1
What is Chemometrics?

Learning objectives

- To define chemometrics
- To learn how to count with bits and how to perform arithmetic or logical operations in a computer
- To understand the principal terminology for computer systems and the meaning of robotics and automation.

The development of the discipline chemometrics is strongly related to the use of computers in chemistry. Some analytical groups in the 1970s were already working with statistical and mathematical methods that are ascribed nowadays to chemometric methods. Those early investigations were connected to the use of mainframe computers.

The notation *chemometrics* was introduced in 1972 by the Swede Svante Wold and the American Bruce R. Kowalski. The foundation of the International Chemometrics Society in 1974 led to the first description of this discipline. In the following years, several conference series were organized, for example, Computer Application in Analytics (COMPANA), Computer-Based Analytical Chemistry (COBAC), and Chemometrics in Analytical Chemistry (CAC). Some journals devoted special sections to papers on chemometrics. Later, novel chemometric journals were started, such as the *Journal of Chemometrics* (Wiley) and *Chemometrics and Intelligent Laboratory Systems* (Elsevier).

An actual definition of chemometrics is:

> *the chemical discipline that uses mathematical and statistical methods, (a) to design or select optimal measurement procedures and experiments, and (b) to provide maximum chemical information by analyzing chemical data.*

Chemometrics: Statistics and Computer Application in Analytical Chemistry, Third Edition. Matthias Otto.

The discipline of chemometrics originates in chemistry. Typical applications of chemometric methods are the development of quantitative structure activity relationships and the evaluation of analytical–chemical data. The data flood generated by modern analytical instrumentation is one reason that analytical chemists in particular develop applications of chemometric methods. Chemometric methods in *analytics* is the discipline that uses mathematical and statistical methods to obtain relevant information on material systems.

With the availability of personal computers at the beginning of the 1980s, a new age commenced for the acquisition, processing, and interpretation of chemical data. In fact, today, every scientist uses software, in one form or another, that is related to mathematical methods or to processing of knowledge. As a consequence, the necessity emerges for a deeper understanding of those methods.

The education of chemists in mathematics and statistics is usually unsatisfactory. Therefore, one of the initial aims of chemometrics was to make complicated mathematical methods practicable. Meanwhile, the commercialized statistical and numerical software simplifies this process, so that all important chemometric methods can be taught in appropriate computer demonstrations.

Apart from the statistical–mathematical methods, the topics of chemometrics are also related to problems of the computer-based laboratory, to methods for handling chemical or spectroscopic databases, and to methods of artificial intelligence.

In addition, chemometricians contribute to the development of all these methods. As a rule, these developments are dedicated to particular practical requirements, such as the automatic optimization of chromatographic separations or in prediction of the biological activity of a chemical compound.

1.1 The Computer-Based Laboratory

Nowadays, the computer is an indispensable tool in research and development. The computer is linked to analytical instrumentation; it serves as a tool for acquiring data, word processing, and handling databases and quality assurance systems. In addition, the computer is the basis for modern communication techniques such as electronic mails or video conferences. In order to understand the important principles of computer usage, some fundamentals are considered here, that is, coding and processing of digital information, the main components of the computer, programming languages, computer networking, and automation processes.

Analog and Digital Data

The use of digital data provides several advantages over the use of analog data. Digital data are less noise sensitive. The only noise arises from round-off errors due to finite representation of the digits of a number. They are less prone to, for instance, electrical interferences, and they are compatible with digital computers.

As a rule, primary data are generated as analog signals either in a discrete or a continuous mode (Figure 1.1). For example, monitoring the intensity of optical radiation by means of a photocell provides a continuous signal. Weak radiation, however, could be monitored by detecting individual photons by a photomultiplier.

Usually, the analog signals generated are converted into digital data by an analog-to-digital converter (ADC) as explained as follows.

Binary versus Decimal Number System

In a digital measurement, the number of pulses occurring within a specified set of boundary conditions is counted. The easiest way to count is to have the pulses represented as binary numbers. In this way, only two electronic states are required. To represent the decimal numbers from 0 to 9, one would need 10 different states. Typically, the binary numbers 0 and 1 are represented electronically by voltage signals of 0.5 and 5 V, respectively. Binary numbers characterize coefficients of the power of 2, so that any number of the decimal system can be described.

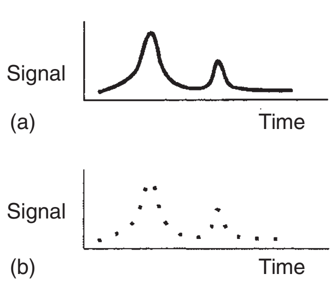

Figure 1.1 Signal dependence on time of an analog (a) and a digital detector (b).

Table 1.1 Relationship between binary and decimal numbers.

Binary number	Decimal number
0	0
1	1
10	2
11	3
100	4
101	5
110	6
111	7
1000	8
1001	9
1010	10
1101	13
10000	16
100000	32
1000000	64

Example 1.1 *Binary Number Representation*

The decimal number 77 is expressed as binary number by 1001101, that is,

1	0	0	1	1	0	1
1×2^6	0×2^5	0×2^4	1×2^3	1×2^2	0×2^1	$1 \times 2^0 =$
64	+0	+0	+8	+4	+0	$+1 = 77$

Table 1.1 summarizes further relationships between binary and decimal numbers. Every binary number is composed of individual *bits* (binary digits). The digit lying farthest to the right is termed the *least significant* digit and the one on the left is the *most significant* digit.

How are calculations done using binary numbers? Arithmetic operations are similar but simpler than those for decimal numbers. In addition, for example, four combinations are feasible:

0	0	1	1
+0	+1	+0	+1
0	1	1	10

Note that for addition of the binary numbers 1 plus 1, a 1 is carried over to the next higher power of 2.

Example 1.2 *Calculation with Binary Numbers*

Consider addition of 21 + 5 in the case of a decimal (a) and of a binary number (b):

a.	21	b.	10101
	+5		101
	26		11010

Apart from arithmetic operations in the computer, logical reasoning is necessary too. This might be in the course of an algorithm or in connection with an expert system. Logical operations with binary numbers are summarized in Table 1.2.

It should be mentioned that a very compact representation of numbers is based on the *hexadecimal number system.* However,

Table 1.2 Truth values for logical connectives of predicates *p* and *q* based on binary numbers.

p	*q*	*p* AND *q*	*p* OR *q*	IF *p* THEN *q*	NOT *p*
1	1	1	1	1	0
1	0	0	1	0	—
0	1	0	1	1	1
0	0	0	0	1	—

1 True and 0 false.

hexadecimal numbers are easily converted into binary data, so the details need not be explored here.

Digital and Analog Converters

Analog-to-Digital Converters (ADCs)

In order to benefit from the advantages of digital data evaluation, the analog signals are converted into digital ones. An analog signal consists of an infinitely dense sequence of signal values in a theoretically infinitely small resolution. The conversion of analog into digital signals in the ADC results in a definite reduction of information. For conversion, signal values are sampled in a predefined time interval and quantified in an *n*-ary raster (Figure 1.2). The output signal is a code word consisting of *n* bits. Using *n* bits, 2^n different levels can be coded, for example, an 8-bit ADC has a resolution of $2^8 = 256$ amplitude levels..

Digital-to-Analog Converters (DACs)

Converting digital into analog information is necessary if an external device is to be controlled or if the data have to be represented by an analog output unit. The resolution of the analog signal is determined by the number of processed bits in the converter. A 10-bit DAC provides $2^{10} = 1024$ different voltage increments. Its resolution is then 1/1024 or approximately 0.1%.

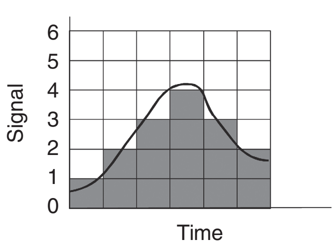

Figure 1.2 Digitization of an analog signal by an analog-to-digital converter (ADC).

Computer Terminology

Representation of numbers in a computer by bits has already been considered. The combination of 8 bits is called a *byte*. A series of bytes arranged in sequence to represent a piece of data is termed a *word*. Typical word sizes are 8, 16, 32, or 64 bits or 1, 2, 4, and 8 bytes.

Words are processed in *registers.* A sequence of operations in a register enables *algorithms* to be performed. One or several algorithms make up a *computer program.*

The physical components of a computer form the *hardware.* Hardware includes the disk and hard drives, clocks, memory units, and registers for arithmetic and logical operations. Programs and instructions for the computer, including the tapes and disks for their storage, represent the *software.*

Components of Computers

Central Processing Units and Buses

A bus consists of a set of parallel conductors that forms a main transition path in a computer.

The heart of a computer is the central processing unit (CPU). In a microprocessor or minicomputer, this unit consists of a highly integrated chip.

The different components of a computer, its memory, and the peripheral devices, such as printers or scanners, are joined by buses. To guarantee rapid communication among the various parts of a computer, information is exchanged on the basis of a definitive word size, for example, 16 bits, simultaneously over parallel lines of the bus. A data bus serves the exchange of data into and out of the CPU. The origin and the destination of the data in the bus are specified by the address bus. For example, an address bus with 16 lines can address $2^{16} = 65536$ different registers or other locations in the computer or in its memory. Control and status information to and from the CPU are administrated in the control bus. The peripheral devices are controlled by an external bus system, for example, an RS-232 interface for serial data transfer or the IEEE-488 interface for parallel transfer of data.

Memory

The microcomputer or microprocessor contains typically two kinds of memory: *random access memory* (RAM) and *read-only memory* (ROM). The term RAM is somewhat misleading and historically reasoned, since random access is feasible for RAM and ROM alike. The RAM can be used to read and write information. In contrast, information in a ROM is written once, so that it can be read, but not reprogrammed. ROMs are needed in microcomputers or pocket calculators in order to perform fixed programs, for example, for calculation of logarithms or standard deviations.

Larger programs and data collections are stored in *bulk storage devices*. In the beginning of the computer age, magnetic tapes were the standard here. Nowadays CD's, DVD' , and Blu-Ray's are used providing a storage capacity of several gigabytes. In addition, every computer is equipped with a hard disk of at least several gigabytes. The access time to retrieve the stored information is in the order of a few milliseconds.

Input/Output Systems

Communication with the computer is carried out by input–output (I/O) operations. Typical input devices are keyboard, magnetic tapes and disks, and the signals of an analytical instrument. Output devices are screens, printers, and plotters, as well as tapes and disks. To convert analog information into digital or vice versa, the aforementioned ADCs or DACs are used.

Programs

Programming a computer at 0 and 1 states or bits is possible using *machine code.* Since this kind of programming is rather time consuming, higher level languages have been developed where whole groups of bit operations are assembled. However, these so-called *assembler languages* are still difficult to handle. Therefore, high-level algorithmic languages, such as FORTRAN, BASIC, PASCAL, or C, are more common in analytical chemistry. With high-level languages, the instructions for performing an algorithm can easily be formulated in a computer program. Thereafter, these instructions are translated into machine code by means of a *compiler.*

For logical programming, additional high-level languages exist, for example, LISP (List Processing language) or PROLOG (Programming in Logic). Further developments are found in the so-called *Shells,* which can be used directly for building expert systems.

Networking

A very effective communication between computers, analytical instruments, and databases is based on networks. There are local nets, for example, within an industrial laboratory as well as

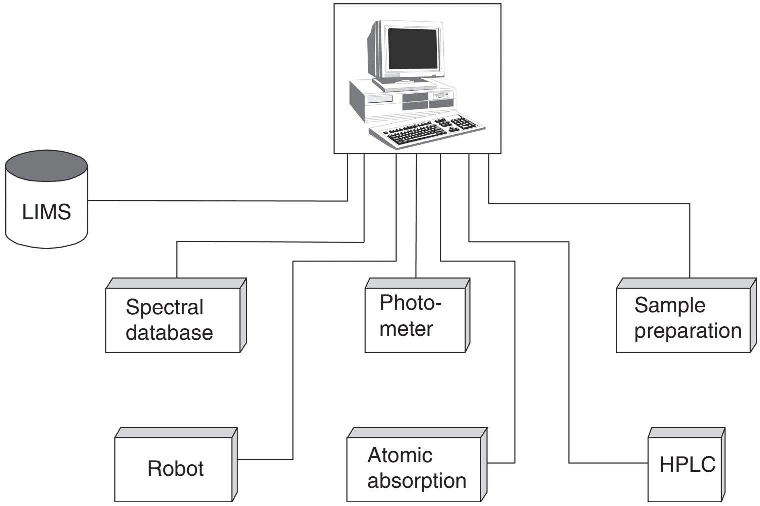

Figure 1.3 Local area network (LAN) to connect analytical instruments, a robot, and a laboratory information management system (LIMS).

national or worldwide networks. Local area networks (LANs) are used to transfer information about analysis samples, measurements, research projects, or in-house databases. A typical LAN is demonstrated in Figure 1.3. It contains a laboratory information management system (LIMS), where all information about the sample or the progresses in a project can be stored and further processed (cf. Section 7.1).

Worldwide networking is feasible, for example, via Internet or CompuServe. These nets are used to exchange electronic mails (e-mail) or data with universities, research institutions, or industry.

Robotics and Automation

Apart from acquiring and processing analytical data, the computer can also be used to control or supervise automatic procedures. To automate manual procedures, a *robot* is applied. A robot is a reprogrammable device that can perform a task more cheaply and effectively than a human being.

Typical geometric shapes of a robot arm are sketched in Figure 1.4. The anthropomorphic geometry (Figure 1.4a) is derived from the human torso, that is, there is a waist, a shoulder, an elbow, and a wrist. Although this type of robot is mainly found in the automobile industry, it can also be used for manipulation of liquid or solid samples.

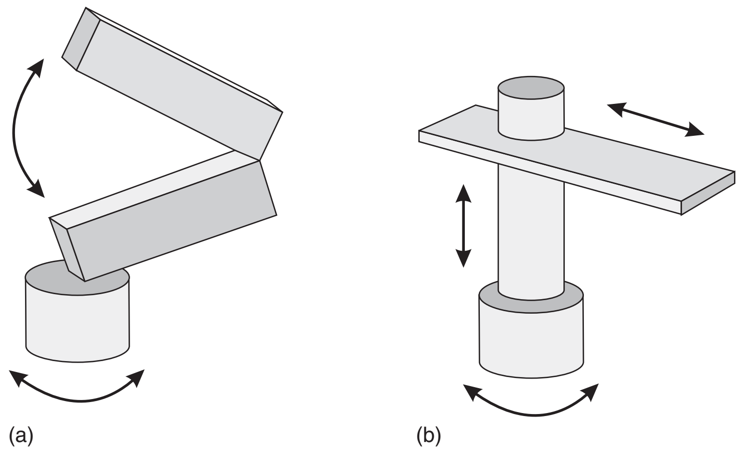

Figure 1.4 Anthropomorphic (a) and cylindrical (b) geometry of robot arms.

In the chemical laboratory, the cylindrical geometry dominates (Figure 1.4b). The revolving robot arm can be moved in horizontal and vertical directions. Typical operations of a robot are as follows:

- *Manipulation* of test tubes or glassware around the robotic work area
- *Weighing,* for the determination of a sample amount or for checking unit operations, for example, addition of solvent
- *Liquid handling,* in order to dilute or add reagent solutions
- *Conditioning* of a sample by heating or cooling
- *Separations* based on filtrations or extractions
- *Measurements* by analytical procedures, such as spectrophotometry or chromatography
- *Control and supervision* of the different analytical steps.

Programming of a robot is based on software dedicated to the actual manufacture. The software consists of elements to control the peripheral devices (robot arm, balance, pumps), to switch the devices on and off, and to provide instructions on the basis of logical structures, for example, IF – THEN rules.

Alternatives for automation in a laboratory are *discrete analyzers* and *flowing systems.* By means of discrete analyzers, unit operations such as dilution, extraction and dialyses can be automated. Continuous flow analyzers or flow injection analyses serve similar objectives for automation, for example, for the determination of clinical parameters in blood serum.

The transfer of manual operations to a robot or an automated system provides the following advantages:

- High productivity and/or minimization of costs
- Improved precision and trueness of results
- Increased assurance for performing laboratory operations
- Easier validation of the different steps of an analytical procedure.

The increasing degree of automation in the laboratory leads to more and more measurements that are available online in the computer and have to be further processed by chemometric data evaluation methods.

1.2 Statistics and Data Interpretation

Table 1.3 provides an overview of chemometric methods. The main emphasis is on statistical – mathematical methods. Random data are characterized and tested by the descriptive and inference methods of statistics, respectively. Their importance increases in connection with the aims of quality control and quality assurance. Signal processing is carried out by means of algorithms for smoothing, filtering, derivation, and integration. Transformation methods such as the Fourier or Hadamard transformations also belong in this area.

Efficient experimentation is based on the methods of experimental design and its quantitative evaluation. The latter can be performed by means of mathematical models or graphical representations. Alternatively, sequential methods are applied, such as the simplex method, instead of these simultaneous methods of experimental optimization. There, the optimum conditions are found by systematic search for the objective criterion, for example, the maximum yield of a chemical reaction, in the space of all experimental variables.

Table 1.3 Chemometric methods for data evaluation and interpretation.

Descriptive and inference statistics
Signal processing
Experimental design
Modeling
Optimization
Pattern recognition
Classification
Artificial intelligence methods
Image processing
Information and system theory

To find patterns in data and to assign samples, materials, or, in general, objects, to those patterns, multivariate methods of data analysis are applied. Recognition of patterns, classes, or clusters is feasible with projection methods, such as principal component analysis or factor analysis, or with cluster analysis. To construct class models for classification of unknown objects, we will introduce discriminant analyses.

To characterize the information content of analytical procedures, information theory is used in chemometrics.

1.3 Computer-Based Information Systems/Artificial Intelligence

A further subject of chemometrics is the computer-based processing of chemical structures and spectra.

There, it might be necessary to extract a complete or partial structure from a collection of molecular structures or to compare an unknown spectrum with the spectra of a spectral library.

For both kinds of queries, methods for representation and manipulation of structures and spectra in databases are needed. In addition, problems of data exchange formats, for example, between a measured spectrum and a spectrum of a database, are to be decided.

If no comparable spectrum is found in a spectral library, then methods for spectra interpretation become necessary. For interpretation of atomic and molecular spectra, in principle, all the statistical methods for pattern recognition are appropriate (cf. Section 1.2). In addition, *methods of artificial intelligence* are used. They include methods of logical reasoning and tools for developing expert systems. Apart from the methods of classical logic in this context, methods of approximate reasoning and of *fuzzy logic* can also be exploited. These interpretation systems constitute methods of *knowledge processing* in contrast to data processing based on mathematical – statistical methods.

Knowledge acquisition is mainly based on expert knowledge, for example, the infrared spectroscopist is asked to contribute his knowledge in the development of an interpretation system for infrared spectra. Additionally, methods are required for automatic knowledge acquisition in the form of *machine learning.*

Methods based on fuzzy theory, neural nets, and evolutionary strategies are denoted as *soft computing.*

The methods of artificial intelligence and machine learning are not restricted to the interpretation of spectra. They can also be used to develop expert systems, for example, for the analysis of drugs or the synthesis of an organic compound.

Novel methods are based on biological analogs, such as neural networks and evolutionary strategies, for example, genetic algorithms. Future areas of research for chemometricians will include the investigation of *fractal structures* in chemistry and of models based on the theory of *chaos.*

General Reading

1. Sharaf, M.A., Illman, D.L., and Kowalski, B.R. (1986) *Chemometrics*, Chemical Analysis Series, vol. **82**, John Wiley & Sons, Inc., New York.
2. Massart, D.L., Vandeginste, B.G.M., Deming, S.N., Mi-chotte, Y., and Kaufmann, L. (1988) *Chemometrics – a Textbook*, Elsevier, Amsterdam.
3. Brown, S.D., Tauler, R., and Walczak, B. (eds) (2009) *Comprehensive Chemometrics – Chemical and Biochemical Data Analysis*, 4 Volumes, Elsevier, Amsterdam.
4. Varmuza, K. and Filzmoser, P. (2009) *Introduction to Multivariate Statistical Analysis in Chemometrics*, CRC Press, Boca Raton, FL, Berlin.

Questions and Problems

1. Calculate the resolution for 10-, 16-, and 20-bit analog-to-digital converters.
2. How many bits are stored in an 8-byte word?
3. What is the difference between procedural and logical programming languages?
4. Discuss typical operations of an analytical robot.