In this chapter, some typical data sets are presented, several of which will occur throughout the book. All data sets are accessible, either through one of the packages mentioned in the text, or in the ChemometricsWithR package. Chemical data sets nowadays are often characterized by a relatively low number of samples and a large number of variables, a result of the predominant spectroscopic measuring techniques enabling the chemist to rapidly acquire a complete spectrum for one sample. Depending on the actual technique employed, the number of variables can vary from several hundreds (typical in infrared measurements) to tens of thousands (e.g., in Nuclear Magnetic Resonance, NMR). A second characteristic is the high correlation between variables: neighbouring spectral variables usually convey very similar information. An example is shown in Figure 2.1, depicting the gasoline data set, one of several data sets that will be used throughout this book. It shows near-infrared (NIR) spectra of sixty gasolines at wavelengths from 900 to 1700 nm in 2 nm intervals [5], and is available in the pls package. The plot is made using the following piece of code:

```r
> data(gasoline, package = "pls")
> wavelengths <- seq(900, 1700, by = 2)
> matplot(wavelengths, t(gasoline$NIR), type = "l",
+          lty = 1, xlab = "Wavelength (nm)", ylab = "1/R")
```

The `matplot` function is used to plot all columns of matrix `t(gasoline$NIR)` (or, equivalently, all rows of matrix `gasoline$NIR`) against the specified wavelengths. Clearly, all samples have very similar features – it is impossible to distinguish individual samples in the plot. NIR spectra are notoriously hard to interpret: they consist of a large number of heavily overlapping peaks which leads to more or less smooth spectra. Nevertheless, the technique has proven to be of immense value in industry: it is a rapid, non-destructive method of analysis requiring almost no sample preprocessing, and it can be used for quantitative predictions of sample properties. The data used here can be used to quantitatively assess the octane number of the gasoline samples, for instance.
Fig. 2.1. Near-infrared spectra of sixty gasoline samples, consisting of 401 reflectance values measured at equally spaced wavelengths between 900 and 1700 nm.

Fig. 2.2. The first gas chromatogram of data set gaschrom from the ptw package.

In other cases, specific variables can be directly related to absolute or relative concentrations. An example is the gaschrom data set from the ptw package, containing gas chromatograms measured for calibration purposes. The first sample is shown in Figure 2.2. Each feature, or peak, corresponds to the elution of a compound, or in more complex cases, a number of overlapping compounds. These peaks can be easily quantified, usually by measuring peak area, but sometimes also by peak height. Since the number of features usually is orders of magnitude smaller than the number of variables in the original data, summarising the chromatograms with a peak table containing position and intensity information can lead to significant data compression.
An example in which most of the variables correspond to concentrations is the wine data set, used throughout the book. It is a set consisting of 177 wine samples, with thirteen measured variables [6]:

```r
> data(wines, package = "kohonen")
> colnames(wines)
[1] "alcohol" "malic acid" "ash"
[4] "ash alkalinity" "magnesium" "tot. phenols"
[7] "flavonoids" "non-flav. phenols" "proanth"
[10] "col. int." "col. hue" "OD ratio"
[13] "proline"
```

Variables are reported in different units. All variables apart from "col. int.", "col. hue" and "OD ratio" are concentrations. The meaning of the variables color intensity and color hue is obvious; the OD ratio is the ratio between the absorbance at wavelengths 280 and 315 nm. All wines are from the Piedmont region in Italy. Three different classes of wines are present: Barolo, Grignolino and Barberas. Barolo wine is made from Nebbiolo grapes; the other two wines have the name of the grapes from which they are made. Production areas are partly overlapping [6].

```r
> table(vintages)

vintages
Barbera Barolo Grignolino
48 58 71
```

The obvious aim in the analysis of such a data set is to see whether there is any structure that can be related to the three cultivars. Possible questions are: “which varieties are most similar?”, “which variables are indicative of the variety?”, “can we discern subclasses within varieties?”, etcetera.

A quick overview of the first few variables can be obtained with a so-called pairs plot:

```r
> pairs(wines[,1:3], pch = wine.classes, col = wine.classes)
```

This leads to the plot shown in Figure 2.3. It is clear that the three classes can be separated quite easily – consider the plot of alcohol against malic acid, for example.

A further data set comes from mass spectrometry. It contains 327 samples from three groups: patients with prostate cancer, benign prostatic hyperplasia, and normal controls [7,8]. The data have already been preprocessed (binned, baseline-corrected, normalized – see Chapter 3). The \textit{m/z} values range from 200 to 2000 Dalton. The data set is available in the R package \texttt{msProstate}:

1 Gray-scale figures such as shown throughout the book are obtained by, e.g., \texttt{col = gray(0:2/4)[wine.classes]}. In the text and the code we will in almost all cases use the default R colour palette.
Fig. 2.3. A pairs plot of the first three variables of the wine data. The three vintages are indicated with different shades of gray and plotting symbols: Barbera wines are indicated with black circles, Barolos with dark gray triangles and Grignolinos with gray plusses.

> data(Prostate2000Raw, package = "msProstate")
> plot(Prostate2000Raw$mz, Prostate2000Raw$intensity[,1],
+   type = "h", xlab = "m/z", ylab = "Intensity",
+   main = "Prostate data")

Figure 2.4 shows the first mass spectrum, that of a healthy control sample. In total, there are 168 tumour samples, 81 controls, and 78 cases of benign prostate enlargement: all samples have been measured in duplicate.

> table(Prostate2000Raw$type)

<table>
<thead>
<tr>
<th></th>
<th>bph</th>
<th>control</th>
<th>pca</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>156</td>
<td>162</td>
<td>336</td>
</tr>
</tbody>
</table>
Such data can serve as diagnostic tools to distinguish between healthy and diseased tissue, or to differentiate between several disease states. The number of samples is almost always very low – for rare diseases, patients are scarce, and stratification to obtain relatively homogeneous groups (age, sex, smoking habits, ...) usually does the rest; and in cases where the measurement is unpleasant or dangerous it may be difficult or even unethical to get data from healthy controls. On the other hand, the number of variables per sample is often huge. This puts severe restrictions on the kind of analysis that can be performed and makes thorough validation even more important.

The final data set in this chapter comes from LC-MS, the combination of liquid chromatography and mass spectrometry. The chromatography step serves to separate the components of a mixture on the basis of properties like polarity, size, or affinity. At specific time points a mass spectrum is recorded, containing the counts of particles with specific mass-to-charge ($m/z$) ratios. Measuring several samples therefore leads to a data cube of dimensions $n_{\text{time}}$, $n_{mz}$, and $n_{\text{sample}}$; the number of timepoints is typically in the order or thousands, whereas the number of samples rarely exceeds one hundred. Mass spectra can be recorded at a very high resolution and to enable statistical analysis, $m/z$ values are typically binned (or “bucketed”). Even then, thousands of variables are no exception. Package \texttt{ptw} provides a data set, \texttt{lcms}, containing data on three tryptic digests of E. coli proteins [9]. Figure 2.5 shows a top view of the first sample, with projections to the top and right of the main plot. The top projection leads to the “Total Ion Current” (TIC) chromatogram, and would be obtained if there would be no separation along the $m/z$ axis;

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{prostate_data.png}
\caption{The first mass spectrum in the prostate MS data set.}
\end{figure}
Fig. 2.5. Top view of the first sample in data set 1cms. The TIC chromatogram is shown on the top, and the direct infusion mass spectrum on the right.

Similarly, if the chromatographic dimension would be absent, the mass spectrum of the whole sample would be very close to the projection on the right (a “direct infusion” spectrum). The whole data set consists of three of such planes, leading to a data cube of size $100 \times 2000 \times 3$. 
Chemometrics with R
Multivariate Data Analysis in the Natural Sciences and Life Sciences
Wehrens, R.
2011, XIV, 286 p. 99 illus., Softcover
ISBN: 978-3-642-17840-5