1

Introduction

In the life sciences, molecular biology in particular, the amount of data has exploded in the last decade. Sequencing a whole genome is becoming routine work, and shortly the amount of money needed to do so will be less than the cost of a medium-sized television set. Rather than focussing on measuring specific predefined characteristics of the sample[1] modern techniques aim at generating a holistic view, sometimes called a "fingerprint". As a result, one analysis of one single sample can easily yield megabytes of data. These physical samples typically are complex mixtures and may, e.g., correspond to body fluids of patients and controls, measured with possibly several different spectroscopic techniques; environmental samples (air, water, soil); measurements on different cell cultures or one cell culture under different treatments; industrial samples from process industry, pharmaceutical industry or food industry; samples of competitor products; quality control samples, and many others. The types of data we will concentrate on are generated by analytical chemical measurement techniques, and are in almost all cases directly related to concentrations or amounts of specific classes of chemicals such as metabolites or proteins. The corresponding research fields are called metabolomics and proteomics, and a host of other -omics sciences with similar characteristics exist. A well-known example from molecular biology is transcriptomics, focussing on the levels of mRNA obtained by transcription from DNA strains. Although we do not include any transcriptomics data, many of the techniques treated in this book are directly applicable – in that sense, the characteristics of data of completely different origins can still be comparable.

These data can be analysed at different levels. The most direct approach is to analyse them as raw data (intensities, spectra, ...), without any prior interpretation other than a suitable pretreatment. Although this has the advantage

[1] The word "sample" will be used both for the physical objects on which measurements are performed (the chemical use of the word) and for the current realization of all possible measurements (the statistical use). Which one is meant should be clear from the context.

R. Wehrens, *Chemometrics with R: Multivariate Data Analysis in the Natural Sciences and Life Sciences*, Use R!, DOI 10.1007/978-3-642-17841-2_1,

that it is completely objective, it is usually also more difficult: typically, the number of variables is huge and the interpretability of the statistical models that are generated to describe the data often is low. A more often used strategy is to apply domain knowledge to convert the raw data into more abstract variables such as concentrations, for example by quantifying a set of compounds in a mixture based on a library of pure spectra. The advantage is that the statistical analysis can be performed on the quantities that really matter, and that the models are simpler and easier to validate and interpret. The obvious disadvantage is the dependence on the interpretation step: not always it is easy to decide which compounds are present and in what amounts. Any error at this stage cannot be corrected in later analysis stages.

The extremely rapid development of analytical techniques in biology and chemistry has left data analysis far behind, and as a result the statistical analysis and interpretation of the data has become a major bottleneck in the pipeline from measurement to information. Academic training in multivariate statistics in the life sciences is lagging. Bioinformatics departments are the primary source of scientists with such a background, but bioinformatics is a very broad field covering many other topics as well. Statistics and machine learning departments are usually too far away from the life sciences to establish joint educational programmes. As a result, scientists doing the data analysis very often have a background in biology or chemistry, and have acquired their statistical skills by training-on-the-job. This can be an advantage, since it makes it easier to interpret results and assess the relevance of certain findings. At the same time, there is a need for easily accessible background material and opportunities for self-study: books like the excellent "The Elements of Statistical Learning" [3] form an invaluable source of information but can also be a somewhat daunting read for scientists without much statistical background.

This book aims to fill the gap, at least to some extent. It is important to combine the sometimes rather abstract descriptions of the statistical techniques with hands-on experience behind a computer screen. In many ways R [1] is the ideal software platform to achieve this – it is extremely powerful, the many add-on packages provide a huge range of functionalities in different areas, and it is freely accessible. As in the other books in this series, the examples can be followed step-by-step by typing or cutting-and-pasting the code, and it is easy to plug in one's own data. To date, there is only one other book specifically focused on the use of R in a similar field of science: "Introduction to Multivariate Statistical Analysis in Chemometrics" [4] which to some extent complements the current volume, in particular in its treatment of robust statistics.

Here, the concepts behind the most important data analysis techniques will be explained using a minimum of mathematics, but in such a way that the book still can be used as a student's text. Its structure more or less follows the steps made in a "classical" data analysis, starting with the *data pretreatment* in Part I. This step is hugely important, yet is often treated only cursorily. An unfortunate choice here can destroy any hope of achieving good results:

background knowledge of the system under study as well as the nature of the measurements should be used in making decisions. This is where science meets art: there are no clear-cut rules, and only by experience we will learn what the best solution is.

The next phase, subject of Part II, consists of *exploratory analysis*. What structure is visible? Are there any outliers? Which samples are very similar, which are different? Which variables are correlated? Questions like these are most easily assessed by eye – the human capacity for pattern recognition in two dimensions is far superior to any statistical method. The methods at this stage all feature strong visualization capabilities. Usually, they are model-free; no model is fitted, and the assumptions about the data are kept to a minimum.

Once we are at the *modelling* phase, described in Part III, we very often do make assumptions: some models work optimally with normally distributed data, for example. The purpose of modelling can be twofold. The first is prediction. Given a set of analytical data, we want to be able to predict properties of the samples that cannot be measured easily. An example is the assessment of whether a specific treatment will be useful for a patient with particular characteristics. Such an application is known as *classification* – one is interested in modelling class membership (will or will not respond). The other major field is *regression*, where the aim is to model continuous real variables (blood pressure, protein content, ...). Such predictive models can mean a big improvement in quality of life, and save large amounts of money. The prediction error is usually taken as a quality measure: a model that is able to predict with high accuracy must have captured some real information about the system under study. Unfortunately, in most cases no analytical expressions can be derived for prediction accuracy, and other ways of estimating prediction accuracy are required in a process called *validation*. A popular example is crossvalidation.

The second aim of statistical modelling is *interpretation*, one of the topics in Part IV. Who cares if the model is able to tell me that this is a Golden Delicious apple rather than a Granny Smith? The label in the supermarket already told me so; but the question of course is why they taste different, feel different and look different. Fitting a predictive model in such a case may still be informative: when we are able to find out why the model makes a particular prediction, we may be able to learn something about the underlying physical, chemical or biological processes. If we know that a particular gene is associated with the process that we are studying, and both this gene and another one show up as important variables in our statistical model, then we may deduce that the second gene is also involved. This may lead to several new hypotheses that should be tested in the lab. Obviously, when a model has little or no predictive ability it does not make too much sense to try and extract this type of information.

Our knowledge of the system can also serve as a tool to assess the quality of our model. A model that fits the data and seems to be able to predict well is not going to be very popular when its parameters contradict what we know about the underlying process. Often, prior knowledge is available (we

expect a peak at a certain position; we know that model coefficients should not be negative; this coefficient should be larger than the other), and we can use that knowledge to assess the relevance of the fitted model. Alternatively, we can constrain the model in the training phase to take prior knowledge into account, which is often done with constraints. In other cases, the model is hard to interpret because of the sheer number of coefficients that have been fitted, and graphical summaries may fail to show what variables contribute in what way. In such cases, *variable selection* can come to the rescue: by discarding the majority of the variables, hopefully without compromising the model quality, one can often improve predictions *and* make the model much more easy to interpret. Unfortunately, variable selection is an NP-complete problem (which in practice means that even for moderate-sized systems it may be impossible to assess all possible solutions) and one never can be sure that the optimal solution has been found. But then again, any improvement over the original, full, model is a bonus.

For each of the stages in this "classical" data analysis pipeline, a plethora of methods is available. It can be hard to assess which techniques should be considered in a particular problem, and perhaps even more importantly, which should not. The view taken here is that the simplest possibilities should be considered first; only when the results are unsatisfactory, one should turn to more complex solutions. Of course, this is only a very crude first approach, and experienced scientists will have devised many shortcuts and alternatives that work better for their types of data. In this book, I have been forced to make choices. It is impossible to treat all methods, or even a large subset, in detail. Therefore the focus is on an ensemble of methods that will give the reader a broad range of possibilities, with enough background information to acquaint oneself with other methods, not mentioned in this book, if needed. In some cases, methods deserve a mention because of the popularity within the bioinformatics or chemometrics communities. Such methods, together with some typical applications, are treated in the final part of the book.

Given the huge number of packages available on CRAN and the speed with which new ones appear, it is impossible to mention all that are relevant to the material in this book. Where possible, I have limited myself to the recommended packages, and those coming with a default R installation. Of course, alternative, perhaps even much simpler, solutions may be available in the packages that this book does not consider. It pays to periodically scan the CRAN and Bioconductor repositories, or, e.g., check the Task Views that provide an overview of all packages available in certain areas – there is one on Physics and Chemistry, too.