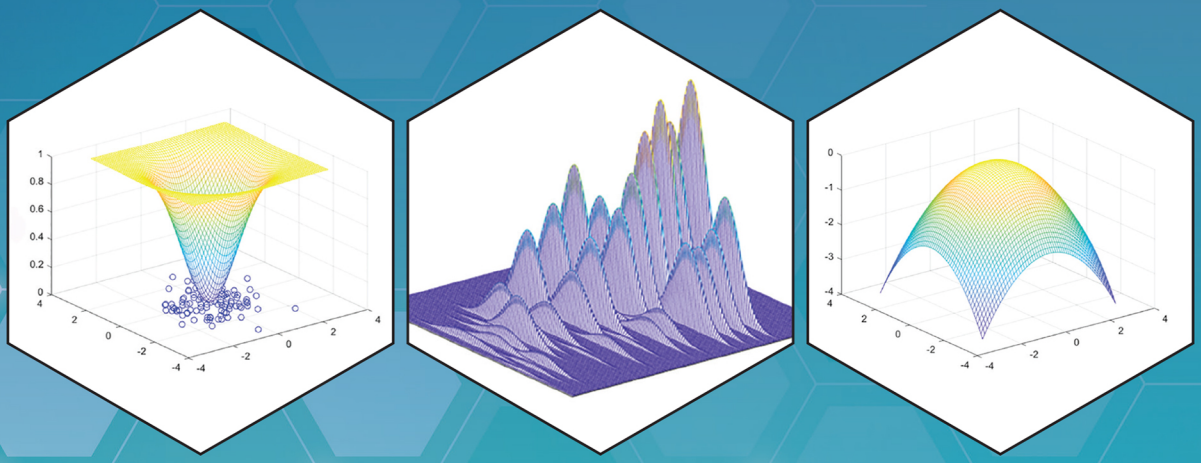
DATA ANALYSIS AND CHEMOMETRICS FOR METABOLOMICS

RICHARD G. BRERETON



with website

WILEY

Data Analysis and Chemometrics for Metabolomics

Data Analysis and Chemometrics for Metabolomics

Richard G. Brereton
University of Bristol
United Kingdom

WILEY

This edition first published 2024

Registered Offices
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Limit of Liability/Disclaimer of Warranty
In view of ongoing research, equipment modifications, changes in governmental regulations, and the constant flow of information relating to the use of experimental reagents, equipment, and devices, the reader is urged to review and evaluate the information provided in the package insert or instructions for each chemical, piece of equipment, reagent, or device for, among other things, any changes in the instructions or indication of usage and for added warnings and precautions. While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional statements for this work. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

Library of Congress Cataloging-in-Publication Data:

Names: Brereton, Richard G., author.
Title: Data analysis and chemometrics for metabolomics / Richard G. Brereton.
Description: Hoboken, NJ : Wiley, 2024. | Includes index.
Identifiers: LCCN 2024002533 (print) | LCCN 2024002534 (ebook) | ISBN 9781119639381 (hardback) | ISBN 9781119639374 (adobe pdf) | ISBN 9781119639404 (epub)
Subjects: LCSH: Metabolites. | Chemometrics.
Classification: LCC QP171 .B74 2024 (print) | LCC QP171 (ebook) | DDC 572/.4—dc23/eng20240307
LC record available at https://lccn.loc.gov/2024002533
LC ebook record available at https://lccn.loc.gov/2024002534
Hardback: 9781119639381

Cover Design: Wiley
Cover Image: © berCheck/Shutterstock; Courtesy of Prof Richard G. Brereton

Set in 10.5/13pt STIX Two Text by Straive, Chennai, India

Contents

Foreword

The term *metabolomics* was first introduced in the early 2000s, when spectroscopic and chromatographic techniques were being developed for profiling of metabolites within cells and tissues. The capabilities of instrumental techniques, primarily NMR, LCMS and GCMS, to obtain large quantities of such data resulted in big datasets. Common to other emerging omics technologies such as genomics, transcriptomics and proteomics, computational methods were required to make sense of this data.

Computational methods for processing data can be applied to two steps of the metabolomics workflow. The first is for resolving and characterising raw instrumental data primarily from coupled chromatography and NMR spectroscopy, to produce peak tables for subsequent statistical processing. We list the main packages currently available in this book and their underlying principles, but this is not the focus of the text. These packages evolve rapidly, and most will change over the lifetime of this text. In addition, many involve proprietary or very complex and ever-evolving algorithms often linked to expanding databases, which would be difficult or impossible to describe in detail. Finally full information is not always easily available from the developers or main users of some of the packages, but there are regularly updated websites with user manuals to which readers should refer. We have tried to gather information on most of these approaches in Chapter 2, all of which are unique, but some have been developed and are maintained by large teams, and it would be unrealistic for a reader of this book to reproduce these methods and not all developers are easily forthcoming about details.

In contrast, approaches for statistical or chemometric processing have a long vintage and are likely to remain available in decades to come in a similar form to now, and all common methods are public domain. A textbook is designed to have a long lifetime, and the focus of this text is on commonly available chemometric methods. Many approaches and concepts used in current metabolomic statistical analysis were first formulated over 100 years ago, such as p values, ANOVA, distributions, least squares regression, PCA and so on. Another set of methods emerged around 50 years ago, such as PLS and SIMCA when there developed a more widespread need to interpret multivariate analytical data. A few are more recent such as ASCA. This text is an aid to modern-day research but not a theoretical text reviewing the latest chemometric methods proposed in the literature.

The choice of methods in this book is based on this author's perception of some of the most widespread in current practice, based on the literature, on talking to practicing colleagues and implemented in widespread packages. Of course there will be many other methods, and a comprehensive description of all chemometric approaches used

in metabolomics would be a series of texts: unfortunately such a series of texts would take many years to compile and probably involve several authors, and with the rapid development of this field the first book would almost certainly be dated when the last one appears. Within a single text and a single author, one can only describe the most common, but the advantage is that there is uniformity in presentation, so the methods are described in similar depth using similar notation and datasets. Most users are only exposed to a few of the most widespread methods.

This book can be used at various levels. At the top level, it is a description of the use and basis of the most common chemometric methods, illustrated by case studies. Readers may discover methods they had not been aware of, or interpretations they had not previously appreciated. This book advocates a hypothesis-based approach, as most of metabolomics is hypothesis driven. At a deeper level, some will want to follow the calculations in the book. This will enable understanding of the methods and can enable comparison with in-house software. There is sufficient description of how the methods are implemented and sample output to allow readers to compare with their own calculations; if there are differences, this may lead to changes in software usage or catalyse additional interpretation of data. Sometimes there are several comparable methods, and users may want to look at their results from different angles.

Although this author performed all calculations with in-house software developed using MATLAB, there is no requirement to use this package to reproduce results. Some numerical output has been compared by myself and by colleagues using other approaches, notably R, Excel (with VBA), PLS Toolbox and SIMCA, to both ensure identical results (except for the sign in PCA) and check numerical accuracy. Readers will be using a wide variety of favoured software environments. This author has over many years performed chemometrics calculations in MATLAB, Fortran, SAS, BASIC, Excel (with VBA), C and PL/1 where appropriate and has co-operated with colleagues who have additionally used R, PLS Toolbox, SIMCA, UNSCRAMBLER, Pirouette, Sirius and Minitab. There are, of course, many other packages available suitable for the statistical analysis of metabolomics data. However, unlike the methods for resolution and characterisation of instrumental data, if correctly used, all these packages should come to identical answers. Most low-level programming languages allow all the steps in an algorithm to be coded in, and many high-level environments allow for scripting or macro commands to swiftly develop applications without complex programming. Some users will not want to do any programming and want more or less automated laboratory-based software, although usually some functionality is available by menu commands – approaches such as PCA and PLS should, if steps such as pre-processing are performed as described in this text, result in identical answers, providing there is sufficient flexibility in the software.

The 10 case studies in this book have been carefully chosen to span a range of applications. We have focussed on the main instrumental workhorses, namely NMR (3 case studies), LCMS (2 case studies) and GCMS (2 case studies). Raman (1 case study), FTIR (1 case study) and single wavelength HPLC (1 case study) are used by some investigators who view themselves as working in metabolomics; therefore to satisfy such readers, we have also included data from these sources. The case studies come from human (4), plant (3), animal (2) and microbial (1) experiments reflecting a range

of applications, with data sources from Europe (6), Asia (2), North America (1) and Africa (1).

Some case studies are used more frequently to illustrate different methods than others, some being just illustrated by PCA, whereas others are mentioned in three or four chapters. There are excellent articles describing the full analysis of each case study referenced in the text, and it is not the aim of this book to copy the existing literature but to describe the main approaches and illustrate them where appropriate with one or more case studies using one or more steps in the analysis. The case studies can be downloaded from the companion Wiley website and are all in Excel format. The case studies are supplemented by a small number of simulations, with the larger simulations also available for download.

Although chemometric methods are widely recognised as essential to the analysis of metabolomic data and there are many texts on general chemometrics methods mainly aimed at analytical chemists, there is a lack of books focussed on their application to metabolomics. It is hoped that this text will be a useful reference.

In addition to a primary focus on metabolomics, this book will also be of interest to the general user of chemometrics in related fields, covering most of the common methods such as PCA, PLS, calibration, classification, experimental design and so on. It will also be of interest to the applied statistician interested in methods used in chemometrics. For these readers, the choice of and relative importance of methods discussed in the text are oriented towards metabolomics, and case studies are also related to data encountered in this field, but the applicability of the statistical approaches can easily be transferred to other fields.

Richard G. Brereton
December 2023 *Bristol, University of Bristol, UK*

Acknowledgments

In the course of preparing this book, over a three-year period, I have been helped by a large number of colleagues around the world.

Several have generously suggested datasets, some after significant discussions looking at which would be most suitable for the purposes of this book. It was necessary to span various applications and techniques so the sources have been likewise varied. Some have generously given their time for personal meetings where I discuss their protocols, and many have carefully commented on relevant sections. This has been particularly important to ensure that the book encompasses a range of different applications.

Several colleagues have checked my calculations by running data from this book through other packages, especially where the results obtained appeared unexpected, and to check the results of my programs in MATLAB agree with those from software in other environments, and that my explanations are sufficiently detailed.

Furthermore, many colleagues have read sections of the text often providing detailed corrections and comments. Some have expressed opinions and explanations for different approaches, often from different viewpoints to my own. Having a wide variety of opinions and experiences has helped shape this text.

A list of collaborators, who have shared knowledge, data and expertise is below.

Trygve	Andreassen	(Norwegian University of Science and Technology, Norway)
Paul	Benton	(Scripps, USA)
Elizabeth	Carter	(University of Sydney, Australia)
Olivier	Cloarec	(Sartorius, France)
Catherine	Deborde	(INRAE, France)
David	Duewer	(NIST, USA)
Oliver	Fiehn	(UC Davis, USA)
Roy	Goodacre	(University of Liverpool, UK)
Andris	Jankevics	(University of Birmingham, UK)
Olav M	Kvalheim	(University of Bergen, Norway)
Gavin	Lloyd	(University of Birmingham, UK)
Gary	Mallard	(NIST, USA)
Annick	Moing	(INRAE, France)
Tomas	Pluskal	(IOCB Prague, Czechia)
Alexey	Pomerantsev	(Federal Research Center for Chemical Physics, Russia)
Francesc	Puig	(INSERM, France)
Rich	Sleeman	(Mass Spec Analytical, UK)
Hans	Stenlund	(University of Umeå, Sweden)

Izabella	Surowiec	(Sartorius, Sweden)
Roma	Tauler	(IDAEA-CSIC, Barcelona, Spain)
Johan	Trygg	(University of Umeå, Sweden)
Yulan	Wang	(Nanyang Technological University, Singapore)
Yun	Xu	(University of Liverpool, UK)

Wiley are also thanked for their patience and encouragement throughout and Jenny Cossham for suggesting this text.

About the Companion Website

This book is accompanied by a companion website:

www.wiley.com/go/Brereton/ChemometricsforMetabolomics 

This website includes:

- Case Study Data
- Simulated Data

The website hosts the datasets in this book as two downloadable Excel files:

1. The file “case study data” contains the data for all 10 case studies in this book.

2. The file “simulated data” contains data for the larger simulations in this book.

These datasets can be used freely for private study or for use in courses. If in courses, please reference the book. The datasets may also be used if required in publications or presentations. If using case study data, please cite both the book and the original source (as cited in Chapter 1), and for the simulated data, just the book.

The datasets can be used to reproduce numerical and graphical results from relevant chapters of the book or used for further exploration.

It is recommended to export the data to an external package for further processing.

CHAPTER 1

Introduction

The subject matter of this book is a synthesis between chemometrics and metabolomics, both relatively recent scientific disciplines. This chapter describes the background to these disciplines and then introduces the background to the case studies which are used to illustrate the chemometric methods and describes some software packages that can be used to obtain results described in this text.

1.1 CHEMOMETRICS

The name chemometrics was first proposed by Svante Wold in 1972 in the context of spline fitting [1]. Together with Bruce Kowalski, they founded the International Chemometrics Society and the term slowly took off in the 1970s. However, the pioneers did not widely use this term for some years, but a major event that catalysed it was a workshop in Cosenza, Italy, in 1983 [2] where many of the early pioneers met. After this time several initiatives took off, including the main niche journals, *Journal of Chemometrics* (Wiley) [3] and *Chemometrics and Intelligent Laboratory Systems* (Elsevier) [4], together with courses and the first textbooks [5, 6] with regular reviews and ACS (American Chemical Society) symposia starting a few years earlier [7].

However, these events primarily concern name recognition and organisation, and the main seeds for the subject were sown many years earlier.

Applied statistics was one of the main influences on chemometrics, although the two approaches have diverged in recent years. The modern framework for applied statistics was developed in the early 20th century and we still use terminology first defined during these decades. Before that, early academic statistics was mainly mathematical and theoretical, often linked to probability theory, game theory, statistical mechanics, distributions etc. and viewed as a subdiscipline of mathematics. Although many early pioneers had already used approaches previously that we would now regard as the

Data Analysis and Chemometrics for Metabolomics, First Edition. Richard G. Brereton.

Companion website: www.wiley.com/go/Brereton/ChemometricsforMetabolomics

forerunners of modern applied statistics, their ideas were not well incorporated into mainstream thinking until the early 20th century.

A problem in the 19th century was partly the division of academic disciplines. Would a mathematician talk to a biologist? They worked in separate institutes and had separate libraries and training. For applied statistics to develop, less insular thinking was required. There also needed to be some level of non-academic contribution as many of the catalysts were at the time linked to industrial, agricultural and medical problems. With core academic disciplines, the application of statistical methods in physics and chemistry, which would eventually progress to quantum mechanics and statistical mechanics, fell outside mainstream applied statistics and has led to specialist statistically-based methods that are largely unrelated to chemometrics.

However, in the first three decades of the 20th century, there was a revolution in thinking. Such changes primarily involved formalising ideas that had been less well established over the previous decades and even centuries. Karl Pearson [8] and William Gossett publishing under the pseudonym 'Student' [9] are recognised as two of the early pioneers. Pearson set up the first statistics department in the world, based in London, and his 1900 paper first introduced the idea of a p value, although historic predecessors can be traced several centuries back [10–12].

It was not until after the First World War that applied statistical methods were properly formalised in their modern incarnation. Ronald Fisher was possibly the most important figure in developing a modern framework for statistical methodology that many people still use today. In 1925 he published *Statistical Methods for Research Workers* [13] and established the concepts of p values, significance tests and ANOVA (analysis of variance). Ten years later he wrote a book that described the basis of almost all statistical experimental designs [14] used even now, and his paper on classification of irises (the plants) [15] is an essential introduction to multivariate classification techniques, with this dataset used even now for demonstrating and comparing new approaches. Other important workers over that period, included Harold Hotelling, who among others was attributed with progressing the widespread use and recognition of PCA (principal components analysis) [16, 17] and Jerzy Neymar and Ergon Pearson who developed alternative approaches to hypothesis tests to those proposed by Fisher [18].

During the interwar period, many of the cornerstones of modern applied statistics were developed, and we continue to use methods first introduced during this era; many approaches used in chemometrics have a hundred-year vintage. However, there were some significant differences from modern practice. There was no capacity to perform intensive computations or generate large quantities of analytical data, so applications were more limited. Agriculture was at the forefront. During this era, the old landowning classes had to modernise to survive: many farm labourers left for the cities and agriculture became more automated. The relationship between landowners and tenants weakened and larger farms were viewed more as an industry rather than the birthright of aristocratic classes. This required a significant change in production, and agricultural statistics was very important, especially to improve the economies of Western Nations. Other important driving forces came from the use of psychology to interpret test scores, and from economics. Common to all these types of data is that experiments involved considerable investment in time, so it was reasonable to spend substantial effort analysing the results, some required weeks of manual calculations, as

data was expensive and precious. In modern days, spectra, in contrast, can be obtained relatively rapidly and quickly, so spending days or weeks performing statistical calculations would be an unbalanced use of resources.

Furthermore, without the aid of computers many of the multivariate methods we now take for granted would involve a large amount of time. Salsburg [19] claims as follows: 'To get some idea of the physical effort involved, consider Table VII that appears on page 123 of *Studies in Crop Variation. I.* [20]. If it took about one minute to complete a single large-digit multiplication, I estimate that Fisher needed about 185 hours of work to generate that table. There are fifteen tables of similar complexity and four large complicated graphs in the article. In terms of physical labor alone, it must have taken at least eight months of 12-hour days to prepare the tables for this article.' Of course, Fisher would have had many assistants to perform calculations, and he would have been very well resourced compared to most workers of the time. Hence, only quite limited statistical studies could be performed routinely. Some algorithms and designs such as Yates' algorithm [21] were developed with simplicity of calculation in mind as the data had special mathematical properties and although still reported in some textbooks even now are not so crucial to know about with the advent of modern computing power. Computers can invert large matrices very quickly, whereas a similar calculation might take days or longer using manual methods. In areas such as quantum chemistry, a calculation that may take up an entire PhD via manual calculations can now be done in seconds or less using modern computing.

The statistician of the first half of the 20th century would be armed with logarithm tables, calculators, slide rules and special types of graph paper, and in many cases would tackle less data-rich problems than nowadays. However, there was a gap between the mathematical literature where quite sophisticated methods could be described, often in intensely theoretical language, and the practical applications of much more limited and in most cases simpler approaches. Many of the more elaborate methods of those early days would not have much widespread practical use, but modern-day multivariate statistics can now take advantage of them. The chemometrician can routinely use methods that on very large spectroscopic or chromatographic datasets that were inconceivable prior to the widespread availability of modern computers.

In the post-war years, chemical manufacturing was of increased importance and multivariate methods were applied by industrial chemical engineers [22]. G.E.P. Box worked with a group in the chemical company ICI in the UK for some years, before moving to the US. His text [23] written together with two co-authors, is considered a classic in modern statistical thinking for applied scientists emphasising experimental design and regression modelling and brings the work of the early 20th century into the modern era.

In the 1970s, mainstream applied statistics started to diverge from chemometrics. In chemometrics, we often come across short fat datasets, where the number of variables may far exceed the number of samples. For example, we may record thousands of mass spectral or NMR or chromatographic data points for each of perhaps 20–100 samples. These sorts of problems were not conceivable to the original statistical pioneers, measurements were expensive, so variables were scarce. Fisher's classic iris data [15] consisted of 150 samples but only four variables. Once sample sizes are less than the number of variables, some classic approaches for multivariate data analysis are no longer directly applicable, an important one is the Mahalanobis distance [24]

and the corresponding method of LDA (linear discriminant analysis) [15] which have to be adapted and cannot be directly applied to data whose variable to sample ratio exceeds 1. More recent methods such as PLS (partial least squares) [25, 26] and SIMCA (Soft Independent Modelling of Class Analogy) [27] were advocated in the 1970s and 1980s very much with the needs of chemometricians in mind and coped well with such data. Chemometricians are often very interested in variables, for example, which are most significant out of possibly hundreds of candidates, whereas statisticians emphasise the significance of factors, in many cases using univariate tests. Although both types of thinking may come to similar types of conclusions, some traditional statistical approaches are invalid, as an example if there are more variables to samples, we cannot use LDA to provide us information as to which variables are the most significant, whereas PLS might provide an answer. As such problems were inconceivable before the 1970s, the impetus to providing niche solutions for the chemometrician is of 50-year vintage. Metabolomics is often a rich source of information where the number of variables far exceeds the number of samples so chemometrics is needed for statistical interpretation.

Although there are still a few workers in the chemometrics field who identify themselves as statisticians, their influence became quite limited after the 1980s, judging by attendance at chemometrics meetings, development of texts and courses and so on. In a way, this divergence is similar to those in areas such as quantum mechanics or statistical mechanics, which are also founded on statistical principles but are primarily led by numerate scientists.

However, the strong statistical parentage is an important cornerstone of chemometrics. As many applications are moving away from the original ones of quantitative analytical and physical chemistry into more hypothesis-based science such as metabolomics, the original aims of measuring more precisely or predicting more accurately are gradually being supplemented by more statistical aims to generate and test hypotheses. In the latter case, chemometrics, whilst once routed in the physical sciences, is being adapted to tackle problems from those in biology, geology, psychology and so on, and requires a return to core statistical thinking. It is questioned for the future whether this will attract more applied statisticians back into the field, or whether niche chemometrics experts will return to the mainstream statistical literature and then incorporate more statistical thinking into their publications and software without the need to bring mainstream statisticians directly into their collaborations.

Another parent of chemometrics was quantitative chemistry. Historically, it is important to understand that interdisciplinary research was not very widespread during most of the formative years in the 20th century. Academia was highly compartmentalised, with students opting for specialist courses in, for example, chemistry, or mathematics or biology. Most university teachings except at the basic levels would have been by staff from a single department. Students might be introduced to statistical concepts at an early stage but if by mathematicians in a somewhat abstract manner. If carried forward in courses such as physical and analytical chemistry, statistical methods would have been strongly oriented towards univariate measurement and estimation or in specialist areas such as quantum chemistry.

At a research level, departments would have their own libraries and staff rooms. In many countries, to obtain academic positions staff would have to be very focused.

An analytical chemist might have to pass a committee for tenure or habilitation, which is narrowly defined in part because subsequent teaching duties will also be highly compartmentalised. Libraries, journals, books, staff rooms, even sports teams and social events would often be organised by departments. A chemist, if working in the right subdiscipline, could snatch some snippets of statistical thinking where needed, but it would mainly be developed independently within their own environment. An analytical chemist would be unlikely to visit a mathematics library or read a mainstream statistics journal, but would gather their statistical knowledge from analytical journals and books; the internet was not yet available so the ability to search for papers from a wide selection of journals was mainly restricted to visits to other departments' libraries. Analytical chemists would be introduced to niche statistics, for example, learning about precision and accuracy of measurements, but most would have limited exposure to other statistical texts except at a basic level.

Industrial cross-over with mainstream academics was also rather limited until the fourth quarter of the 20th century. Thus, the work, for example, G.E.P. Box and colleagues were doing in ICI in the 1950s would be unknown to chemists in many prestigious universities. Formal statistical design of experiments, so important for improved industrial productivity, would not be adopted by mainstream synthetic chemists in an academic environment until the last decade of the 20th century.

Hence, the development of statistical thinking within mainstream chemistry, such as multivariate analysis and statistical experimental design, developed in quite unique ways. A few brave workers did, however try to introduce statistical and computational ideas to the chemistry community. Statistical methods such as univariate linear regression, determining accuracy and precision of measurements and so on, have been part of the analytical and physical chemists' toolbox for more than a century. However, many techniques now recognised as part of chemometrics were shown to be recognised within the mainstream chemistry community. In 1949, Mandel draws together a number of statistical techniques, including design of experiments and ANOVA [28], in a paper which has been cited just seven times at time of writing (December 2023), an almost forgotten paper. W.J. Youden, a pioneering statistician, wrote 37 articles for *Industrial and Engineering Chemistry* between 1950 and 1957, many on experimental designs and very relevant to analytical chemists and chemical engineers, which in turn have been cited only 60 times in total. In 1952, he published a review in the journal *Analytical Chemistry* [29] citing 154 references, which has, in turn, only received seven citations. Yet a paper in the journal *Cancer* by the same author published in 1950 [30] has received over 7000 citations at the time of writing. It would be a value judgement as to whether these different publications contained more in-depth or original information, and the difference in reception would primarily relate to the difference in readership.

Hence, although analytical chemists were aware of traditional statistical methods, for example, how to fit a straight line or how to determine the 95% confidence in a mean, they were relatively uninterested at the time in the statistical revolution of Fisher and colleagues, and a traditional course in chemistry would not cover systematic experimental design, ANOVA, multilinear regression, multivariate pattern recognition, etc. It took until the 1970s before the importance of these approaches, well established 50 years ago in the mainstream statistical literature, started to slowly become recognised

within the chemistry community. Chemometrics by name was fundamentally born within the chemical sciences, where the first applications of for example chromatography and spectroscopy for the analysis of complex mixtures were developed. As time has progressed, although the first uses of these instruments were within the analytical and physical community, their application to data-rich sciences such as in the biomedical sciences where large mixtures of chemical compounds have a metabolic significance, has led to a much wider applicability of these instrumental techniques, and so the concepts from chemometrics. The original applications of NMR and MS in chemistry were primarily for the structural elucidation of individual molecules and it took several decades before they became established for the studies of mixtures.

It was not until the 1970s that analytical chemists started widely recognising the potential of statistical principles for experimental design. Stan Deming, whose first paper was published in 1967, published a well-cited paper in analytical chemistry on simplex optimisation in 1973 [31] followed up by a more comprehensive statistically based and well-regarded book in 1987 [32].

A parallel development happened within the physical chemistry community. Spectroscopists studied problems involving overlapping peaks of mixtures. They probably did not have much access to the statistical literature at the time, but did read the physical literature for example about eigenanalysis, which is related to PCA. There were several papers in the 1960s of which two are cited [33, 34]. Physical/analytical chemists of the time developed their methods in isolation to statisticians and, due to the limited availability of computers, applied their approaches to what we now regard as quite simple problems.

The prolific pioneer Ed Malinowski put together many of the first approaches to multivariate resolution of mixtures in the 1970s with statistical and algorithmic descriptions, culminating in his classic book published in 1980 [35]. Many of the methods now recognised as multivariate curve resolution that have a high profile in chemometrics, emerge from Malinowski's early work. Unlike most chemometrics methods, for example for classification or exploratory data analysis or regression, factor analysis (as defined by Malinowski) or multivariate curve resolution have a very specific role in chemometrics, so play a unique role; even now many approaches for resolving peaks in coupled chromatography incorporated in elaborate software have their origin in concepts first advocated by Malinowski.

By the 1980s, ideas from quantitative analytical and physical chemistry had started to become formalised both in the area of multivariate analysis and experimental design, and were slowly recognised initially by a rather small group primarily of analytical chemists. In the 2000s, basic chemometric concepts were starting to be introduced in general analytical chemistry courses and books and rapidly developed into software that was essential for burgeoning applications of instrumental analysis such as metabolomics as discussed in Section 1.4.

The third catalyst was scientific computing. Many of the approaches described in the first half of the 20th century remained theoretical without the availability of good computer power, and prior to the 1970s, only really quite simple problems could be solved by chemometric methods.

In the 1950s, very few scientists and engineers had access to computers for daily calculations. Computers were large and expensive institutional machines often developed

for defence (as the initiative originally came out of the Second World War). Individual researchers rarely had direct access, programs had to be sent in using punch cards or paper tape to trained operators, who would then feed the instructions to a mainframe and output would usually be in the form of printer paper sent back to the user. If there was a mistake in a program, this was very costly.

Early programs were in assembly language or machine code, which would be very hard for non-experts and take time to develop simple sets of instructions. In the 1950s, IBM developed a high-level language, called Fortran (Formula Translation) [36], which was the basis of most scientific software for the next two decades. The language continues to be developed with regular updates. Most classical scientific software was written using Fortran. The vast majority of quantum chemistry packages are still written in Fortran, and some build on subroutines over 50 years old. The 1960s was a particularly fruitful time for quantum mechanics computing. Prior to this era, it could take a graduate student almost an entire PhD just to calculate a few quantum mechanical integrals, so the use of scientific computers revolutionised this field.

However, access to mainframes capable of running large scientific programs was very limited until the late 1960s and early 1970s. Scientists who were not viewed as hard-core physical scientists had limited possibilities, and it took some years before access to significant scientific computer power was broadened.

In the very late 1960s, a few chemists did get access to significant computing power. Peter Jurs and his co-worker Bruce Kowalski were some of the first writing a series of papers in the analytical literature [37]. The impetus had been from Djerassi and coworkers who had pioneered the use of computers in structural elucidation [38] leading to the field of artificial intelligence. The Arthur program [39], developed for both mainframes and VAX minicomputers was one of the first widespread chemometrics packages but still one had to be very expert to install and use it.

In the 1980s, two important developments happened in computing that would move chemometrics from a specialist discipline to one that had the potential to be more widespread. The first and most important was the development of microcomputers in the late 1970s to early 1980s, the most successful scientific models of the type based on the original IBM PC [40]. Once micros became widespread and relatively user-friendly, chemometrics software could extend to a far wider user base and allow laboratory-based scientists rather than primarily specialists with access to expensive communal mainframes, to access chemometrics software.

Another important development of the time was MATLAB, originally developed by Cleve Moler [41] in 1981. Most chemometricians like to think in terms of matrices, and languages such as Fortran and BASIC were not naturally oriented towards matrix operations at the time. MATLAB, however, allows the programmer to develop code using matrix and vector algebra directly and contains fast algorithms for operations such as inverting large matrices. Operations such as PCA are also built-in. Since its early days, the software has been substantially expanded with GUIs (graphic user interfaces) and extensive graphics. MATLAB had a significant role in the development of chemometrics because workers could quickly develop matrix-oriented algorithms and often swapped code. Although many developing chemometrics methods now prefer R or Python, there still is an important group of MATLAB users, and this environment strongly catalysed the fundamental developments from the 1980s onwards.

The three catalysts, namely applied statistics, analytical/physical chemistry and scientific computing, converged in the 1980s to provide a fertile environment for the establishment of the discipline. Several events were responsible to create a specific launching pad for what is now well regarded as a coherent body of knowledge.

The NATO-sponsored meeting in Cosenza, Italy, in 1983 [2] brought together many of the experts working in this field at the time, not all would have regarded themselves as fundamental chemometricians until then. The journal *Analytical Chemistry* started publishing biennial fundamental reviews under the name 'chemometrics' as from 1980 [7] bringing together the last two years of papers in the field. Several regular workshops were established. The first comprehensive texts were published [5, 6] although some more specialist books had been published covering specific areas, but without the name 'chemometrics' in the title, in the years before. Two journals were established, published by Wiley [3] and Elsevier [4]. Several series of conferences were started at this period. These attracted mainly niche workers, most of whom had a back in computing within a chemistry environment and were not yet attracting biologists. There was a separate and very well-established area of biological statistics, mainly oriented towards univariate data.

The applications at this phase were fairly simple, with NIR (near infrared) calibration and HPLC (high-performance liquid chromatography) deconvolution predominating, and relatively small sample sizes. Some approaches such as PLS [26, 27] and SIMCA (self independent modelling of class analogy) [27] as well as Malinowski's extensive methods which he called factor analysis [35] were very oriented towards and widely reported within the chemometrics community rather than using a general statistics toolbox, although PLS (originating in economics) has found a home more generally since. The emphasis at the time was primarily to be able to measure and estimate accurately, often in areas such as pharmaceutical and food science, where the concentration of an ingredient or of a reactant had to be estimated by spectroscopic or chromatographic means accurately and quickly. Pattern recognition that has a high profile in modern chemometrics and metabolomics was not a very large part of the original literature. Many dedicated packages were developed over this period, most of which are still in existence now, these will be described in the section on software (1.4) below.

Hence, most of the tools now recognised as chemometrics were available in the 1980s. However, the early promise did not materialise at the time, unlike many other data-rich areas such as bioinformatics and QSAR (quantitative structure–activity relationship) which had a similar vintage. As from the mid-1990s, there was a tremendous interest in the application of chemometrics to fundamental analytical chemistry, for example the resolution of overlapping peaks in HPLC, which generated a large number of technically sophisticated but in most cases not particularly widespread papers and a very introspective number of groups and conferences. There was less scope for huge innovation compared to the 1970s when scientific computing was rapidly evolving and multivariate statistical methods were quite new to numerate chemists. In 2008, Paul Geladi and Phil Hopke ask 'Is there a future for chemometrics?' [42]. Many viewed the subject as rather a technical niche area without much general applicability. The number of specialist chemometrics groups in the world had hardly changed over the decades, just with a few new faces replacing those that had retired or left for other fields. It was not a particularly attractive area for researchers, without much funds, and appeared to flatten off.

In contrast, during the last 15 years or so, there has been a substantial renaissance. This is because chemometrics has moved out of its original comfort zone of quantitative analytical and physical chemistry and been embraced in areas such as metabolomics, among others. Big datasets are now readily available using instruments, such as LCMS (liquid chromatography mass spectrometry), NMR (nuclear magnetic resonance), GCMS (gas chromatography mass spectrometry) and so on. The capability of instruments to analyse many samples, each in turn containing large quantities of information, means a new and urgent need to correctly interpret these immense datasets and also to understand how to correctly design the experiments and perform the sampling. Thus, over more than a decade, there has been a renewed urgent need for chemometrics expertise. The sort of problems tackled nowadays often differs from the traditional analytical chemistry problems. The latter often involved measuring more accurately. For example, can a spectroscopic technique estimate the concentration of an analyte in a mixture without chromatography, and if so how well? We might create some reference standards or perform some independent and slower method of analysis, such as HPLC, and use multivariate analysis of a series of mixtures to create a calibration model. However, in many metabolomics problems, we are not certain of the answer in advance. For example, we may obtain the LCMS of donors' serum with and without a disease. How certain there is enough information in the serum to distinguish each group? It may depend on whether the disease was correctly diagnosed and how far it has progressed. There will be confounding factors such as age or diet or genetics – how representative were the samples? And then if we think we can separate groups, which metabolites are most likely to be markers for the disease? How confident are we? In such a situation we do not know the answer in advance and are generating and testing hypotheses. In fact, most of science outside the core physical sciences is primarily based on hypothesis testing. Much of early chemometrics was developed by programmers good at algorithms and matrices, whereas the needs of biological scientists are more hypothesis testing. This text aligns chemometrics methods primarily from the point of view of hypothesis formulation, which communicates more closely with the language of clinicians and biologists.

Chemometrics has had a renaissance because its methods are being applied to many scientific problems outside core quantitative chemistry. Nevertheless, most of the original methods, first pioneered over 50 years ago, are still relevant, and techniques such as molecular spectroscopy or chromatography, once the domain of chemists, are now widely used in many scientific fields. Understanding the fundamental statistical basis of chemometrics is an essential aid to safely and usefully employing these techniques, which have become widely available due to a plethora of software and datasets.

1.2 METABOLOMICS

The central dogma of biology is illustrated in Figure 1.1. In its simplest form DNA makes RNA, which makes proteins which make metabolites. The metabolic profile influences phenotype and so the characteristics of all organisms.

The study of systems biology is to connect these steps.

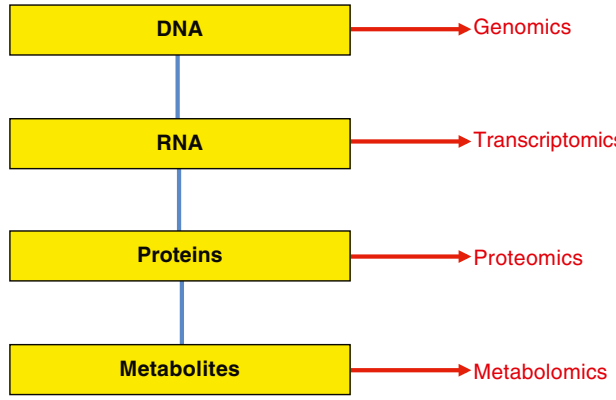


FIGURE 1.1 Central dogma of biology.

The concept of a genotype is the oldest and was first defined early in the 20th century by a number of papers by Johannsen [43]. The concept of a genome was first described by Winkler in 1920 [44]. There is in fact no universally agreed definition of these two terms, despite widespread usage; however, a common distinction is as follows. A genome is an organism's complete DNA, which includes all of its genes (coding/non-coding) and intergenic regions. The genotype refers to the genetic information for a particular trait. In humans, the genome, for example, includes the portion of human DNA that codes for hair colour and decides the genotype for that trait.

Genomics therefore studies the entirety of an organism's DNA. With improvements in high-throughput sequences, the first whole genomes were sequenced in 1980s and 1990s. The first complete genome sequence of a eukaryotic organelle, the human mitochondrion, was reported in 1981 [45] and the first chloroplast genomes followed in 1986 [46]. In 1992, the first eukaryotic chromosome, chromosome III of brewer's yeast *Saccharomyces cerevisiae,* was sequenced [47]. The first free-living organism to be sequenced was that of *Haemophilus influenzae* in 1995 [48]. It was however the human genome project that catalysed this scientific discipline with the complete sequencing announced in 2003 [49], although a small number of sequences still remained.

Whereas the concept of a genome was well developed, the name genomics and the concept of this as a discipline is rumoured to have been proposed in 1986, with a birthplace in the journal *Genomics* in 1987 [50]. This was the first recognised omics discipline and great-grandparent of metabolomics. With genomics arrived a large amount of data and this catalysed the arrival of computationally intense disciplines such as bioinformatics to mine these large datasets, placing computing at the centre of modern biological research.

Next up was transcriptomics, this time concerned with RNA. The earliest known use of the noun transcriptome is in the 1990s. The earliest known use of this term in the scientific literature is from 1997 [51], which was the first key work to report the transcriptome of an organism describing 60,633 transcripts expressed in *S. cerevisiae* using serial analysis of gene expression. With the rise of high-throughput technologies and bioinformatics and the subsequent increased computational