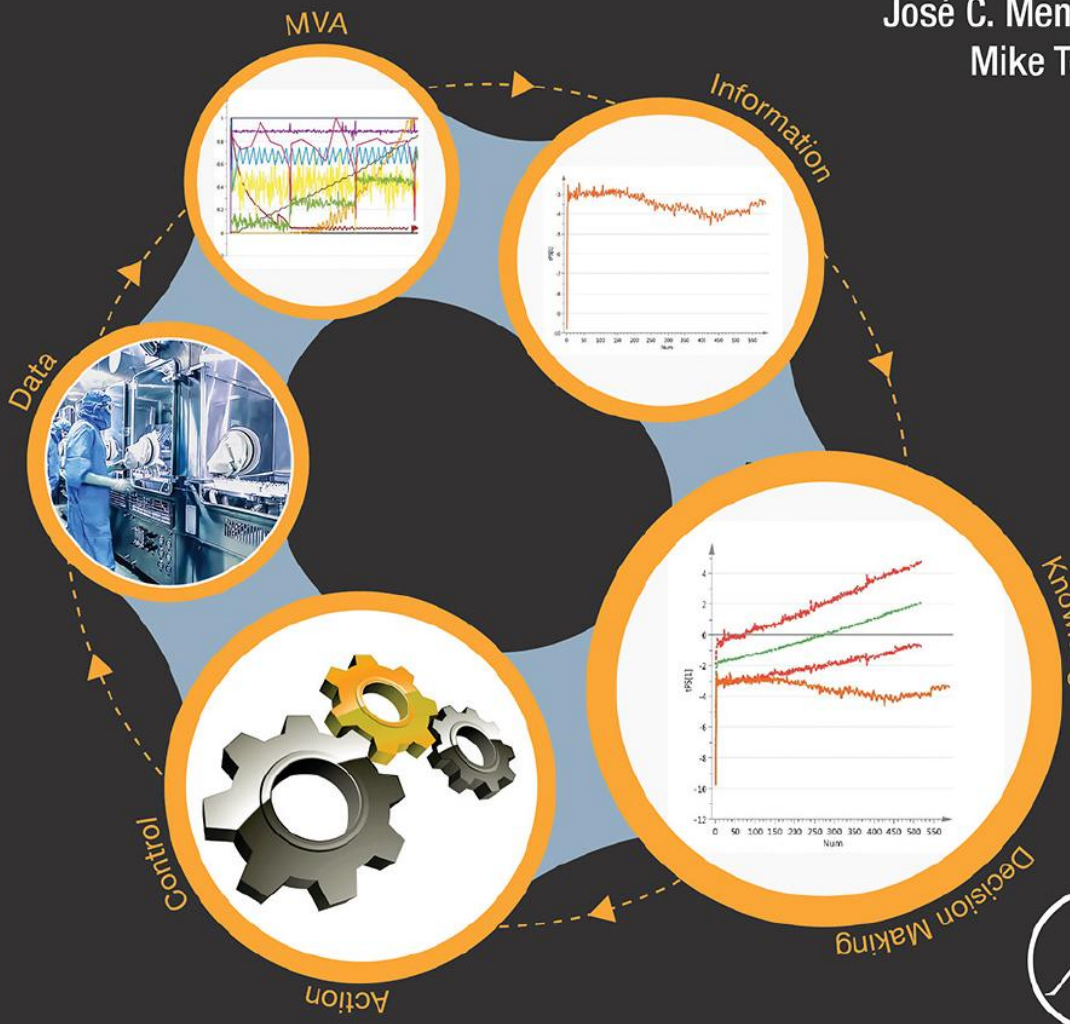
MULTIVARIATE ANALYSIS IN THE PHARMACEUTICAL INDUSTRY

Edited by
Ana Patricia Ferreira
José C. Menezes
Mike Tobyn



MULTIVARIATE ANALYSIS IN THE PHARMACEUTICAL INDUSTRY

This page intentionally left blank

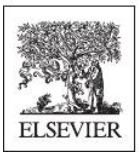
MULTIVARIATE ANALYSIS IN THE PHARMACEUTICAL INDUSTRY

Edited by

ANA PATRICIA FERREIRA
Bristol-Myers Squibb, Moreton, United Kingdom

JOSÉ C. MENEZES
University of Lisbon, Lisbon, Portugal

MIKE TOBYN
Bristol-Myers Squibb, Moreton, United Kingdom



ACADEMIC PRESS
An imprint of Elsevier

Academic Press is an imprint of Elsevier
125 London Wall, London EC2Y 5AS, United Kingdom
525 B Street, Suite 1800, San Diego, CA 92101-4495, United States
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

ISBN: 978-0-12-811065-2

Publisher: John Fedor
Acquisition Editor: Erin Hill-Parks
Editorial Project Manager: Kristi Anderson
Production Project Manager: Poulouse Joseph
Cover Designer: Christian J. Bilbow

Typeset by MPS Limited, Chennai, India

MT: For Sally, Joe, Ben, Ted and Meg. For being everything.

APF: To my family. Thank you for your support.

JCM: To Helena.

This page intentionally left blank

Contents

List of Contributors

Christian Airiau GlaxoSmithKline, King of Prussia, PA, United States
Chris Antoniou Biogen, Cambridge, MA, United States
Massimiliano Barolo University of Padova, Padova, Italy
Fabrizio Bezzo University of Padova, Padova, Italy
Robert W. Bondi Jr. GlaxoSmithKline, King of Prussia, PA, United States
Johan Bøtker University of Copenhagen, Copenhagen, Denmark
Richard D. Braatz Massachusetts Institute of Technology, Cambridge, MA, United States
Dongsheng Bu Bristol-Myers Squibb, New Brunswick, NJ, United States
Chunsheng Cai Office of Pharmaceutical Quality, Center of Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, Maryl and, United States
Graham Cook Global Quality Intelligence and Compendial Affairs, Pfizer, United Kingdom
Claudia C. Corredor Bristol-Myers Squibb, New Brunswick, NJ, United States
Kim H. Esbensen KHE Consulting, Copenhagen, Denmark; Geological Survey of Denmark and Greenland (GEUS), Copenhagen, Denmark; University of Aalborg, Aalborg, Denmark
Pierantonio Facco University of Padova, Padova, Italy
Pedro M. Felizardo 4Tune Engineering Ltd., Lisbon, Portugal
Ana Patricia Ferreira Bristol-Myers Squibb, Moreton, United Kingdom
Geir Rune Fla ˚ten CAMO Softwar e AS, Oslo, Norway
John Gamble Bristol-Myers Squibb, Moreton, United Kingdom
Joerg Gampfer Hovione FarmaCiencia SA, Lisbon, Portugal
Paul Geladi Swedish University of Agricultural Sciences, Umea ˚, Sw eden
Francisca F. Gouveia 4Tune Engineering Ltd., Lisbon, Portugal; University of Copenhagen, Frederiksberg, Denmark
Hans Grahn Corpus Data & Image Analysis AB, Stockholm, Sweden
Mark J. Henson Shire, Exton, United States
Benoıt Igne GlaxoSmithKline, King of Prussia, PA, United States
Dominique S. Kummli Novartis Pharma AG, Basel, Switzerland
Lorenz Liesum Novartis Pharma AG, Basel, Switzerland
João A. Lopes Universidade de Lisboa, Lisboa, Portugal
David Lovett Perceptive Engineering Ltd, Sci-Tech Daresbury, Cheshire, United Kingdom
John Mack Perceptive Engineering Ltd, Sci-Tech Daresbury, Cheshire, United Kingdom
Nuno Matos Hovione FarmaCiencia, Lisbon, Portugal
Neil M ᶜDowall Novartis Pharma AG, Basel, Switzerland
Gary McGeorge Bristol-Myers Squibb, New Brunswick, NJ, United States
Natascia Meneghetti University of Padova, Padova, Italy

José C. Menezes University of Lisbon, Lisbon, Portugal
Ewan Mercer Perceptive Engineering Ltd, Sci-Tech Daresbury, Cheshire, United Kingdom
Charles E. Miller Merck & Co., Inc., West Point, PA, United States
Chris Morris Science and Technology Facilities Council, Warrington, United Kingdom
Venkatesh Natarajan Biogen, Cambridge, MA, United States
Sarah Nicholson Bristol-Myers Squibb, Moreton , United Kingdom
Julia O'Neill Tunnell Consulting, Inc., Glenside, PA, United States
Antonio Peinado Novartis Pharma AG, Basel, Switzerland
Alan R. Potts Patheon, Greenville, United States
Jukka Rantanen University of Copenhagen, Copenhagen, Denmark
Clare Frances Rawlinson-Malone Bristol-Myers Squibb, Moreton, United Kingdom
Rodolfo J. Romañach Recinto Universitario de Mayagu ¨ez, Mayagu¨ez, Puerto Rico
Andre´s D. Roma´n-Ospino Rutgers University, New Brunswick, NJ, United States
Mafalda C. Sarraguça Universidade do Porto, Porto, Portugal
Kristen A. Severson Massachusetts Institute of Technology, Cambridge, MA, United States
Zhenqi Shi Eli Lilly and Company, Indianapolis, United States
Christopher M. Sinko Bristol-Myers Squibb, Lawrenceville, NJ, United States
Brad Swarbrick Quality by Design Consultancy, Emu Heights, NSW, Australia
Furqan Tahir Perceptive Engineering Ltd, Sci-Tech Daresbury, Ch eshire, United Kingdom
Jörg Tho¨mmes Biogen, Cambridge, MA, United States
Mike Tobyn Bristol-Myers Squibb, Moreton, United Kingdom
Jeremy G. VanAntwerp Massachusetts Institute of Technology , Cambridge, MA, United States; Calvin College, Grand Rapids, MI, United States

About the Editors

Ana Patricia Ferreira Principal Scientist, Bristol-Myers Squibb. She has over 10 years of experience in the application of multivariate analysis in the pharmaceutical industry both in R&D and manufacturing, spanning both small- and large-molecule applications. She has published papers on the use of multivariate analysis for extraction of information from large data sets spanning diverse topics such as near-infrared spectroscopy, process analysis, and material characterization.

José C. Menezes Professor at University of Lisbon. He has over 20 years of experience in academia and pharma/biopharma industries in which he has conducted multiple projects. He is a pioneer of the application of PAT and QbD principles to the bioengineering field. He is the coeditor of three books and has published more than 75 papers and several book chapters on MVA, PAT, QbD, data, and knowledge management.

Mike Tobyn Rese arch Fellow, Bristol-Myers Squibb. After training as a Pharmacist and obtaining his PhD, he joined the faculty in the University of Bath, where he studied and worked under Prof. John Staniforth. He has worked for, or consulted for, large pharmaceutical companies, small companies and University spinouts, as well as excipient suppliers. His fascination with materials has led him to believe that the properties of materials in processes are governed more by their faults than their intrinsic perfect properties, but that these are more difficult to detect than conventional analysis will allow. He has over 20 years of ex perience in academia and the pharmaceutical industry, and has published more than 75 papers in the fields of oral drug delivery, inhalation drug delivery, and MVA.

This page intentionally left blank

Foreword

"Big data," the popular term to describe data sets that are both extensive and complex, is the result of data generated by numerous information-sensing devices. A similar situation arose back in the 1940s and 1950s with the advent of the computer. The ability of a machine to process addition and subtraction at a rate of 5000 calculations per second far outpaced man's ability to compute manually. With ever-inc reasing processing power, the ability to measure practically anything and the advent of cheap data storage systems, we are now facing the point where one may believe we're drowning in data. This is where mathematical techniques such as multivariate analysis (MVA) come to play and, as you'll discover, will help extract what we really need—knowledge. The timing is nearly perfect. With an international effort to modernize and standardize our quality systems through the International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human Use (ICH), knowledge is a core component of what is being requested from pharmaceutical manufacturers by regulatory authorities. More importantly, manufacturers can use techniques like MVA, to begin to understand their manufacturing processes more completely. And as you'll see, the knowledge gained can be used to provide better and more precise control. The story doesn't end here though. Before us are advanced process control systems that will allow manufacturing to become more adap table to raw material and environmental variability. Beyond this are artificial intelligence-based systems that learn and adapt. One category of artificial intelligence is pattern recognition systems that use data about a problem to generate conclusions. MVA is certainly knocking at the door. But let's first start at the basics and manage the mountain of data that we're currently generating. Knowledge awaits!

Christopher M. Sinko
Head of Product Development, Bristol-Myers Squibb

This page intentionally left blank

SECTION I

BACKGROUND AND METHODOLOGY

This page intentionally left blank

CHAPTER

1

The Preeminence of Multivariate Data Analysis as a Statistical Data Analysis Technique in Pharmaceutical R&D and Manufacturing

Mike Tobyn[1], Ana Patricia Ferreira[1], Chris Morris[2] and JoséC. Menezes[3]

[1] Bristol-Myers Squibb, Moreton, United Kingdom [2]Science and Technology Facilities Council, Warrington, United Kingdom [3]University of Lisbon, Lisbon, Portugal

1.1 DATA SIZE GLOSSARY (TABLE 1.1)

TABLE 1.1 Common Data Descriptors

Data Descriptor	Short Descriptor	Metric Bytes Equivalent
Kilobyte	kB	1000
Megabyte	MB	1000^2
Gigabyte	GB	1000^3
Terabyte	TB	1000^4
Petabyte	PB	1000^5
Exabyte	EB	1000^6

Multivariate Analysis in the Pharmaceutical Industry

DOI: https://doi-org.passerelle.univ-rennes1.fr/10.1016/B978-0-12-811065-2.00001-1

1.2 BIG DATA—OVERALL VIEW

It is difficult to work out how much data is being generated, and stored, in the world, as a lot of it is held locally or privately and/or securely. Factors such as online sto rage, data traffic, and more specific data (e.g., uploads to YouTube) can all be used as estimates or surrogates.

It is nevertheless estimated that, from 2005 to 2020, the digital universe will grow significantly, per haps to 40 EB, which represents a doubling every 2 years. This represents an average of 5 GB per person on the planet (although this is not evenly spread), and many people carry a device which can carry 50 GB of information in their pocket.

Of the purposes that are proposed for analyzing this data, some of the most prominent are security, science/medical, political, and business related. Each of these has different requirements, different levels of input to rationalize and analyze the outcomes from the techniques, and different consequences for a failure to analyze or utilize correctly.

An organized, even very large, data set can be examined if enough computing power is made available, and there is a will to do so. For instance, the large Hadron Collider at CERN has the capability to capture 600 million events per second, and then use algorithms to filter and then store the data. CERN currently has 200 PB of da ta on tape, however network restrictions have "restricted " the openly available data to 300 TB of information.

Due to a lack of structure, it is estimated that ,1% of the world's data is "analyzable." That may improve in future years but only to higher single-digit levels, even if bottlenecks in the networks and infrastructure can be relieved.

A wide range of techniques can be used to analyze the data that is being made available. New technologies to analyze them are now coming into pro minence in commercial and academic circles, and new organizations and individuals are developing them into usable technologies. Developments in artificial intelligence and cognitive computing have received prominence, and achieved successes. Large training sets, improved algorithms, and the availability of highly performant engines for linear algebra in the form of GPUs have led to notable successes for artificial neu ral nets (Krizhevsky, Sutskever, & Hinton, 2017). The availability of large streams of data has led to the development of algorithms capable of processing them, notably stochastic gradient descent (Bottou, 2010).

However, although many of these techniques are successful, they remain as "black box" approaches, where the rationale for their outputs is not always clear to the end user, process controller, or potential regulator.

1.3 BIG DATA—PHARMACEUTICAL CONTEXT

The data from pharmaceutical companies represents a small proportion of the overall data set, but pharmaceutical companies generate and store large amounts of data, for business and science. For one large pharmaceutical company, estimates of the data under storage, for all purposes, are 6 PB of which 2 PB are from the R&D function. This data encompasses a wide range of business, manufacturing, medical, and technical fields. Even larger amounts of data in healthcare organization s and in the public domain are relevant

to adverse event detection and the understanding of risk factors and pathophysiology leading to novel targets.

Pharmaceutical companies rely on their data to run their businesses, and to manufacture their products. When pharmaceutical giant Merck was hit by a malware attack in June 2017, it was unable to manufacture some products for a number of months, resulting in losses to the business estimated in hundreds of millions of dollars.

Pharmaceutical companies have access to the latest technology, which may mean that they become subject to ongoing ballooning of data, as the abili ty of instruments to gather data continues to increase, and new high-performance analytical equipment and methods, field instruments, and manufacturing equipment generate considerable data for the organization, to analyze and wareho use. However, although the costs of local storage to house that data may have reached an inflexio n point, cloud computing storage is becoming a viable solution (Alyass, Turcotte, & Meyre, 2015; Dalpe ´ & Joly, 2014; Fukunishi et al., 2016; Korb, Finn, & Jones, 2014; Milian, Spinola, & Carvalho, 2017; Nowotka et al., 2017; Ranjith, Balajee, & Kumar, 2016), even for proprietary data requiring highly secure warehousing.

Pharmaceutical companies are becoming better at structuring their R&D data so that it is amenable to analysis to establish findings leading to better insights (Dali et al., 2014), and manufacturing data is inherently structured and traceable. However, generating and using clinically relevant specifications fed back to the R&D and development programs inside companies—and doing that across the entire industry—will perhaps be an even bigger challenge that is about to enter our lives. Or using a multitude of data about our well-being and our environment for advisory health prescriptions.

At the moment there is not much pooling of data between companies, but some steps are being taken (Giovani, 2017), and may result in developments until now unheard of. It has been suggested that genomic analysis is a challenge as big as any of the common "Big Data" challenges that currently exist (Stephens et al., 2015).

Pharmaceutical companies also make use of publically available data sets, from genomics data, medical literature, and prescribing data, to make business and R&D decisions. Some of these applications are amenable to black box techniques to generate leads which are then vetted by experienced staff and according to protocols, but others require real-time control of processes, with large amounts of data to be analyzed efficie ntly and transparently. The techniques that are available for lead generation and trends in prescribing may not be the same ones that will operate in compliant environments, and "black box" techniques are unlikely to be compliant with regulatory requirements and the need to be demonstrably robust.

1.4 STATISTICAL DATA ANALYSIS METHODS IN THE PHARMACEUTICAL INDUSTRY

A range of statistical modeling techniques are theoretically available to the pharmaceutical industry, and they have the resources to develop them further, if they meet the needs of the regulators and the industry.

While there are applications for Monte Carlo simulat ions and random walk models in drug discovery and method development, the most prominent and directly applicable

methods are those for neural networks, closed-form Bayesian analysis, and multivariate data analysis.

Bayesian analysis has been shown to be a very powerful technique in understanding systems and building robust models, and there are innovative organizations wh ich can support the use of these systems within the pharmaceutical industry (Altan & Schofield, 2012), but they have not yet found as wide a use in controlled environments.

Similarly, there have been useful developments using neural networks as tools for specific purposes (Buket et al., 2012; Colbourn & Rowe, 2003; Landin & Rowe, 2013; Mansa, Bridson, Greenwood, Barker, & Seville, 2008; Tan & Degim, 2012), and there are products and services to support these developments.

Multivariate data analysis (MVA), specifically multivariate projection (or latent variable) method, is a set of data analysis methods long established throughout many sectors in industry. Beginning in the 1990s, the application of these techniques to pharmaceutical systems (Kourti & MacGregor, 1995, 1996) became widely reported and the techniques have evolved since. Every challengi ng statistical analysis has multiple independent (input) variables. By multivariate analysis, we mean a statistical analysis in which there are more than one independent (input) and/or dependent (output) variable of interest. This is usually the case in process control, where yield, cost, and purity are all of concern.

In the early days of development each of these methods had adherents and applications. Then, as now, they could be used in a complementary fashion to elucidate problems. Fig. 1.1 demonstrates the reported use of these techniques within a large journal database, Scopus, using the relevant term associated with the search term “pharmaceutical.” In recent years, it has become clear that multivariate analysis techniques are becoming predominant. They also form the bulk of validated methods that are used in regulat ory filings. It is worthwhile examining why these MVA techniques now predominate (Ferreira & Tobyn, 2015).

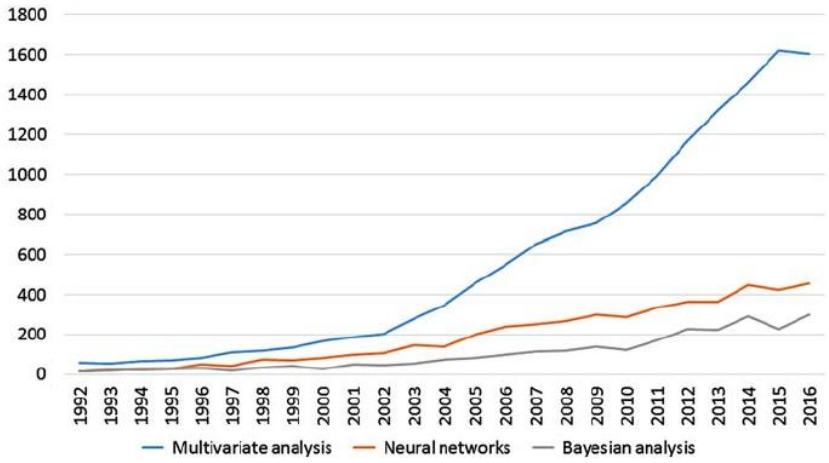


FIGURE 1.1 Number of publications with the term “Pharmaceutical” per year, by method of statistical analysis

1.5 DEVELOPMENT OF MULTIVARIATE DATA ANALYSIS AS A DATA ANALYSIS TECHNIQUE WITHIN THE PHARMACEUTICAL INDUSTRY

Early reports on the potential (Bohi dar, Restaino, & Schwartz, 1979; Chowhan & Chi, 1981) of multivariate techniques to illuminate problems in the pharm aceutical industry, to help with localized problems in development and research, focused on the potential of the technique. However, at that time, the physical instruments to gather and war ehouse the data were not sufficiently evolved to allow wide use in production and development, and there were no guiding principles to allow their use in the industry, and no data analytical tools to utilize for that work.

By the mid-1990s pioneers in the field, primarily John McGregor and Theodora Kourti of McMaster University, Svante Wold and his group at Umea ˚ University, and Julian Morris and Elaine Martin at the University of Newcastle, had begun to descri be the rationale of using MVA to address manufacturing issues, beginning to note the beneficial fit between the challenges that pharmaceutical production had and the power that MVA brought to those challenges. Frequently, the successes of other industries, for instance, steel, paper and pulp, and petrochemicals, provided the basis for believing that MVA could be a tool that could meet the needs of the pharmaceutical industry. Nevertheless, by 2002, while the potential of the techniques remained clear, their broad application was not yet fully demonstrated (Gabrielsson, Lindberg, & Lundstedt, 2002).

When a company opera tes in a regulated environment, particularly a closely regulated one such as the pharmaceutical industry, there is often the imperative of "first to be second," i.e. one does not want to be the one that faces a regulatory question for the first time, in case the outcome is not favorable and this leads to delay on a product. This is why guidance from regulatory agencies is key in establishing principles, and indicating that the door can be pushed open, if the right systems and safeguards are in place. Once the door has been opened, many others can rush through it.

A paradigm shift in the regulators' attitude toward product control and understanding, and the use of multivariate analysis came in 2004. FDA's guidance for industry document "PAT—A Framework for Innovative Pharmaceutical Development, Manufacturing a nd Quality Assurance" (FDA, 2004) described for the first time in detail the steps that the FDA would like industry to take to assure quality. This document mentioned multivariate analysis specifically as a suitable technique, if used appropriately.

Perhaps even more important in the development of multivariate analysis in the pharmaceutical industry was the 2008 ICH Q8 guideline (ICH, 2009). This will be discussed extensively in a subsequent (Chapter 19, Multivariate Anal ysis and the Pharmaceutical Regulatory Framework), but once again it became clear that MVA would be a key enabling tool for quality by design (QbD) for pharmaceuticals, and this, and the refinements, worked example s and case studies that have followed have borne out the optimism that QbD could be enabled by MVA.

Once an apparent regulatory framework for multivariate analysis was in place for companies and practitioners to use in the development and utilization of models, the number of applications began to increase significantly.

1.6 CURRENT STATUS OF THE USE OF MULTIVARIATE DATA ANALYSIS IN THE PHARMACEUTICAL SPACE

The purpose of this chapter is to capture the current state of the use of MVA in pharmaceutical R&D and production, across small molecules and large molecules. Both by the number of publications, as illustrated in Fig. 1.1, and the depth and breadth of these applications this is indicative of wide, and growing, use of these methods.

The ability of MVA techniques to build models that are transparent (at least to the regulators who will assess their utility, they are not always open for scrutiny by others) , compliant, and robust mak es them preeminent in the pharmaceutical industry and it will be necessary for development and production of vital pharmaceutical products going forward.

In addition to the suitability of the techniques, elucidating the challenges faced by the development and production organizations requires a statistical toolbox which not only meets the technical needs of the user, but also their subsequent regulatory needs. For instance, if a software tool is being relied upon for record keeping and/or decision making, it is likely that it will have to be compliant with the applicable software regulations, for instance 21 CFR Part 11, from the FDA . This code is "the part of Title 21 of the Code of Federal Regulations that establishes the United States Food and Drug Administrat ion (FDA) regulations on electronic records and electronic signatures (ERES)". Part 11, as it is commonly called, defines the criteria under which electronic records and electronic signatures are considered trustworthy, reliable, and equivalent to paper records (FDA CRF Title 21 Part 11 Section 11.1 (a)) (FDA, 2003, 2017; "Title 21 CFR Part 11," 2017).

Compliance with this regulation is not trivial, and any software tool that did not have the requisite comp liance would have to invest significantly in achieving it. This would require significant inves tments in time and money for any data analysis or model software provider to incorporate into their systems. Pharmaceutical companies may be reluctant to use the software until the compliance was achieved, particularly if the application relates to activities in the good manufacturing practice area, but that would mean the payoff from any investment by the software company would take some time to achieve. However, there are a number of multivariate and chemometric software packages that have already achieved 21 CFR Title 11 Compliance, which means that the regulatory barriers are lower.

Currently, there are developments involving different aspects of the "data generation and usage lifecycle" of which information extraction and knowledge management are part of.

One is related to data integrity, that has been enforced since 2015 by regulatory authorities in an effort that will become mandatory and that aims to ensure that the data on which companies based all their critical decisions both prelaunch and during the commercial life of their products, has the required quality and authenticity (ALCOA: attributable , legible, contemporaneously recorded, original or a true copy, and accurate).

The second major development is at the other extreme. It deals with establishing knowledge excellence at companies that give integrity and the associated quality-culture, precedence over behaviors that will over the lifecycle compromise a company being class A in quality and overall performance including all its stakeholders (people , supply chain, etc.). To balance these two extremes, we need to better understand why the pharma industry is

a data-rich environment, how this data is generated, how and why the information therein can be extracted, and finally how that information can lead to platform knowledge across multiple products and the "wisdom" levels needed to design and develop disruptive new therapeutic concepts and modalities (Calnan, Lipa, Paige, & Menezes, 2017).

When making a decision on which methods to use to support development or control of product man ufacture, this would be an important factor. There are, of course, risks associated with model development and data analysis, and no program can ever be sure of reaching the goal, but an avoidable risk of regulatory noncompliance when technical success has been achieved is one which companies wish to avoid.

The number of technical and regulatory successes associated with MVA means that risks can be assessed as being lower, and the chances of overall success are higher.

1.7 WHAT MVA CAN BE USED FOR/WHAT IT CANNOT BE USED FOR

While MVA methods have demonstrated benefits in addressing many problem s faced by the industry, it is important to assess if they are the best solution for each specific problem statement. In addition, before initiating any project, it is important to consider whether all requirements for successful application of the methods are met. Spending time upfront considering these two points and addressing any gaps identified will increase the likelihood of success and ensure faster results. The decision tree provided in Fig. 1.2 provides guidance on what should be considered when planning to use MVA to address a specific problem statement.

1.8 CURRENT LIMITATIONS AND FUTURE DEVELOPMENTS

While other techniques currently lack the regulatory "endorsement" and successful ecosystem of hardware, software, and applications they may, 1 day, reach a similar level of utility and even ubiquity, if they can address challenges that are not met with MVA. Current industrial practice is to mainly rely on linear methods such as principal component analysis and partial squares regression and its regularized relatives. For many problems, these methods are successful. In particular, a stable process is operated in a region where responses are indeed linear.

A wide range of other modeling techniques can be applied to multivariate problems, for example, support vector machines (Gawehn, Hiss, & Schneider, 2016; Lima et al., 2016; Radzol, Khuan, Mansor, & Twon Tawi, 2016; Wang, Wu, Lin, & Yang, 2015). A multioutcome problem can also be collapsed into a univariate one, by choosing a suitable loss function, e.g., the sum of square errors. This is the usual approach when applying artificial neural nets to such problems.

It is not always appropriate to collapse the output variables into a single loss function. Sometimes "Pareto optimization" is required.

Bayesian methods are eminently suitable for this. Unfortunately, they are not available on a "click and go" basis, but require some mathematical sophistication to use them. More

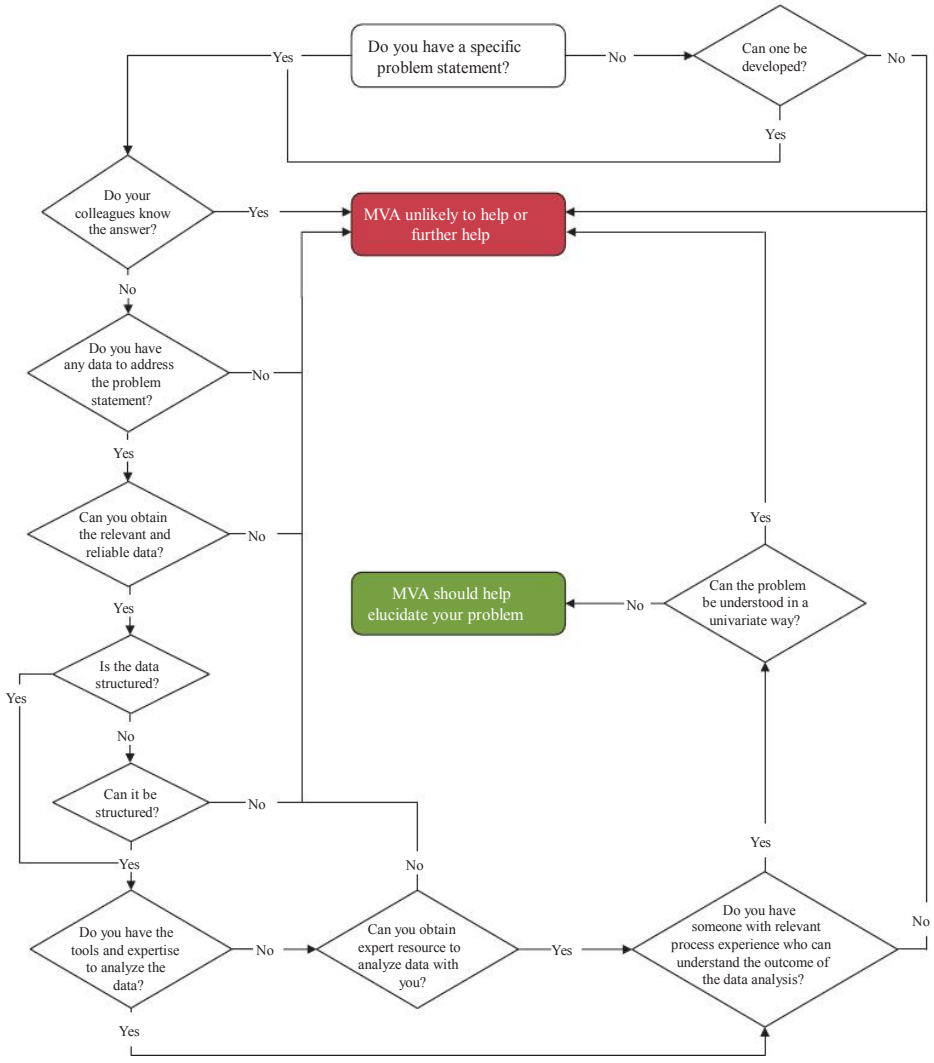


FIGURE 1.2 A workflow for the use of MVA in the Pharmaceutical Industry

seriously, some sophistication is needed to validate the results. It is possible to produce a powerful graphic visualization of a posterior distribution, illustrating the conclusion that it meets the multiple performance objectives. However, it is not necessarily easy for regulators to convince themselves that this indeed the appropriate posterior distribution, and so it could be difficult to validate.

The full range of methods can be used at the research stage, while accepting that a narrower set will be acceptable in production. The financial industry faces a similar challenge: they investigate using the most sophisticated methods available, but for actual use approximate the research model using a more transparent method, e.g., a decision tree, is required for transparency and reproducibility.

We suggest that there are a range of opportunities where nonlinear methods have not yet been tried and could yield valuable results, and that this may become standard practice in years to come.

However, for the current state of the art, MVA is the method which leads the industry.

Acknowledgments

The helpful support of Peter Webster, Jason Bronfeld, Richard Hammerstone, and Nelson Novoa is gratefully acknowledged.

References

Altan, S., & Schofield, T. L. (2012). Introduction to special issue on nonclinical biopharmaceutical statistics. Statistics in Biopharmaceutical Research, 4(2), 100□101. Available from https://doi-org.passerelle.univ-rennes1.fr/10.1080/19466315.2012.707561.

Alyass, A., Turcotte, M., & Meyre, D. (2015). From big data analysis to personalized medicine for all: Challenges and opportunities. BMC Medical Genomics, 8(1). Available from https://doi-org.passerelle.univ-rennes1.fr/10.1186/s12920-015-0108-y.

Bohidar, N. R., Restaino, F. A., & Schwartz, J. B. (1979). Selecting key pharmaceutical formulation factors by regression analysis. Drug Development and Industrial Pharmacy, 5(2), 175□216. Available from https://doi-org.passerelle.univ-rennes1.fr/10.3109/03639047909055671.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. Paper presented at the Proceedings of COMPSTAT 2010—19th international conference on computational statistics, Keynote, Invited and Contributed Papers.

Buket, A., de Matas, M., Cevher, E., Ozsoy, Y., Guneri, T., & York, P. (2012). Quality by design approach for tablet formulations containing spray coated ramipril by using artificial intelligence techniques. International Journal of Drug Delivery, 4(1), 59□69.

Calnan, N., Lipa, M. J., Paige, E. K., & Menezes, J. C. (2017). A lifecycle approach to knowledge excellence in the biopharmaceutical industry. Boca Raton, FL: Taylor & Francis.

Chowhan, Z. T., & Chi, L. H. (1981). Mixing of pharmaceutical solids III: Multivariate statistical analysis of multicomponent mixing. Journal of Pharmaceutical Sciences, 70(3), 247□251. Available from https://doi-org.passerelle.univ-rennes1.fr/10.1002/jps.2600700305.

Colbourn, E. A., & Rowe, R. C. (2003). Neural computing boosts formulation productivity. Pharmaceutical Technology, 27(Suppl. 11), 22□25.

Dali, M., Stewart, A., Behling, R. W., Raglione, T., Stamato, H. J., & Tom, J. W. (2014). Optimizing knowledge creation at Bristol-Myers Squibb—A case study within pharmaceutical development. Journal of Pharmaceutical Innovation, 10(1), 1□12. Available from https://doi-org.passerelle.univ-rennes1.fr/10.1007/s12247-014-9209-y.

Dalpé, G., & Joly, Y. (2014). Opportunities and challenges provided by cloud repositories for bioinformatics-enabled drug discovery. Drug Development Research, 75(6), 393□401. Available from https://doi-org.passerelle.univ-rennes1.fr/10.1002/ddr.21211.

FDA. (2003). Guidance for industry Part 11, electronic records; electronic signatures—Scope and application.

FDA. (2004). Guidance for industry PAT—A framework for innovative pharmaceutical development, manufacturing, and quality assurance. , http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm070305.pdf . .

FDA. (2017). CFR—Code of federal regulations title 21. Part 11, Electronic Records; Electronic Signatures — Scope and Application. , https://www.fda.gov/RegulatoryInformation/Guidances/ucm125067.htm . .

Ferreira, A. P., & Tobyn, M. (2015). Multivariate analysis in the pharmaceutical industry: Enabling process understanding and improvement in the PAT and QbD era. Pharmaceutical Development and Technology, 20(5), 513–527. Available from https://doi-org.passerelle.univ-rennes1.fr/10.3109/10837450.2014.898656.

Fukunishi, Y., Mashimo, T., Misoo, K., Wakabayashi, Y., Miyaki, T., Ohta, S., … Ikeda, K. (2016). Miscellaneous topics in computer-aided drug design: Synthetic accessibility and GPU computing, and other topics. Current Pharmaceutical Design, 22(23), 3555–3568.

Gabrielsson, J., Lindberg, N. O., & Lundstedt, T. (2002). Multivariate methods in pharmaceutical applications. Journal of Chemometrics, 16(3), 141–160. Available from https://doi-org.passerelle.univ-rennes1.fr/10.1002/cem.697.

Gawehn, E., Hiss, J. A., & Schneider, G. (2016). Deep learning in drug discovery. Molecular Informatics, 35(1), 3–14. Available from https://doi-org.passerelle.univ-rennes1.fr/10.1002/minf.201501008.

Giovani, B. (2017). Open data for research and strategic monitoring in the pharmaceutical and biotech industry. Data Science Journal, 16. Available from https://doi-org.passerelle.univ-rennes1.fr/10.5334/dsj-2017-018.

ICH. (2009). Q8(R2)—Pharmaceutical development. , http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Quality/Q8_R1/Step4/Q8_R2_Guideline.pdf. .

Korb, O., Finn, P. W., & Jones, G. (2014). The cloud and other new computational methods to improve molecular modelling. Expert Opinion on Drug Discovery, 9(10), 1121–1131. Available from https://doi-org.passerelle.univ-rennes1.fr/10.1517/17460441.2014.941800

Kourti, T., & MacGregor, J. F. (1995). Process analysis, monitoring and diagnosis, using multivariate projection methods. Chemometrics and Intelligent Laboratory Systems, 28(1), 3–21. Available from https://doi-org.passerelle.univ-rennes1.fr/10.1016/0169-7439(95)80036-9.

Kourti, T., & MacGregor, J. F. (1996). Multivariate SPC methods for process and product monitoring. Journal of Quality Technology, 28(4), 409–428.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural net works. Communications of the ACM, 60(6), 84–90. Available from https://doi.org/10.1145/3065386.

Landin, M., & Rowe, R.C. (2013). Artificial neural networks technology to model, understand, and optimize drug formulations. In Formulation tools for pharmaceutical development (pp. 7–37). Available from https://doi-org.passerelle.univ-rennes1.fr/10.1533/9781908818508.7.

Lima, A. N., Philot, E. A., Trossini, G. H. G., Scott, L. P. B., Maltarollo, V. G., & Honorio, K. M. (2016). Use of machine learning approaches for novel drug discovery. Expert Opinion on Drug Discovery, 11(3), 225–239. Available from https://doi-org.passerelle.univ-rennes1.fr/10.1517/17460441.2016.1146250.

Mansa, R. F., Bridson, R. H., Greenwood, R. W., Barker, H., & Seville, J. P. K. (2008). Using intelligent software to predict the effects of formulation and processing parameters on roller compaction. Powder Technology, 181(2), 217–225. Available from https://doi-org.passerelle.univ-rennes1.fr/10.1016/j.powtec.2007.02.011.

Milian, E. Z., Spinola, M. M., & Carvalho, M. M. (2017). Risks and uncertainties in cloud computing: Literature review, trends and gaps. IEEE Latin America Transactions, 15(2), 349–357. Available from https://doi-org.passerelle.univ-rennes1.fr/10.1109/TLA.2017.7854632.

Nowotka, M. M., Gaulton, A., Mendez, D., Bento, A. P., Hersey, A., & Leach, A. (2017). Using ChEMBL web services for building applications and data processing workflows relevant to drug discovery. Expert Opinion on Drug Discovery, 12(8), 757–767. Available from https://doi-org.passerelle.univ-rennes1.fr/10.1080/17460441.2017.1339032.

Radzol, A. R. M., Khuan, L. Y., Mansor, W., & Twon Tawi, F. M. (2016). Signal processing for Raman spectra for disease detection. International Journal of Pharmacy and Pharmaceutical Sciences, 8(6), 4–10.

Ranjith, D., Balajee, J., & Kumar, C. (2016). In premises of cloud computing and models. International Journal of Pharmacy and Technology, 8(3), 4685–4695.

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., … Robinson, G. E. (2015). Big data: Astronomical or genomical? PLoS Biology, 13(7). Available from https://doi-org.passerelle.univ-rennes1.fr/10.1371/journal.pbio.1002195

Tan, C., & Degim, I. T. (2012). Development of sustained release formulation of an antithrombotic drug and application of fuzzy logic. Pharmaceutical Development and Technology, 17(2), 242–250. Available from https://doi.org/10.3109/10837450.2010.531739.

Title 21 CFR Part 11. (2017). ,https://en.wikipedia.org/wiki/Title_21_CFR_Part_11. .

Wang, T., Wu, M. B., Lin, J. P., & Yang, L. R. (2015). Quantitative structure-activity relationship: Promising advances in drug discovery platforms. Expert Opinion on Drug Discovery, 10(12), 1283–1300. Available from https://doi-org.passerelle.univ-rennes1.fr/10.1517/17460441.2015.1083006.

CHAPTER

2

The Philosophy and Fundamentals of Handling, Modeling, and Interpreting Large Data Sets—the Multivariate Chemometrics Approach

Paul Geladi[1] and Hans Grahn[2]

[1] Swedish University of Agricultural Sciences, Umea °, Sweden [2]Corpus Data & Image Analysis AB, Stockholm, Sweden

2.1 INTRODUCTION

2.1.1 The Nature of this Chapter

This chapter is a very general introduction to data sets and large data sets. Therefore a number of strict definitions of different data structures are needed. It is also impossible to talk about data without giving a slight hint on the models that these data are subjected to and on the ways of checking the models for correctness and qua lity. After that some practical philosophical thoughts and hints can be given.

The title of this chapter mentions the term “philosophy” and this is quite a big word. The authors do not claim to be real professional philosophers, nor do they claim to compare chemometrics to the works of the great philosophers. What is me ant is that the chapter is about thinking in a wider perspective than just technical or applied ways and trying to define the fundamental questions.

The structure of the chapter is as follows: (1) introduction, (2) univariate data and classical statistics, (3) multivariate data in all kinds of 2D blocks, 3D and higher-D blocks and multiblock data, (4) modeling of the data by looking inside blocks and between blocks, and (5) conclusions. Many aspects of data and modeling, such as transformation or

Multivariate Analysis in the Pharmaceutical Industry

DOI: https://doi-org.passerelle.univ-rennes1.fr/10.1016/B978-0-12-811065-2.00002-5

preprocessing, are too technical to fit into this chapter. The choice between philosophical and technical aspects is quite difficult because there is a huge overlap, but one may assume that really technical aspects are dealt with in other chapters of this book.

Data may be produced by different analytical (clinical, physical, psychological, etc.) techniques but the purpose of this chapter is to be general. If any examples are given, they only serve to illustrate a general principle and they are not there for their own sake. The topic of this chapter is huge. For this reason, some methods are shown in detail, while others are just mentioned and the reade r is referred to some good books or review articles on the topic.

2.1.2 The History of Metrics

An important part of the philosophical aspect of chemometrics is its history and relations to other metrics disciplines. The history of a metrics discipline can be traced by the appearance of dedicated sc ientific journals. Biometrika (Oxford University Press) (Cox, 2001) started in 1901 and was in the beginning mostly about univariate statistics and discrete distributions applied to large biological populations. Biometrics (Wiley-Blackwell) was started in 1947. Psychometrika (Psychometrics Society) started in 1936 and handled taking care of multivariate and correlated data from patient populations. An interesting article in the first volume was that of Harold Hotelling on principal component calculation (Hotelling, 1936). In 1959, Technometrics (American Statistical Association) was founded. It was concerned with statistics applied to industrial applications and engineering sciences. A high emphasis in Technometrics was on experimental design. A general comment is that these journals were started because just using t- and F-tests was no longer sufficient for the data available.

The history of chemometrics began in the late 1960s to early 1970s. The reason was that analytical instruments suddenly started producing multivariate and correlated data. The time was ready for dedicated publications on chemometrics in the mid-1980s. The two journals were Journal of Chemometrics (Wiley 1987) and Chemometrics and Intelligent Laboratory Systems (Elsevier 1986), but also before that Analytica Chimica Acta (Elsevier) had started a chemometrics section. Although chemometrics has many data set structures and modeling techniques in common with the other metrics disciplines, the interaction has not been very extensive. Some literature references can be found with more details (Brereton, 2012; Esbensen & Geladi, 1990; Geladi & Esbensen, 1990; Kvalheim, 2012).

2.2 UNIVARIATE DATA AND HOW IT IS HANDLED

2.2.1 Data Vectors and Some Definitions

Given the title of this chapter , which contains the term multivariate, the part on univariate data is kept brief, but some concepts and definitions must be given. The important concepts are presented as a table with comments for each one because there is no room for longer explanations. Table 2.1 gives some data types to be considered. The table is important because sometimes errors occur by not really understanding the data type and just presenting it to the used software in a sloppy way. Complex numbers occur rarely in

TABLE 2.1 Data Types that are Sometimes Encountered in Measurement

Data Type	Comment
Numerical	Measured data are represented by numbers, e.g., 10.5, □2.3, or 0.55783, 113, the most usual data type
Categorical	Sometimes measured data are categorical A or B, or A, B, C, D, and so on. Representation by 0,1 or 0,1,2,3 is also possible. Questionnaire, psychological, and sensory data are categorical most of the time
Integer	An integer is a number without decimals, sometimes used to express order (see later)
Floating point	A number with decimals
Interval floating point	Positive and negative values are possible
Ratio floating point	Only positive values and zero are possible (mass, concentration, intensity, etc.)
Scalar	Any single number can be called a scalar
Vector	A vector is a collection (population) of scalars
Missing data	For univariate data, the missing data concept is not important For multivariate data, missing data becomes a major issue
Ordered data	Sometimes numbers are ordered according to size and then the number is replaced by an integer expressing order
Computer representation	Computer representation of data can be binary, octal, hexadecimal, 8 bit, 16 bit, 32 bit, and many other forms. Human interpretation is almost always decimal

multivariate data analysis and are therefore left out from the general discussion (Geladi, Nelson, & Lindholm-Sethson, 2007).

2.2.2 Some Statistics on Vectors

Also, for statistical concepts a table is given of names and short comments. The readers are referred to the internet or literature for more extensive treatments of this subject. Data that are measured form a collection of numbers, also called a measured population. In this measured population, one or more true populations can be hidden. A very incomplete list is given in Table 2.2.

Statistics can be ab out measured discrete populations, or about theoretical probability density function (pdf), and continuous functions.

It is important to have a basic knowledge of the statistical concepts that can apply to one's data. It is also good to always have some tables of t-test and F-test available, on paper or on a screen (most stati stical tables are found on the Internet). When it comes to large data sets, visualization usually replaces the statistics that are normally used. It is also very likely that subsets are more interesting for testing than the whole data set. To give a hint of visualization, see Fig. 2.1. This is a histogram in 50 bins of 10,000 randomly generated numbers with a normal distribution of mean 0 and standard deviation 1. The figure shows

TABLE 2.2 Some Univariate Statistical Terms and Comments

Statistical Term	Comment
Measured population	The ensemble of measured data, a bunch of numbers filling a vector
Population	A theoretical population assumed to fit the measured population. The theoretical shape could be normal or t-distribution (or one of many others)
Subpopulation	A subset of the measured data that has its own (assumed?) population model
Normal or Gaussian distribution	The most popular statistical distribution that is symmetrical and assumed to describe a natural process with random errors
Discrete distribution	A distribution that fits discrete measured values (Table 2.1), e.g., Poisson and many others
Mean	Mean can be defined for a measured or theoretical population. Means could be more interesting for comparing subpopulations
Median	The middle point of a measured population or the 50th percentile (see later). In symmetrical populations mean and median are identical
Standard deviation	A measure of the spread of a population (measured or theoretical)
Variance	The square of the standard deviation. Required for F-testing
Degrees of freedom	A very important concept for statistical testing. Confusing to all newcomers and even to seasoned users
Percentile	A point p with p% of the population below it and (100 2 p)% above it is the pth percentile
Quartile	Quartiles are 25th, 50th, and 75th percentiles
Interquartile range	A robust measure of spread. Difference between the 75th and 25th percentiles
Outlier [a]	A measurement that does not fit in with the other data. The reason may not be clear
Skewness	An expression of the fact that not all distributions are symmetrical around some central point
Histogram	For a sampled population, the histogram is a visualization of the pdf function
Robust measures	Extreme values are removed from populations to avoid too high an influence from nonidentified or hard to find outliers. This is highly recommended
t-Test	The normal distribution is only valid for a large number of degrees of freedom. The t-distribution is a wider and flatter normal distribution. Smaller data sets use the t-test
F-test	A test for comparing two variances, each with their own number of degrees of freedom
Order statistics	Statistics on ordered data
Nonparametric testing	Testing by not taking into account the numerical values but an order determined for the samples

[a]Some of the most interesting industrial products started their life as outliers. Champagne was an outlier in a failed series of experiments for making bubble-free wine, so outliers are not always bad. Also, sticky notes were an outlier in an experiment for making a very strong glue.

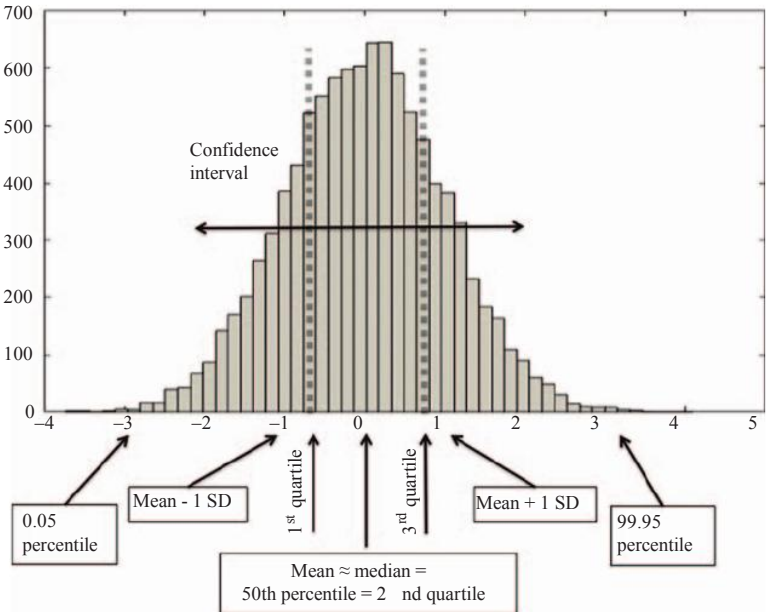


FIGURE 2.1 A histogram of 10,000 randomly generated numbers belongs to a normal distribution of mean 0 and standard deviation 1. Some interesting statistics for indicating central tendency and spread are shown. Skewness is close to zero.

some important statistics. Two very important concepts are accuracy and precision because they also come back for multivariate cases. Fig. 2.2 shows these two definitions. A remedy for low precision is to mak e averages from many replicates. There is no easy remedy for low accuracy, but measuring reference standards is one possible way to go.

2.2.3 Some General Thoughts about Univariate Thinking

With many years of multivariate statistics teaching, we have experienced that a basic introduction of univariate examples and concepts is needed, also as a refresher for those who have taken statistics courses earlier . Another nice aspect of univariate statistics is that it can be used on latent variables that will be introduced later in this chapter. There are already many things that can be done with univariate data and it is good to have these in mind when dealing with multivariate data. Maybe the most important aspect of univariate statistics, especially for large data sets, is that almost everything can be visualized. A small disadvantage of visualization in plots is that these can sometimes be misinterpreted. One should be aware of that.

The distinction between a sampled population and a theoretical one is importan t. The histogram in Fig. 2.1 is for the sampled population. Under the assumption that the data follow a normal distribution, some properties of that distribution (from tables) can be

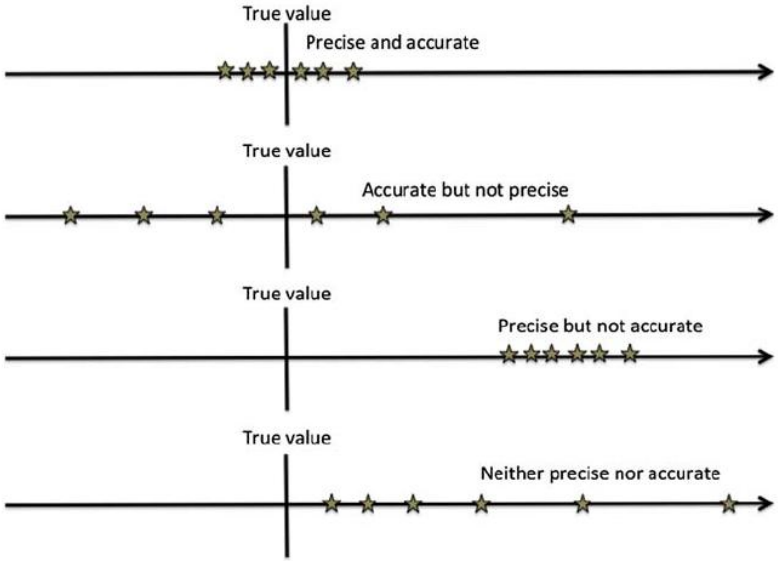


FIGURE 2.2 A simple illustration of accuracy and precision.

used. The question is always whether such an assumption is valid. The assumption of validity of some distributions is often used in a very untidy way. The authors have, in their more than 30 years of experience, never seen a "real" norm al distribution.

An important thing to consider is that almost all large data sets have meaningful subsets. The subsets may be defined in many ways: (1) sampling considerations, (2) measurement consid erations, (3) external knowledge of the data set and how it came about, (4) checking of histograms, and (5) many more criteria. These things are done a priori. Then a posteriori it is possible to check whether the splits that were introduced make sense by using statistical tests.

The statistics literature has evolved from a few very theoretical books in the previous century to more software-based and applied books, which makes it easier for statistical thinking and practice to reach larger numbers of users. An exhaustive list is not given, but the chapter authors provide some personal favorites that have not gone out of print.

Quite a few books have been written on univariate statistics and on all levels of the topic from very theoretical to very applied. Some statistics books are classical (Agresti, 2007; Devore, 2014; Forbes, Evans, Hastings, & Peacock, 2011; Wonnacott & Wonnacott, 1990) (the example s are theoretical or can be calculated by using a calculator) and some are based on a software program (Crawley, 2005; Goos & Meintrup, 2015; Haslwanter, 2016; James, Witten, Hastie, & Tibshirani, 2013) (the examples are intended for the software program introduced in the book). There are also applied statistics books where the examples are given in specific fields (Grafen & Hails, 2002; Hawkins, 2014; Riffenburgh, 2012) (statistics in medicine, in pharmacy, biology, psychology, etc.), which

may be an advantage for the newcomer who fears integrals. There are nowadays also many Internet books and Wikis available. The readers of this chapter are as good as the authors in finding Internet books and Wikis so they are not placed in the literature list.

2.3 MULTIVARIATE DATA WITH DEFINITIONS

2.3.1 Data Matrices, Two-Way Arrays

Multivariate data is presented as an array of size I × K (I objects, K variables), see Fig. 2.3. The presence of K variables, where K can be rather large, makes most tools of univariate statistics useless. If some variables are correlated with each other having many variables becomes extra complicated.

A few facts can be mentioned about data arrays. The K variables could be a homogeneous or a heterogeneous set. Homogeneous means that all K variables are given in the same units and come from the same instrument. They could be wavelengths, wavenumber, and energies. One may also note that changing the position of these variables is not recommended. Wavelength, energies, etc., have a natural order from lower to higher and the variables are put in the array in that order. Heterogeneous variables do not have a natural order; they could be pH, thermal conductivity, electrical conductivity, temperature, blood pressure, etc.

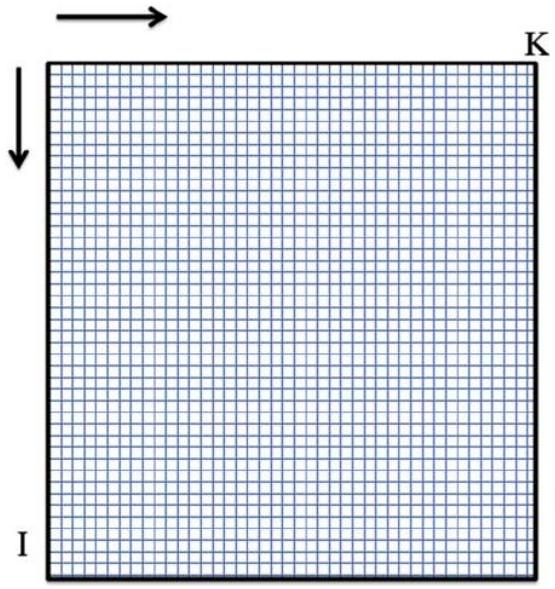


FIGURE 2.3 The data matrix or array of size I × K.

With arrays, as in Fig. 2.4, the concepts of missing data suddenly become important because some of the variables may have missing data, while their neighbors have none. Many things can be said about the pattern of holes in a data matrix and of how meaningful it is to fill the holes. There are two philosophies around missing data: fill the holes in a meaningful way or leave them and write algorithms that somehow ignore them. Especially, heterogeneous variables can have different missing data issues. There have been quite some discussions about how much mis sing data can be allowed. The truth is that miss ing data may occur and that there is no point in throwing away a large data set just because some parts of it are emp ty (Bro & Smilde, 2014; Camacho, 2010; Folguera, Zupan, Cicerone, & Magellanes, 2015).

The array in Fig. 2.3 is too large to be inspected by just looking at the numbers. There needs to be a reduction. This is done as shown in Fig. 2.5. This figure shows how A latent variables are created from the K measured variables. At the same time A latent objects are created. These new matrices can then be used for making further investigations. The principles and techniques for doing this are explained more in detail in Section 2.4.

More on data matrices and their content may be found in the chemometrics literature. Selected paper-based books in print are mentioned here (Brereton, 2003; Gemperline, 2006; Tauler, Walczak, & Brown, 2009; Varmuza & Filzmoser 2009). e-Books are readily available on the Internet.

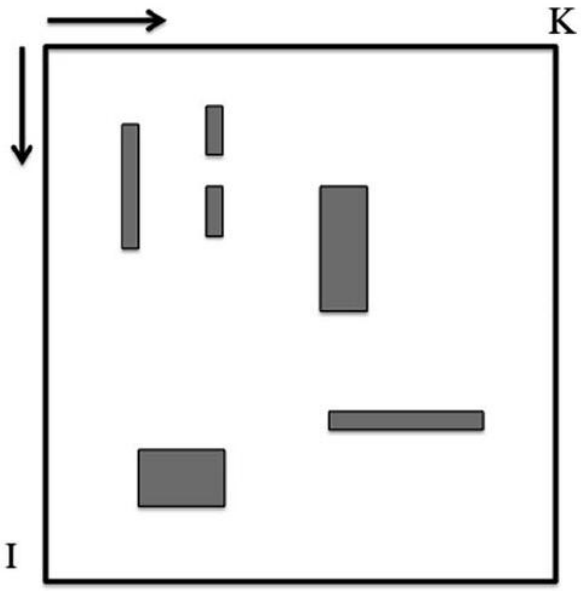


FIGURE 2.4 Missing data can come in different shapes. Some are easier dealt with than others.

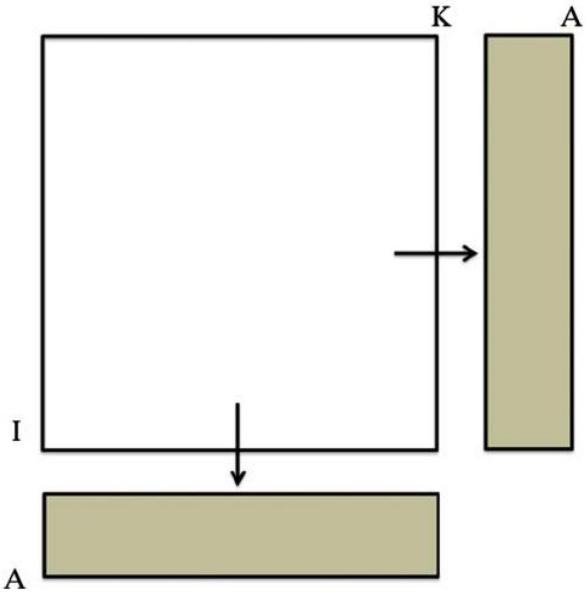


FIGURE 2.5 A data array may be decomposed in a meaningful way into latent variable-based arrays that are much smaller.

2.3.2 Three- and More-Way Arrays

Three-way arrays are also possible to generate. Even four-way, five-way, etc., can be imagined and constructed. Fig. 2.6 shows a three-way array and possible decomposition (Smilde, Bro, & Geladi, 2004). Three- and higher-way arrays contain a lot of data and are even more difficult to overview than two-way arrays. Therefore extra care must be taken to simplify them properly. Some exampl es may help in understanding Fig. 2.6. The 3D array could be samples 3 chromatograms 3 mass spectra or batch process number 3 spectra 3 evolved time in batch. Or it could be I judges 3 J products (e.g., different breads) 3 K quality parameters (mechanical properties, hardness, taste, smell, etc.) judged. The latent variables obtained (see Fig. 2.6) would describe the I judges, J products, and K quality parameters. The most used data analysis methods are Tucker, Parafac/Candecomp, and Parafac 2. This 3D array modeling is too extensive and technical in its description to fit in this chapter, but more can be found in the literature (Cichoski, Zdunek, & Anh, 2009; Coppi & Bolasco, 1989 ; Law et al., 1984). Extensions of most three-way methods to arrays with more than three modes are easy.

A different type of three-way array is the multivariate or hyperspectral image. In this case, two of the ways are image (spatial) dimensions (pixel coordinates) and the third way is variable (see Fig. 2.7). Pixel coordinates could be microscopic up to astronomical. All types of microscopy and macroscopic imaging can create image variables: electron energy levels, gamma or X-ray energies, UV or visible wavelengths, NIR wavelengths or Raman/FTIR wavenumbers. Many more imaging variables can be found, e.g., in ultrasound,

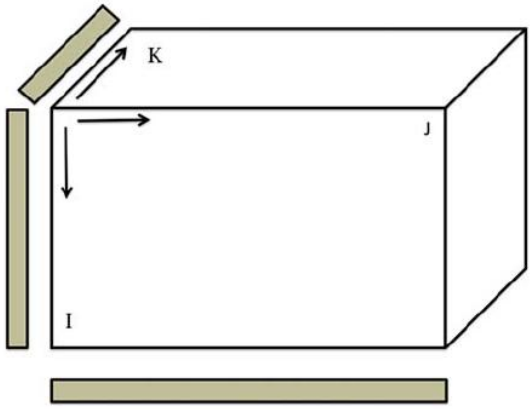


FIGURE 2.6 A three-way array of size I 3 J 3 K with a possible latent variable simplification.

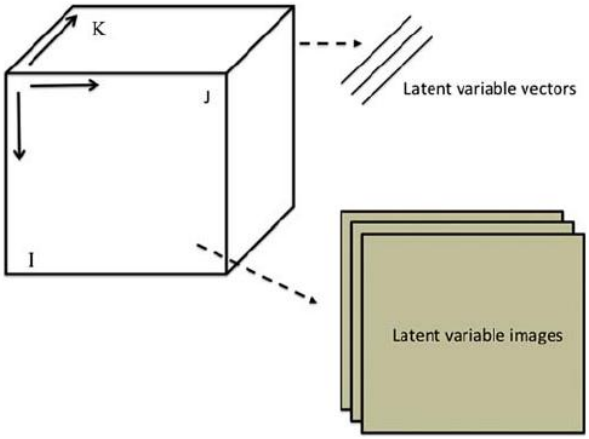


FIGURE 2.7 A hyperspectral image with image size I 3 J and K variables. The simplification is by making latent variable vectors and latent variable images. This is different from the decomposition of a three-way array as in Fig. 2.6.

microwave, or magnetic imaging. Many developed applications are clinical for obvious reasons. Some useful literature is Chang (2003) and Grahn and Geladi (2007). Extensions to four-way images would be: 3D images in many variables or hyperspectral images taken over time. Fig. 2.8 shows a representation of a four-way array.

2.3.3 Multiblock Data

Data may also occur in many blocks instead of a single data array (see Fig. 2.9). In some cases, it is not meaningful to make one big block of several smal ler ones, but the variables

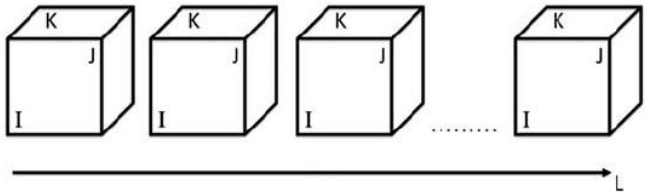


FIGURE 2.8 Four-way data array or image of size I 3 J 3 K 3 L.

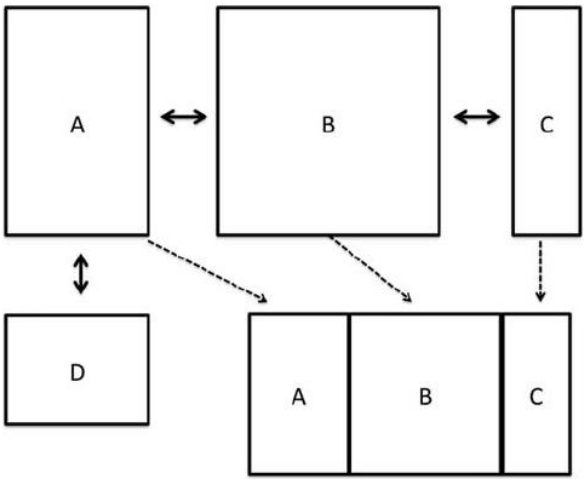


FIGURE 2.9 Data may occur in data matrices A to D that are kept as separate blocks. Relationships between the blocks may exist. Sometimes certain blocks can be used to build a bigger block (dashed arrows).

in the blocks need to be kept apart. Especially in metabolomics, it is easy to encounter more than five blocks with tens to hundreds of variables in each block. Often these data are available for only a few objects.

One extremely important case of multiblock (two block) is in regression modeling. This situation is shown in Fig. 2.10. A block of many variables is related to one of one or a few variables by a model. The reason for this model is that the data in one block are cheap, fast, and easy to measure, while the data in the other block are expensive, slow, produce waste, etc. Especially for the process industry this is an important situation. An example would be measurement of a polymer in a polymer ization process as shown in this generic polymerization reaction:

monomer 1 catalyst 1 temperature 5 . polymer

Classical laboratory chemical analysi s of samples taken from the reactor would be too infrequent and too slow, and also very expensive. But the monomer and polymer may have their own near-infrared, infrared, or Raman spectra. These can be measured fast and

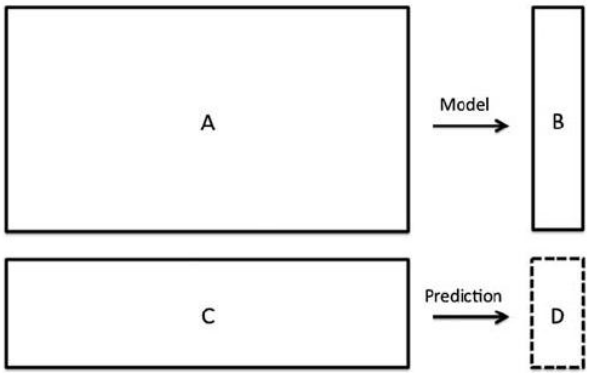


FIGURE 2.10 If a model can be made between a block A and a single variable B, then it can be used in predicting the data in D from those in C.

frequently. If a model can be made between the spectra block and the polymer concentration block as in Fig. 2.10, then polymer concentrations can be calculated from the spectra and be made available almost instantly, allowing the operators to follow the process in real time (Gurden, Martin, & Morris, 1989; Wise & Gallagher, 1996).

Besides regression modeling, there is path modeling of the connecti ons between many blocks. This is often done hierarchically: A - B - C - D, where A to D are symbols of the different blocks. Most literature in path analysis is in psychology, sociology, and economics (Gelman & Hill, 2007; Loehlin, 2004; Lohmo ¨ller, 1989). A popular technique is LISREL: linear structural relations (Byrne, 1989).

2.3.4 General Thoughts About Multivariate Thinking

Multivariate data comes in arrays instead of vectors (univariate data). These are often shown symbolically as rectangles (matrices) or parallelepipeds (three-way arrays or hyperspectral images) or multiblock structures. In all cases, the amount of numbers makes just looking at the data impossible and univariate statistics is not useful anymore. Because most measured data has noise (errors, etc.) in it, the first thing to think about is a decomposition into simpler structures (latent variables) that remove some of the noise and make interpretation easier. The next section gives some specifics on how to model multivariate data.

Multivariate data is almost never used in the raw data form. There is very often a preprocessing, such as variable-wise mean-center ing or variable-wise scaling. There may also be method-specific transformations for baseline removal, etc. Besides transformations, it is also possible to model subsets separately. This can be done in two ways, either based on external knowledge or by doing an internal clustering of the data to get the subsets. For multiblock data, the alternatives are often putting all data in one big block or keeping the subblocks separate (see Fig. 2.9).

A general observation is that multivariate data often has many variables for only a few objects. This makes these data quite different from the univariate case.

2.4 MODELING

2.4.1 General Factor Models

Factor models are the basis of chemometrics. They are used and useful everywhere for multivariate data. The section gives a general factor model and then there are two sections with more specific models: principal component analysis (PCA) and multivariate curve resolution.

Fig. 2.5 shows a general factor model. A data array is represented by several factors . The way these factors are calculated may vary, depending on the demands of the analyst. First some nomenclature from linear algebra is needed.

A data array of size I objects and K variables is written as boldface uppercase: X. Vectors are written as boldface lowercase: a, b, c. All vectors are column vectors. Making them into a row vector is done by transposition, using the symbol superscript uppercase T. a^T is a row vector.

A factor model can be written as:

$$X = AB^T + E \tag{2.1}$$

X: a data array of size I × K
E: a residual of size I × K
A: a matrix I × M with M latent variables as columns
B: a matrix K × M with M latent variables as columns.

Eq. 2.1 can also be written as:

$$X = a_1 b_1^T + a_2 b_2^T + \ldots + a_M b_M^T + E \tag{2.2}$$

where $a_1, a_2 \ldots$ are column vectors in A and $b_1, b_2 \ldots$ are column vectors in B. This is also shown graphically in Fig. 2.11. Another similar decomposition for three-way arrays is also given in Fig. 2.11. The three-way decomposition is known as parallel factor analysis

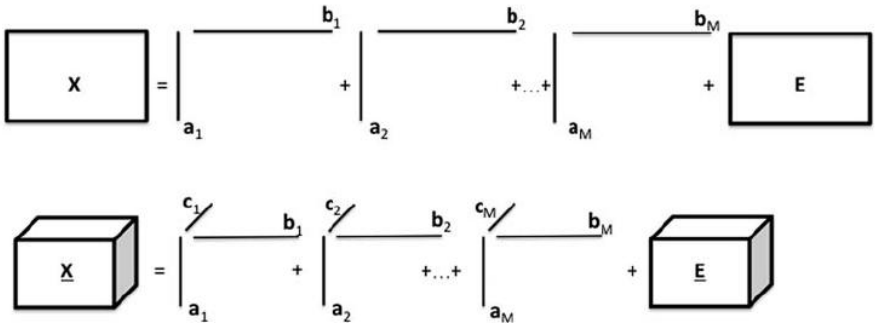


FIGURE 2.11 A graphical view of the factor decomposition for two-way (upper) and three-way (lower) data arrays.

(PARAFAC) (Smilde et al., 2004). Technical details and variations of three-way analysis exist but are not expanded upon because of space limitations.

A general rule, also for three- and more-way arrays, is that AB^T contains meaningful information about the objects or systems studied and that E contains mainly noise.

$$SS(X) = SS(AB^T) + SS(E) \quad (2.3)$$

SS() stands for the operation of calculating the sum of squares of what is within the parentheses, often for mean-centered data. The difference between the algorithms for doing the split as in Eq. 2.1 is what constraints are put on the data in AB^T or on E. The split in Eq. 2.1 should be made with a purpose in mind. There are three main purposes: finding outliers, exploration, and clustering/classification. Finding outliers is using the information in A or B to find outliers and to explain and remove them. Exploration is about finding out what the structured (A and B) and unstructured (noise) parts of the data are and making sense of the structures found.

Clustering is finding out how objects in the data set belong together (are similar) or not. After a meaningful clustering is found it can be used for classification. Probably the most difficult part of Eqs. 2.1 and 2.2 is to find a correct value for M, the number of latent variables to be extracted. The following sections describe PCA and multivariate curve resolution.

2.4.2 Principal Component Analysis

PCA is a technique that is based on finding directions in multivariate space (see Fig. 2.12). A direction is found that has the highest possible sum of squares when all

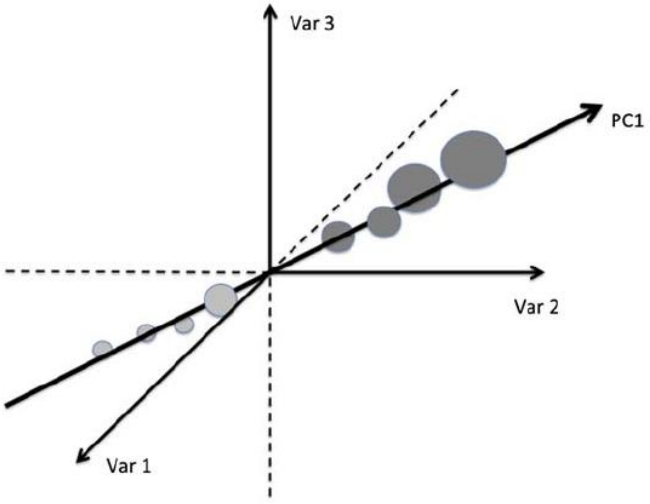


FIGURE 2.12 Illustration of a multivariate space containing two groups, two clusters, shown in darker and lighter gray.

points are projected on it. The assumption is that systematic information has a high sum of squares and that a lower sum of squares is noise or noisy information.

For PCA, it is good to first mean-center the data in X variable-wise. Other pretreatments may be useful.

For homogeneous data, variance-based scaling is not recommended. For heterogeneous data, variance scaling is a must because all variables are measured in different units.

$$X = TP^{T} + E \tag{2.4}$$

T: score matrix (I × A) for A components, the columns of T are orthogonal
P: loading matrix (K × A) for A components, the columns of P are orthogonal.

Eq. 2.4 can be rewritten just like Eq. 2.2:

$$X = t_1 p_1^{T} + t_2 p_2^{T} + \ldots + t_M p_M^{T} + E \tag{2.5}$$

In that case, one can write:

$$SS(X) = SS(t_1 p_1^{T}) + SS(t_2 p_2^{T}) + \ldots + SS(t_M p_M^{T}) + SS(E) \tag{2.6}$$

In this sum, $SS(t_1 p_1^{T})$ is the highest possible sum of squares, $SS(t_2 p_2^{T})$ is the next highest one, and so on. If the SS values are expressed in %, then this would amount to (in an example):

$$100\% = 55\% + 33\% + 10\% + 2\%$$

The first three components explain 98% together and the residual explains 2%, which is probably noise. This is quite a simplification of the data. The usual problem is that it is not easy to select the correct number of components to be calculated.

More can be found in the works of Jackson (1991) and Joliffe (2002). Jackson is out of print but can be bought electronically.

2.4.3 Multivariate Curve Resolution

A completely different situation from PCA is the curve resolution equation. Here variable-wise mean-centering and standard deviation-based scaling are not to be used. They make the constraints invalid. Other pretreatments such as a priori baseline correction may be fine, as long as they do not result in values below zero.

$$X = CS^{T} + E \tag{2.7}$$

Eq. 2.7 is solved based on constraints. It is assumed that the columns of C are concentrations of chemicals and that the columns of S are their pure spectra. This then leads to non-negativity constraints because concentrations cannot go below zero and spectral absorbances cannot go below zero. Eq. 2.7 is solved by iterating between C and S, each time setting any negative number to zero. When these iterations converge, the elements in C are all positive and the elements in S are all positive. Two problems are finding how many components should be calculated and how to check for convergence. It is always good to first do a PCA model for finding how many components are needed.

Eq. 2.6 does not work for curve resolution. The reason is that the extracted components are not orthogonal. Eq. 2.3 is valid. The big advantage of curve resolution is that the interpretation is very easy because spectra and concentrations of pure chemicals are obtained. More can be found in the work of Mulaik (2009).

2.4.4 Clustering—Classification

A huge field of study for multivariate data is clustering. Clustering has been used in anything from psychology, sociology, economics, medicine, biology, and more topics even before chemometrics existed. There are many methods for clustering multivariate data; too many to be mentioned here. Clustering is the act of finding clusters in a calibration data set.

Fig. 2.13 can be thou ght of as a multivariate space containing two groups, two clusters, shown in darker and lighter gray. The activity of clustering is to find:

- whether there are clusters;
- how clusters can be delineated; and
- how valid the clusters are.

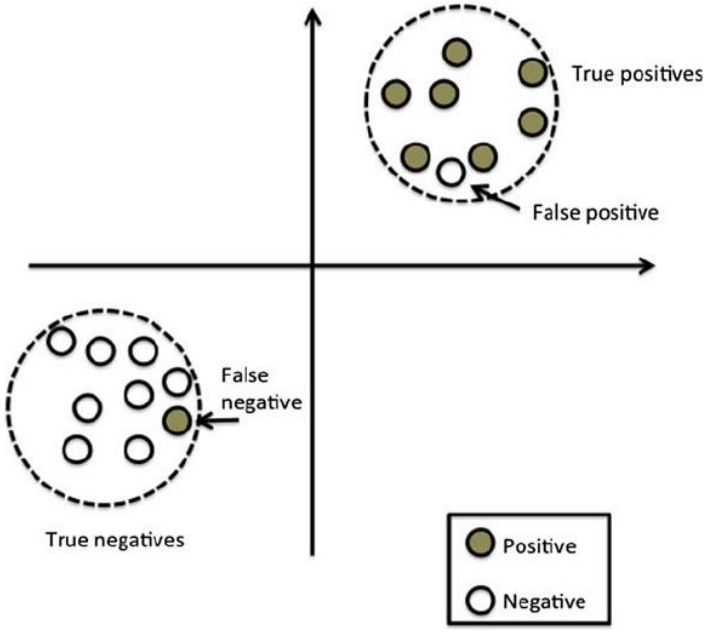


FIGURE 2.13 The figure shows axes in multivariate space (only two variables here). The objects are shown as dark and white circles are positives and negatives. The dashed circles are cluster delineations. This figure is the simplest one possible. There can be more than two clusters and distances between clusters may vary. Some situations can get really complicated.

Here one always encounters outliers, multiple cluster memberships, and other issues that must be dealt with.

Clustering can be done unsupervised: the data form clusters based on some distance or sum of squares criterion. The distance criterion means that being close together in multivariate space is belonging to the same cluster, being far away in multivariate space is belonging to different clusters. With the sum of squares criterion, a small sum of squares means a dense cluster of objects lying close together. Clustering can often be done in a supervised manner: the cluster membership of the objects is known in advance, e.g., sick and healthy patients.

Clusters need to be delineated. All kinds of geometrical figures have been used for this: ellipses, rectangles, bags, border surfaces, etc. After a satisfactory clustering is obtained, this can be applied to new objects (data sets). This is classification. There are a number of books that describe this (Aggarwal & Reddy, 2014; Everitt, Landau , Leese, & Stahl, 2011; Hennig, Meila, Murta gh, & Rocci, 2016; Johnson & Wichern, 2013; Kaufman & Rousse euw, 2005; Kogan, Nicholas, & Teboulle, 2006; Mardia, Kent, & Bibby, 1982; Xu &Wunsch, 2009).

Two important classification concepts are sensitivity (true positive ratio) and selectivity (true negatives correctly class ified). True positives, false positives, true negatives, and false negatives are shown in Fig. 2.13. The terms sensitivity and selectivity are confusing but they were introduced long ago and they have become a traditional ingredient.

In Fig. 2.13, the dark circles could represent patients with a certain disease and the white circles could be healthy patients. The circles draw n around the classes are the cluster models. There are many ways of making such a delineation. Fig. 2.13 also shows one false positive and one false negative. Fig. 2.13 therefore has a true positive ratio of 7/8 (87.5%) and a true negative ration of 8/9 (88.9%). For many clinical applications, true positive and negative ratios below 75% are considered useless.

2.4.5 Regression Models

Fig. 2.10 shows a regress ion situation. The equation is given as:

$$y_{cal} 5 X_{cal} b 1 f_{cal} \quad (2.8)$$

y_{cal} : dependent variable data values (I 3 1) mean-center ed (B in Fig. 2.9);
X_{cal}: independent variable data values (I 3 K) mean-centered (A in Fig. 2.9);
b: a column vector of K regression coefficients;
f_{cal}: a column vector (I 3 1) of residuals.

Eq. 2.8 can be explained by:

$$SSðy_{cal}Þ 5 SSðX_{cal} bÞ 1 SSðf_{cal} Þ \quad (2.9)$$

The SS operator means calculating the sum of squares of a vector. The regression equation's purpose is to make the explainable (structured) part SS(X $_{cal}$ b) large and the unexplainable (noise) part SS(f $_{cal}$) small. This is done by finding an appropriate value for b.

Most of the time Eq. 2.8 cannot be solved unless X $_{cal}$ is replaced by a limited number of latent variables:

$$y_{cal} = T_{cal} b + f_{cal} \quad (2.10)$$

T_{cal}: (I3 A) A latent variables calculated from X_{cal}.

It needs to be pointed out that b and f_{cal} in Eq. 2.10 are different from those in Eq. 2.8.

Methods such as principal component and partial least squares regression are used to calculate T_{cal} from X_{cal}.

There are some advantages to using latent variables in Eq. 2.10. The first is that the calculation of b is possible at all. Another is that the obtained b is better at predicting. Another aspect is again the question of how many latent variables should be used. Too few latent variables are an underfit (not all systematic variation is used for building the model) and too many is an overfit (noise in X_{cal} is used to build the model).

The real purpose of the models in Eqs. 2.8 and 2.10 is prediction of unknowns as shown in Fig. 2.10. Fig. 2.14 shows two situations. In Fig. 2.14A the model works re asonably well and the prediction of test objects is good. There are several statistics inspired from univariate statistics to test model and predictions quality. Fig. 2.14B shows a model that works extremely well, but the prediction is bad. For Eq. 2.10, a good model with bad predictions (Fig. 2.14B) can be obtained by taking too many latent variables, which is called overfitting. Taking the correct number of latent variables leads to the situation in Fig. 2.14A.

2.4.6 Model Diagnostics

The above sections have explained multivariate data in its different guises (Section 2.3) and the most important models that can be made (Section 2.4.1 to 2.4.5). There are three important tasks remaining. The first is finding whether the created models make any sense. The second is checking how good the models are for handling new data. The third important task is to find out why models work. This amounts to studying how the different

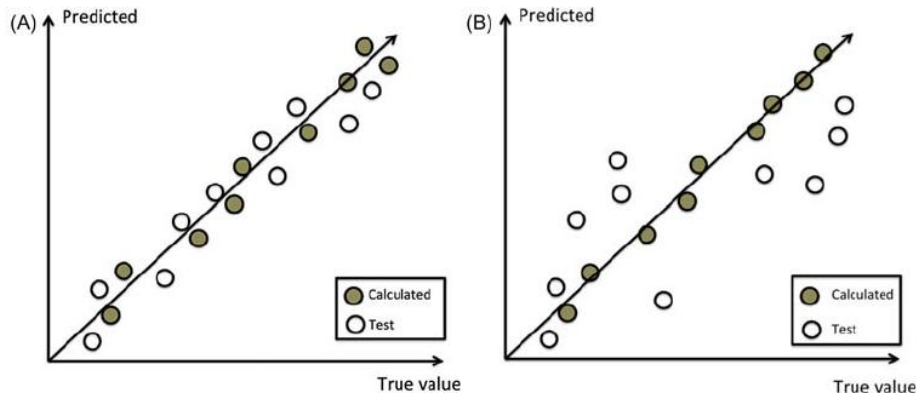


FIGURE 2.14 Testing of a regression. (A) A good regression model with good predictions and (B) an overfitted regression model gives bad predictions.

variables or variable combinations influence the model. This is a very technical topic and the reader can look in general chemometrics books and literature for more detail.

When it comes to diagnostics for the different methods presented earlier, it is obvious that the least squares criterion is very often used for making the models. Therefore the diagnostics are also based on lowering variation and that brings us back to univariate statistics and significance testing.

2.4.7 Some General Thoughts About Modeling

Models can only be made with a purpose in mind. The purposes discussed are outlier detection, exploration, clustering/classification, regression, and multiblock (path model) relationships.

A principal or independent component model is very general. The calculation is mathematically straightforward. Every PCA model of the same data should give the same results with six significant decimals. A curve resolution model is based on specific constraints. The curve resolution algorithms are iterative and the stopping criterion is not very strict. Out of 10 calcul ations made, 9 could be almost identical and one could be totally diffe rent.

An important thing to consider is the number of latent variables needed. This is a difficult choice and a big discussion issue. Already for PCA, the choice of the number of components to be used is quite difficult and no method for finding that number is perfect. For curve resolution, it is impossible to find the correct number of components to be used. Therefore a PCA is done in advance.

The choice between curve resolution and PCA is a difficult one and should be based on background knowledge about the data and how they were obtained.

A huge field of study for multivariate data is clustering. Concepts to be learned and remembered are: (1) unsupervised, (2) supervised, (3) classification modeling, and (4) classification prediction. The re is a huge amount of methods for finding clusters and for delineating them; too much to fit here.

Regression modeling between two blocks of data is very important for any data collected on chemical constituents of mixtures. Regr ession is also very important for modeling of batch or continuous processes.

Preprocessing combined with using a simpler model is often as good as not prepro cessing and using an advanced algorithm. For regression modeling, sometimes just taking logarithms or square root of some variables does the trick for making the model work. Also, modeling on well-chosen subse ts works very well, while trying to model the whole data set just causes confusion.

2.5 CONCLUSIONS

The major conclusion is that too many variables can be dealt with by calculating a reduced number of latent variables and too many objects are dealt with by creating subsets. Making latent variables can be done in many ways and some background information

and a priori thinking are needed to select a few good ones. Also, taking subsets can be done in different ways, as the hundreds of clustering method variations explain.

There is also the topic of asking the right questions. Are the objects studied the ones that should be studied, or a haphazard or in some way limited collection? Are the variables obtained the ones needed, or are they just based on available instrumentation? Chemometrics is always the last process done in a project. All sampling and measurement mistakes were made before any data get analyzed. A remedy here is careful planning in cooperation with all persons involved in a project.

A visually oriented conclusion may be of interest. Some of the important univariate aspects are seen in Figs. 2.1 and 2.2. Figs. 2.3–2.10 show different types of multivariate data sets and how they can be decomposed to achieve less data and more information. Technical details of how a decomposition may be achieved are in Figs. 2.11 and 2.12. Fig. 2.10 shows the essence of regression modeling. Some important aspects of clustering and classification are shown in Fig. 2.13. The figures are not a complete explanation of all chemometrics, but they inspire further thinking.

The experimenter has always to ask the question: What is the truth? What is the difference between subjective and objective knowledge? Already the decision to collect data on a system, or number of samples, is subjective. Then somebody selects a measurement method because of availability or low price. Once these choices are made, the resulting data may have been collected with regard to correct sampling and almost error-free measurement. This data set may be called objective, but then comes the difficult choice of selecting a modeling technique. Here again, subjectivity creeps in. Every chemometrician has his/her own personal favorites and many modeling techniques give widely different results, although the calculation algorithm works correctly up to 10 decimals. A possible way of becoming more objective is by consensus. Many research groups get to study the same data by their own subjective (favorite) methods, but whatever conclusions are made in common are the most objective ones and widely deviating conclusions may be less interesting or more subjective.

It is important to ask the question why data are collected and analyzed. Real chemical, biological, physical, etc., basic properties need to be treated more objectively than parameters for running an industrial process to give maximum efficiency and profit. The recommendation to all chemometricians and users of chemometrics is to think carefully and to define why they are measuring and analyzing data. There is a whole range of whys from universal basic knowledge to "let's set the machine parameters to produce the maximum amount of a desired product." Defining the "why?" may help in posing constraints on sampling, analysis, and data analysis methods.

As a final thought: sometimes the most unwanted results may be the most valuable ones! It just requires some unusual thinking.

References

Aggarwal, C., & Reddy, C. (Eds.), (2014). Data clustering, algorithms and applications. Boca Raton, FL: CRC Press.
Agresti, A. (2007). An introduction to categorical data analysis. Hoboken, NJ: Wiley.
Brereton, R. (2003). Chemometrics: Data analysis for the laboratory and the chemical plant. Chichester: Wiley.

Brereton, R. (2012). A short history of chemometrics: A personal view. Journal of Chemometrics, 28, 749□760.
Bro, R., & Smilde, A. (2014). Principal component analysis. Analytical Methods, 6, 2812□2831.
Byrne, B. (1989). A primer of LISREL: Basic applications and programming for confirmatory factor analytic models. New York, NY: Springer.
Camacho, J. (2010). Missing-data theory in the context of exploratory data analysis. Chemometrics and Intelligent Laboratory Systems, 103, 8□18.
Chang, C. (2003). Hyperspectral imaging. Techniques for spectral detection and classification. New York, NY: Springer.
Cichoski, A., Zdunek, R., & Anh, H. (2009). Non-negative matrix and tensor factorizations: Applications to exploratory multi-way data analysis and blind source separation. Hoboken, NJ: Wiley.
Coppi, R., & Bolasco, S. (Eds.), (1989). Multiway data analysis. North Holland, Amsterdam, The Netherlands: Elsevier.
Cox, D. (2001). Biometrika: The first 100 years. Biometrika, 88, 3□11.
Crawley, M. (2005). Statistics: An introduction using R. Chichester: Wiley.
Devore, J. (Ed.), (2014). Probability and statistics for engineering and the sciences 9th. Boston, MA: Congage Learning.
Esbensen, K., & Geladi, P. (1990). The start and early history of chemometrics—Selected interviews 2. Journal of Chemometrics, 4, 389□412.
Everitt, B., Landau, S., Leese, M., & Stahl, D. (2011). Cluster Analysis (5th ed.). Chichester: Wiley.
Forbes, C., Evans, M., Hastings, N., & Peacock, B. (2011). Statistical distributions (4th ed.). New York, NY: Wiley.
Folguera, L., Zupan, J., Cicerone, D., & Magellanes, J. (2015). Self-organizing maps for imputation of missing data in incomplete data matrices. Chemometrics and Intelligent Laboratory Systems, 143, 146□151.
Geladi, P., & Esbensen, K. (1990). The start and early history of chemometrics—Selected interviews 1. Journal of Chemometrics, 4, 337□354.
Geladi, P., Nelson, A., & Lindholm-Sethson, B. (2007). Complex numbers in chemometrics: Examples from multivariate impedance measurements on lipid monolayers. Analytica Chimica Acta, 595, 152□159.
Gelman, A., & Hill, J. (2007). Data analysis using regression and multilevel hierarchical models. New York, NY: Cambridge University Press.
Gemperline, P. (Ed.), (2006). Practical guide to chemometrics (2nd ed.). Boca Raton, FL: CRC Press.
Goos, P., & Meintrup, D. (2015). Statistics with JMP: Graphs, descriptive statistics and probability. Hoboken, NJ: Wiley.
Grafen, A., & Hails, R. (2002). Modern statistics for the life sciences . Oxford: Oxford University Press.
Grahn, H., & Geladi, P. (Eds.), (2007). Techniques and applications of hyperspectral image analysis. Chichester: Wiley.
Gurden, S., Martin, E., & Morris, A. (1989). The introduction of process chemometrics into an industrial pilot plant laboratory. Chemometrics and Intelligent Laboratory Systems, 14, 319□330.
Haslwanter, T. (2016). An introduction to statistics with python. New York, NY: Springer Verlag.
Hawkins, D. (2014). Biomeasurement. A student's guide to biological statistics. Oxford: Oxford University Press.
Hennig, C., Meila, M., Murtagh, F., & Rocci, R. (Eds.), (2016). Handbook of cluster analysis. Boca Raton, FL: CRC Press.
Hotelling, H. (1936). Simplified calculation of principal components. Psychometrika, 1, 27□35.
Jackson, J. (1991). A user's guide to principal components. New York, NY: Wiley.
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning with applications in R. New York, NY: Springer Verlag.
Johnson, R., & Wichern, D. (2013). Applied multivariate statistical analysis (6th ed). Upper Saddle River, NJ: Pearson Education.
Joliffe, I. (2002). Principal component analysis (2nd ed.). New York, NY: Springer.
Kaufman, L., & Rousseeuw, P. (2005). Finding groups in data: An introduction to cluster analysis. Hoboken, NJ: Wiley.
Kogan, J., Nicholas, C., & Teboulle, M. (Eds.), (2006). Grouping multidimensional data, recent advances in clustering. Berlin: Springer.
Kvalheim, O. (2012). History, philosophy and mathematical basis of the latent variable approach—from a peculiarity in psychology to a general method for analysis of multivariate data. Journal of Chemometrics, 26, 210□217.
Law, H., Snyder, C., Hattie, J., & McDonald, R. (1984). Research methods for multimode data analysis. Santa Barbara, CA: Praeger.

Loehlin, J. (2004). Latent variable models (4th ed.). An introduction to factor, path and structural equation analysis. Mahwah, NJ: Lawrence Erlbaum Association.
Lohmo¨ller, J. B. (1989). Latent variable path modeling with partial least squares. Berlin: Springer.
Mardia, K., Kent, J., & Bibby, J. (1982). Multivariate analysis. London: Academic Press.
Mulaik, S. (2009). Foundations of factor analysis (2nd ed.). Boca Raton, FL: CRC Press.
Riffenburgh, R. (2012). Statistics in medicine (3rd ed.). Amsterdam, The Netherlands: Academic Press Elsevier.
Smilde, A., Bro, R., & Geladi, P. (2004). Multi-way data analysis, applications in the chemical sciences. Chichester: Wiley.
Tauler, R., Walczak, B., & Brown, S. (2009). Comprehensive chemometrics (1st ed). Amsterdam, The Netherlands: Elsevier.
Varmuza, K., & Filzmoser, P. (2009). Introduction to multivariate statistical analysis in chemometrics. Boca Raton, FL: CRC Press.
Wise, B., & Gallagher, N. (1996). The process chemometrics approach to process monitoring and fault detection. Journal of Process Control, 6, 329□348.
Wonnacott, T., & Wonnacott, R. (1990). Introductory statistics (5th ed.). Hoboken, NJ: Wiley.
Xu, R., & Wunsch, D. (2009). Clustering. Hoboken, NJ: Wiley.

CHAPTER

3

Data Processing in Multivariate Analysis of Pharmaceutical Processes

João A. Lopes[1] and Mafalda C. Sarraguça[2]

[1]Universidade de Lisboa, Lisboa, Portugal [2]Universidade do Porto, Porto, Portugal

3.1 INTRODUCTION

Pharmaceutical processes are typically divided into primary and secondary: primary pharmaceutical processes are those intended for the production of active pharmaceutical ingredients (APIs; or eventually excipients) that will be included in pharmaceutical forms. Secondary pharmaceutical processing typically means the incorporation of one or more APIs or drugs on ph armaceutical dosage forms, that can be of a different nature.

These processes, as in any other industry, generate a substantial amount of data. However, it is known that these data, although containing enormou s potential, are often neglected in terms of their intrinsic value. As there are many reasons for this situation, some connected to the traditional way these processes are "optimized" and also regulatory bottlenecks, the situation today is changing dramatically. With the recent guidelines for pharmaceutical development (ICH-Q8), risk assessment (ICH-Q9), quality systems (ICH-Q10), development and manufacture of drug substances (ICH-Q11), pro cess data and data processing methods gain a new importance in this field, in what concerns both the primary and secondary industries (International Conference on Harmonization, I, 2009; International Conference on Harmonization, II, 2005; International Conference on Harmonization, III, 2008; International Conference on Harmonization, IV, 2012).

Pharmaceutical data can be understood from the very early stages of the development of a drug product. Massive amounts of data (including clinical data) are already generated and these need appropriate recording, storage, and consequent analysis. Pharmaceutical data might come additionally from different stages of the drug product development: drug discovery, preclinical research, and clinical research with its multiple stages (Palgon, 2017). What happens often is that the nature of the organization of these data is sparse and

DOI: https://doi-org.passerelle.univ-rennes1.fr/10.1016/B978-0-12-811065-2.00003-3

difficult to use from a broad perspective. Very often these data are unstructured and stored in separate databases with no possible linkage. At this level, it is still a challenge for the pharma industry to take the most out of these data of diverse nature and eventually to use them appropriately on "big-data" management systems (Cattell, Chilukuri, & Levy, 2013). After regulatory approval of a pharma product, pharmaceutical process data (considering here data from R&D to production scale) are in turn more prone for consistent analysis and to establish relationships that may be used for process optimization (e.g., quality, yield, etc.), process scale-up, and continuous improvement over the product's life-cycle. The key for appropriately handling process data and taking the most out of them is to produce consistent, reliable, and linked data. This requires powerful data processing and data management tools (software tools) able to collect, process, store, and integrate data, coming from different stages of the value chain (from discovery to after approval, such as process data, market data, pharmacovigilance, etc.). Regarding process (or manufacturing) data, data integration is also very important so that appropriate process quality control can be achieved. This is especially relevant in the context of new pharmaceuti cal production trends such as the continuous manufacturing of drug products (Mascia et al., 2013) or real-time-release- testing (Pawar et al., 2016). These paradigms of production, somehow connected, require a new way of handling process data which no longer can be considered as advantageous, but is simply a fundamental key for the entire production paradigm to succeed. The novel approaches to medicinal products manufacturing involve the utilization of in-process analytics, sensors, algorithms combined and integrated in data management platforms, that validated under current good manufacturing practice (cGMP) requirements can operate in real-time and output consistent and valuable outcomes, not only for process engineers/pharmaceutic scientists, but for all levels of a pharmaceutical company structure (from production to business). Moreover, the increasing number of approvals of biologics and nanostructured pharmaceutical products presents a novel challenge for the industry in regards to data generation and use. New analytical tools that better address the need for appropriate identification of critical quality attributes (CQAs) for these products (originating from the definition of the target product profile) are needed. Additionally, alternative and more efficient ways of characterizing these products (even (re)defining quality issues in the context of submission to re gulatory authorities) is needed as product's structural properties are highly related to their clinical efficacy.

3.1.1 Pharmaceutical Process Data

Pharmaceutical process data is subjected to special requirements as companies operate under cGMP, which implies several require ments, some related to data quality and management. In a recent article, Ruth (2017) stated that by now, data integrity has become of utmost importance as increased attention to shortc omings by global regulatory agencies and emergent less stringent regulatory environments are here to stay. These data include any data flowing in the contex t of pharmaceutical production and naturally includes data coming from product manufacturing. A tendency to move in the direction of sharply decreased regulation is emerging. Hence, a lower burden on the regulatory structure should not be reflected in a decrease in the safety and efficacy of drug products. A recent

guidance for industry reinforces the need for ensuring data integrity in the pharmaceutical industrial sector (Food and Drug Administration, 2016). Data integrity is absolutely critical as it is the industry's responsibility to ensure the safety, efficacy, and quality of drugs. cGMP regulations are there to provide guidance on how to implement flexible ways based on risk assessment to prevent data integrity pro blems that should be adopted by companies. Data integrity in the context specified in the aforementioned guidelines means the completeness, consistency, and accuracy of data. Therefore, any data processing methodology applied to pharmaceutical data, independently of their origin, should not corrupt or alter the fundamental nature of the data, which is often collected by multiple sensors. Besides data coming from processes or products, metadata is also of crucial importance as it allows the generated data to be properly used for any needed reconstruction within the cGMP environment under which pharma companies are expected to operate. A major (but not the only) component of data integrity is accuracy. However, mathematical processing methods should not change data accuracy and should be used with knowledge and parsimony. Data scientists in this context assume especial importance so that valid relationships are always established on the basis of conclusions taken from data.

3.1.2 The Quality-by-Design Principle

Over the last decade, a change in the manufacturing paradigm in the pharmaceutical industry has evolved: from quality-by-testing (QbT) to quality-by-design (QbD). QbD was framed by the International Conference on Harmonization (ICH) Q8 guideline and is based on science-based approaches designed to create a more flexible production environment. With QbD, the product quality is assured by understanding and controlling the manufacturing process and formulation. The QbD approach starts with the identification of the quality target product profile (QTPP) which determines the design criteria for the intended product (Pramod, Abu Tahir, Cha roo, Ansari, & Ali, 2016). This profile is the basis for the definition of the CQAs and critical process parameters (CPPs). A CQA is defined as a property or characteristic (physical, chemical, biological, or microbiological) that should be within an appropriate limit, range, or distribution to ensure product quality (International Conference on Harmonization, I, 2009). CQ As are applicable in both in-process and finished product and are dependent on the process itself. To identify CQAs it is necessary to consider all quality attributes, physical attributes, identification, assay, content uniformity, dissolution and drug release, degradation products, residual solvents, moisture, microbial limits, among others. The CPP is defined as a parameter whose variability has an impact in a CQA and therefore should be monitored and controlled to ensure product quality (International Conference on Harmonization, I, 2009). Some examples of the definition of CQAs and CP P in unit operations are given here. In a fluid-bed granulation to produce tablets, the CQAs were defined as particle size and particle size distribution, powder densities, angle of repose, and flowability. Regarding the CPP, the inlet air temperature, binder spray rate, and air flow rate were considered important (Lourenço et al., 2012). In a roller compaction unit for tablet production, the estimated CQAs were tablet weight, tablet dissolution, hardness, and ribbon density. The respective CPPs were API flow rate, lubricant flow rate, and precompression pressure (Teckoe,

Mascaro, Farrell, & Rajabi-Siahboomi, 2013). These are some examples, but for a more comprehensive review on examples of CQAs and CPPs please refer to Yu et al. (2014) and the references therein.

A successful QbD strategy applied to one product, involves multiple stages such as the definition of a QTPP, identifying CQAs, evaluating risk, building a design space, implementing a control strategy, and identifying a strategy for the life-cycle management. All steps require appropriate tools. In particular, the definition of the design space typically requires the adoption of process models (first principles when available or empirical data based). Additionally, collecting process data is normally needed, and for that the process behavior must be tested against different manufacturing settings for the process variables. In this context, design of experiments (DoEs) is required. DoEs is not an optimization technique in itself. It is rather a way of choosing samples in the design space in order to get the maxim um amount of information using the minimum amount of resources, i.e., with a lower number of samples (Armstrong, 2006). DoE is an exceptional tool that when applied to pharmaceutical processes allows the systematic manipulation of factors according to a predefined design. DoE can be used to determine the relationship between the input and output parameters of a process; therefore, it can be used to help in the identification of CPPs and to understand design space as defined in the ICH-Q8 guideline. The type of DoE depends on the specific need of the user. The methodology can be used for comparative or screening experiments, for defining response surfaces and regression modeling. Screening experiments involves the selection of key factors affecting a response. Normally, for screening experiments, a relatively small number of experiments are needed. After selecting the target factors, response surface modeling can be used to optimize the response, in this way reducing variability and making the process more robust. Within this classification, there is a large number of experimental designs that can be used and their choice is based on the objective and also previous knowledge (Armstrong, 2006).

Process analytical technology (PAT) is also an important part of QbD. The ICH-Q8 (International Conference on Harmonization, I, 2009) identifies the use of PAT tools to monitor and control the process and ensure that it remains within the design space. The use of multivariate sensors, such as near-infrared spectroscopy (NIRS) as a PAT tool combined with multivariate modeling and analysis will allow for the defining of control strategies.

3.2 CONTINUOUS VERSUS BATCH PROCESSES

Traditionally, production in the pharmaceutical industry is in the batch mode. Nowadays, there is an increase interest in switching from batch to continuous manufacturing due mainly to economic factors as, e.g., the reduction of production costs and increase in production flexibility (Tezyk, Milanowski, Ernst, & Lulek, 2016).

Batch manufacturing is composed of isolated processes in which the materials (APIs and excipients) are introduced before the process startup and are discharged at the end. Because of this configuration, raw materials and intermediate products are typically tested offline before each subsequent unit operation. In continuous manufacturing, the starting materials are fed continuously and intermediates are also processed in a continuous mode without breaks. In a fully continuous manufactur ing process, the unit operations