Chapter 2
Data

In this chapter some data sets are presented that will be used throughout the book. In a couple of places (in particular in Chap. 11) other sets will be discussed focusing on particular analysis aspects. All data sets are accessible, either through one of the packages mentioned in the text, or in the **ChemometricsWithR** package. In addition to a short description, the data will be visualized to get an idea of their form and characteristics—one cannot stress enough how important it is to eyeball the data, not only through convenient summaries but also in their raw form!

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: neighboring spectral variables usually convey very similar information. An example is shown in Fig. 2.1, depicting the gasoline data set. It contains near-infrared (NIR) spectra of sixty gasolines at wavelengths from 900 to 1700 nm in 2 nm intervals (Kalivas 1997), and is available in the **pls** package. Clearly, the spectra are very smooth: there is very high correlation between neighboring wavelengths. This implies that the actual dimensionality of the data is lower than the number of variables.

The plot is made using the following piece of code:

```
> data(gasoline)
> 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 wave-

R. Wehrens, *Chemometrics with R*, Use R!,
https://doi.org/10.1007/978-3-662-62027-4_2

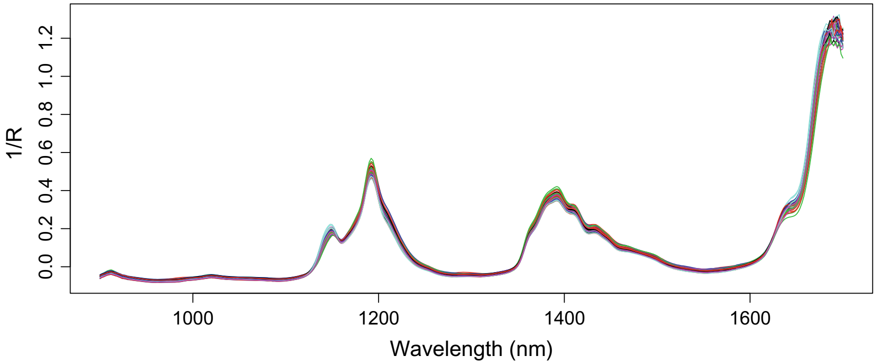

Fig. 2.1 Near-infrared spectra of sixty gasoline samples, consisting of 401 reflectance values measured at equally spaced wavelengths between 900 and 1700 nm

lengths. 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.

In other cases, specific variables can be directly related to absolute or relative concentrations. An example in which is the case for most variables is the wine data set from the **kohonen** package, used throughout the book. It is a set consisting of 177 wine samples, with thirteen measured variables (Forina et al. 1986):

```
> data(wines)
> 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 (Forina et al. 1986).

```
> 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:

```
> wine.classes <- as.integer(vintages)
> pairs(wines[, 1:3], pch = wine.classes, col = wine.classes)
```

This leads to the plot shown in Fig. 2.2. 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 the field of mass-spectrometry-based proteomics.[1] Figure 2.3, showing the first mass spectrum (a healthy control sample) is generated by:

```
> data(Prostate2000Raw)
> plot(Prostate2000Raw$mz, Prostate2000Raw$intensity[, 1],
+      type = "h", main = "Prostate data",
+      xlab = bquote(italic(.("m/z"))~.("(Da)")),
+      ylab = "Intensity")
```

Each peak in the chromatogram corresponds to the elution of a compound, or in more complex cases, a number of overlapping compounds. In a process called peak picking (see next chapter) these peaks can be easily quantified, usually by measuring peak area, but sometimes also by peak height. Since the number of peaks 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. Mass spectra, containing intensities for different mass-to-charge ratios (indicated by m/z), can be recorded at a very high resolution. To enable statistical analysis, m/z values are typically *binned* (or “bucketed”). Even then, thousands of variables are no exception.

The data set contains 327 samples from three groups: patients with prostate cancer, benign prostatic hyperplasia, and normal controls (Adam et al. 2002; Qu et al. 2002). All samples have been measured in duplicate:

[1] Originally from the R package **msProstate**, which is no longer available.

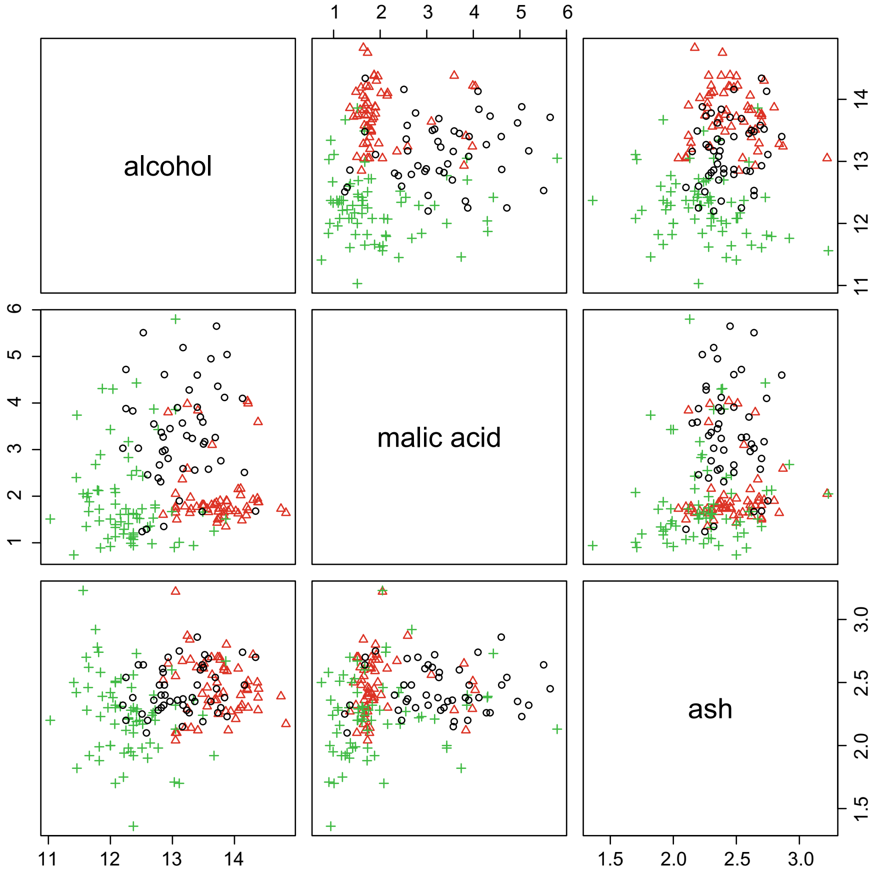

Fig. 2.2 A pairs plot of the first three variables of the wine data. The three vintages are indicated with different colors and plotting symbols: Barbera wines are indicated with black circles, Barolos with red triangles and Grignolinos with green plusses

```
> table(Prostate2000Raw$type)

     bph control      pca
     156     162      336
```

The data have already been preprocessed (binned, baseline-corrected, normalized—see Chap. 3); m/z values range from 200 to 2000 Dalton.

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

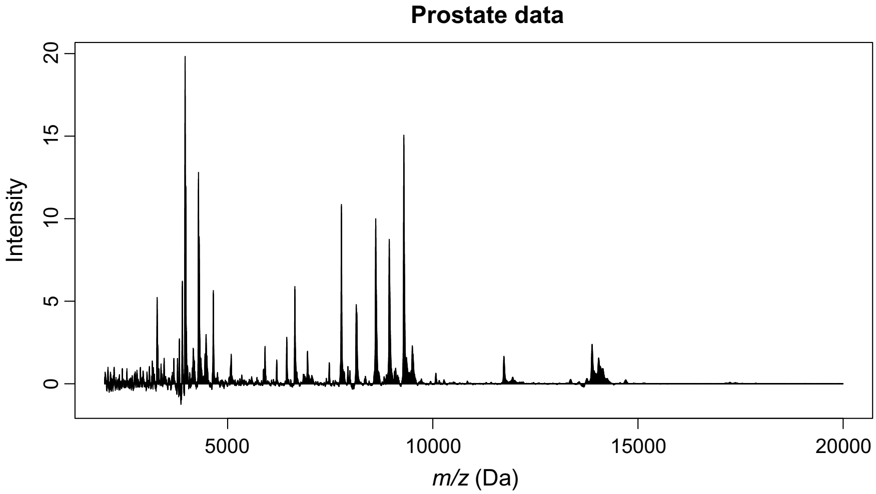

Fig. 2.3 The first mass spectrum in the prostate MS data set

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 also comes from proteomics and is measured with 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 m/z values. Measuring several samples therefore leads to a data cube of dimensions `ntime`, `nmz`, and `nsample`; the number of time points is typically in the order or thousands, whereas the number of samples rarely exceeds one hundred. Package **ptw** provides a data set, `lcms`, containing data on three tryptic digests of E. coli proteins (Bloemberg et al. 2010).

Figure 2.4 shows a top view of the first sample. The projection to the top of the figure, effectively summing over all m/z values, leads to the "Total Ion Current" (TIC) chromatogram. 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

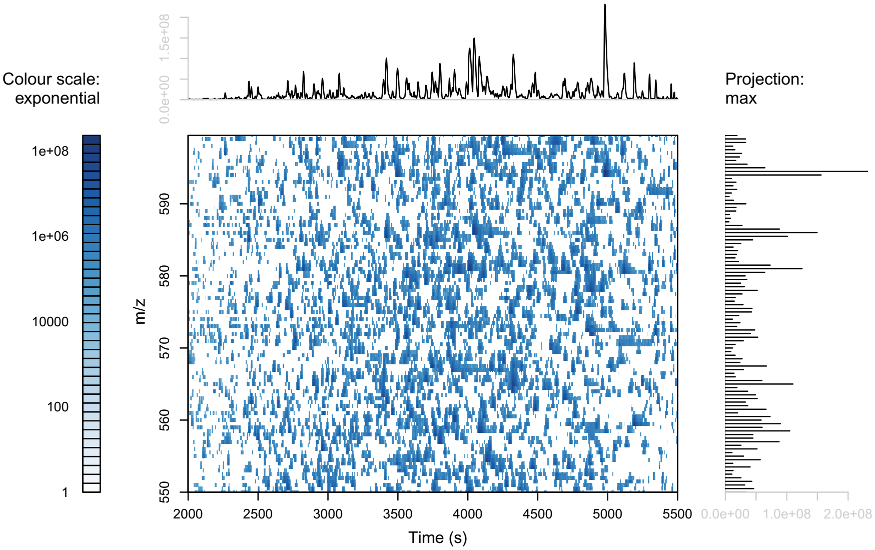

Fig. 2.4 Top view of the first sample in data set `lcms`. The TIC chromatogram is shown on the top, and the direct infusion mass spectrum on the right

planes, leading to a data cube of size 100 × 2000 × 3. Similar data sets are seen in the field of metabolomics, where the chemical entities that are sampled are not peptides (small fragments of proteins) as in proteomics, but small chemical molecules called metabolites. Because of the high dimensionality and general complexity of such data sets, chemometric methods have caught on very well in the -omics sciences.