1 Introduction

1.1 Past, Present and Future

Chemometrics has been with us, as a recognizable name, for around 35 years, its origins being traced back to the early 1970s. In those heydays many of the pioneers were programmers, good at writing software in Fortran, or Basic, that would be used to perform, what in those days were considered intensive calculations, often on large mainframe computers. The concept of the laboratory based chemist sitting at his or her desk manipulating data from an NMR or HPLC instrument in real time on a cheap micro was probably not even contemplated at the time. Indeed in the early 1970s most people had little idea that computers would shrink and become user friendly and on everyone’s desk, the future of computers was ‘big’, to be developed and used by large institutes such as for defence and for space exploration. People who used computers were in the most part programmers and so a computationally based discipline that can look back to the early 1970s is a discipline that was founded primarily by people that could program computers.

It is my view that the development and aspiration of chemometricians and most computer based disciplines changed during the late 1970s, when the economic catalyst for the development of computing started to move in new directions. With the oil crisis and global inflation in Western countries of the 1970s, there was no longer an inexorable march to larger and more expensive computing. In the early 1970s many people might have envisaged that we would have established colonies on the moon by now, powered by huge, but somewhat impersonal and clunky, mainframes. View any science fiction film from this era, and computer screens are full of numbers and letters, not ergonomic mice and windows, but although the vision of this future computer interface was primitive, in our modern view, the computer power envisaged was enormous, way beyond what we might normally come across nowadays. In science fiction films and television programmes of that era, most space ships and cities had a large central computer control room filled with big cables, consoles, rather old fashioned keyboards, and lineprinters churning out printouts on large perforated sheets of paper: that is what people thought the future would be. So with early chemometrics,
did the pioneers realize that chemometric methods would end up on a desktop, and that the
click of a mouse and the ability to rotate a plot in real time would be considered more
important than the development of computer intensive pattern recognition algorithms? The
answer is probably not.

The first methods to be marketed specifically for chemometrics were reported in those
eyears, methods such as NIPALS (for performing Principal Components Analysis), k
Near Neighbours (for pattern recognition), and SIMCA (also for pattern recognition).
Early texts such as Sharaf et al. from the mid 1980s still illustrated diagrams by using
lineprinter output. That was the norm.

One of the planks of early chemometrics was undoubtedly pattern recognition. This area
in itself also is very broad, and includes numerous journals and conferences associated
with it. Pattern recognition can mean a lot of different things to different people, and in the
context of chemometrics there are two broadly based sets of approaches, one influenced by
the machine learning community, that emphasizes Neural Networks, Kernel Methods, Self
Organising Maps, Support Vector Machines etc. and another influenced by the statistical
community emphasizing Discriminant Analysis, Method Validation, Bayesian approaches
and so on. Whereas both traditions are important ones they tend not to meet much, almost
existing as some parallel universes that occasionally collide: to the chemist this seems
strange almost looking as an outsider, but both types of community really feels they are
separate. Each ‘tribe’ has different beliefs probably based on prior education and experi-
cence. It is well known that people with different underlying psychologies choose to work
(or succeed) in different disciplines. Hence someone successful as a statistician will
probably have a different psychological makeup to someone that is successful in machine
learning, and both will approach similar problems often from very different perspectives.
They will often be reluctant to cross into each others’ territories, and prefer to discuss
scientific problems with their own ‘tribe’ who will organize their own conferences and
journals. Chemometrics, too, has developed its own ‘tribe’, the foundations being set in the
1970s. I remember when starting my academic career, being asked to attend a course on
teaching, and being told that people choosing chemistry are in themselves very much
outliers against all other sciences. If one performs certain psychological tests and then
performs some form of multivariate analysis, I was told, biologists, physicists and math-
ematicians would cluster together, with chemists as a separate outlying group. This at first
may seem strange but walk into any department of biology or mathematics. The ‘big
professor’ often does hands-on work himself or herself: there are less research students and
smaller groups but more long term junior researchers; research students often pursue quite
independent lines with encouragement; research groups are often less well defined, or at
least are not defined around the individual academic but more as a ‘sub-department’ with
common themes and several leaders. In a chemistry department there tend to be more
research students and larger groups; the job of the research professor is primarily that of a
manager with very few working hands-on in the laboratory; a group is primarily connected
to the group leader and research students feel their strongest affiliation to their professor
and his or her fellow students. So, some of the pioneering chemometricians, by psychology
chemists, were good organizers as well as having strong mathematical ability. In contrast
in different disciplines, such as computer science or statistics, senior scientists often
publish sole author papers throughout their career, and in many cases their research
students also publish independently; in chemistry most senior scientists publish mainly
collaborative papers often with large teams. Hence many of the early chemometrics pioneers had the aptitude of mathematicians and/or computer scientists but the psychology of chemists. This probably flavoured the early development.

So what has this to do with pattern recognition? Over the past decades there have been enormous developments in statistical and computational pattern recognition, but despite many of the original seeds of chemometrics being within the application of pattern recognition to chemistry these theoretical developments are rarely understood in modern chemometrics practice. What happened is that there were indeed pioneering ideas in the 1970s and early 1980s but at the time chemometrics was developing as a recognizable subject. Good organizers and managers look for developing their teams and their funding. Only a few of those prominent in the 1970s regularly published books or papers as sole authors after the initial seeds in the 1970s: their job was to build up students, departments or companies to market their ideas because that was what they were very good at doing. Although in fact there were (and continue to be) people that have not moved in this direction within chemometrics, several of those ‘feeling the water’ in the 1970s and 1980s have moved back to other disciplines such as statistics, where one can be of the highest distinction without having set up a company or a large research group, leaving behind a somewhat large gap between a growing and excellent theoretical base, but hardly understood or even acknowledged by analytical chemistry users of chemometrics, and a slower intellectually developing base primarily focused on software that is much better known by users of chemometrics techniques.

Chemometrics in the 1980s onwards focused on dollars. Newcomers to the field were successful newcomers if their ideas could be quickly converted to funds. This meant that some areas moved faster than others. Areas in particular where there were specific challenges to the chemist, particularly in industry, and where chemometrics could help, developed fast. Multivariate calibration, especially in the context of NIR (Near Infrared) spectroscopy, is a definite success story from this era, catalysing the development of methods especially of benefit to the food and the pharmaceutical industry, metamorphosing over the last few years into initiatives such as the PAT (Process Analytical Technology) initiative. The more theoretical chemometrics still moved forward, but mainly in areas that are specific to chemistry such as multivariate curve resolution (MCR), e.g. in spectroscopy or coupled chromatography, encompassing approaches such as Factor Analysis or Alternating Least Squares, mainly applied in the interface of physical and analytical chemistry. Although there were many papers published applying pattern recognition to analytical chemistry problems over this period, look at any major chemometrics conference, and the topics emphasized tend to be areas such as calibration, multiway methods, signal resolution, experimental design, etc. – and NIR spectroscopy of course.

So pattern recognition has got left behind within the chemometrics context. Pattern recognition though provided one of the main original motivations of chemometrics. What happened? Some of those publishing in the early days of chemometrics pattern recognition have moved back into core disciplines such as statistics or machine learning. The reason for this ‘divorce’ is two fold. The first is that the ‘audience’, primarily of practicing chemists, found (and still find) many of the more intense statistical papers very tough to understand. Theoreticians mainly have a very intense way of writing – which they are trained in and which referees will often insist on – but that makes many of their papers
inaccessible to all but a small handful of readers outside the field, so although their ideas may be fantastic they are not communicated in a way that the mainstream user of chemometrics methods might like. The second is that many of the chemists that took over chemometrics in the 1970s and 1980s were often very good at selling themselves – to get funds and develop their own groups and companies – and because they were more in tune with the psychology of their market, the way they communicated ideas won out, and being able to communicate ideas is a very important skill especially in a subject such as chemometrics which involves talking to a wide variety of collaborators. Hence much of the accepted ‘philosophy’ of chemometric pattern recognition is based on principles that have hardly changed for nearly thirty years, whilst other areas have progressed much faster. One reason why areas such as calibration and multivariate curve resolution could develop at such a pace is that the methods were very specific to chemical problems and so could thrive within an ‘island’ and a fairly isolated community. Plus, as we will see below, for many of the problems encountered by the early chemometrics pioneers, such as in NIR spectroscopy, the existing methods were good enough at the time. Finally, there were many more dollars at the time associated with development of multivariate calibration and signal analysis compared to pattern recognition within an analytical chemical context, although with the growing interface to metabolic profiling this is now starting to change.

Since the 1980s, though, there have been huge changes in scientific data analysis. The first and foremost is the continual explosion of computer power. In several places in this text we will cite Moore’s law. This can be formulated in a variety of ways, but is based in an exponential increase in computing power per unit cost, variously put at doubling every one to two years (dependent on the yardstick). If we say it doubles every 18 months, then it increases around 100 times every 10 years, or has increased 10 million times in the 35 years or so that chemometrics has been around. Naturally ‘at the beginning’ people could not foresee the power of desktop PCs and used mainframes, so if one estimates a thousand fold difference between a mainframe and a desktop, this still results in an impressive 100 000 fold increase in capabilities since the ‘beginning’. This means that computationally intense methods that were impossible several years ago can now be applied routinely. Many of the earliest chemometrics packages were based around being a very economical use of computer power: that is what a good programmer and algorithm developer will learn; still these skills are needed for good systems programmers but now it is possible to contemplate more computationally intense approaches – if these are useful. Why use cross-validation if one can use the bootstrap even if the latter has to be repeated 200 times and so is 200 times slower? People tell me they remember when they would go out for a long cup of coffee to wait for their ‘cross-validation’ to creek through their micro – they now sit in front of a PC for a few seconds and the bootstrap is completed. Why divide into a single test set when one can repeatedly divide into 100 different test sets and average the models? In the past each test set split was computationally expensive and so took time – one could not wait a week for the answer. However there is now no need to work within the straightjacket of methods developed thirty years ago. The old methods were good for their time, and indeed there is nothing wrong with them in their days. We may use energy efficient light bulbs rather than old wasteful ones now, but for one hundred years the old light bulbs served their purpose. We use DVDs now rather than VHS tapes – the latter technology seems very primitive in this day and age but in its time VHS was a revolution and served people well for twenty years. We buy CDs or download music rather than
purchase vinyl gramophone records but gramophone records served an important purpose in civilisation for a century. Hence there is always a need to evaluate what methods are most appropriate according to the needs of the time: we no longer live in caves – give a caveman a flatscreen TV and they would not know what to do with it, but chemometrics should adapt with the times even though the pioneering work of the 1970s and 1980s made the subject what it is.

There is however a problem in chemometrics in that lots of people, often without a good mathematical or computational background, want to ‘use’ it. Often I am surprised that people without any prior knowledge of this subject feel that they can pick it up in a workshop that ‘should not last too long’. They want to walk in, then walk out and understand how to do pattern recognition in a couple of afternoons. This desire, unfortunately, is an important economic driving force in this subject. I say to my students that it may take a year or so just working through examples to learn the basis and gain sufficient feel for the subject. They accept this, but that is why they are giving up so much of their time to learn chemometrics. If they did not accept this, they would not be my students. The dilemma though is that for chemometrics to become widespread there should be a big user base. This is where the subject, and especially the application of pattern recognition, has a problem – almost like a split personality. Keep the subject theoretical and to an elite who are really good at maths and computing, and it is not widespread. Tell an analytical chemist in the lab that he or she cannot do any pattern recognition and many will turn round, download a package, and put some data through and go away, even if he or she cannot understand the results. A few will get interested and learn but then they need to be in an environment where they have a lot of time – and many employers or even research supervisors will not allow this. So most will either give up or try to cut corners. They will pay money for chemometrics, but not to spend a year or two learning the ropes, but rather to buy a package, that they believe does what they want, and go on a course that will teach them how to enter data and print out results in a couple of afternoons. The people that market these packages will make it easy for someone to take a series of spectra, import them into a package, view a graph on a screen, change the appearance with a mouse or a menu and incorporate into a nice report in Word that will be on their boss’s (or their sponsor’s) desk within a few hours. They won’t gain much insight, but they will spread the word that chemometrics is a useful discipline. The course they go on will not really give them an insight into chemometrics (how can one in an afternoon?) but will teach them how to put data through a package and learn to use software and will catalyse the wider name recognition of the subject.

Many of the early successes in chemometrics were in quite narrowly defined areas, NIR spectroscopy being one of them. In fact in NIR spectroscopy many of the challenges are not so much with the multivariate statistics, but with the preparation of data. Baseline correction, multiplicative scatter correction, derivatives, smoothing etc. are all tools of the trade and have been developed to a high degree and incorporated into most commercially available NIR chemometrics software. The user spends more time on this than on the front end multivariate analysis – and indeed NIR spectroscopy is a technique for which, if the data are correctly processed, excellent results can be obtained, but this is more in the area of visualization of multivariate trends that could usually be determined using other methods. Still this provided an impetus, and most NIR software contains some chemometrics capability, particularly in calibration. NIR spectroscopists often feel they
understand chemometrics but maybe this is not so, particularly in pattern recognition. However since many NIR problems are quite straightforward from the chemometrics point of view (see in this text Case Study 2 – NIR of Food), in many cases no harm has come, and this at least demonstrates the tremendous power of multivariate methods for simplifying and visualizing data, even if the difficult part is the spectroscopic data handling, and so NIR spectroscopy can be considered correctly as an early success story and an important historic driving force of the subject.

The problem is that over the past decade new sources of data have come on-stream, and this is particularly confusing many analytical chemists. The development of metabolic profiling, e.g. using coupled chromatography, mass spectrometry and nuclear magnetic resonance spectroscopy, has had a very fast development, with improved, more sensitive, and automated instruments. It looks easy, but it is not. The problem is that datasets have now become vastly more difficult to handle, and are no longer the easy NIR spectra that have comforted chemometricians for a generation. The potential application of chemometrics to analytical data arising from problems in biology and medicine is enormous, but often the experimentalists have little understanding of how to acquire and handle these data. They want to learn but have only the odd afternoon or downloaded package with which to learn. They are funded to obtain data not to spend a year learning about Matlab. They usually want quick fixes. A few turn to collaborators, but sometimes the collaborator will say their data are not good enough, or the experiments need to be designed differently, or that it will take a long time and lots of resources and be very expensive, and so in many cases these collaborations don’t develop or stop at a small pilot study or a paper. The biologists are anxious to be first to publish their ‘marker compounds’ and to claim that their work is a success and see data analysis as the afterthought that can be done on a Friday afternoon once all experiments are complete. So they will turn to the user-friendly packages and afternoon workshops and learn how to use the mouse and the menu and get a graph for incorporation into their report and then move on to the next project.

Many do not realize that the methods they are using probably were developed for different purposes. Most chemometrics methods have their origins in traditional analytical chemistry, where there are often underlying certainties, for example in calibration we know what answer we are aiming for and as such just want to get our multivariate method as close as possible to the known answer. In some of the original applications of chemical pattern recognition such as spectroscopy we know what the underlying groups of compounds are and want our methods to classify spectra as effectively as possible into these groupings. We aim for 100 % accuracy and the original algorithms were considered to be better the more accurate the answer. With nice reproducible spectra, a known solution, and no hidden factors, this was possible. But there often is no certain answer in biology, for example, we cannot be sure that by measuring some compounds in a patient’s serum that we can predict whether they will develop kidney disease within the next five years: we are uncertain whether there will be an answer or not. We are testing hypotheses as well as trying to obtain accurate predictions, and now do not just want to predict properties with a high degree of accuracy, but also to determine whether there really is sufficient information in the analytical data to detect the desired trend. Overfitting involves overinterpreting data and seeing trends that are not really there. Many biologists do not have a feel for whether data are overfitted or not. One can start with purely random data and by a judicious choice of variables end up with graphs that look as if two arbitrarily selected groups are
separate. Most people when submitting a paper for publication will actively seek out the graph that ‘looks better’ even if it is misleading. Variable selection prior to pattern recognition is common in chemometrics even though it can be dangerous. A reason is that if we know there is a certain answer we want the graphs to look their best. If we have a series of NMR spectra of compounds differing in only a small region, for clarity we often show only that part of the spectrum that differs, to emphasize why we are confident in our assignment – this is not dishonest, it is normal scientific practice as the rest of the spectrum may be irrelevant. However in pattern recognition it is like wiping evidence that we disagree with so that the final picture confirms our prejudice (which, if correct, may result in publishing a paper, getting another grant or obtaining a PhD). Many people do this unwittingly because they are not aware of the problem and many packages try their best to present the data in a way they think the user likes – and people pay money for this. With an inadequate background it is often hard to assess what comes out of the ‘black box’. With an inadequate background it is often hard to assess what comes out of the ‘black box’. A final problem is that there often is enormous variability in biological samples. In traditional analytical chemistry the variability is often much more limited, and so groups of samples are much easier to define. In biomedical studies there are often a large number of factors that can influence the chemical signal, and so groups can be much harder to distinguish. This often means one may need tens or hundreds or even thousands more samples to obtain an adequate result. Many traditional papers on chemical pattern recognition are published on quite small sample sets – in some cases this is justified – for example it is probably quite reasonable to try to distinguish ten ketones from ten esters spectroscopically – but once translated into metabolomic applications (or indeed in other related areas such as environmental or forensic investigations), there are so many other factors that can influence variability that some papers using limited sample sizes are quite misleading. If one has many variables, e.g. hundreds of GCMS peaks, on small sample sizes, it is always possible to find some variables that appear good discriminators – just by chance – like tossing a coin ten times and repeating this experiment over and again hundreds of times – there will always be a few occasions where we obtain eight or nine ‘heads’ and if these are selected in the absence of the other tosses, it will look like the coin is biased.

How to solve this is not easy as the problem is not technical but about persuasion and education. In areas such as bioinformatics and chemoinformatics it is now generally accepted that there are specially trained informaticians who do nothing all day except process data and mine databases. The need for these specialists is accepted by the market. One reason why chemometrics has to function differently, is that in the informatics areas, most data analysis is done post event, over five, ten or even twenty years after the data have been obtained. There are many databases available, and the aim of the informatician is to interpret trends that might have been lurking in public (or company) databases accumulated over many years. Chemical structures, pKas, quantum mechanical descriptors, do not change much over time, and so can be gradually built up: and many sponsors insist that these data are made available to the scientific community, and so there are not too many barriers. But most chemometrics is a ‘here and now’ subject. The experiments and analysis are performed in real time, and often answers are required during or immediately after the project, examples being process analysis or clinical experiments. One cannot afford to wait years for the results. Plus although there are databases available, often there are incompatibilities between instrumental techniques, for example chromatography varies according to instrument and conditions, and also over the years methods improve, old machines are
discarded and so on; therefore really large international and compatible databases are often hard to set up, and most databases relate to specific applications performed by particular groups. However, with the ‘here and now’ need for chemometrics expertise and often the commercially sensitive nature of large databases, only limited information is available in the public domain and there is much less emphasis on long term data mining and interpretation, and more on integrating chemometrics into the daily life of a laboratory. Hence most chemometrics groups or companies or institutes are primarily geared up to solving other people’s problems in real time as this is where the funding comes from, and there is comparably little long term investment in development, in contrast to areas such as bioinformatics where sponsors are patient and it is accepted that the product may be many years down the line, so long as the product is a good one, of course. So although there are several groups that do train chemometricians, and give them time to pick up fundamental skills, the market emphasis is on producing something that can be used on the spot in the laboratory.

The problem is that software development costs money and takes time, especially user friendly software. It can take a hundred times longer to produce a graphical user interface than to develop the underlying algorithm. Many companies are reluctant to invest such resources. The market likes colour graphics and menu driven screens, and prefers this to some advanced statistical output that probably cannot easily be understood. Many of the main chemometrics software companies were founded in the 1980s, where, as we have seen, there were quite different needs, and are locked into methods that were excellent then but maybe not so appropriate now. The cost of software development can be enormous, and the market for chemometrics software is not enormous. If we estimate that one billion copies of Windows have been sold, we might estimate the total global market in chemometrics software to be between 10 000 and 100 000 or between 1/(100 000) and 1/(10 000) of that for Windows. This limited market means that to sell one has either to have a cheap product that is not very sophisticated or a front end product that is extremely expensive but with a narrow market. The problem here is that this limits capacity for radical redevelopment of the underlying philosophy of a software package. To produce a new package that trashes the old one is often not very good for the established market, but also a huge amount of work and a big risk. Twenty years ago the emphasis on a good user interface was much less – the market would be for people that in themselves would tolerate teleprinter output and awkward ways of typing in data and the occasional crash, and so the cost of getting something accepted on the market would be much lower. Nowadays, even with the best algorithms in the world, if one wants to sell, especially to the upper end of the market, the interface must be excellent. So there have been only modest changes in the underlying algorithms in the best established packaged software over many years, with just a slow and conservative evolution in the nature of the underlying algorithms, despite a revolution in the way chemometrics pattern recognition can be applied, in the quantity of data and breadth of problems, and the capability of computers.

So who should be doing chemometrics? One additional feature of chemometrics is that the number of variables often far exceeds the number of samples – something traditional statisticians rarely encountered. Plus often there is a lot of interest in interpreting the variables and finding out which are important – after all, variables relate to chemicals and so provide insight into the chemical process. Many traditional statistical and machine learning procedures were not designed for this – mainly we are interested in distinguishing
samples (or objects or subjects) and not why they can be distinguished. Hence we cannot just slot into the existing data analysis community, as these people are not experienced in dealing with chemometrics projects that have become increasingly multivariate over time.

Of course ideally practitioners of chemometrics should have good programming ability and be able to understand the basis of the algorithms and statistical methods and this would ideally solve the dilemma of the occasional user of packaged software trying to navigate through the minefield of pattern recognition. But this is never likely to be and the community will always be very diverse. A major aim of this text is to bridge the gap between application scientists, and the more theoretical literature that is excellent but often inaccessible to the practicing user of chemometrics methods, and illustrate the principles primarily graphically. This text is therefore aimed at gaining insight into the use and ideas behind pattern recognition within a chemometrics context, without delving too far into the theory. It is not however a software package, although we have developed a Matlab code for the methods in this text, but not in the form of a user friendly GUI (Graphical User Interface). The need to bridge the gap between theory and practice in the area of chemometric pattern recognition is particularly pressing with the rapid growth of applications in the area of biology and medicine, and perhaps was not so evident ten years ago when the focus was more on process monitoring and on spectroscopy which pose different challenges.

It is hoped that readers of this text will gain much more understanding of the ‘whats and whys’ of pattern recognition within a modern chemometric context. Some readers will decide to continue this self learning and turn into programmers themselves (this book is not a programming manual though), developing and applying methods to their own and other people’s data. For a graduate student or consultant or independent researcher, this text will provide an important springboard. Other readers may feel they have not the time, or cannot cope with, the necessary technical programming, but still will gain from the insight and use pattern recognition methods with care in the future. They may decide to team up with a colleague who is good at developing numerical methods, or form a collaboration, or go to a consultant, or even hire someone to do this for them if there is enough work. There will be a group of people from disciplines such as statistics, machine learning and chemical engineering that will gain insight into which methods are most applicable in chemometrics and how these can be applied in real world practice, and it is hoped that it will attract some such people back into the subject, even at a periphery as many have unfortunately deserted over the years. Finally there will be some who are applications scientists and not maybe at the moment able to spend the time learning about chemometrics, who will have heard of pattern recognition in a conference or read a paper or be thinking whether chemometrics can help their work, who will benefit – as a highly visual case study based text, skimming the chapters and looking at the diagrams and diving in and out of different sections will give a flavour of what is possible and allow planning as to whether to invest time or resources into applying methods to their research in the future.

1.2 About this Book

There have been a few books about pattern recognition in chemometrics, as listed in the Bibliography, but most of these are quite old, and are primarily algorithmic descriptions. Much of the development of chemometrics in the 1980s and 1990s involved cataloguing
methods – formally describing algorithms, listing them and often applying them to a few benchmark datasets. Many of those using chemometrics would describe themselves first and foremost as analytical chemists, and indeed, this is where the majority of papers in chemometrics have been published: indeed the two subjects have developed hand in hand with some saying that chemometrics is the theoretical basis of analytical chemistry. But analytical chemists tend to like to list things. Previously we discussed the unusual psychology of chemists, and many chemists, whilst having many other skills, are scientific ‘stamp collectors’, and so a lot of the early pattern recognition books read like a list of methods, often though without the statistical motivation.

Surprisingly, despite a huge increase in the number of chemometrics texts over the past few years, with several well established ones in niche areas such as calibration, signal resolution and experimental design, and the large growth of pattern recognition texts in areas such as machine learning, there has been very limited development of such texts in the subject of chemometrics. Some of the historic reasons why this may be so are discussed above.

This current book does not aim to be an encyclopaedia. There are probably a hundred or more named ‘methods’ in the chemometrics literature for classification in the chemometrics literature. In order to publish a paper, win a grant or obtain a PhD it is often useful to propose a ‘named’ method. Some such methods become widespread, others are reported in one or two papers, cited half a dozen times and forgotten. Most named methods do indeed involve one or more steps that are different to existing published methods, but if every defined method for pattern recognition contained ten steps, and each step had five alternatives, there would be ten million combinations of alternatives, and so plenty of possibilities of announcing a new set of alternatives, for example, one could study all existing methods and find the combination of alternatives that is most different and publish a paper on this, and so on. This plethora of ‘named’ methods can be quite confusing, and it is not the intention of this author to describe all in detail. However there are only a few underlying principles and so it has been decided in this book to focus on a smaller number of generic methods and approaches and to discuss the motivations behind the methods rather than focus on a comprehensive list of methods. All the widespread ones including Principal Components Analysis, Linear Discriminant Analysis, Partial Least Squares Discriminant Analysis, Support Vector Machines, SIMCA and Self Organizing Maps are discussed, plus some others, but the choice is very much based on illustrating generic principles rather than being comprehensive.

The methods are illustrated by their use in analysing a variety of case studies chosen for their applicability in many branches of science, including biology, medicine, materials characterization, food, environment, pharmaceuticals and forensics, and many analytical techniques including MS, LCMS, GCMS, NIR Spectroscopy, NMR Spectroscopy, Thermal Analysis and Atomic Spectroscopy. For readers unfamiliar with these techniques (such as those from the machine learning or statistics community) it is not necessary to have an in depth appreciation of the analytical techniques to understand this text. The techniques and applications are biased a little towards this author’s experience but attempt to be broad enough to be of interest to a wide variety of readers. For reasons of brevity only the most common techniques (such as PCA) are illustrated on all case studies, and in most cases we choose specific examples where it is most useful to illustrate the technique.
There also needs to be a decision as to the depth with which each method is discussed. Some, such as Partial Least Squares, can quite easily be described in adequate detail in a few pages, but others such as Support Vector Machines, could occupy a full text in its own right to adequately describe the maths. So the policy in this book is to allocate approximately equal space to techniques according to a judgement as to their relative importance, whilst in all cases trying to aid understanding using diagrams. In most cases statistically based methods are described in sufficient detail that a good programmer can reproduce the results if need be. All results in this text were obtained using in-house software written in Matlab which aided our understanding of the methods.

Decisions have also been made as to which topics to discuss in detail. The main emphasis of this book is on classification methods which are the focus of Chapters 5 to 11, although exploratory data analysis and data preprocessing are also discussed in some detail. Areas that could be expanded include Multiblock approaches (Chapter 12) and ANOVA based methods for looking at the significance of variables (Chapter 9). Both these could form books in their own rights, and for specific groups of chemometricians are considered very important as they pose interesting potential development areas. But practicing laboratory based chemists or biologists only rarely come across the need for such methods and so they are introduced only in a limited form although in sufficient detail to provide readers with a springboard and complete description of methods where needed. Multiway methods which are hotly topical in certain circles are not described in this book as they are already the subject of another excellent and dedicated text, and this book could not do justice to them except in a very peripheral way.

The references associated with each chapter are not aimed to be comprehensive but to allow readers a springboard should they wish to explore certain topics further. They are primarily chosen as references that my coworkers and I have found useful when trying to understand topics in more depth. They also include precursor papers published from the Bristol group which often expand on explanations and examples in this text around certain key topics.

Readers of a book such as this are unlikely to work through in a linear fashion (from beginning to end), and will probably come from different backgrounds, dipping in and out of the book. One good way of understanding methods (if you are a programmer) is to try to reproduce results or apply methods to your favourite datasets and see what happens. Some methods are iterative and not completely reproducible and so it is not always possible to obtain identical results each time, but the general principles should be comprehensible and it is hoped that this book offers insights into what is possible and what the fundamental assumptions are behind this.

Finally whereas we recommend certain protocols (for example repeatedly splitting datasets into test and training sets many times), this is not prescriptive, and a good scientist may come up with his or her own combination of techniques: this book gives important information about the building blocks and what we recommend, but other people may piece together their house differently. Furthermore some of the illustrations using PC scores plots are for purpose of visualization to understand the basis of methods and in practice one may use more PCs for a model but then not be able to visualize the results; however by experience a simple visual representation of a method makes it much easier to understand than by describing it in several pages of equations, although of course, we do not neglect the necessary algebra where appropriate.
Bibliography

General Chemometrics Texts


Chemometrics Texts Focused on Pattern Recognition

New York, NY, USA, 1975.
D. Coomans, I. Broeckaert, *Potential Pattern Recognition in Chemical and Medical Decision