Part I

Classical Methods

1

Multidimensional Data

Denken ist interessanter als Wissen, aber nicht als Anschauen (Johann Wolfgang von Goethe, Werke – Hamburger Ausgabe Bd. 12, Maximen und Reflexionen, 1749–1832). *Thinking is more interesting than knowing, but not more interesting than looking at.*

1.1 Multivariate and High-Dimensional Problems

Early in the twentieth century, scientists such as Pearson (1901), Hotelling (1933) and Fisher (1936) developed methods for analysing multivariate data in order to

- understand the structure in the data and summarise it in simpler ways;
- understand the relationship of one part of the data to another part; and
- make decisions and inferences based on the data.

The early methods these scientists developed are *linear*; their conceptual simplicity and elegance still strike us today as natural and surprisingly powerful. Principal Component Analysis deals with the first topic in the preceding list, Canonical Correlation Analysis with the second and Discriminant Analysis with the third. As time moved on, more complex methods were developed, often arising in areas such as psychology, biology or economics, but these linear methods have not lost their appeal. Indeed, as we have become more able to collect and handle very large and high-dimensional data, renewed requirements for linear methods have arisen. In these data sets essential structure can often be obscured by noise, and it becomes vital to

> reduce the original data in such a way that informative and interesting structure in the data is preserved while noisy, irrelevant or purely random variables, dimensions or features are removed, as these can adversely affect the analysis.

Principal Component Analysis, in particular, has become indispensable as a dimension-reduction tool and is often used as a first step in a more comprehensive analysis.

The data we encounter in this book range from two-dimensional samples to samples that have thousands of dimensions or consist of continuous functions. Traditionally one assumes that the dimension d is small compared to the sample size n, and for the asymptotic theory, n increases while the dimension remains constant. Many recent data sets do not fit into this framework; we encounter

- data whose dimension is comparable to the sample size, and both are large;
- high-dimension low sample size (HDLSS) data whose dimension d vastly exceeds the sample size n, so $d \gg n$; and
- functional data whose observations are functions.

High-dimensional and functional data pose special challenges, and their theoretical and asymptotic treatment is an active area of research. Gaussian assumptions will often not be useful for high-dimensional data. A deviation from normality does not affect the applicability of Principal Component Analysis or Canonical Correlation Analysis; however, we need to exercise care when making inferences based on Gaussian assumptions or when we want to exploit the normal asymptotic theory.

The remainder of this chapter deals with a number of topics that are needed in subsequent chapters. Section 1.2 looks at different ways of displaying or visualising data, Section 1.3 introduces notation for random vectors and data and Section 1.4 discusses Gaussian random vectors and summarises results pertaining to such vectors and data. Finally, in Section 1.5 we find results from linear algebra, which deal with properties of matrices, including the spectral decomposition. In this chapter, I state results without proof; the references I provide for each topic contain proofs and more detail.

1.2 Visualisation

Before we analyse a set of data, it is important to *look* at it. Often we get useful clues such as skewness, bi- or multi-modality, outliers, or distinct groupings; these influence or direct our analysis. Graphical displays are exploratory data-analysis tools, which, if appropriately used, can enhance our understanding of data. The insight obtained from graphical displays is more *subjective* than quantitative; for most of us, however, visual cues are easier to understand and interpret than numbers alone, and the knowledge gained from graphical displays can complement more quantitative answers.

Throughout this book we use graphic displays extensively and typically in the examples. In addition, in the introduction to non-Gaussian analysis, Section 9.1 of Part III, I illustrate with the simple graphical displays of Figure 9.1 the difference between *interesting* and purely random or *non-informative* data.

1.2.1 Three-Dimensional Visualisation

Two-dimensional scatterplots are a natural – though limited – way of looking at data with three or more variables. As the number of variables, and therefore the dimension, increases, sequences of two-dimensional scatterplots become less feasible to interpret. We can, of course, still display three of the d dimensions in scatterplots, but it is less clear how one can look at more than three dimensions in a single plot.

We start with visualising three data dimensions. These arise as three-dimensional data or as three specified variables of higher-dimensional data. Commonly the data are displayed in a *default* view, but rotating the data can better reveal the structure of the data. The scatterplots in Figure 1.1 display the 10,000 observations and the three variables *CD3*, *CD8* and *CD4* of the five-dimensional HIV$^+$ and HIV$^-$ data sets, which contain measurements of blood cells relevant to HIV. The left panel shows the HIV$^+$ data and the right panel the

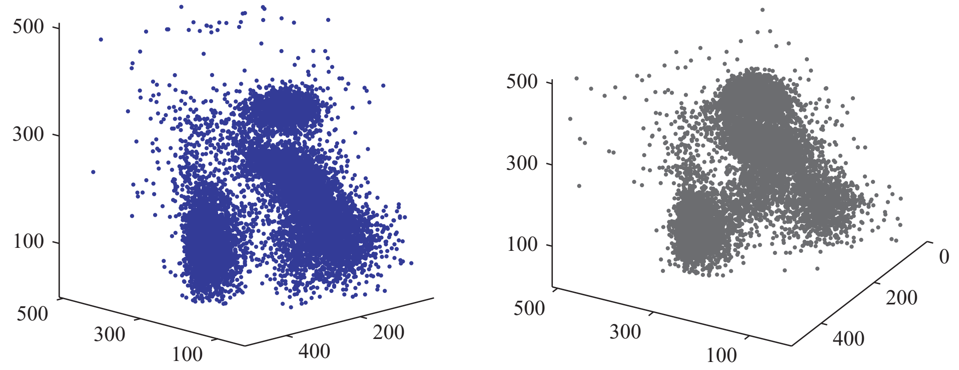

Figure 1.1 HIV$^+$data (*left*) and HIV$^-$data (*right*) of Example 2.4 with variables *CD3*, *CD8* and *CD4*.

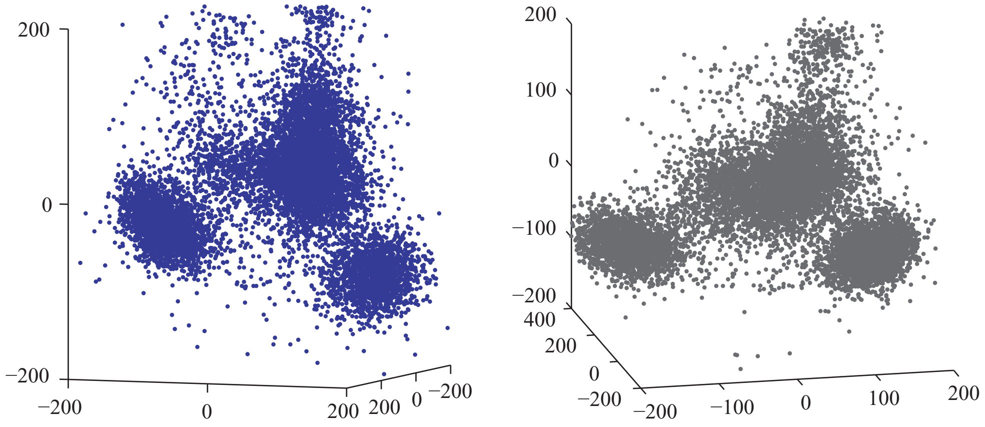

Figure 1.2 Orthogonal projections of the five-dimensional HIV$^+$data (*left*) and the HIV$^-$data (*right*) of Example 2.4.

HIV$^-$ data. There are differences between the point clouds in the two figures, and an important task in the analysis of such data is to exhibit and quantify the differences. The data are described in Example 2.4 of Section 2.3.

It may be helpful to present the data in the form of movies or combine a series of different views of the same data. Other possibilities include projecting the five-dimensional data onto a smaller number of orthogonal directions and displaying the lower-dimensional projected data as in Figure 1.2. These figures, again with HIV$^+$ in the left panel and HIV$^-$ in the right panel, highlight the cluster structure of the data in Figure 1.1. We can see a smaller fourth cluster in the top right corner of the HIV$^-$ data, which seems to have almost disappeared in the HIV$^+$ data in the left panel.

We return to these figures in Section 2.4, where I explain how to find *informative* projections. Many of the methods we explore use projections: Principal Component Analysis, Factor Analysis, Multidimensional Scaling, Independent Component Analysis and Projection Pursuit. In each case the projections focus on different aspects and properties of the data.

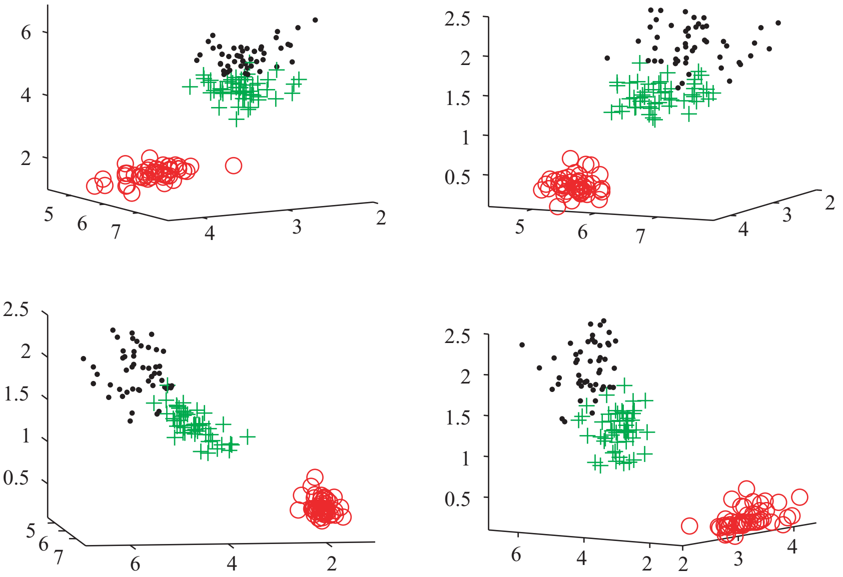

Figure 1.3 Three species of the iris data: dimensions 1, 2 and 3 (*top left*), dimensions 1, 2 and 4 (*top right*), dimensions 1, 3 and 4 (*bottom left*) and dimensions 2, 3 and 4 (*bottom right*).

Another way of representing low-dimensional data is in a number of three-dimensional scatterplots – as seen in Figure 1.3 – which make use of colour and different plotting symbols to enhance interpretation. We display the four variables of Fisher's *iris* data – *sepal length, sepal width, petal length* and *petal width* – in a sequence of three-dimensional scatterplots. The data consist of three species: red refers to *Setosa*, green to *Versicolor* and black to *Virginica*. We can see that the red observations are well separated from the other two species for all combinations of variables, whereas the green and black species are not as easily separable. I describe the data in more detail in Example 4.1 of Section 4.3.2.

1.2.2 Parallel Coordinate Plots

As the dimension grows, three-dimensional scatterplots become less relevant, unless we know that only some variables are important. An alternative, which allows us to *see* all variables at once, is to follow Inselberg (1985) and to present the data in the form of **parallel coordinate plots**. The idea is to present the data as two-dimensional graphs. Two different versions of parallel coordinate plots are common. The main difference is an interchange of the axes. In **vertical parallel coordinate plots** – see Figure 1.4 – the variable numbers are represented as values on the y-axis. For a vector $\mathbf{X} = [X_1, \ldots, X_d]^\mathsf{T}$ we represent the first variable X_1 by the point $(X_1, 1)$ and the jth variable X_j by (X_j, j). Finally, we connect the d points by a line which goes from $(X_1, 1)$ to $(X_2, 2)$ and so on to (X_d, d). We apply the same rule to the next d-dimensional datum. Figure 1.4 shows a vertical parallel coordinate plot for Fisher's *iris* data. For easier visualisation, I have used the same colours for the three species as in Figure 1.3, so red refers to the observations of species 1, green to those of species 2 and black to those of species 3. The parallel coordinate plot of the iris data shows that the data fall into two distinct groups – as we have also seen in Figure 1.3, but unlike the previous figure it tells us that dimension 3 separates the two groups most strongly.

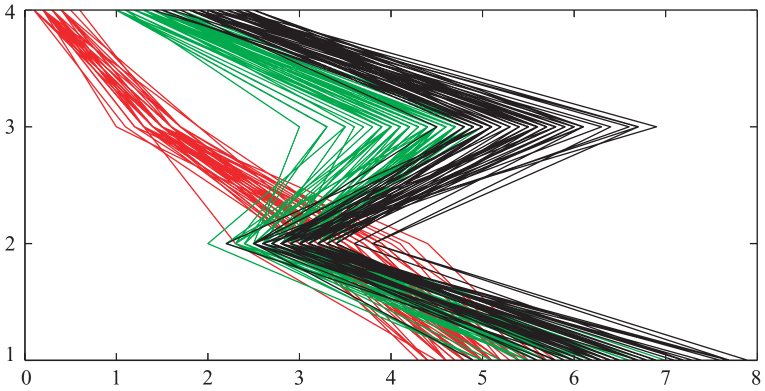

Figure 1.4 Iris data with variables represented on the y-axis and separate colours for the three species as in Figure 1.3.

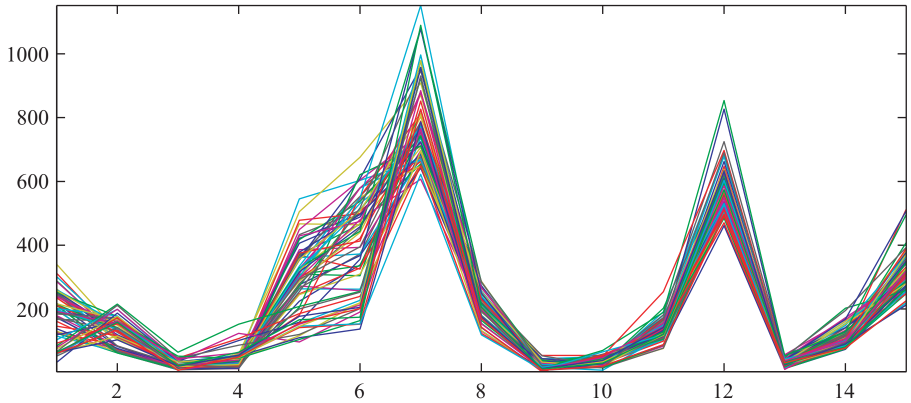

Figure 1.5 Parallel coordinate view of the illicit drug market data of Example 2.14.

Instead of the three colours shown in the plot of the iris data, different colours can be used for each observation, as in Figure 1.5. In a **horizontal parallel coordinate plot**, the x-axis represents the variable numbers $1,\ldots,d$. For a datum $\mathbf{X} = [X_1 \cdots X_d]^{\mathsf{T}}$, the first variable gives rise to the point $(1, X_1)$ and the jth variable X_j to (j, X_j). The d points are connected by a line, starting with $(1, X_1)$, then $(2, X_2)$, until we reach (d, X_d). Because we typically identify variables with the x-axis, we will more often use horizontal parallel coordinate plots. The differently coloured lines make it easier to trace particular observations.

Figure 1.5 shows the 66 monthly observations on 15 features or variables of the *illicit drug market* data which I describe in Example 2.14 in Section 2.6.2. Each observation (month) is displayed in a different colour. I have excluded two variables, as these have much higher values and would obscure the values of the remaining variables. Looking at variable 5, heroin overdose, the question arises whether there could be two groups of observations corresponding to the high and low values of this variable. The analyses of these data throughout the book will allow us to look at this question in different ways and provide answers.

Interactive graphical displays and movies are also valuable visualisation tools. They are beyond the scope of this book; I refer the interested reader to Wegman (1991) or Cook and Swayne (2007).

1.3 Multivariate Random Vectors and Data

Random vectors are vector-valued functions defined on a sample space. We consider a single random vector and collections of random vectors. For a single random vector we assume that there is a model such as the first few moments or the distribution, or we might assume that the random vector satisfies a ‘signal plus noise’ model. We are then interested in deriving properties of the random vector under the model. This scenario is called the **population** case.

For a collection of random vectors, we assume the vectors to be independent and identically distributed and to come from the same model, for example, have the same first and second moments. Typically we do not know the true moments. We use the collection to construct estimators for the moments, and we derive properties of the estimators. Such properties may include how ‘good’ an estimator is as the number of vectors in the collection grows, or we may want to draw inferences about the appropriateness of the model. This scenario is called the **sample** case, and we will refer to the collection of random vectors as the **data** or the **(random) sample**.

In applications, specific values are measured for each of the random vectors in the collection. We call these values the **realised** or **observed values** of the data or simply the **observed data**. The observed values are no longer random, and in this book we deal with the observed data in examples only. If no ambiguity exists, I will often refer to the observed data as data throughout an example.

Generally, I will treat the population case and the sample case separately and start with the population. The distinction between the two scenarios is important, as we typically have to switch from the population parameters, such as the mean, to the sample parameters, in this case the sample mean. As a consequence, the definitions for the population and the data are similar but not the same.

1.3.1 The Population Case

Let

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_d \end{bmatrix}$$

be a random vector from a distribution $F\colon \mathbb{R}^d \to [0,1]$. The individual X_j, with $j \leq d$, are **random variables**, also called the **variables**, **components** or **entries** of $\mathbf{X}$, and $\mathbf{X}$ is ***d*-dimensional** or ***d*-variate**. We assume that $\mathbf{X}$ has a finite d-dimensional mean or expected value $\mathbb{E}\mathbf{X}$ and a finite $d \times d$ covariance matrix $\operatorname{var}(\mathbf{X})$. We write

$$\boldsymbol{\mu} = \mathbb{E}\mathbf{X} \qquad \text{and} \qquad \Sigma = \operatorname{var}(\mathbf{X}) = \mathbb{E}\left[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^{\top}\right].$$

The entries of $\boldsymbol{\mu}$ and Σ are

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_d \end{bmatrix} \qquad \text{and} \qquad \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{pmatrix}, \tag{1.1}$$

where $\sigma_j^2 = \text{var}(X_j)$ and $\sigma_{jk} = \text{cov}(X_j, X_k)$. More rarely, I write σ_{jj} for the diagonal elements σ_j^2 of Σ.

Of primary interest in this book are the mean and covariance matrix of a random vector rather than the underlying distribution F. We write

$$\mathbf{X} \sim (\boldsymbol{\mu}, \Sigma) \tag{1.2}$$

as shorthand for a random vector $\mathbf{X}$ which has mean $\boldsymbol{\mu}$ and covariance matrix Σ.

If $\mathbf{X}$ is a d-dimensional random vector and A is a $d \times k$ matrix, for some $k \geq 1$, then $A^{\mathsf{T}}\mathbf{X}$ is a k-dimensional random vector. Result 1.1 lists properties of $A^{\mathsf{T}}\mathbf{X}$.

Result 1.1 *Let $\mathbf{X} \sim (\boldsymbol{\mu}, \Sigma)$ be a d-variate random vector. Let A and B be matrices of size $d \times k$ and $d \times \ell$, respectively.*

1. *The mean and covariance matrix of the k-variate random vector $A^{\mathsf{T}}\mathbf{X}$ are*
$$A^{\mathsf{T}}\mathbf{X} \sim \left(A^{\mathsf{T}}\boldsymbol{\mu}, A^{\mathsf{T}}\Sigma A\right). \tag{1.3}$$
2. *The random vectors $A^{\mathsf{T}}\mathbf{X}$ and $B^{\mathsf{T}}\mathbf{X}$ are uncorrelated if and only if $A^{\mathsf{T}}\Sigma B = \mathbf{0}_{k\times\ell}$, where $\mathbf{0}_{k\times\ell}$ is the $k \times n$ matrix all of whose entries are* 0.

Both these results can be strengthened when $\mathbf{X}$ is Gaussian, as we shall see in Corollary 1.6.

1.3.2 The Random Sample Case

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be d-dimensional random vectors. Unless otherwise stated, we assume that the $\mathbf{X}_i$ are independent and from the same distribution $F: \mathbb{R}^d \to [0,1]$ with finite mean $\boldsymbol{\mu}$ and covariance matrix Σ. We omit reference to F when knowledge of the distribution is not required.

In statistics one often identifies a random vector with its observed values and writes $\mathbf{X}_i = \mathbf{x}_i$. We explore properties of random samples but only encounter observed values of random vectors in the examples. For this reason, I will typically write

$$\mathbb{X} = \left[\mathbf{X}_1\ \mathbf{X}_2 \cdots \mathbf{X}_n\right], \tag{1.4}$$

for the sample of independent random vectors $\mathbf{X}_i$ and call this collection a **random sample** or **data**. I will use the same notation in the examples as it will be clear from the context whether I refer to the random vectors or their observed values. If a clarification is necessary, I will provide it. We also write

$$\mathbb{X} = \begin{bmatrix} X_{11} & X_{21} & \cdots & X_{n1} \\ X_{12} & X_{22} & \cdots & X_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1d} & X_{2d} & \cdots & X_{nd} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{\bullet 1} \\ \mathbf{X}_{\bullet 2} \\ \vdots \\ \mathbf{X}_{\bullet d} \end{bmatrix}. \tag{1.5}$$

The ith column of $\mathbb{X}$ is the ith random vector $\mathbf{X}_i$, and the jth row $\mathbf{X}_{\bullet j}$ is the jth variable across all n random vectors. Throughout this book the first subscript i in X_{ij} refers to the ith vector $\mathbf{X}_i$, and the second subscript j refers to the jth variable.

A Word of Caution. The data $\mathbb{X}$ are $d \times n$ matrices. This notation differs from that of some authors who write data as an $n \times d$ matrix. I have chosen the $d \times n$ notation for one main reason: The population random vector is regarded as a column vector, and it is therefore more natural to regard the random vectors of the sample as column vectors. An important consequence of this notation is the fact that the population and sample cases can be treated the same way, and no additional transposes are required.[1]

For data, the mean $\boldsymbol{\mu}$ and covariance matrix Σ are usually not known; instead, we work with the sample mean $\overline{\mathbf{X}}$ and the sample covariance matrix S and sometimes write

$$\mathbb{X} \sim \mathsf{Sam}(\overline{\mathbf{X}}, S) \tag{1.6}$$

in order to emphasise that we refer to the sample quantities, where

$$\overline{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \qquad \text{and} \tag{1.7}$$

$$S = \frac{1}{n-1}\sum_{i=1}^{n} (\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^{\mathsf{T}}. \tag{1.8}$$

As before, $\mathbb{X} \sim (\boldsymbol{\mu}, \Sigma)$ refers to the data with population mean $\boldsymbol{\mu}$ and covariance matrix Σ.

The sample mean and sample covariance matrix depend on the sample size n. If the dependence on n is important, for example, in the asymptotic developments, I will write S_n instead of S, but normally I will omit n for simplicity.

Definitions of the sample covariance matrix use n^{-1} or $(n-1)^{-1}$ in the literature. I use the $(n-1)^{-1}$ version. This notation has the added advantage of being compatible with software environments such as MATLAB and R.

Data are often centred. We write $\mathbb{X}_{\text{cent}}$ for the centred data and adopt the (*unconventional*) notation

$$\mathbb{X}_{\text{cent}} \equiv \mathbb{X} - \overline{\mathbf{X}} = \left[\mathbf{X}_1 - \overline{\mathbf{X}} \ \ldots \ \mathbf{X}_n - \overline{\mathbf{X}}\right]. \tag{1.9}$$

The centred data are of size $d \times n$. Using this notation, the $d \times d$ sample covariance matrix S becomes

$$S = \frac{1}{n-1}\left(\mathbb{X} - \overline{\mathbf{X}}\right)\left(\mathbb{X} - \overline{\mathbf{X}}\right)^{\mathsf{T}}. \tag{1.10}$$

In analogy with (1.1), the entries of the sample covariance matrix S are s_{jk}, and

$$s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n} (X_{ij} - m_j)(X_{ik} - m_k), \tag{1.11}$$

with $\overline{\mathbf{X}} = [m_1, \ldots, m_d]^{\mathsf{T}}$, and m_j is the sample mean of the jth variable. As for the population, we write s_j^2 or s_{jj} for the diagonal elements of S.

[1] Consider $\mathbf{a} \in \mathbb{R}^d$; then the projection of $\mathbf{X}$ onto $\mathbf{a}$ is $\mathbf{a}^{\mathsf{T}}\mathbf{X}$. Similarly, the projection of the matrix $\mathbb{X}$ onto $\mathbf{a}$ is done elementwise for each random vector $\mathbf{X}_i$ and results in the $1 \times n$ vector $\mathbf{a}^{\mathsf{T}}\mathbb{X}$. For the $n \times d$ matrix notation, the projection of $\mathbb{X}$ onto $\mathbf{a}$ is $\mathbb{X}^{\mathsf{T}}\mathbf{a}$, and this notation differs from the population case.