *single* numerical measure of variability is provided by the determinant of the sample variance–covariance matrix. When p is greater than 3, this scatter plot representation cannot actually be graphed. Yet the consideration of the data as n points in p dimensions provides insights that are not readily available from algebraic expressions. Moreover, the concepts illustrated for $p = 2$ or $p = 3$ remain valid for the other cases.

Example 3.1 (Computing the mean vector) Compute the mean vector $\bar{\mathbf{x}}$ from the data matrix.

$$\mathbf{X} = \begin{bmatrix} 4 & 1 \\ -1 & 3 \\ 3 & 5 \end{bmatrix}$$

Plot the $n = 3$ data points in $p = 2$ space, and locate $\bar{\mathbf{x}}$ on the resulting diagram.

The first point, $\mathbf{x}_1$, has coordinates $\mathbf{x}_1' = [4, 1]$. Similarly, the remaining two points are $\mathbf{x}_2' = [-1, 3]$ and $\mathbf{x}_3' = [3, 5]$. Finally,

$$\bar{\mathbf{x}} = \begin{bmatrix} \dfrac{4-1+3}{3} \\ \dfrac{1+3+5}{3} \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

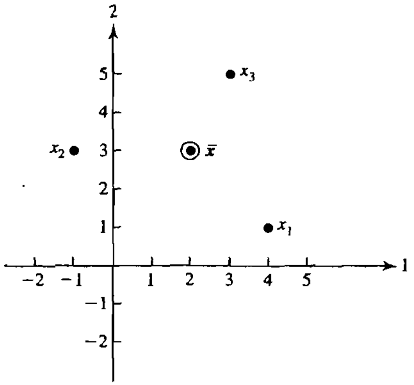


Figure 3.1 A plot of the data matrix $\mathbf{X}$ as $n = 3$ points in $p = 2$ space.

Figure 3.1 shows that $\bar{\mathbf{x}}$ is the balance point (center of gravity) of the scatter plot. ■

The alternative geometrical representation is constructed by considering the data as p vectors in n-dimensional space. Here we take the elements of the *columns* of the data matrix to be the coordinates of the vectors. Let

$$\underset{(n\times p)}{\mathbf{X}} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = [\mathbf{y}_1 \,\vdots\, \mathbf{y}_2 \,\vdots\, \cdots \,\vdots\, \mathbf{y}_p] \tag{3-2}$$

Then the coordinates of the first point $\mathbf{y}_1' = [x_{11}, x_{21}, \ldots, x_{n1}]$ are the n measurements on the first variable. In general, the ith point $\mathbf{y}_i' = [x_{1i}, x_{2i}, \ldots, x_{ni}]$ is determined by the n-tuple of all measurements on the ith variable. In this geometrical representation, we depict $\mathbf{y}_1, \ldots, \mathbf{y}_p$ as vectors rather than points, as in the p-dimensional scatter plot. We shall be manipulating these quantities shortly using the algebra of vectors discussed in Chapter 2.

Example 3.2 (Data as *p* vectors in *n* dimensions) Plot the following data as $p = 2$ vectors in $n = 3$ space:

$$\mathbf{X} = \begin{bmatrix} 4 & 1 \\ -1 & 3 \\ 3 & 5 \end{bmatrix}$$

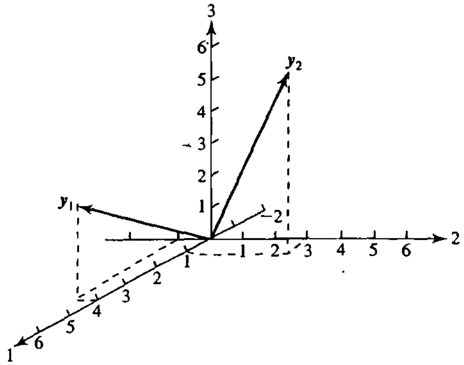


Figure 3.2 A plot of the data matrix $\mathbf{X}$ as $p = 2$ vectors in $n = 3$ space.

Here $\mathbf{y}_1' = [4, -1, 3]$ and $\mathbf{y}_2' = [1, 3, 5]$. These vectors are shown in Figure 3.2. ■

Many of the algebraic expressions we shall encounter in multivariate analysis can be related to the geometrical notions of length, angle, and volume. This is important because geometrical representations ordinarily facilitate understanding and lead to further insights.

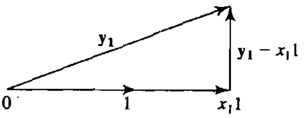
Unfortunately, we are limited to visualizing objects in three dimensions, and consequently, the n-dimensional representation of the data matrix $\mathbf{X}$ may not seem like a particularly useful device for $n > 3$. It turns out, however, that geometrical relationships and the associated statistical concepts depicted for any three vectors remain valid regardless of their dimension. This follows because three vectors, even if n dimensional, can span no more than a three-dimensional space, just as two vectors with any number of components must lie in a plane. By selecting an appropriate three-dimensional perspective—that is, a portion of the n-dimensional space containing the three vectors of interest—a view is obtained that preserves both lengths and angles. Thus, it is possible, with the right choice of axes, to illustrate certain algebraic statistical concepts in terms of only two or three vectors of any dimension n. Since the specific choice of axes is not relevant to the geometry, we shall always label the coordinate axes 1, 2, and 3.

It is possible to give a geometrical interpretation of the process of finding a sample mean. We start by defining the $n \times 1$ vector $\mathbf{1}_n' = [1, 1, \ldots, 1]$. (To simplify the notation, the subscript n will be dropped when the dimension of the vector $\mathbf{1}_n$ is clear from the context.) The vector $\mathbf{1}$ forms equal angles with each of the n coordinate axes, so the vector $(1/\sqrt{n})\mathbf{1}$ has unit length in the equal-angle direction. Consider the vector $\mathbf{y}_i' = [x_{1i}, x_{2i}, \ldots, x_{ni}]$. The projection of $\mathbf{y}_i$ on the unit vector $(1/\sqrt{n})\mathbf{1}$ is, by (2-8),

$$\mathbf{y}_i'\left(\frac{1}{\sqrt{n}}\mathbf{1}\right)\frac{1}{\sqrt{n}}\mathbf{1} = \frac{x_{1i} + x_{2i} + \cdots + x_{ni}}{n}\mathbf{1} = \bar{x}_i\mathbf{1} \qquad (3\text{-}3)$$

That is, the sample mean $\bar{x}_i = (x_{1i} + x_{2i} + \cdots + x_{ni})/n = \mathbf{y}_i'\mathbf{1}/n$ corresponds to the multiple of $\mathbf{1}$ required to give the projection of $\mathbf{y}_i$ onto the line determined by $\mathbf{1}$.

Further, for each $\mathbf{y}_i$, we have the decomposition



where $\bar{x}_i\mathbf{1}$ is perpendicular to $\mathbf{y}_i - \bar{x}_i\mathbf{1}$. The deviation, or mean corrected, vector is

$$\mathbf{d}_i = \mathbf{y}_i - \bar{x}_i\mathbf{1} = \begin{bmatrix} x_{1i} - \bar{x}_i \\ x_{2i} - \bar{x}_i \\ \vdots \\ x_{ni} - \bar{x}_i \end{bmatrix} \tag{3-4}$$

The elements of $\mathbf{d}_i$ are the deviations of the measurements on the ith variable from their sample mean. Decomposition of the $\mathbf{y}_i$ vectors into mean components and deviation from the mean components is shown in Figure 3.3 for $p = 3$ and $n = 3$.

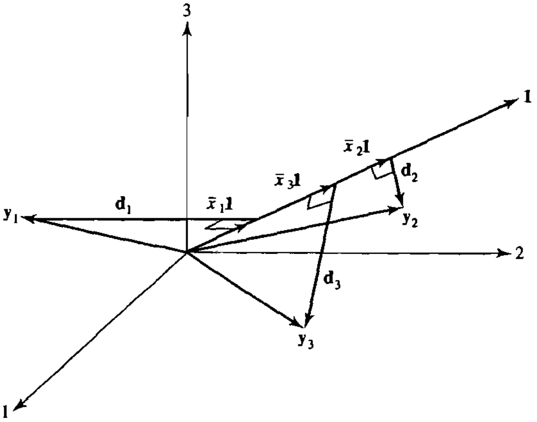


Figure 3.3 The decomposition of $\mathbf{y}_i$ into a mean component $\bar{x}_i\mathbf{1}$ and a deviation component $\mathbf{d}_i = \mathbf{y}_i - \bar{x}_i\mathbf{1}$, $i = 1, 2, 3$.

Example 3.3 (Decomposing a vector into its mean and deviation components) Let us carry out the decomposition of $\mathbf{y}_i$ into $\bar{x}_i\mathbf{1}$ and $\mathbf{d}_i = \mathbf{y}_i - \bar{x}_i\mathbf{1}$, $i = 1, 2$, for the data given in Example 3.2:

$$\mathbf{X} = \begin{bmatrix} 4 & 1 \\ -1 & 3 \\ 3 & 5 \end{bmatrix}$$

Here, $\bar{x}_1 = (4 - 1 + 3)/3 = 2$ and $\bar{x}_2 = (1 + 3 + 5)/3 = 3$, so

$$\bar{x}_1\mathbf{1} = 2\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix} \qquad \bar{x}_2\mathbf{1} = 3\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix}$$

Consequently,

$$\mathbf{d}_1 = \mathbf{y}_1 - \bar{x}_1\mathbf{1} = \begin{bmatrix} 4 \\ -1 \\ 3 \end{bmatrix} - \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 2 \\ -3 \\ 1 \end{bmatrix}$$

and

$$\mathbf{d}_2 = \mathbf{y}_2 - \bar{x}_2\mathbf{1} = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix} - \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix} = \begin{bmatrix} -2 \\ 0 \\ 2 \end{bmatrix}$$

We note that $\bar{x}_1\mathbf{1}$ and $\mathbf{d}_1 = \mathbf{y}_1 - \bar{x}_1\mathbf{1}$ are perpendicular, because

$$(\bar{x}_1\mathbf{1})'(\mathbf{y}_1 - \bar{x}_1\mathbf{1}) = [2 \quad 2 \quad 2]\begin{bmatrix} 2 \\ -3 \\ 1 \end{bmatrix} = 4 - 6 + 2 = 0$$

A similar result holds for $\bar{x}_2\mathbf{1}$ and $\mathbf{d}_2 = \mathbf{y}_2 - \bar{x}_2\mathbf{1}$. The decomposition is

$$\mathbf{y}_1 = \begin{bmatrix} 4 \\ -1 \\ 3 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix} + \begin{bmatrix} 2 \\ -3 \\ 1 \end{bmatrix}$$

$$\mathbf{y}_2 = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix} + \begin{bmatrix} -2 \\ 0 \\ 2 \end{bmatrix}$$ ■

For the time being, we are interested in the deviation (or residual) vectors $\mathbf{d}_i = \mathbf{y}_i - \bar{x}_i\mathbf{1}$. A plot of the deviation vectors of Figure 3.3 is given in Figure 3.4.

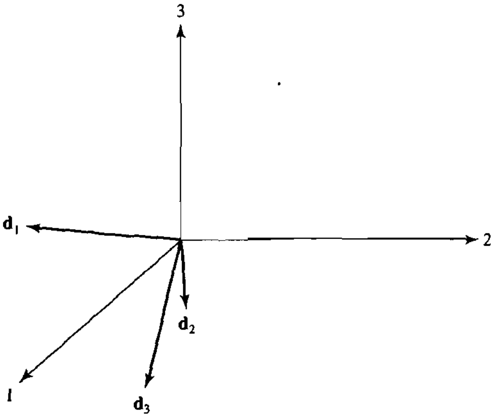


Figure 3.4 The deviation vectors $\mathbf{d}_i$ from Figure 3.3.

We have translated the deviation vectors to the origin without changing their lengths or orientations.

Now consider the squared lengths of the deviation vectors. Using (2-5) and (3-4), we obtain

$$L_{\mathbf{d}_i}^2 = \mathbf{d}_i'\mathbf{d}_i = \sum_{j=1}^{n} (x_{ji} - \bar{x}_i)^2 \tag{3-5}$$

$$(\text{Length of deviation vector})^2 = \text{sum of squared deviations}$$

From (1-3), we see that the squared length is proportional to the variance of the measurements on the ith variable. Equivalently, the *length* is proportional to the *standard deviation*. Longer vectors represent more variability than shorter vectors.

For any two deviation vectors $\mathbf{d}_i$ and $\mathbf{d}_k$,

$$\mathbf{d}_i'\mathbf{d}_k = \sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k) \tag{3-6}$$

Let θ_{ik} denote the angle formed by the vectors $\mathbf{d}_i$ and $\mathbf{d}_k$. From (2-6), we get

$$\mathbf{d}_i'\mathbf{d}_k = L_{\mathbf{d}_i} L_{\mathbf{d}_k} \cos(\theta_{ik})$$

or, using (3-5) and (3-6), we obtain

$$\sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k) = \sqrt{\sum_{j=1}^{n} (x_{ji} - \bar{x}_i)^2}\,\sqrt{\sum_{j=1}^{n} (x_{jk} - \bar{x}_k)^2}\,\cos(\theta_{ik})$$

so that [see (1-5)]

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\,\sqrt{s_{kk}}} = \cos(\theta_{ik}) \tag{3-7}$$

The *cosine* of the angle is the sample *correlation coefficient*. Thus, if the two deviation vectors have nearly the same orientation, the sample correlation will be close to 1. If the two vectors are nearly perpendicular, the sample correlation will be approximately zero. If the two vectors are oriented in nearly opposite directions, the sample correlation will be close to -1.

Example 3.4 (Calculating $\mathbf{S}_n$ and $\mathbf{R}$ from deviation vectors) Given the deviation vectors in Example 3.3, let us compute the sample variance–covariance matrix $\mathbf{S}_n$ and sample correlation matrix $\mathbf{R}$ using the geometrical concepts just introduced.

From Example 3.3,

$$\mathbf{d}_1 = \begin{bmatrix} 2 \\ -3 \\ 1 \end{bmatrix} \quad \text{and} \quad \mathbf{d}_2 = \begin{bmatrix} -2 \\ 0 \\ 2 \end{bmatrix}$$

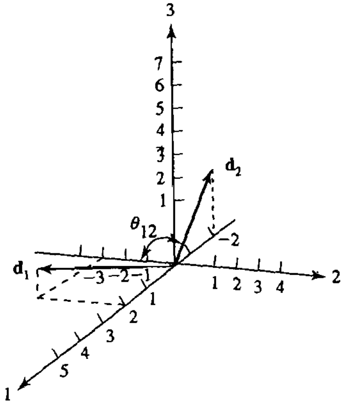


Figure 3.5 The deviation vectors $\mathbf{d}_1$ and $\mathbf{d}_2$.

These vectors, translated to the origin, are shown in Figure 3.5. Now,

$$\mathbf{d}_1'\mathbf{d}_1 = [2 \quad -3 \quad 1]\begin{bmatrix}2\\-3\\1\end{bmatrix} = 14 = 3s_{11}$$

or $s_{11} = \frac{14}{3}$. Also,

$$\mathbf{d}_2'\mathbf{d}_2 = [-2 \quad 0 \quad 2]\begin{bmatrix}-2\\0\\2\end{bmatrix} = 8 = 3s_{22}$$

or $s_{22} = \frac{8}{3}$. Finally,

$$\mathbf{d}_1'\mathbf{d}_2 = [2 \quad -3 \quad 1]\begin{bmatrix}-2\\0\\2\end{bmatrix} = -2 = 3s_{12}$$

or $s_{12} = -\frac{2}{3}$. Consequently,

$$r_{12} = \frac{s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}} = \frac{-\frac{2}{3}}{\sqrt{\frac{14}{3}}\sqrt{\frac{8}{3}}} = -.189$$

and

$$\mathbf{S}_n = \begin{bmatrix}\frac{14}{3} & -\frac{2}{3}\\ -\frac{2}{3} & \frac{8}{3}\end{bmatrix}, \qquad \mathbf{R} = \begin{bmatrix}1 & -.189\\ -.189 & 1\end{bmatrix}$$ ■

The concepts of length, angle, and projection have provided us with a geometrical interpretation of the sample. We summarize as follows:

Geometrical Interpretation of the Sample

1. The projection of a column $\mathbf{y}_i$ of the data matrix $\mathbf{X}$ onto the equal angular vector $\mathbf{1}$ is the vector $\bar{x}_i\mathbf{1}$. The vector $\bar{x}_i\mathbf{1}$ has length $\sqrt{n}\,|\,\bar{x}_i\,|$. Therefore, the ith sample mean, $\bar{x}_i$, is related to the length of the projection of $\mathbf{y}_i$ on $\mathbf{1}$.
2. The information comprising $\mathbf{S}_n$ is obtained from the deviation vectors $\mathbf{d}_i = \mathbf{y}_i - \bar{x}_i\mathbf{1} = [x_{1i} - \bar{x}_i, \bar{x}_{2i} - \bar{x}_i, \ldots, x_{ni} - \bar{x}_i]'$. The square of the length of $\mathbf{d}_i$ is ns_{ii}, and the (inner) product between $\mathbf{d}_i$ and $\mathbf{d}_k$ is ns_{ik}.[1]
3. The sample correlation r_{ik} is the cosine of the angle between $\mathbf{d}_i$ and $\mathbf{d}_k$.

3.3 Random Samples and the Expected Values of the Sample Mean and Covariance Matrix

In order to study the sampling variability of statistics such as $\bar{\mathbf{x}}$ and $\mathbf{S}_n$ with the ultimate aim of making inferences, we need to make assumptions about the variables whose observed values constitute the data set $\mathbf{X}$.

Suppose, then, that the data have not yet been observed, but we *intend* to collect n sets of measurements on p variables. Before the measurements are made, their values cannot, in general, be predicted exactly. Consequently, we treat them as random variables. In this context, let the (j, k)-th entry in the data matrix be the random variable X_{jk}. Each set of measurements $\mathbf{X}_j$ on p variables is a random vector, and we have the random matrix

$$\underset{(n\times p)}{\mathbf{X}} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \\ \vdots \\ \mathbf{X}_n' \end{bmatrix} \tag{3-8}$$

A *random sample* can now be defined.

If the row vectors $\mathbf{X}_1', \mathbf{X}_2', \ldots, \mathbf{X}_n'$ in (3-8) represent *independent* observations from a *common* joint distribution with density function $f(\mathbf{x}) = f(x_1, x_2, \ldots, x_p)$, then $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ are said to form a *random sample* from $f(\mathbf{x})$. Mathematically, $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ form a random sample if their joint density function is given by the product $f(\mathbf{x}_1)f(\mathbf{x}_2)\cdots f(\mathbf{x}_n)$, where $f(\mathbf{x}_j) = f(x_{j1}, x_{j2}, \ldots, x_{jp})$ is the density function for the jth row vector.

Two points connected with the definition of random sample merit special attention:

1. The measurements of the p variables in a *single* trial, such as $\mathbf{X}_j' = [X_{j1}, X_{j2}, \ldots, X_{jp}]$, will usually be correlated. Indeed, we expect this to be the case. The measurements from *different* trials must, however, be independent.

[1] The square of the length and the inner product are $(n - 1)s_{ii}$ and $(n - 1)s_{ik}$, respectively, when the divisor $n - 1$ is used in the definitions of the sample variance and covariance.

2. The independence of measurements from trial to trial may not hold when the variables are likely to drift over time, as with sets of p stock prices or p economic indicators. Violations of the tentative assumption of independence can have a serious impact on the quality of statistical inferences.

The following examples illustrate these remarks.

Example 3.5 (Selecting a random sample) As a preliminary step in designing a permit system for utilizing a wilderness canoe area without overcrowding, a natural-resource manager took a survey of users. The total wilderness area was divided into subregions, and respondents were asked to give information on the regions visited, lengths of stay, and other variables.

The method followed was to select persons randomly (perhaps using a random number table) from all those who entered the wilderness area during a particular week. All persons were equally likely to be in the sample, so the more popular entrances were represented by larger proportions of canoeists.

Here one would expect the sample observations to conform closely to the criterion for a random sample from the population of users or potential users. On the other hand, if one of the samplers had waited at a campsite far in the interior of the area and interviewed only canoeists who reached that spot, successive measurements would not be independent. For instance, lengths of stay in the wilderness area for different canoeists from this group would all tend to be large. ■

Example 3.6 (A nonrandom sample) Because of concerns with future solid-waste disposal, an ongoing study concerns the gross weight of municipal solid waste generated per year in the United States (Environmental Protection Agency). Estimated amounts attributed to x_1 = paper and paperboard waste and x_2 = plastic waste, in millions of tons, are given for selected years in Table 3.1. Should these measurements on $\mathbf{X}' = [X_1, X_2]$ be treated as a random sample of size $n = 7$? No! In fact, except for a slight but fortunate downturn in paper and paperboard waste in 2003, *both* variables are increasing over time.

Table 3.1 Solid Waste

Year	1960	1970	1980	1990	1995	2000	2003
x_1 (paper)	29.2	44.3	55.2	72.7	81.7	87.7	83.1
x_2 (plastics)	.4	2.9	6.8	17.1	18.9	24.7	26.7

■

As we have argued heuristically in Chapter 1, the notion of statistical independence has important implications for measuring distance. Euclidean distance appears appropriate if the components of a vector are independent and have the same variances. Suppose we consider the location of the kth column $\mathbf{Y}'_k = [X_{1k}, X_{2k}, \ldots, X_{nk}]$ of $\mathbf{X}$, regarded as a point in n dimensions. The location of this point is determined by the joint probability distribution $f(\mathbf{y}_k) = f(x_{1k}, x_{2k}, \ldots, x_{nk})$. When the measurements $X_{1k}, X_{2k}, \ldots, X_{nk}$ are a random sample, $f(\mathbf{y}_k) = f(x_{1k}, x_{2k}, \ldots, x_{nk}) = f_k(x_{1k})f_k(x_{2k}) \cdots f_k(x_{nk})$ and, consequently, each coordinate x_{jk} contributes equally to the location through the identical marginal distributions $f_k(x_{jk})$.

If the n components are not independent or the marginal distributions are not identical, the influence of individual measurements (coordinates) on location is asymmetrical. We would then be led to consider a distance function in which the coordinates were weighted unequally, as in the "statistical" distances or quadratic forms introduced in Chapters 1 and 2.

Certain conclusions can be reached concerning the sampling distributions of $\overline{\mathbf{X}}$ and $\mathbf{S}_n$ without making further assumptions regarding the form of the underlying joint distribution of the variables. In particular, we can see how $\overline{\mathbf{X}}$ and $\mathbf{S}_n$ fare as point estimators of the corresponding population mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

Result 3.1. Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be a random sample from a joint distribution that has mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Then $\overline{\mathbf{X}}$ is an *unbiased* estimator of $\boldsymbol{\mu}$, and its covariance matrix is

$$\frac{1}{n}\boldsymbol{\Sigma}$$

That is,

$$\begin{aligned} E(\overline{\mathbf{X}}) &= \boldsymbol{\mu} && \text{(population mean vector)} \\ \operatorname{Cov}(\overline{\mathbf{X}}) &= \frac{1}{n}\boldsymbol{\Sigma} && \left(\begin{array}{c}\text{population variance–covariance matrix} \\ \text{divided by sample size}\end{array}\right) \end{aligned} \tag{3-9}$$

For the covariance matrix $\mathbf{S}_n$,

$$E(\mathbf{S}_n) = \frac{n-1}{n}\boldsymbol{\Sigma} = \boldsymbol{\Sigma} - \frac{1}{n}\boldsymbol{\Sigma}$$

Thus,

$$E\left(\frac{n}{n-1}\mathbf{S}_n\right) = \boldsymbol{\Sigma} \tag{3-10}$$

so $[n/(n-1)]\mathbf{S}_n$ is an *unbiased* estimator of $\boldsymbol{\Sigma}$, while $\mathbf{S}_n$ is a *biased* estimator with (bias) $= E(\mathbf{S}_n) - \boldsymbol{\Sigma} = -(1/n)\boldsymbol{\Sigma}$.

Proof. Now, $\overline{\mathbf{X}} = (\mathbf{X}_1 + \mathbf{X}_2 + \cdots + \mathbf{X}_n)/n$. The repeated use of the properties of expectation in (2-24) for two vectors gives

$$\begin{aligned} E(\overline{\mathbf{X}}) &= E\left(\frac{1}{n}\mathbf{X}_1 + \frac{1}{n}\mathbf{X}_2 + \cdots + \frac{1}{n}\mathbf{X}_n\right) \\ &= E\left(\frac{1}{n}\mathbf{X}_1\right) + E\left(\frac{1}{n}\mathbf{X}_2\right) + \cdots + E\left(\frac{1}{n}\mathbf{X}_n\right) \\ &= \frac{1}{n}E(\mathbf{X}_1) + \frac{1}{n}E(\mathbf{X}_2) + \cdots + \frac{1}{n}E(\mathbf{X}_n) = \frac{1}{n}\boldsymbol{\mu} + \frac{1}{n}\boldsymbol{\mu} + \cdots + \frac{1}{n}\boldsymbol{\mu} \\ &= \boldsymbol{\mu} \end{aligned}$$

Next,

$$\begin{aligned} (\overline{\mathbf{X}} - \boldsymbol{\mu})(\overline{\mathbf{X}} - \boldsymbol{\mu})' &= \left(\frac{1}{n}\sum_{j=1}^{n}(\mathbf{X}_j - \boldsymbol{\mu})\right)\left(\frac{1}{n}\sum_{\ell=1}^{n}(\mathbf{X}_\ell - \boldsymbol{\mu})\right)' \\ &= \frac{1}{n^2}\sum_{j=1}^{n}\sum_{\ell=1}^{n}(\mathbf{X}_j - \boldsymbol{\mu})(\mathbf{X}_\ell - \boldsymbol{\mu})' \end{aligned}$$

so

$$\text{Cov}(\overline{\mathbf{X}}) = E(\overline{\mathbf{X}} - \boldsymbol{\mu})(\overline{\mathbf{X}} - \boldsymbol{\mu})' = \frac{1}{n^2}\left(\sum_{j=1}^{n}\sum_{\ell=1}^{n} E(\mathbf{X}_j - \boldsymbol{\mu})(\mathbf{X}_\ell - \boldsymbol{\mu})'\right)$$

For $j \neq \ell$, each entry in $E(\mathbf{X}_j - \boldsymbol{\mu})(\mathbf{X}_\ell - \boldsymbol{\mu})'$ is zero because the entry is the covariance between a component of $\mathbf{X}_j$ and a component of $\mathbf{X}_\ell$, and these are independent. [See Exercise 3.17 and (2-29).]

Therefore,

$$\text{Cov}(\overline{\mathbf{X}}) = \frac{1}{n^2}\left(\sum_{j=1}^{n} E(\mathbf{X}_j - \boldsymbol{\mu})(\mathbf{X}_j - \boldsymbol{\mu})'\right)$$

Since $\boldsymbol{\Sigma} = E(\mathbf{X}_j - \boldsymbol{\mu})(\mathbf{X}_j - \boldsymbol{\mu})'$ is the common population covariance matrix for each $\mathbf{X}_j$, we have

$$\begin{aligned}\text{Cov}(\overline{\mathbf{X}}) &= \frac{1}{n^2}\left(\sum_{j=1}^{n} E(\mathbf{X}_j - \boldsymbol{\mu})(\mathbf{X}_j - \boldsymbol{\mu})'\right) = \frac{1}{n^2}\underbrace{(\boldsymbol{\Sigma} + \boldsymbol{\Sigma} + \cdots + \boldsymbol{\Sigma})}_{n \text{ terms}} \\ &= \frac{1}{n^2}(n\boldsymbol{\Sigma}) = \left(\frac{1}{n}\right)\boldsymbol{\Sigma}\end{aligned}$$

To obtain the expected value of $\mathbf{S}_n$, we first note that $(X_{ji} - \overline{X}_i)(X_{jk} - \overline{X}_k)$ is the (i, k)th element of $(\mathbf{X}_j - \overline{\mathbf{X}})(\mathbf{X}_j - \overline{\mathbf{X}})'$. The matrix representing sums of squares and cross products can then be written as

$$\begin{aligned}\sum_{j=1}^{n}(\mathbf{X}_j - \overline{\mathbf{X}})(\mathbf{X}_j - \overline{\mathbf{X}})' &= \sum_{j=1}^{n}(\mathbf{X}_j - \overline{\mathbf{X}})\mathbf{X}_j' + \left(\sum_{j=1}^{n}(\mathbf{X}_j - \overline{\mathbf{X}})\right)(-\overline{\mathbf{X}})' \\ &= \sum_{j=1}^{n}\mathbf{X}_j\mathbf{X}_j' - n\overline{\mathbf{X}}\,\overline{\mathbf{X}}'\end{aligned}$$

since $\sum_{j=1}^{n}(\mathbf{X}_j - \overline{\mathbf{X}}) = \mathbf{0}$ and $n\overline{\mathbf{X}}' = \sum_{i=1}^{n}\mathbf{X}_j'$. Therefore, its expected value is

$$E\left(\sum_{j=1}^{n}\mathbf{X}_j\mathbf{X}_j' - n\overline{\mathbf{X}}\overline{\mathbf{X}}'\right) = \sum_{j=1}^{n}E(\mathbf{X}_j\mathbf{X}_j') - nE(\overline{\mathbf{X}}\overline{\mathbf{X}}')$$

For any random vector $\mathbf{V}$ with $E(\mathbf{V}) = \boldsymbol{\mu}_V$ and $\text{Cov}(\mathbf{V}) = \boldsymbol{\Sigma}_V$, we have $E(\mathbf{V}\mathbf{V}') = \boldsymbol{\Sigma}_V + \boldsymbol{\mu}_V\boldsymbol{\mu}_V'$. (See Exercise 3.16.) Consequently,

$$E(\mathbf{X}_j\mathbf{X}_j') = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}' \quad \text{and} \quad E(\overline{\mathbf{X}}\overline{\mathbf{X}}') = \frac{1}{n}\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}'$$

Using these results, we obtain

$$\sum_{j=1}^{n}E(\mathbf{X}_j\mathbf{X}_j') - nE(\overline{\mathbf{X}}\overline{\mathbf{X}}') = n\boldsymbol{\Sigma} + n\boldsymbol{\mu}\boldsymbol{\mu}' - n\left(\frac{1}{n}\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}'\right) = (n-1)\boldsymbol{\Sigma}$$

and thus, since $\mathbf{S}_n = (1/n)\left(\sum_{j=1}^{n}\mathbf{X}_j\mathbf{X}_j' - n\overline{\mathbf{X}}\overline{\mathbf{X}}'\right)$, it follows immediately that

$$E(\mathbf{S}_n) = \frac{(n-1)}{n}\boldsymbol{\Sigma} \qquad \blacksquare$$

Result 3.1 shows that the (i, k)th entry, $(n - 1)^{-1} \sum_{j=1}^{n} (X_{ji} - \bar{X}_i)(X_{jk} - \bar{X}_k)$, of $[n/(n - 1)]\mathbf{S}_n$ is an unbiased estimator of σ_{ik}. However, the individual sample standard deviations $\sqrt{s_{ii}}$, calculated with either n or $n - 1$ as a divisor, are not unbiased estimators of the corresponding population quantities $\sqrt{\sigma_{ii}}$. Moreover, the correlation coefficients r_{ik} are *not* unbiased estimators of the population quantities ρ_{ik}. However, the bias $E\left(\sqrt{s_{ii}}\right) - \sqrt{\sigma_{ii}}$, or $E(r_{ik}) - \rho_{ik}$, can usually be ignored if the sample size n is moderately large.

Consideration of bias motivates a slightly modified definition of the sample variance–covariance matrix. Result 3.1 provides us with an unbiased estimator $\mathbf{S}$ of $\mathbf{\Sigma}$:

(Unbiased) Sample Variance–Covariance Matrix

$$\mathbf{S} = \left(\frac{n}{n-1}\right)\mathbf{S}_n = \frac{1}{n-1}\sum_{j=1}^{n}(\mathbf{X}_j - \bar{\mathbf{X}})(\mathbf{X}_j - \bar{\mathbf{X}})' \qquad (3\text{-}11)$$

Here S, without a subscript, has (i, k)th entry $(n - 1)^{-1} \sum_{j=1}^{n} (X_{ji} - \bar{X}_i)(X_{jk} - \bar{X}_k)$. This definition of sample covariance is commonly used in many multivariate test statistics. Therefore, it will replace $\mathbf{S}_n$ as the sample covariance matrix in most of the material throughout the rest of this book.

3.4 Generalized Variance

With a single variable, the sample variance is often used to describe the amount of variation in the measurements on that variable. When p variables are observed on each unit, the variation is described by the sample variance–covariance matrix

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{12} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{1p} & s_{2p} & \cdots & s_{pp} \end{bmatrix} = \left\{ s_{ik} = \frac{1}{n-1}\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k) \right\}$$

The sample covariance matrix contains p variances and $\frac{1}{2}p(p - 1)$ potentially different covariances. Sometimes it is desirable to assign a *single* numerical value for the variation expressed by $\mathbf{S}$. One choice for a value is the determinant of $\mathbf{S}$, which reduces to the usual sample variance of a single characteristic when $p = 1$. This determinant[2] is called the *generalized sample variance:*

$$\text{Generalized sample variance} = |\mathbf{S}| \qquad (3\text{-}12)$$

[2] Definition 2A.24 defines "determinant" and indicates one method for calculating the value of a determinant.

Example 3.7 (Calculating a generalized variance) Employees (x_1) and profits per employee (x_2) for the 16 largest publishing firms in the United States are shown in Figure 1.3. The sample covariance matrix, obtained from the data in the April 30, 1990, *Forbes* magazine article, is

$$\mathbf{S} = \begin{bmatrix} 252.04 & -68.43 \\ -68.43 & 123.67 \end{bmatrix}$$

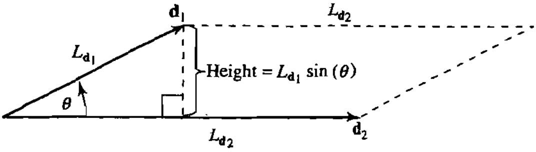
Evaluate the generalized variance.

In this case, we compute

$$|\mathbf{S}| = (252.04)(123.67) - (-68.43)(-68.43) = 26{,}487$$ ■

The generalized sample variance provides one way of writing the information on all variances and covariances as a single number. Of course, when $p > 1$, some information about the sample is lost in the process. A geometrical interpretation of $|\mathbf{S}|$ will help us appreciate its strengths and weaknesses as a descriptive summary.

Consider the area generated within the plane by two deviation vectors $\mathbf{d}_1 = \mathbf{y}_1 - \bar{x}_1\mathbf{1}$ and $\mathbf{d}_2 = \mathbf{y}_2 - \bar{x}_2\mathbf{1}$. Let $L_{\mathbf{d}_1}$ be the length of $\mathbf{d}_1$ and $L_{\mathbf{d}_2}$ the length of $\mathbf{d}_2$. By elementary geometry, we have the diagram



and the area of the trapezoid is $| L_{\mathbf{d}_1} \sin(\theta) | L_{\mathbf{d}_2}$. Since $\cos^2(\theta) + \sin^2(\theta) = 1$, we can express this area as

$$\text{Area} = L_{\mathbf{d}_1} L_{\mathbf{d}_2} \sqrt{1 - \cos^2(\theta)}$$

From (3-5) and (3-7),

$$L_{\mathbf{d}_1} = \sqrt{\sum_{j=1}^{n} (x_{j1} - \bar{x}_1)^2} = \sqrt{(n-1)s_{11}}$$

$$L_{\mathbf{d}_2} = \sqrt{\sum_{j=1}^{n} (x_{j2} - \bar{x}_2)^2} = \sqrt{(n-1)s_{22}}$$

and

$$\cos(\theta) = r_{12}$$

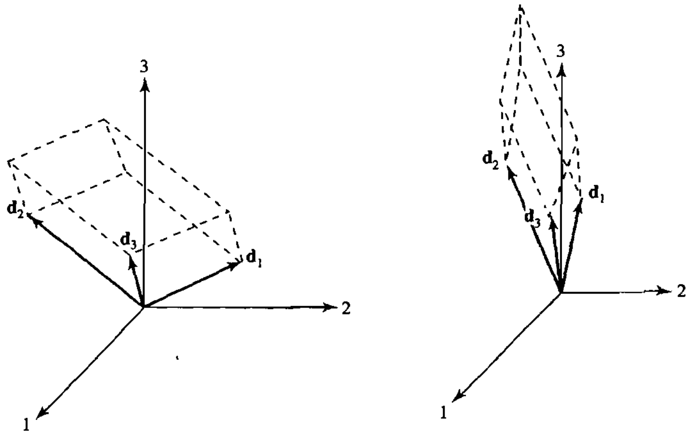
Therefore,

$$\text{Area} = (n-1)\sqrt{s_{11}}\sqrt{s_{22}}\sqrt{1 - r_{12}^2} = (n-1)\sqrt{s_{11}s_{22}(1 - r_{12}^2)} \qquad (3\text{-}13)$$

Also,

$$|\mathbf{S}| = \left| \begin{bmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{bmatrix} \right| = \left| \begin{bmatrix} s_{11} & \sqrt{s_{11}}\sqrt{s_{22}}r_{12} \\ \sqrt{s_{11}}\sqrt{s_{22}}r_{12} & s_{22} \end{bmatrix} \right|$$

$$= s_{11}s_{22} - s_{11}s_{22}r_{12}^2 = s_{11}s_{22}(1 - r_{12}^2) \qquad (3\text{-}14)$$



(a) (b)

Figure 3.6 (a) "Large" generalized sample variance for $p = 3$. (b) "Small" generalized sample variance for $p = 3$.

If we compare (3-14) with (3-13), we see that

$$|\mathbf{S}| = (\text{area})^2/(n-1)^2$$

Assuming now that $|\mathbf{S}| = (n-1)^{-(p-1)}(\text{volume})^2$ holds for the volume generated in n space by the $p - 1$ deviation vectors $\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_{p-1}$, we can establish the following general result for p deviation vectors by induction (see [1], p. 266):

$$\text{Generalized sample variance} = |\mathbf{S}| = (n-1)^{-p}(\text{volume})^2 \qquad (3\text{-}15)$$

Equation (3-15) says that the generalized sample variance, for a fixed set of data, is proportional to the square of the volume generated by the p deviation vectors[3] $\mathbf{d}_1 = \mathbf{y}_1 - \bar{x}_1\mathbf{1}$, $\mathbf{d}_2 = \mathbf{y}_2 - \bar{x}_2\mathbf{1}, \ldots, \mathbf{d}_p = \mathbf{y}_p - \bar{x}_p\mathbf{1}$. Figures 3.6(a) and (b) show trapezoidal regions, generated by $p = 3$ residual vectors, corresponding to "large" and "small" generalized variances.

For a fixed sample size, it is clear from the geometry that volume, or $|\mathbf{S}|$, will increase when the length of any $\mathbf{d}_i = \mathbf{y}_i - \bar{x}_i\mathbf{1}$ (or $\sqrt{s_{ii}}$) is increased. In addition, volume will increase if the residual vectors of fixed length are moved until they are at right angles to one another, as in Figure 3.6(a). On the other hand, the volume, or $|\mathbf{S}|$, will be small if just one of the s_{ii} is small or one of the deviation vectors lies nearly in the (hyper) plane formed by the others, or both. In the second case, the trapezoid has very little height above the plane. This is the situation in Figure 3.6(b), where $\mathbf{d}_3$ lies nearly in the plane formed by $\mathbf{d}_1$ and $\mathbf{d}_2$.

[3] If generalized variance is defined in terms of the sample covariance matrix $\mathbf{S}_n = [(n-1)/n]\mathbf{S}$, then, using Result 2A.11, $|\mathbf{S}_n| = |[(n-1)/n]\mathbf{I}_p\mathbf{S}| = |[(n-1)/n]\mathbf{I}_p||\mathbf{S}| = [(n-1)/n]^p|\mathbf{S}|$. Consequently, using (3-15), we can also write the following: Generalized sample variance $= |\mathbf{S}_n| = n^{-p}(\text{volume})^2$.

Generalized variance also has interpretations in the p-space scatter plot representation of the data. The most intuitive interpretation concerns the spread of the scatter about the sample mean point $\bar{\mathbf{x}}' = [\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p]$. Consider the measure of distance given in the comment below (2-19), with $\bar{\mathbf{x}}$ playing the role of the fixed point $\boldsymbol{\mu}$ and $\mathbf{S}^{-1}$ playing the role of $\mathbf{A}$. With these choices, the coordinates $\mathbf{x}' = [x_1, x_2, \ldots, x_p]$ of the points a constant distance c from $\bar{\mathbf{x}}$ satisfy

$$(\mathbf{x} - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}) = c^2 \tag{3-16}$$

[When $p = 1$, $(\mathbf{x} - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}) = (x_1 - \bar{x}_1)^2/s_{11}$ is the squared distance from x_1 to $\bar{x}_1$ in standard deviation units.]

Equation (3-16) defines a hyperellipsoid (an ellipse if $p = 2$) centered at $\bar{\mathbf{x}}$. It can be shown using integral calculus that the volume of this hyperellipsoid is related to $|\mathbf{S}|$. In particular,

$$\text{Volume of } \{\mathbf{x}: (\mathbf{x} - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}) \le c^2\} = k_p|\mathbf{S}|^{1/2}c^p \tag{3-17}$$

or

$$(\text{Volume of ellipsoid})^2 = (\text{constant})\,(\text{generalized sample variance})$$

where the constant k_p is rather formidable.[4] A large volume corresponds to a large generalized variance.

Although the generalized variance has some intuitively pleasing geometrical interpretations, it suffers from a basic weakness as a descriptive summary of the sample covariance matrix $\mathbf{S}$, as the following example shows.

Example 3.8 (Interpreting the generalized variance) Figure 3.7 gives three scatter plots with very different patterns of correlation.

All three data sets have $\bar{\mathbf{x}}' = [2, 1]$, and the covariance matrices are

$$\mathbf{S} = \begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix}, r = .8 \quad \mathbf{S} = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}, r = 0 \quad \mathbf{S} = \begin{bmatrix} 5 & -4 \\ -4 & 5 \end{bmatrix}, r = -.8$$

Each covariance matrix $\mathbf{S}$ contains the information on the variability of the component variables and also the information required to calculate the correlation coefficient. In this sense, $\mathbf{S}$ captures the orientation and size of the pattern of scatter.

The eigenvalues and eigenvectors extracted from $\mathbf{S}$ further describe the pattern in the scatter plot. For

$$\mathbf{S} = \begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix}, \quad \text{the eigenvalues satisfy} \quad \begin{aligned} 0 &= (\lambda - 5)^2 - 4^2 \\ &= (\lambda - 9)(\lambda - 1) \end{aligned}$$

[4] For those who are curious, $k_p = 2\pi^{p/2}/p\,\Gamma(p/2)$, where $\Gamma(z)$ denotes the gamma function evaluated at z.

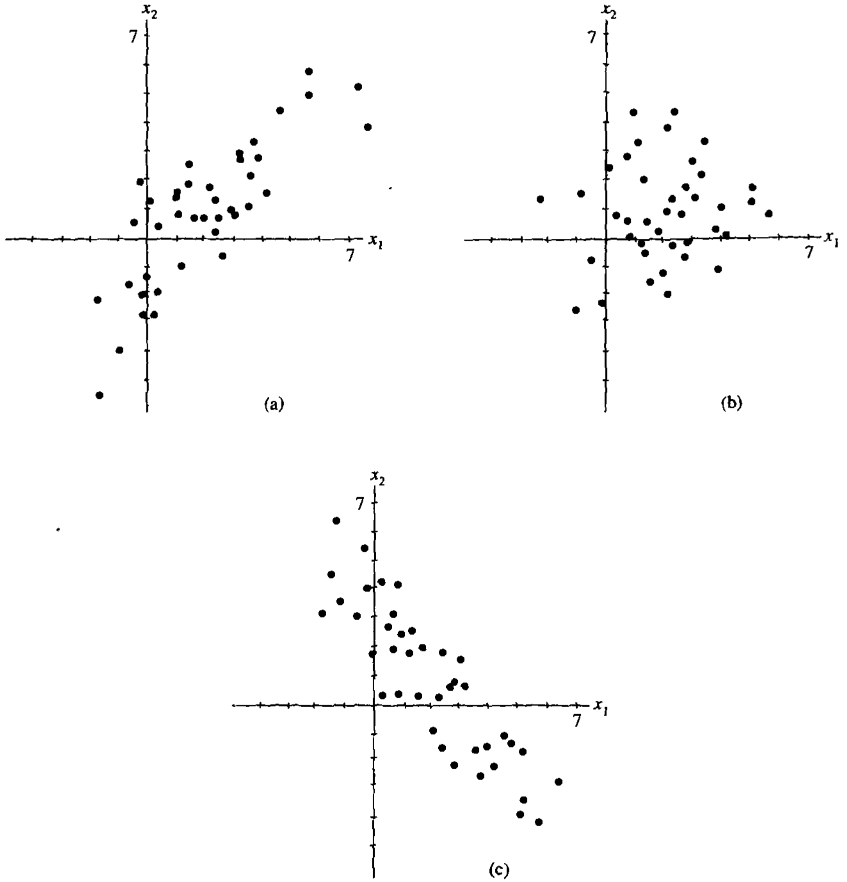


Figure 3.7 Scatter plots with three different orientations.

and we determine the eigenvalue–eigenvector pairs $\lambda_1 = 9$, $\mathbf{e}_1' = \left[1/\sqrt{2}, 1/\sqrt{2}\right]$ and $\lambda_2 = 1$, $\mathbf{e}_2' = \left[1/\sqrt{2}, -1/\sqrt{2}\right]$.

The mean-centered ellipse, with center $\bar{\mathbf{x}}' = [2, 1]$ for all three cases, is

$$(\mathbf{x} - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}) \leq c^2$$

To describe this ellipse, as in Section 2.3, with $\mathbf{A} = \mathbf{S}^{-1}$, we notice that if $(\lambda, \mathbf{e})$ is an eigenvalue–eigenvector pair for $\mathbf{S}$, then $(\lambda^{-1}, \mathbf{e})$ is an eigenvalue–eigenvector pair for $\mathbf{S}^{-1}$. That is, if $\mathbf{S}\mathbf{e} = \lambda\mathbf{e}$, then multiplying on the left by $\mathbf{S}^{-1}$ gives $\mathbf{S}^{-1}\mathbf{S}\mathbf{e} = \lambda\mathbf{S}^{-1}\mathbf{e}$, or $\mathbf{S}^{-1}\mathbf{e} = \lambda^{-1}\mathbf{e}$. Therefore, using the eigenvalues from $\mathbf{S}$, we know that the ellipse extends $c\sqrt{\lambda_i}$ in the direction of $\mathbf{e}_i$ from $\bar{\mathbf{x}}$.

In $p = 2$ dimensions, the choice $c^2 = 5.99$ will produce an ellipse that contains approximately 95% of the observations. The vectors $3\sqrt{5.99}\,\mathbf{e}_1$ and $\sqrt{5.99}\,\mathbf{e}_2$ are drawn in Figure 3.8(a). Notice how the directions are the natural axes for the ellipse, and observe that the lengths of these scaled eigenvectors are comparable to the size of the pattern in each direction.

Next, for

$$\mathbf{S} = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}, \qquad \text{the eigenvalues satisfy} \qquad 0 = (\lambda - 3)^2$$

and we arbitrarily choose the eigenvectors so that $\lambda_1 = 3$, $\mathbf{e}_1' = [1, \quad 0]$ and $\lambda_2 = 3$, $\mathbf{e}_2' = [0, \quad 1]$. The vectors $\sqrt{3}\,\sqrt{5.99}\,\mathbf{e}_1$ and $\sqrt{3}\,\sqrt{5.99}\,\mathbf{e}_2$ are drawn in Figure 3.8(b).

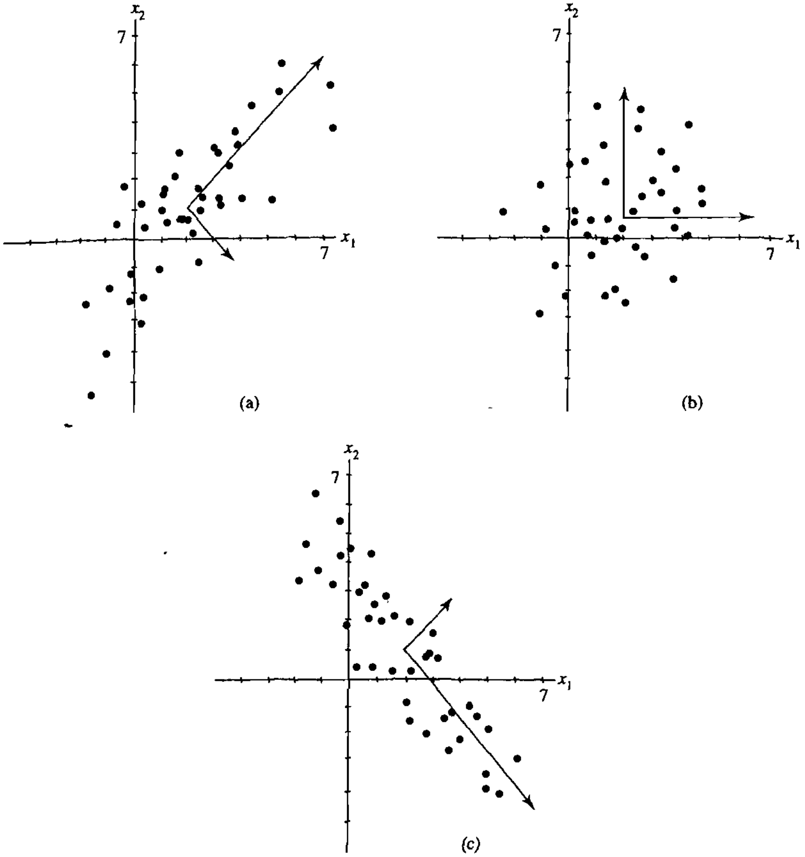


Figure 3.8 Axes of the mean-centered 95% ellipses for the scatter plots in Figure 3.7.

Finally, for

$$\mathbf{S} = \begin{bmatrix} 5 & -4 \\ -4 & 5 \end{bmatrix}, \quad \text{the eigenvalues satisfy} \quad \begin{aligned} 0 &= (\lambda - 5)^2 - (-4)^2 \\ &= (\lambda - 9)(\lambda - 1) \end{aligned}$$

and we determine the eigenvalue–eigenvector pairs $\lambda_1 = 9$, $\mathbf{e}_1' = [1/\sqrt{2}, \ -1/\sqrt{2}]$ and $\lambda_2 = 1$, $\mathbf{e}_2' = [1/\sqrt{2}, \ 1/\sqrt{2}]$. The scaled eigenvectors $3\sqrt{5.99}\,\mathbf{e}_1$ and $\sqrt{5.99}\,\mathbf{e}_2$ are drawn in Figure 3.8(c).

In two dimensions, we can often sketch the axes of the mean-centered ellipse by eye. However, the eigenvector approach also works for high dimensions where the data cannot be examined visually.

Note: Here the generalized variance $|\mathbf{S}|$ gives the same value, $|\mathbf{S}| = 9$, for all three patterns. But generalized variance does not contain any information on the orientation of the patterns. Generalized variance is easier to interpret when the two or more samples (patterns) being compared have nearly the same orientations.

Notice that our three patterns of scatter appear to cover approximately the same area. The ellipses that summarize the variability

$$(\mathbf{x} - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}) \le c^2$$

do have exactly the same area [see (3-17)], since all have $|\mathbf{S}| = 9$. ■

As Example 3.8 demonstrates, different correlation structures are not detected by $|\mathbf{S}|$. The situation for $p > 2$ can be even more obscure.

Consequently, it is often desirable to provide more than the single number $|\mathbf{S}|$ as a summary of $\mathbf{S}$. From Exercise 2.12, $|\mathbf{S}|$ can be expressed as the product $\lambda_1\lambda_2\cdots\lambda_p$ of the eigenvalues of $\mathbf{S}$. Moreover, the mean-centered ellipsoid based on $\mathbf{S}^{-1}$ [see (3-16)] has axes whose lengths are proportional to the square roots of the λ_i's (see Section 2.3). These eigenvalues then provide information on the variability in all directions in the p-space representation of the data. It is useful, therefore, to report their individual values, as well as their product. We shall pursue this topic later when we discuss principal components.

Situations in which the Generalized Sample Variance Is Zero

The generalized sample variance will be zero in certain situations. A generalized variance of zero is indicative of extreme degeneracy, in the sense that at least one column of the matrix of deviations,

$$\begin{bmatrix} \mathbf{x}_1' - \bar{\mathbf{x}}' \\ \mathbf{x}_2' - \bar{\mathbf{x}}' \\ \vdots \\ \mathbf{x}_n' - \bar{\mathbf{x}}' \end{bmatrix} = \begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2p} - \bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{np} - \bar{x}_p \end{bmatrix}$$

$$= \underset{(n\times p)}{\mathbf{X}} - \underset{(n\times 1)}{\mathbf{1}}\underset{(1\times p)}{\bar{\mathbf{x}}'} \tag{3-18}$$

can be expressed as a linear combination of the other columns. As we have shown geometrically, this is a case where one of the deviation vectors—for instance, $\mathbf{d}_i' = [x_{1i} - \bar{x}_i, \ldots, x_{ni} - \bar{x}_i]$—lies in the (hyper) plane generated by $\mathbf{d}_1, \ldots, \mathbf{d}_{i-1}, \mathbf{d}_{i+1}, \ldots, \mathbf{d}_p$.

Result 3.2. The generalized variance is zero when, and only when, at least one deviation vector lies in the (hyper) plane formed by all linear combinations of the others—that is, when the columns of the matrix of deviations in (3-18) are linearly dependent.

Proof. If the columns of the deviation matrix $(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')$ are linearly dependent, there is a linear combination of the columns such that

$$\begin{aligned} \mathbf{0} &= a_1 \operatorname{col}_1(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}') + \cdots + a_p \operatorname{col}_p(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}') \\ &= (\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')\mathbf{a} \qquad \text{for some } \mathbf{a} \neq \mathbf{0} \end{aligned}$$

But then, as you may verify, $(n-1)\mathbf{S} = (\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')'(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')$ and

$$(n-1)\mathbf{S}\mathbf{a} = (\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')'(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')\mathbf{a} = \mathbf{0}$$

so the same $\mathbf{a}$ corresponds to a linear dependency, $a_1 \operatorname{col}_1(\mathbf{S}) + \cdots + a_p \operatorname{col}_p(\mathbf{S}) = \mathbf{S}\mathbf{a} = \mathbf{0}$, in the columns of $\mathbf{S}$. So, by Result 2A.9, $|\mathbf{S}| = 0$.

In the other direction, if $|\mathbf{S}| = 0$, then there is some linear combination $\mathbf{S}\mathbf{a}$ of the columns of $\mathbf{S}$ such that $\mathbf{S}\mathbf{a} = \mathbf{0}$. That is, $\mathbf{0} = (n-1)\mathbf{S}\mathbf{a} = (\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')'(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')\mathbf{a}$. Premultiplying by $\mathbf{a}'$ yields

$$0 = \mathbf{a}'(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')'(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')\mathbf{a} = L^2_{(\mathbf{X}-\mathbf{1}\bar{\mathbf{x}}')\mathbf{a}}$$

and, for the length to equal zero, we must have $(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')\mathbf{a} = \mathbf{0}$. Thus, the columns of $(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')$ are linearly dependent. ■

Example 3.9 (A case where the generalized variance is zero) Show that $|\mathbf{S}| = 0$ for

$$\underset{(3\times 3)}{\mathbf{X}} = \begin{bmatrix} 1 & 2 & 5 \\ 4 & 1 & 6 \\ 4 & 0 & 4 \end{bmatrix}$$

and determine the degeneracy.

Here $\bar{\mathbf{x}}' = [3, 1, 5]$, so

$$\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}' = \begin{bmatrix} 1-3 & 2-1 & 5-5 \\ 4-3 & 1-1 & 6-5 \\ 4-3 & 0-1 & 4-5 \end{bmatrix} = \begin{bmatrix} -2 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \end{bmatrix}$$

The deviation (column) vectors are $\mathbf{d}_1' = [-2, 1, 1]$, $\mathbf{d}_2' = [1, 0, -1]$, and $\mathbf{d}_3' = [0, 1, -1]$. Since $\mathbf{d}_3 = \mathbf{d}_1 + 2\mathbf{d}_2$, there is column degeneracy. (Note that there is row degeneracy also.) This means that one of the deviation vectors—for example, $\mathbf{d}_3$—lies in the plane generated by the other two residual vectors. Consequently, the *three*-dimensional volume is zero. This case is illustrated in Figure 3.9 and may be verified algebraically by showing that $|\mathbf{S}| = 0$. We have

$$\underset{(3\times 3)}{\mathbf{S}} = \begin{bmatrix} 3 & -\frac{3}{2} & 0 \\ -\frac{3}{2} & 1 & \frac{1}{2} \\ 0 & \frac{1}{2} & 1 \end{bmatrix}$$

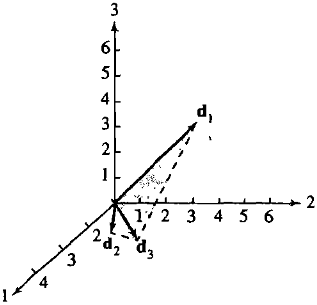


Figure 3.9 A case where the three-dimensional volume is zero ($|\mathbf{S}| = 0$).

and from Definition 2A.24,

$$|\mathbf{S}| = 3\begin{vmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{vmatrix}(-1)^2 + \left(-\tfrac{3}{2}\right)\begin{vmatrix} -\frac{3}{2} & \frac{1}{2} \\ 0 & 1 \end{vmatrix}(-1)^3 + (0)\begin{vmatrix} -\frac{3}{2} & 1 \\ 0 & \frac{1}{2} \end{vmatrix}(-1)^4$$

$$= 3\left(1 - \tfrac{1}{4}\right) + \left(\tfrac{3}{2}\right)\left(-\tfrac{3}{2} - 0\right) + 0 = \tfrac{9}{4} - \tfrac{9}{4} = 0 \qquad \blacksquare$$

When large data sets are sent and received electronically, investigators are sometimes unpleasantly surprised to find a case of zero generalized variance, so that **S** does not have an inverse. We have encountered several such cases, with their associated difficulties, before the situation was unmasked. A singular covariance matrix occurs when, for instance, the data are test scores and the investigator has included variables that are sums of the others. For example, an algebra score and a geometry score could be combined to give a total math score, or class midterm and final exam scores summed to give total points. Once, the total weight of a number of chemicals was included along with that of each component.

This common practice of creating new variables that are sums of the original variables and then including them in the data set has caused enough lost time that we emphasize the necessity of being alert to avoid these consequences.

Example 3.10 (Creating new variables that lead to a zero generalized variance) Consider the data matrix

$$\mathbf{X} = \begin{bmatrix} 1 & 9 & 10 \\ 4 & 12 & 16 \\ 2 & 10 & 12 \\ 5 & 8 & 13 \\ 3 & 11 & 14 \end{bmatrix}$$

where the third column is the sum of first two columns. These data could be the number of successful phone solicitations per day by a part-time and a full-time employee, respectively, so the third column is the total number of successful solicitations per day.

Show that the generalized variance $|\mathbf{S}| = 0$, and determine the nature of the dependency in the data.

We find that the mean corrected data matrix, with entries $x_{jk} - \bar{x}_k$, is

$$\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}' = \begin{bmatrix} -2 & -1 & -3 \\ 1 & 2 & 3 \\ -1 & 0 & -1 \\ 2 & -2 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

The resulting covariance matrix is

$$\mathbf{S} = \begin{bmatrix} 2.5 & 0 & 2.5 \\ 0 & 2.5 & 2.5 \\ 2.5 & 2.5 & 5.0 \end{bmatrix}$$

We verify that, in this case, the generalized variance

$$|\mathbf{S}| = 2.5^2 \times 5 + 0 + 0 - 2.5^3 - 2.5^3 - 0 = 0$$

In general, if the three columns of the data matrix $\mathbf{X}$ satisfy a linear constraint $a_1x_{j1} + a_2x_{j2} + a_3x_{j3} = c$, a constant for all j, then $a_1\bar{x}_1 + a_2\bar{x}_2 + a_3\bar{x}_3 = c$, so that

$$a_1(x_{j1} - \bar{x}_1) + a_2(x_{j2} - \bar{x}_2) + a_3(x_{j3} - \bar{x}_3) = 0$$

for all j. That is,

$$(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')\mathbf{a} = \mathbf{0}$$

and the columns of the mean corrected data matrix are *linearly dependent*. Thus, the inclusion of the third variable, which is linearly related to the first two, has led to the case of a zero generalized variance.

Whenever the columns of the mean corrected data matrix are linearly dependent,

$$(n - 1)\mathbf{S}\mathbf{a} = (\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')'(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')\mathbf{a} = (\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')\mathbf{0} = \mathbf{0}$$

and $\mathbf{S}\mathbf{a} = \mathbf{0}$ establishes the linear dependency of the columns of $\mathbf{S}$. Hence, $|\mathbf{S}| = 0$.

Since $\mathbf{S}\mathbf{a} = \mathbf{0} = 0\,\mathbf{a}$, we see that $\mathbf{a}$ is a scaled eigenvector of $\mathbf{S}$ associated with an eigenvalue of zero. This gives rise to an important diagnostic: If we are unaware of any extra variables that are linear combinations of the others, we can find them by calculating the eigenvectors of $\mathbf{S}$ and identifying the one associated with a zero eigenvalue. That is, if we were unaware of the dependency in this example, a computer calculation would find an eigenvalue proportional to $\mathbf{a}' = [1, 1, -1]$, since

$$\mathbf{S}\mathbf{a} = \begin{bmatrix} 2.5 & 0 & 2.5 \\ 0 & 2.5 & 2.5 \\ 2.5 & 2.5 & 5.0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = 0 \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}$$

The coefficients reveal that

$$1(x_{j1} - \bar{x}_1) + 1(x_{j2} - \bar{x}_2) + (-1)(x_{j3} - \bar{x}_3) = 0 \quad \text{for all } j$$

In addition, the sum of the first two variables minus the third is a constant c for all n units. Here the third variable is actually the sum of the first two variables, so the columns of the original data matrix satisfy a linear constraint with $c = 0$. Because we have the special case $c = 0$, the constraint establishes the fact that the columns of the data matrix are linearly dependent. ■

Let us summarize the important equivalent conditions for a generalized variance to be zero that we discussed in the preceding example. Whenever a nonzero vector $\mathbf{a}$ satisfies one of the following three conditions, it satisfies all of them:

$$\underbrace{(1)\ \mathbf{S}\mathbf{a}=\mathbf{0}}_{\substack{\mathbf{a}\text{ is a scaled}\\ \text{eigenvector of }\mathbf{S}\\ \text{with eigenvalue } 0.}} \qquad \underbrace{(2)\ \mathbf{a}'(\mathbf{x}_j-\bar{\mathbf{x}})=0 \text{ for all } j}_{\substack{\text{The linear combination}\\ \text{of the mean corrected}\\ \text{data, using }\mathbf{a}\text{, is zero.}}} \qquad \underbrace{(3)\ \mathbf{a}'\mathbf{x}_j=c \text{ for all } j\ (c=\mathbf{a}'\bar{\mathbf{x}})}_{\substack{\text{The linear combination of}\\ \text{the original data, using }\mathbf{a}\text{,}\\ \text{is a constant.}}}$$

We showed that if condition (3) is satisfied—that is, if the values for one variable can be expressed in terms of the others—then the generalized variance is zero because $\mathbf{S}$ has a zero eigenvalue. In the other direction, if condition (1) holds, then the eigenvector $\mathbf{a}$ gives coefficients for the linear dependency of the mean corrected data.

In any statistical analysis, $|\mathbf{S}| = 0$ means that the measurements on some variables should be removed from the study as far as the mathematical computations are concerned. The corresponding reduced data matrix will then lead to a covariance matrix of full rank and a nonzero generalized variance. The question of which measurements to remove in degenerate cases is not easy to answer. When there is a choice, one should retain measurements on a (presumed) causal variable instead of those on a secondary characteristic. We shall return to this subject in our discussion of principal components.

At this point, we settle for delineating some simple conditions for $\mathbf{S}$ to be of full rank or of reduced rank.

Result 3.3. If $n \le p$, that is, (sample size) $\le$ (number of variables), then $|\mathbf{S}| = 0$ for all samples.

Proof. We must show that the rank of $\mathbf{S}$ is less than or equal to p and then apply Result 2A.9.

For any fixed sample, the n row vectors in (3-18) sum to the zero vector. The existence of this linear combination means that the rank of $\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}'$ is less than or equal to $n - 1$, which, in turn, is less than or equal to $p - 1$ because $n \le p$. Since

$$(n-1)\underset{(p\times p)}{\mathbf{S}} = \underset{(p\times n)}{(\mathbf{X}-\mathbf{1}\bar{\mathbf{x}})'}\underset{(n\times p)}{(\mathbf{X}-\mathbf{1}\bar{\mathbf{x}}')}$$

the kth column of $\mathbf{S}$, $\text{col}_k(\mathbf{S})$, can be written as a linear combination of the columns of $(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')'$. In particular,

$$\begin{aligned}(n-1)\,\text{col}_k(\mathbf{S}) &= (\mathbf{X}-\mathbf{1}\bar{\mathbf{x}}')'\,\text{col}_k(\mathbf{X}-\mathbf{1}\bar{\mathbf{x}}')\\ &= (x_{1k}-\bar{x}_k)\,\text{col}_1(\mathbf{X}-\mathbf{1}\bar{\mathbf{x}}')' + \cdots + (x_{nk}-\bar{x}_k)\,\text{col}_n(\mathbf{X}-\mathbf{1}\bar{\mathbf{x}}')'\end{aligned}$$

Since the column vectors of $(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')'$ sum to the zero vector, we can write, for example, $\text{col}_1(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')'$ as the negative of the sum of the remaining column vectors. After substituting for $\text{row}_1(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')'$ in the preceding equation, we can express $\text{col}_k(\mathbf{S})$ as a linear combination of the at most $n - 1$ linearly independent row vectors $\text{col}_2(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')', \ldots, \text{col}_n(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')'$. The rank of $\mathbf{S}$ is therefore less than or equal to $n - 1$, which—as noted at the beginning of the proof—is less than or equal to $p - 1$, and $\mathbf{S}$ is singular. This implies, from Result 2A.9, that $|\mathbf{S}| = 0$. ■

Result 3.4. Let the $p \times 1$ vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, where $\mathbf{x}_j'$ is the jth row of the data matrix $\mathbf{X}$, be realizations of the independent random vectors $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$. Then

1. If the linear combination $\mathbf{a}'\mathbf{X}_j$ has positive variance for each constant vector $\mathbf{a} \neq \mathbf{0}$, then, provided that $p < n$, $\mathbf{S}$ has full rank with probability 1 and $|\mathbf{S}| > 0$.
2. If, with probability 1, $\mathbf{a}'\mathbf{X}_j$ is a constant (for example, c) *for all* j, then $|\mathbf{S}| = 0$.

Proof. (Part 2). If $\mathbf{a}'\mathbf{X}_j = a_1X_{j1} + a_2X_{j2} + \cdots + a_pX_{jp} = c$ with probability 1, $\mathbf{a}'\mathbf{x}_j = c$ for all j, and the sample mean of this linear combination is $c = \sum_{j=1}^{n} (a_1x_{j1} + a_2x_{j2} + \cdots + a_px_{jp})/n = a_1\bar{x}_1 + a_2\bar{x}_2 + \cdots + a_p\bar{x}_p = \mathbf{a}'\bar{\mathbf{x}}$. Then

$$(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')\mathbf{a} = a_1\begin{bmatrix} x_{11} - \bar{x}_1 \\ \vdots \\ x_{n1} - \bar{x}_1 \end{bmatrix} + \cdots + a_p\begin{bmatrix} x_{1p} - \bar{x}_p \\ \vdots \\ x_{np} - \bar{x}_p \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{a}'\mathbf{x}_1 - \mathbf{a}'\bar{\mathbf{x}} \\ \vdots \\ \mathbf{a}'\mathbf{x}_n - \mathbf{a}'\bar{\mathbf{x}} \end{bmatrix} = \begin{bmatrix} c - c \\ \vdots \\ c - c \end{bmatrix} = \mathbf{0}$$

indicating linear dependence; the conclusion follows from Result 3.2.

The proof of Part (1) is difficult and can be found in [2]. ■

Generalized Variance Determined by |R| and Its Geometrical Interpretation

The generalized sample variance is unduly affected by the variability of measurements on a single variable. For example, suppose some s_{ii} is either large or quite small. Then, geometrically, the corresponding deviation vector $\mathbf{d}_i = (\mathbf{y}_i - \bar{x}_i\mathbf{1})$ will be very long or very short and will therefore clearly be an important factor in determining volume. Consequently, it is sometimes useful to scale all the deviation vectors so that they have the same length.

Scaling the residual vectors is equivalent to replacing each original observation x_{jk} by its standardized value $(x_{jk} - \bar{x}_k)/\sqrt{s_{kk}}$. The sample covariance matrix of the standardized variables is then $\mathbf{R}$, the sample correlation matrix of the original variables. (See Exercise 3.13.) We define

$$\begin{pmatrix} \text{Generalized sample variance} \\ \text{of the standardized variables} \end{pmatrix} = |\mathbf{R}| \tag{3-19}$$

Since the resulting vectors

$$[(x_{1k} - \bar{x}_k)/\sqrt{s_{kk}}, (x_{2k} - \bar{x}_k)/\sqrt{s_{kk}}, \ldots, (x_{nk} - \bar{x}_k)/\sqrt{s_{kk}}] = (\mathbf{y}_k - \bar{x}_k\mathbf{1})'/\sqrt{s_{kk}}$$

all have length $\sqrt{n-1}$, the generalized sample variance of the standardized variables will be large when these vectors are nearly perpendicular and will be small when two or more of these vectors are in almost the same direction. Employing the argument leading to (3-7), we readily find that the cosine of the angle θ_{ik} between $(\mathbf{y}_i - \bar{x}_i\mathbf{1})/\sqrt{s_{ii}}$ and $(\mathbf{y}_k - \bar{x}_k\mathbf{1})/\sqrt{s_{kk}}$ is the sample correlation coefficient r_{ik}. Therefore, we can make the statement that $|\mathbf{R}|$ is large when all the r_{ik} are nearly zero and it is small when one or more of the r_{ik} are nearly $+1$ or -1.

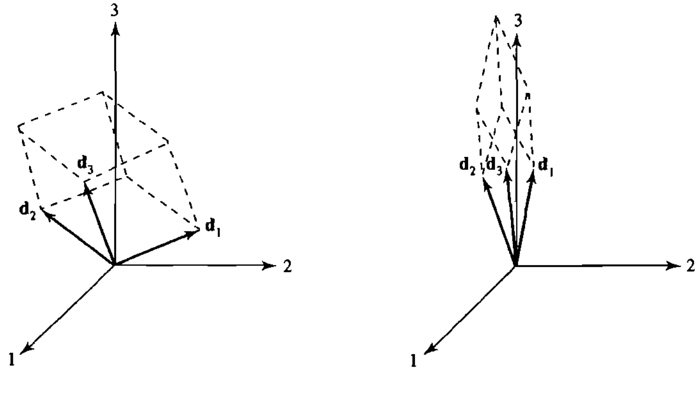
In sum, we have the following result: Let

$$\frac{(\mathbf{y}_i - \bar{x}_i\mathbf{1})}{\sqrt{s_{ii}}} = \begin{bmatrix} \dfrac{x_{1i} - \bar{x}_i}{\sqrt{s_{ii}}} \\ \dfrac{x_{2i} - \bar{x}_i}{\sqrt{s_{ii}}} \\ \vdots \\ \dfrac{x_{ni} - \bar{x}_i}{\sqrt{s_{ii}}} \end{bmatrix}, \qquad i = 1, 2, \dots, p$$

be the deviation vectors of the standardized variables. The ith deviation vectors lie in the direction of $\mathbf{d}_i$, but all have a squared length of $n - 1$. The volume generated in p-space by the deviation vectors can be related to the generalized sample variance. The same steps that lead to (3-15) produce

$$\begin{pmatrix}\text{Generalized sample variance}\\ \text{of the standardized variables}\end{pmatrix} = |\mathbf{R}| = (n - 1)^{-p}(\text{volume})^2 \tag{3-20}$$

The volume generated by deviation vectors of the standardized variables is illustrated in Figure 3.10 for the two sets of deviation vectors graphed in Figure 3.6. A comparison of Figures 3.10 and 3.6 reveals that the influence of the $\mathbf{d}_2$ vector (large variability in x_2) on the squared volume $|\mathbf{S}|$ is much greater than its influence on the squared volume $|\mathbf{R}|$.



(a) (b)

Figure 3.10 The volume generated by equal-length deviation vectors of the standardized variables.

The quantities $|\mathbf{S}|$ and $|\mathbf{R}|$ are connected by the relationship

$$|\mathbf{S}| = (s_{11}s_{22}\cdots s_{pp})|\mathbf{R}| \tag{3-21}$$

so

$$(n-1)^p|\mathbf{S}| = (n-1)^p(s_{11}s_{22}\cdots s_{pp})|\mathbf{R}| \tag{3-22}$$

[The proof of (3-21) is left to the reader as Exercise 3.12.]

Interpreting (3-22) in terms of volumes, we see from (3-15) and (3-20) that the squared volume $(n-1)^p|\mathbf{S}|$ is proportional to the squared volume $(n-1)^p|\mathbf{R}|$. The constant of proportionality is the product of the variances, which, in turn, is proportional to the product of the squares of the lengths $(n-1)s_{ii}$ of the $\mathbf{d}_i$. Equation (3-21) shows, algebraically, how a change in the measurement scale of X_1, for example, will alter the relationship between the generalized variances. Since $|\mathbf{R}|$ is based on standardized measurements, it is unaffected by the change in scale. However, the relative value of $|\mathbf{S}|$ will be changed whenever the multiplicative factor s_{11} changes.

Example 3.11 (Illustrating the relation between $|\mathbf{S}|$ and $|\mathbf{R}|$) Let us illustrate the relationship in (3-21) for the generalized variances $|\mathbf{S}|$ and $|\mathbf{R}|$ when $p = 3$. Suppose

$$\underset{(3\times 3)}{\mathbf{S}} = \begin{bmatrix} 4 & 3 & 1 \\ 3 & 9 & 2 \\ 1 & 2 & 1 \end{bmatrix}$$

Then $s_{11} = 4$, $s_{22} = 9$, and $s_{33} = 1$. Moreover,

$$\mathbf{R} = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 1 & \frac{2}{3} \\ \frac{1}{2} & \frac{2}{3} & 1 \end{bmatrix}$$

Using Definition 2A.24, we obtain

$$\begin{aligned} |\mathbf{S}| &= 4\begin{vmatrix} 9 & 2 \\ 2 & 1 \end{vmatrix}(-1)^2 + 3\begin{vmatrix} 3 & 2 \\ 1 & 1 \end{vmatrix}(-1)^3 + 1\begin{vmatrix} 3 & 9 \\ 1 & 2 \end{vmatrix}(-1)^4 \\ &= 4(9-4) - 3(3-2) + 1(6-9) = 14 \end{aligned}$$

$$\begin{aligned} |\mathbf{R}| &= 1\begin{vmatrix} 1 & \frac{2}{3} \\ \frac{2}{3} & 1 \end{vmatrix}(-1)^2 + \frac{1}{2}\begin{vmatrix} \frac{1}{2} & \frac{2}{3} \\ \frac{1}{2} & 1 \end{vmatrix}(-1)^3 + \frac{1}{2}\begin{vmatrix} \frac{1}{2} & 1 \\ \frac{1}{2} & \frac{2}{3} \end{vmatrix}(-1)^4 \\ &= \left(1-\frac{4}{9}\right) - \left(\frac{1}{2}\right)\left(\frac{1}{2}-\frac{1}{3}\right) + \left(\frac{1}{2}\right)\left(\frac{1}{3}-\frac{1}{2}\right) = \frac{7}{18} \end{aligned}$$

It then follows that

$$14 = |\mathbf{S}| = s_{11}s_{22}s_{33}|\mathbf{R}| = (4)(9)(1)\left(\tfrac{7}{18}\right) = 14 \qquad \text{(check)}$$

■

Another Generalization of Variance

We conclude-this discussion by mentioning another generalization of variance. Specifically, we define the *total sample variance* as the sum of the diagonal elements of the sample variance–covariance matrix **S**. Thus,

$$\text{Total sample variance} = s_{11} + s_{22} + \cdots + s_{pp} \tag{3-23}$$

Example 3.12 (Calculating the total sample variance) Calculate the total sample variance for the variance–covariance matrices **S** in Examples 3.7 and 3.9.

From Example 3.7.

$$\mathbf{S} = \begin{bmatrix} 252.04 & -68.43 \\ -68.43 & 123.67 \end{bmatrix}$$

and

$$\text{Total sample variance} = s_{11} + s_{22} = 252.04 + 123.67 = 375.71$$

From Example 3.9,

$$\mathbf{S} = \begin{bmatrix} 3 & -\frac{3}{2} & 0 \\ -\frac{3}{2} & 1 & \frac{1}{2} \\ 0 & \frac{1}{2} & 1 \end{bmatrix}$$

and

$$\text{Total sample variance} = s_{11} + s_{22} + s_{33} = 3 + 1 + 1 = 5$$ ■

Geometrically, the total sample variance is the sum of the squared lengths of the p deviation vectors $\mathbf{d}_1 = (\mathbf{y}_1 - \bar{x}_1\mathbf{1}), \ldots, \mathbf{d}_p = (\mathbf{y}_p - \bar{x}_p\mathbf{1})$, divided by $n - 1$. The total sample variance criterion pays no attention to the orientation (correlation structure) of the residual vectors. For instance, it assigns the same values to both sets of residual vectors (a) and (b) in Figure 3.6.

3.5 Sample Mean, Covariance, and Correlation as Matrix Operations

We have developed geometrical representations of the data matrix **X** and the derived descriptive statistics $\bar{\mathbf{x}}$ and **S**. In addition, it is possible to link algebraically the calculation of $\bar{\mathbf{x}}$ and **S** directly to **X** using matrix operations. The resulting expressions, which depict the relation between $\bar{\mathbf{x}}$, **S**, and the full data set **X** concisely, are easily programmed on electronic computers.

We have it that $\bar{x}_i = (x_{1i} \cdot 1 + x_{2i} \cdot 1 + \cdots + x_{ni} \cdot 1)/n = \mathbf{y}_i'\mathbf{1}/n$. Therefore,

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix} = \begin{bmatrix} \dfrac{\mathbf{y}_1'\mathbf{1}}{n} \\ \dfrac{\mathbf{y}_2'\mathbf{1}}{n} \\ \vdots \\ \dfrac{\mathbf{y}_p'\mathbf{1}}{n} \end{bmatrix} = \frac{1}{n} \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pn} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

or

$$\bar{\mathbf{x}} = \frac{1}{n}\mathbf{X}'\mathbf{1} \tag{3-24}$$

That is, $\bar{\mathbf{x}}$ is calculated from the transposed data matrix by postmultiplying by the vector $\mathbf{1}$ and then multiplying the result by the constant $1/n$.

Next, we create an $n \times p$ *matrix of means* by transposing both sides of (3-24) and premultiplying by $\mathbf{1}$; that is,

$$\mathbf{1}\bar{\mathbf{x}}' = \frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{X} = \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \\ \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \end{bmatrix} \tag{3-25}$$

Subtracting this result from $\mathbf{X}$ produces the $n \times p$ *matrix of deviations* (residuals)

$$\mathbf{X} - \frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{X} = \begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2p} - \bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{np} - \bar{x}_p \end{bmatrix} \tag{3-26}$$

Now, the matrix $(n-1)\mathbf{S}$ representing sums of squares and cross products is just the transpose of the matrix (3-26) times the matrix itself, or

$$(n-1)\mathbf{S} = \begin{bmatrix} x_{11} - \bar{x}_1 & x_{21} - \bar{x}_1 & \cdots & x_{n1} - \bar{x}_1 \\ x_{12} - \bar{x}_2 & x_{22} - \bar{x}_2 & \cdots & x_{n2} - \bar{x}_2 \\ \vdots & \vdots & \ddots & \vdots \\ x_{1p} - \bar{x}_p & x_{2p} - \bar{x}_p & \cdots & x_{np} - \bar{x}_p \end{bmatrix}$$

$$\times \begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2p} - \bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{np} - \bar{x}_p \end{bmatrix}$$

$$= \left(\mathbf{X} - \frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{X}\right)'\left(\mathbf{X} - \frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{X}\right) = \mathbf{X}'\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)\mathbf{X}$$

since

$$\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)'\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right) = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}' - \frac{1}{n}\mathbf{1}\mathbf{1}' + \frac{1}{n^2}\mathbf{1}\mathbf{1}'\mathbf{1}\mathbf{1}' = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'$$

To summarize, the matrix expressions relating $\bar{\mathbf{x}}$ and $\mathbf{S}$ to the data set $\mathbf{X}$ are

$$\begin{aligned} \bar{\mathbf{x}} &= \frac{1}{n}\mathbf{X}'\mathbf{1} \\ \mathbf{S} &= \frac{1}{n-1}\mathbf{X}'\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)\mathbf{X} \end{aligned} \tag{3-27}$$

The result for $\mathbf{S}_n$ is similar, except that $1/n$ replaces $1/(n-1)$ as the first factor.

The relations in (3-27) show clearly how matrix operations on the data matrix $\mathbf{X}$ lead to $\bar{\mathbf{x}}$ and $\mathbf{S}$.

Once $\mathbf{S}$ is computed, it can be related to the sample correlation matrix $\mathbf{R}$. The resulting expression can also be "inverted" to relate $\mathbf{R}$ to $\mathbf{S}$. We first define the $p \times p$ *sample standard deviation matrix* $\mathbf{D}^{1/2}$ and compute its inverse, $(\mathbf{D}^{1/2})^{-1} = \mathbf{D}^{-1/2}$. Let

$$\underset{(p\times p)}{\mathbf{D}^{1/2}} = \begin{bmatrix} \sqrt{s_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{s_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{s_{pp}} \end{bmatrix} \tag{3-28}$$

Then

$$\underset{(p\times p)}{\mathbf{D}^{-1/2}} = \begin{bmatrix} \frac{1}{\sqrt{s_{11}}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{s_{22}}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sqrt{s_{pp}}} \end{bmatrix}$$

Since

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{1p} & s_{2p} & \cdots & s_{pp} \end{bmatrix}$$

and

$$\mathbf{R} = \begin{bmatrix} \frac{s_{11}}{\sqrt{s_{11}}\sqrt{s_{11}}} & \frac{s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}} & \cdots & \frac{s_{1p}}{\sqrt{s_{11}}\sqrt{s_{pp}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{s_{1p}}{\sqrt{s_{11}}\sqrt{s_{pp}}} & \frac{s_{2p}}{\sqrt{s_{22}}\sqrt{s_{pp}}} & \cdots & \frac{s_{pp}}{\sqrt{s_{pp}}\sqrt{s_{pp}}} \end{bmatrix} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p} & r_{2p} & \cdots & 1 \end{bmatrix}$$

we have

$$\mathbf{R} = \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2} \tag{3-29}$$

Postmultiplying and premultiplying both sides of (3-29) by $\mathbf{D}^{1/2}$ and noting that $\mathbf{D}^{-1/2}\mathbf{D}^{1/2} = \mathbf{D}^{1/2}\mathbf{D}^{-1/2} = \mathbf{I}$ gives

$$\mathbf{S} = \mathbf{D}^{1/2}\mathbf{R}\mathbf{D}^{1/2} \tag{3-30}$$

That is, $\mathbf{R}$ can be obtained from the information in $\mathbf{S}$, whereas $\mathbf{S}$ can be obtained from $\mathbf{D}^{1/2}$ and $\mathbf{R}$. Equations (3-29) and (3-30) are sample analogs of (2-36) and (2-37).

3.6 Sample Values of Linear Combinations of Variables

We have introduced linear combinations of p variables in Section 2.6. In many multivariate procedures, we are led naturally to consider a linear combination of the form

$$\mathbf{c}'\mathbf{X} = c_1X_1 + c_2X_2 + \cdots + c_pX_p$$

whose observed value on the jth trial is

$$\mathbf{c}'\mathbf{x}_j = c_1x_{j1} + c_2x_{j2} + \cdots + c_px_{jp}, \qquad j = 1, 2, \ldots, n \tag{3-31}$$

The n derived observations in (3-31) have

$$\begin{aligned}\text{Sample mean} &= \frac{(\mathbf{c}'\mathbf{x}_1 + \mathbf{c}'\mathbf{x}_2 + \cdots + \mathbf{c}'\mathbf{x}_n)}{n}\\ &= \mathbf{c}'(\mathbf{x}_1 + \mathbf{x}_2 + \cdots + \mathbf{x}_n)\frac{1}{n} = \mathbf{c}'\bar{\mathbf{x}}\end{aligned} \tag{3-32}$$

Since $(\mathbf{c}'\mathbf{x}_j - \mathbf{c}'\bar{\mathbf{x}})^2 = (\mathbf{c}'(\mathbf{x}_j - \bar{\mathbf{x}}))^2 = \mathbf{c}'(\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})'\mathbf{c}$, we have

$$\text{Sample variance} = \frac{(\mathbf{c}'\mathbf{x}_1 - \mathbf{c}'\bar{\mathbf{x}})^2 + (\mathbf{c}'\mathbf{x}_2 - \mathbf{c}'\bar{\mathbf{x}})^2 + \cdots + (\mathbf{c}'\mathbf{x}_n - \mathbf{c}'\bar{\mathbf{x}})^2}{n-1}$$

$$= \frac{\mathbf{c}'(\mathbf{x}_1 - \bar{\mathbf{x}})(\mathbf{x}_1 - \bar{\mathbf{x}})'\mathbf{c} + \mathbf{c}'(\mathbf{x}_2 - \bar{\mathbf{x}})(\mathbf{x}_2 - \bar{\mathbf{x}})'\mathbf{c} + \cdots + \mathbf{c}'(\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})'\mathbf{c}}{n-1}$$

$$= \mathbf{c}'\left[\frac{(\mathbf{x}_1 - \bar{\mathbf{x}})(\mathbf{x}_1 - \bar{\mathbf{x}})' + (\mathbf{x}_2 - \bar{\mathbf{x}})(\mathbf{x}_2 - \bar{\mathbf{x}})' + \cdots + (\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})'}{n-1}\right]\mathbf{c}$$

or

$$\text{Sample variance of } \mathbf{c}'\mathbf{X} = \mathbf{c}'\mathbf{S}\mathbf{c} \tag{3-33}$$

Equations (3-32) and (3-33) are sample analogs of (2-43). They correspond to substituting the sample quantities $\bar{\mathbf{x}}$ and $\mathbf{S}$ for the "population" quantities $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively, in (2-43).

Now consider a second linear combination

$$\mathbf{b}'\mathbf{X} = b_1X_1 + b_2X_2 + \cdots + b_pX_p$$

whose observed value on the jth trial is

$$\mathbf{b}'\mathbf{x}_j = b_1x_{j1} + b_2x_{j2} + \cdots + b_px_{jp}, \qquad j = 1, 2, \ldots, n \tag{3-34}$$

It follows from (3-32) and (3-33) that the sample mean and variance of these derived observations are

$$\text{Sample mean of } \mathbf{b}'\mathbf{X} = \mathbf{b}'\bar{\mathbf{x}}$$

$$\text{Sample variance of } \mathbf{b}'\mathbf{X} = \mathbf{b}'\mathbf{S}\mathbf{b}$$

Moreover, the sample covariance computed from pairs of observations on $\mathbf{b}'\mathbf{X}$ and $\mathbf{c}'\mathbf{X}$ is

$$\begin{aligned}
&\text{Sample covariance}\\
&= \frac{(\mathbf{b}'\mathbf{x}_1 - \mathbf{b}'\bar{\mathbf{x}})(\mathbf{c}'\mathbf{x}_1 - \mathbf{c}'\bar{\mathbf{x}}) + (\mathbf{b}'\mathbf{x}_2 - \mathbf{b}'\bar{\mathbf{x}})(\mathbf{c}'\mathbf{x}_2 - \mathbf{c}'\bar{\mathbf{x}}) + \cdots + (\mathbf{b}'\mathbf{x}_n - \mathbf{b}'\bar{\mathbf{x}})(\mathbf{c}'\mathbf{x}_n - \mathbf{c}'\bar{\mathbf{x}})}{n-1}\\
&= \frac{\mathbf{b}'(\mathbf{x}_1 - \bar{\mathbf{x}})(\mathbf{x}_1 - \bar{\mathbf{x}})'\mathbf{c} + \mathbf{b}'(\mathbf{x}_2 - \bar{\mathbf{x}})(\mathbf{x}_2 - \bar{\mathbf{x}})'\mathbf{c} + \cdots + \mathbf{b}'(\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})'\mathbf{c}}{n-1}\\
&= \mathbf{b}'\left[\frac{(\mathbf{x}_1 - \bar{\mathbf{x}})(\mathbf{x}_1 - \bar{\mathbf{x}})' + (\mathbf{x}_2 - \bar{\mathbf{x}})(\mathbf{x}_2 - \bar{\mathbf{x}})' + \cdots + (\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})'}{n-1}\right]\mathbf{c}
\end{aligned}$$

or

$$\text{Sample covariance of } \mathbf{b}'\mathbf{X} \text{ and } \mathbf{c}'\mathbf{X} = \mathbf{b}'\mathbf{S}\mathbf{c} \tag{3-35}$$

In sum, we have the following result.

Result 3.5. The linear combinations

$$\begin{aligned}
\mathbf{b}'\mathbf{X} &= b_1X_1 + b_2X_2 + \cdots + b_pX_p\\
\mathbf{c}'\mathbf{X} &= c_1X_1 + c_2X_2 + \cdots + c_pX_p
\end{aligned}$$

have sample means, variances, and covariances that are related to $\bar{\mathbf{x}}$ and $\mathbf{S}$ by

$$\begin{aligned}
\text{Sample mean of } \mathbf{b}'\mathbf{X} &= \mathbf{b}'\bar{\mathbf{x}}\\
\text{Sample mean of } \mathbf{c}'\mathbf{X} &= \mathbf{c}'\bar{\mathbf{x}}\\
\text{Sample variance of } \mathbf{b}'\mathbf{X} &= \mathbf{b}'\mathbf{S}\mathbf{b} \qquad\qquad (3\text{-}36)\\
\text{Sample variance of } \mathbf{c}'\mathbf{X} &= \mathbf{c}'\mathbf{S}\mathbf{c}\\
\text{Sample covariance of } \mathbf{b}'\mathbf{X} \text{ and } \mathbf{c}'\mathbf{X} &= \mathbf{b}'\mathbf{S}\mathbf{c}
\end{aligned}$$

■

Example 3.13 (Means and covariances for linear combinations) We shall consider two linear combinations and their derived values for the $n = 3$ observations given in Example 3.9 as

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & x_{13}\\ x_{21} & x_{22} & x_{23}\\ x_{31} & x_{32} & x_{33}\end{bmatrix} = \begin{bmatrix} 1 & 2 & 5\\ 4 & 1 & 6\\ 4 & 0 & 4\end{bmatrix}$$

Consider the two linear combinations

$$\mathbf{b}'\mathbf{X} = [2 \quad 2 \quad -1]\begin{bmatrix} X_1\\ X_2\\ X_3\end{bmatrix} = 2X_1 + 2X_2 - X_3$$

and

$$\mathbf{c}'\mathbf{X} = [1 \quad -1 \quad 3]\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = X_1 - X_2 + 3X_3$$

The means, variances, and covariance will first be evaluated directly and then be evaluated by (3-36).

Observations on these linear combinations are obtained by replacing X_1, X_2, and X_3 with their observed values. For example, the $n = 3$ observations on $\mathbf{b}'\mathbf{X}$ are

$$\begin{aligned} \mathbf{b}'\mathbf{x}_1 &= 2x_{11} + 2x_{12} - x_{13} = 2(1) + 2(2) - (5) = 1 \\ \mathbf{b}'\mathbf{x}_2 &= 2x_{21} + 2x_{22} - x_{23} = 2(4) + 2(1) - (6) = 4 \\ \mathbf{b}'\mathbf{x}_3 &= 2x_{31} + 2x_{32} - x_{33} = 2(4) + 2(0) - (4) = 4 \end{aligned}$$

The sample mean and variance of these values are, respectively,

$$\text{Sample mean} = \frac{(1 + 4 + 4)}{3} = 3$$

$$\text{Sample variance} = \frac{(1 - 3)^2 + (4 - 3)^2 + (4 - 3)^2}{3 - 1} = 3$$

In a similar manner, the $n = 3$ observations on $\mathbf{c}'\mathbf{X}$ are

$$\begin{aligned} \mathbf{c}'\mathbf{x}_1 &= 1x_{11} - 1x_{12} + 3x_{13} = 1(1) - 1(2) + 3(5) = 14 \\ \mathbf{c}'\mathbf{x}_2 &= 1(4) - 1(1) + 3(6) = 21 \\ \mathbf{c}'\mathbf{x}_3 &= 1(4) - 1(0) + 3(4) = 16 \end{aligned}$$

and

$$\text{Sample mean} = \frac{(14 + 21 + 16)}{3} = 17$$

$$\text{Sample variance} = \frac{(14 - 17)^2 + (21 - 17)^2 + (16 - 17)^2}{3 - 1} = 13$$

Moreover, the sample covariance, computed from the pairs of observations $(\mathbf{b}'\mathbf{x}_1, \mathbf{c}'\mathbf{x}_1)$, $(\mathbf{b}'\mathbf{x}_2, \mathbf{c}'\mathbf{x}_2)$, and $(\mathbf{b}'\mathbf{x}_3, \mathbf{c}'\mathbf{x}_3)$, is

Sample covariance

$$= \frac{(1 - 3)(14 - 17) + (4 - 3)(21 - 17) + (4 - 3)(16 - 17)}{3 - 1} = \frac{9}{2}$$

Alternatively, we use the sample mean vector $\bar{\mathbf{x}}$ and sample covariance matrix $\mathbf{S}$ derived from the original data matrix $\mathbf{X}$ to calculate the sample means, variances, and covariances for the linear combinations. Thus, if only the descriptive statistics are of interest, we do not even need to calculate the observations $\mathbf{b}'\mathbf{x}_j$ and $\mathbf{c}'\mathbf{x}_j$.

From Example 3.9,

$$\bar{\mathbf{x}} = \begin{bmatrix} 3 \\ 1 \\ 5 \end{bmatrix} \quad \text{and} \quad \mathbf{S} = \begin{bmatrix} 3 & -\frac{3}{2} & 0 \\ -\frac{3}{2} & 1 & \frac{1}{2} \\ 0 & \frac{1}{2} & 1 \end{bmatrix}$$

Consequently, using (3-36), we find that the two sample means for the derived observations are

$$\text{Sample mean of } \mathbf{b}'\mathbf{X} = \mathbf{b}'\bar{\mathbf{x}} = \begin{bmatrix} 2 & 2 & -1 \end{bmatrix} \begin{bmatrix} 3 \\ 1 \\ 5 \end{bmatrix} = 3 \qquad \text{(check)}$$

$$\text{Sample mean of } \mathbf{c}'\mathbf{X} = \mathbf{c}'\bar{\mathbf{x}} = \begin{bmatrix} 1 & -1 & 3 \end{bmatrix} \begin{bmatrix} 3 \\ 1 \\ 5 \end{bmatrix} = 17 \qquad \text{(check)}$$

Using (3-36), we also have

$$\begin{aligned} \text{Sample variance of } \mathbf{b}'\mathbf{X} &= \mathbf{b}'\mathbf{S}\mathbf{b} \\ &= \begin{bmatrix} 2 & 2 & -1 \end{bmatrix} \begin{bmatrix} 3 & -\frac{3}{2} & 0 \\ -\frac{3}{2} & 1 & \frac{1}{2} \\ 0 & \frac{1}{2} & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 2 \\ -1 \end{bmatrix} \\ &= \begin{bmatrix} 2 & 2 & -1 \end{bmatrix} \begin{bmatrix} 3 \\ -\frac{3}{2} \\ 0 \end{bmatrix} = 3 \qquad \text{(check)} \end{aligned}$$

$$\begin{aligned} \text{Sample variance of } \mathbf{c}'\mathbf{X} &= \mathbf{c}'\mathbf{S}\mathbf{c} \\ &= \begin{bmatrix} 1 & -1 & 3 \end{bmatrix} \begin{bmatrix} 3 & -\frac{3}{2} & 0 \\ -\frac{3}{2} & 1 & \frac{1}{2} \\ 0 & \frac{1}{2} & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ 3 \end{bmatrix} \\ &= \begin{bmatrix} 1 & -1 & 3 \end{bmatrix} \begin{bmatrix} \frac{9}{2} \\ -1 \\ \frac{5}{2} \end{bmatrix} = 13 \qquad \text{(check)} \end{aligned}$$

$$\begin{aligned} \text{Sample covariance of } \mathbf{b}'\mathbf{X} \text{ and } \mathbf{c}'\mathbf{X} &= \mathbf{b}'\mathbf{S}\mathbf{c} \\ &= \begin{bmatrix} 2 & 2 & -1 \end{bmatrix} \begin{bmatrix} 3 & -\frac{3}{2} & 0 \\ -\frac{3}{2} & 1 & \frac{1}{2} \\ 0 & \frac{1}{2} & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ 3 \end{bmatrix} \\ &= \begin{bmatrix} 2 & 2 & -1 \end{bmatrix} \begin{bmatrix} \frac{9}{2} \\ -1 \\ \frac{5}{2} \end{bmatrix} = \frac{9}{2} \qquad \text{(check)} \end{aligned}$$

As indicated, these last results check with the corresponding sample quantities computed directly from the observations on the linear combinations. ■

The sample mean and covariance relations in Result 3.5 pertain to any number of linear combinations. Consider the q linear combinations

$$a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{ip}X_p, \qquad i = 1, 2, \ldots, q \tag{3-37}$$

These can be expressed in matrix notation as

$$\begin{bmatrix} a_{11}X_1 & + & a_{12}X_2 & + \cdots + & a_{1p}X_p \\ a_{21}X_1 & + & a_{22}X_2 & + \cdots + & a_{2p}X_p \\ \vdots & & \vdots & \vdots & \vdots \\ a_{q1}X_1 & + & a_{q2}X_2 & + \cdots + & a_{qp}X_p \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{q1} & a_{q2} & \cdots & a_{qp} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = \mathbf{AX} \tag{3-38}$$

Taking the ith row of $\mathbf{A}$, $\mathbf{a}_i'$, to be $\mathbf{b}'$ and the kth row of $\mathbf{A}$, $\mathbf{a}_k'$, to be $\mathbf{c}'$, we see that Equations (3-36) imply that the ith row of $\mathbf{AX}$ has sample mean $\mathbf{a}_i'\bar{\mathbf{x}}$ and the ith and kth rows of $\mathbf{AX}$ have sample covariance $\mathbf{a}_i'\mathbf{S}\,\mathbf{a}_k$. Note that $\mathbf{a}_i'\mathbf{S}\,\mathbf{a}_k$ is the (i, k)th element of $\mathbf{ASA}'$.

Result 3.6. The q linear combinations $\mathbf{AX}$ in (3-38) have sample mean vector $\mathbf{A}\bar{\mathbf{x}}$ and sample covariance matrix $\mathbf{ASA}'$. ■

Exercises

3.1. Given the data matrix

$$\mathbf{X} = \begin{bmatrix} 9 & 1 \\ 5 & 3 \\ 1 & 2 \end{bmatrix}$$

(a) Graph the scatter plot in $p = 2$ dimensions. Locate the sample mean on your diagram.

(b) Sketch the $n = 3$-dimensional representation of the data, and plot the deviation vectors $\mathbf{y}_1 - \bar{x}_1\mathbf{1}$ and $\mathbf{y}_2 - \bar{x}_2\mathbf{1}$.

(c) Sketch the deviation vectors in (b) emanating from the origin. Calculate the lengths of these vectors and the cosine of the angle between them. Relate these quantities to $\mathbf{S}_n$ and $\mathbf{R}$.

3.2. Given the data matrix

$$\mathbf{X} = \begin{bmatrix} 3 & 4 \\ 6 & -2 \\ 3 & 1 \end{bmatrix}$$

(a) Graph the scatter plot in $p = 2$ dimensions, and locate the sample mean on your diagram.

(b) Sketch the $n = 3$-space representation of the data, and plot the deviation vectors $\mathbf{y}_1 - \bar{x}_1\mathbf{1}$ and $\mathbf{y}_2 - \bar{x}_2\mathbf{1}$.

(c) Sketch the deviation vectors in (b) emanating from the origin. Calculate their lengths and the cosine of the angle between them. Relate these quantities to $\mathbf{S}_n$ and $\mathbf{R}$.

3.3. Perform the decomposition of $\mathbf{y}_1$ into $\bar{x}_1\mathbf{1}$ and $\mathbf{y}_1 - \bar{x}_1\mathbf{1}$ using the first column of the data matrix in Example 3.9.

3.4. Use the six observations on the variable X_1, in units of millions, from Table 1.1.

(a) Find the projection on $\mathbf{1}' = [1, 1, 1, 1, 1, 1]$.

(b) Calculate the deviation vector $\mathbf{y}_1 - \bar{x}_1\mathbf{1}$. Relate its length to the sample standard deviation.

(c) Graph (to scale) the triangle formed by $\mathbf{y}_1, \bar{x}_1\mathbf{1}$, and $\mathbf{y}_1 - \bar{x}_1\mathbf{1}$. Identify the length of each component in your graph.

(d) Repeat Parts a–c for the variable X_2 in Table 1.1.

(e) Graph (to scale) the two deviation vectors $\mathbf{y}_1 - \bar{x}_1\mathbf{1}$ and $\mathbf{y}_2 - \bar{x}_2\mathbf{1}$. Calculate the value of the angle between them.

3.5. Calculate the generalized sample variance $|\mathbf{S}|$ for (a) the data matrix $\mathbf{X}$ in Exercise 3.1 and (b) the data matrix $\mathbf{X}$ in Exercise 3.2.

3.6. Consider the data matrix

$$\mathbf{X} = \begin{bmatrix} -1 & 3 & -2 \\ 2 & 4 & 2 \\ 5 & 2 & 3 \end{bmatrix}$$

(a) Calculate the matrix of deviations (residuals), $\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}'$. Is this matrix of full rank? Explain.

(b) Determine $\mathbf{S}$ and calculate the generalized sample variance $|\mathbf{S}|$. Interpret the latter geometrically.

(c) Using the results in (b), calculate the total sample variance. [See (3-23).]

3.7. Sketch the solid ellipsoids $(\mathbf{x} - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}) \le 1$ [see (3-16)] for the three matrices

$$\mathbf{S} = \begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} 5 & -4 \\ -4 & 5 \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$$

(Note that these matrices have the *same* generalized variance $|\mathbf{S}|$.)

3.8. Given

$$\mathbf{S} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{S} = \begin{bmatrix} 1 & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & 1 & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & 1 \end{bmatrix}$$

(a) Calculate the total sample variance for each $\mathbf{S}$. Compare the results.

(b) Calculate the generalized sample variance for each $\mathbf{S}$, and compare the results. Comment on the discrepancies, if any, found between Parts a and b.

3.9. The following data matrix contains data on test scores, with x_1 = score on first test, x_2 = score on second test, and x_3 = total score on the two tests:

$$\mathbf{X} = \begin{bmatrix} 12 & 17 & 29 \\ 18 & 20 & 38 \\ 14 & 16 & 30 \\ 20 & 18 & 38 \\ 16 & 19 & 35 \end{bmatrix}$$

(a) Obtain the mean corrected data matrix, and verify that the columns are linearly dependent. Specify an $\mathbf{a}' = [a_1, a_2, a_3]$ vector that establishes the linear dependence.

(b) Obtain the sample covariance matrix $\mathbf{S}$, and verify that the generalized variance is zero. Also, show that $\mathbf{Sa} = \mathbf{0}$, so $\mathbf{a}$ can be rescaled to be an eigenvector corresponding to eigenvalue zero.

(c) Verify that the third column of the data matrix is the sum of the first two columns. That is, show that there is linear dependence, with $a_1 = 1$, $a_2 = 1$, and $a_3 = -1$.

3.10. When the generalized variance is zero, it is the columns of the mean corrected data matrix $\mathbf{X}_c = \mathbf{X} - \mathbf{1}\bar{\mathbf{x}}'$ that are linearly dependent, not necessarily those of the data matrix itself. Given the data

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

(a) Obtain the mean corrected data matrix, and verify that the columns are linearly dependent. Specify an $\mathbf{a}' = [a_1, a_2, a_3]$ vector that establishes the dependence.

(b) Obtain the sample covariance matrix $\mathbf{S}$, and verify that the generalized variance is zero.

(c) Show that the columns of the data matrix are linearly independent in this case.

3.11. Use the sample covariance obtained in Example 3.7 to verify (3-29) and (3-30), which state that $\mathbf{R} = \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}$ and $\mathbf{D}^{1/2}\mathbf{R}\mathbf{D}^{1/2} = \mathbf{S}$.

3.12. Show that $|\mathbf{S}| = (s_{11}s_{22}\cdots s_{pp})|\mathbf{R}|$.

Hint: From Equation (3-30), $\mathbf{S} = \mathbf{D}^{1/2}\mathbf{R}\mathbf{D}^{1/2}$. Taking determinants gives $|\mathbf{S}| = |\mathbf{D}^{1/2}||\mathbf{R}||\mathbf{D}^{1/2}|$. (See Result 2A.11.) Now examine $|\mathbf{D}^{1/2}|$.

3.13. Given a data matrix $\mathbf{X}$ and the resulting sample correlation matrix $\mathbf{R}$, consider the standardized observations $(x_{jk} - \bar{x}_k)/\sqrt{s_{kk}}$, $k = 1, 2, \ldots, p$, $j = 1, 2, \ldots, n$. Show that these standardized quantities have sample covariance matrix $\mathbf{R}$.

3.14. Consider the data matrix $\mathbf{X}$ in Exercise 3.1. We have $n = 3$ observations on $p = 2$ variables X_1 and X_2. Form the linear combinations

$$\mathbf{c}'\mathbf{X} = [-1 \quad 2]\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = -X_1 + 2X_2$$

$$\mathbf{b}'\mathbf{X} = [2 \quad 3]\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = 2X_1 + 3X_2$$

(a) Evaluate the sample means, variances, and covariance of $\mathbf{b}'\mathbf{X}$ and $\mathbf{c}'\mathbf{X}$ from first principles. That is, calculate the observed values of $\mathbf{b}'\mathbf{X}$ and $\mathbf{c}'\mathbf{X}$, and then use the sample mean, variance, and covariance formulas.

(b) Calculate the sample means, variances, and covariance of $\mathbf{b}'\mathbf{X}$ and $\mathbf{c}'\mathbf{X}$ using (3-36). Compare the results in (a) and (b).

3.15. Repeat Exercise 3.14 using the data matrix

$$\mathbf{X} = \begin{bmatrix} 1 & 4 & 3 \\ 6 & 2 & 6 \\ 8 & 3 & 3 \end{bmatrix}$$

and the linear combinations

$$\mathbf{b}'\mathbf{X} = [1 \quad 1 \quad 1]\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$$

and

$$\mathbf{c}'\mathbf{X} = [1 \quad 2 \quad -3]\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$$

3.16. Let $\mathbf{V}$ be a vector random variable with mean vector $E(\mathbf{V}) = \boldsymbol{\mu}_{\mathrm{v}}$ and covariance matrix $E(\mathbf{V} - \boldsymbol{\mu}_{\mathrm{V}})(\mathbf{V} - \boldsymbol{\mu}_{\mathrm{V}})' = \boldsymbol{\Sigma}_{\mathrm{v}}$. Show that $E(\mathbf{V}\mathbf{V}') = \boldsymbol{\Sigma}_{\mathrm{V}} + \boldsymbol{\mu}_{\mathrm{V}}\boldsymbol{\mu}_{\mathrm{V}}'$.

3.17. Show that, if $\underset{(p\times 1)}{\mathbf{X}}$ and $\underset{(q\times 1)}{\mathbf{Z}}$ are independent, then each component of $\mathbf{X}$ is independent of each component of $\mathbf{Z}$.

Hint: $P[X_1 \le x_1, X_2 \le x_2, \ldots, X_p \le x_p \text{ and } Z_1 \le z_1, \ldots, Z_q \le z_q]$

$$= P[X_1 \le x_1, X_2 \le x_2, \ldots, X_p \le x_p] \cdot P[Z_1 \le z_1, \ldots, Z_q \le z_q]$$

by independence. Let $x_2, \ldots, x_p$ and $z_2, \ldots, z_q$ tend to infinity, to obtain

$$P[X_1 \le x_1 \text{ and } Z_1 \le z_1] = P[X_1 \le x_1] \cdot P[Z_1 \le z_1]$$

for all x_1, z_1. So X_1 and Z_1 are independent. Repeat for other pairs.

3.18. Energy consumption in 2001, by state, from the major sources

x_1 = petroleum $\qquad$ x_2 = natural gas

x_3 = hydroelectric power $\qquad$ x_4 = nuclear electric power

is recorded in quadrillions (10^{15}) of BTUs (Source: *Statistical Abstract of the United States 2006*).

The resulting mean and covariance matrix are

$$\bar{\mathbf{x}} = \begin{bmatrix} 0.766 \\ 0.508 \\ 0.438 \\ 0.161 \end{bmatrix} \qquad \mathbf{S} = \begin{bmatrix} 0.856 & 0.635 & 0.173 & 0.096 \\ 0.635 & 0.568 & 0.128 & 0.067 \\ 0.173 & 0.127 & 0.171 & 0.039 \\ 0.096 & 0.067 & 0.039 & 0.043 \end{bmatrix}$$

(a) Using the summary statistics, determine the sample mean and variance of a state's total energy consumption for these major sources.

(b) Determine the sample mean and variance of the excess of petroleum consumption over natural gas consumption. Also find the sample covariance of this variable with the total variable in part a.

3.19. Using the summary statistics for the first three variables in Exercise 3.18, verify the relation

$$|\mathbf{S}| = (s_{11}\, s_{22}\, s_{33})\, |\mathbf{R}|$$

3.20. In northern climates, roads must be cleared of snow quickly following a storm. One measure of storm severity is x_1 = its duration in hours, while the effectiveness of snow removal can be quantified by x_2 = the number of hours crews, men, and machine, spend to clear snow. Here are the results for 25 incidents in Wisconsin.

Table 3.2 Snow Data

x_1	x_2	x_1	x_2	x_1	x_2
12.5	13.7	9.0	24.4	3.5	26.1
14.5	16.5	6.5	18.2	8.0	14.5
8.0	17.4	10.5	22.0	17.5	42.3
9.0	11.0	10.0	32.5	10.5	17.5
19.5	23.6	4.5	18.7	12.0	21.8
8.0	13.2	7.0	15.8	6.0	10.4
9.0	32.1	8.5	15.6	13.0	25.6
7.0	12.3	6.5	12.0		
7.0	11.8	8.0	12.8		

(a) Find the sample mean and variance of the difference $x_2 - x_1$ by first obtaining the summary statistics.

(b) Obtain the mean and variance by first obtaining the individual values $x_{j2} - x_{j1}$, for $j = 1, 2, \ldots, 25$ and then calculating the mean and variance. Compare these values with those obtained in part a.

References

1. Anderson, T. W. *An Introduction to Multivariate Statistical Analysis* (3rd ed.). New York: John Wiley, 2003.

2. Eaton, M., and M. Perlman. "The Non-Singularity of Generalized Sample Covariance Matrices." *Annals of Statistics*, **1** (1973), 710–717.