Xiaoli Chu · Yue Huang · Yong-Huan Yun · Xihui Bian

Chemometric Methods in Analytical Spectroscopy Technology

Springer

Chemometric Methods in Analytical Spectroscopy Technology

Xiaoli Chu · Yue Huang · Yong-Huan Yun ·
Xihui Bian

Chemometric Methods in Analytical Spectroscopy Technology

Springer

Xiaoli Chu
Analytical Research Department
Sinopec Research Institute of Petroleum Processing
Beijing, China

Yong-Huan Yun
School of Food Science and Engineering
Hainan University
Haikou, China

Yue Huang
College of Food Science and Nutritional Engineering
China Agricultural University
Beijing, China

Xihui Bian
School of Chemical Engineering and Technology
Tiangong University
Tianjin, China

ISBN 978-981-19-1624-3 ISBN 978-981-19-1625-0 (eBook)
https://doi.org/10.1007/978-981-19-1625-0

Translation from the Chinese language edition: "现代光谱分析技术中的化学计量学方法" by Xiaoli Chu et al., © Chemical Industry Press 2022. Published by Chemical Industry Press.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

In recent years, modern spectroscopic analysis techniques (such as near-infrared, mid-infrared, ultraviolet-visible, molecular fluorescence, Raman, terahertz, laser-introduced breakdown spectroscopy, etc.) have been tremendously developed at high speed. The main feature of these technologies is the involvement of chemometric methods to process spectral data, so as to obtain as much quantitative and qualitative information as possible, and significantly improve the robustness and accuracy of the spectral analysis. Specifically, modern spectroscopies can directly perform qualitative and quantitative analyses of various complex such as gases, liquids, and solids, exhibiting the advantages of high speed, high efficiency, non-destruction, and online feasibility. It has been widely applied in fields of agriculture, food, pharmaceuticals, petroleum, chemical industry, tobacco, environmental protection and medicine, etc., playing an increasingly important role in scientific research and industries.

In recent decades, with the rapid development of artificial intelligence, data mining, and cloud computing, new chemometric methods have sprung up and become one of the fastest-growing branches in spectroscopic analysis technology, which is also a research hotspot for scholars all around the world. This book mainly discusses the chemometric methods used for spectral analysis, including spectral preprocessing, variable selection, data dimensionality reduction, linear or nonlinear multivariate calibrations, pattern recognition, calibration sample selection, outlier recognition, model update and maintenance, multi-spectral data fusion, calibration transfer, and deep learning algorithms, etc.

Considering the comprehensiveness and systematic reviewing, this book summarizes and reviews the latest research progresses of chemometrics in the spectral analysis, particularly, which are closely combined with scientific researches and practical applications, as well as, many algorithm improvements and strategy extensions. The authors believe this book will provide new aspects and ideas for researchers and users in this field. From the perspective of practicability, this book provides as much as possible the complete framework of several kinds of algorithm so that readers can initially understand the mainstream knowledge and context of chemometrics. If readers are interested in the details of certain algorithms, they can find out more knowledge according to the reference documents of this book.

This book was planned by Prof. Xiaoli Chu. He wrote the Chinese version of this book, which is widely praised by readers in the field of chemometrics and spectral analysis in China. For this book of English version, Dr. Yue Huang wrote the Chaps. 1, 7, 11, 17, 19, and 20. Dr. Yong-Huan Yun wrote the Chaps. 2, 3, 5, 9, 10, 13, and 15. Dr. Xihui Bian wrote the Chaps. 4, 6, 8, 12, 14, 16, and 18. At last, Prof. Xiaoli Chu made the final revision and proofread of the book. Due to the rapid development of chemometrics and the limitation of the authors' knowledge and English writing level, there must be some unavoidable omissions, errors, and inadequate interpretation in the book. Please feel free to criticize and correct them, and the e-mail of the corresponding author is cxlyuli@sina.com.

Beijing, China — Xiaoli Chu, Ph.D.
Beijing, China — Yue Huang, Ph.D.
Haikou, China — Yong-Huan Yun, Ph.D.
Tianjin, China — Xihui Bian, Ph.D.

Contents

Chapter 1
Chemometric Methods in Analytical Spectroscopy Technology

Summary

In recent decades, with the rapid development of artificial intelligence, data mining, and cloud computing, new chemometric methods have sprung up and become one of the fastest-growing branches in spectroscopic analysis technology, which is also a research hotspot for scholars all around the world. This book mainly discusses the chemometric methods used for spectral analysis, including spectral preprocessing, variable selection, data dimensionality reduction, linear or nonlinear multivariate calibrations, pattern recognition, calibration sample selection, outlier recognition, model update and maintenance, multi-spectral fusion, model transfer, and deep learning algorithms, etc. Considering the comprehensiveness and systematic reviewing, this book summarizes and reviews the latest research progress in the world, particularly, which are closely combined with scientific researches and practical applications, as well as, many algorithm improvements and strategy extensions. The authors believe this book will provide new aspects and ideas for researchers and users in this field.

1.1 Introduction

Chemometrics was born in the early 1970s. It is usually defined as "Chemometrics is a branch of chemistry, which uses mathematical and statistical methods with computer technology, designs and selects the best measurement procedures and experimental methods, in order to obtain the maximum information by interpreting chemical data". Change with development, the definition of chemometrics has many expressions, but its goal is very clear, that is, to extract the most useful information from the measured data. Kant once said "Among the branches of natural sciences, only those that can be expressed in mathematics are true sciences". The feature of chemometrics is to construct the chemical measurement as a mathematical model that can be expressed

X. Chu et al., *Chemometric Methods in Analytical Spectroscopy Technology*,
https://doi.org/10.1007/978-981-19-1625-0_1

through mathematical formula. Different from other branches of theoretical mathematic, chemometrics is a discipline of all the theories and methods based on the chemical experimental data [1–3].

Spectral analysis technology, including molecular spectroscopy and atomic spectroscopy, such as mid-infrared, ultraviolet-visible, molecular fluorescence, Raman, terahertz, laser-induced breakdown, nuclear magnetic resonance, etc., has the advantages of simple sample processing, non-destructive, fast and real-time monitoring, and on-site online analysis [4]. With regard to the quantitative and qualitative analyses of complex samples (such as petroleum, grain, traditional Chinese medicine, tobacco, food, soil, etc.), traditional experimental methods cannot extract very useful information from spectra with serious matrix effects and obtain quantitative or qualitative results. The popularity of computers and the rise of chemometrics have brought lots of new ideas and methods to the development of spectroscopic analysis, because the significant contribution is to awaken the sleeping "analytical giant" of near-infrared spectroscopy (NIR) technology [5, 6]. Subsequently, chemometrics was gradually combined with other spectroscopies like LIBS, which greatly improves the accuracy and robustness of spectral analysis. Nowadays, chemometrics has become a common method for spectrum discrimination and simultaneous determination of multiple components in the complex systems, and also become an important part of the interdisciplinary process analytical technology (PAT) [7].

This book mainly introduces the chemometric methods commonly used in modern spectroscopic analysis, calibration strategies, and their latest developments.

1.1.1 Overview of Chemometrics

1.1.1.1 Origin, Definition, and Development History

Chemometrics was born in the early 1970s. In 1971, when the Swedish chemist S. Wold was naming a fund project from three concepts as **chemical data analysis**, **computer in chemistry**, and **chemometrics**, and finally he chose the last one, the moment from then on it was officially announced the birth of the emerging discipline of chemometrics. Three years later, he and Professor Kowalski of the University of Washington established the International Chemometrics Society (ICS) in Seattle, USA. In fact, the early chemometrics were mostly from the classical statistical methods. For example, the concept of principal component analysis (PCA) was proposed by British statistician K. Pearson early as in 1901, and was later developed and popularized by American statistician H. Hotelling in 1933. Till 1972, PCA was then used for deconvolution of chromatographic overlapping peaks. The famous partial least squares (PLS) was proposed by H. Wold, an econometric statistician in Sweden, for processing economic data in the 1960s. Later, his son S. Wold developed it in 1983 to solve the difficult chemical data regression problem, and obtained very satisfactory results. Currently, PLS algorithm has become a standard multivariate modeling method. Another example, early as in 1953, Hammond et al., proposed

derivative spectrophotometry, which is now widely used in molecular spectroscopies, to improve spectral resolution and reduce interference.

The flourishing period of chemometrics was in the 1980s. The popularity of computers, drive of industrial interests (pharmaceutical development and process analysis), upgrade of analytical instruments, together made the research of chemometrics reach an unprecedented depth and breadth. In fact, some of the foremost methods now widely used are mostly created or perfected at that time. In general, development of chemometrics can be roughly divided into four stages [8].

(1) Pre-establishment

The characteristic of this stage is the application of mathematical statistics in chemistry, especially analytical chemistry. Analysts discussed the standard deviation, confidence interval, least square regression, and other issues of the analysis results. Organic chemists studied the structure-activity relationship of linear free energy, which can be considered the predecessor of chemical quantitative structure-activity relationship (QSAR). In general, the mathematical methods used by analysts during this period are basically descriptive. However, in other disciplines such as engineering science, psychology and other behavioral sciences, factor analysis, pattern recognition, and other methods have been used for higher-level data processing. In 1920, some economists had tried to introduce methods such as principal component analysis, factor analysis and canonical correlation analysis in mathematics to process massive amounts of information such as economic trends and stock prices. They achieved great success and proposed the Econometrics.

(2) Birth of chemometrics

According to the specific requirements of chemistry, analysts developed and created a series of data processing, classification, prediction, and analysis methods. Chemometrics had become a major branch of analytical chemistry. This development includes two factors. One is the gradual popularization of computers, including the instrumentation of analytical chemistry that can accurately provide chemists with a large amount of reliable data. How to efficiently convert the data of these instruments into useful information naturally became the original drive for developing chemometrics. The second one is that various powerful mathematical methods can be applied in analytical chemistry with the help of faster computing. The rise of chemometrics can be regarded as the main manifestation of modern technological changes in chemistry marked by computer applications.

(3) 1980s

The unique multivariate calibration, multivariate discrimination, and chemical pattern recognition methods, such as partial least squares, soft independent modeling of class analogy (SIMCA), rank annihilation factor analysis, evolving factor analysis, etc., had been greatly developed in theory and algorithm research. During this period, the professional journals as “Journal of Chemometrics” (1987, Wiley) and “Chemometrics and Intelligence Laboratory Systems” (1988, Elsevier) were established, along with many classic chemometrics monographs published. These

publications played an important role in disseminating knowledge of chemometrics, introducing development trends, and guiding scientific research topics. In 1984, American Mathwork Company officially launched MATLAB software, by which many complex mathematical calculations used in chemometrics can be realized with only one coding expression, making it almost a standard programming language for chemometrics research. When a new algorithm was published, usually MATLAB codes were attached, that greatly promoted the development of the discipline.

(4) 1990s

Chemometrics had truly entered the stage of practical applications, such as near-infrared spectroscopy, sensors, medicine and pharmacy, etc. Almost all modern analytical instruments had a computer or microprocessor containing the chemometrics software. Chemometrics was becoming an indispensable tool in the daily work of chemistry or analytical chemistry. Furthermore, series of new methods such as artificial neural networks, wavelet transforms, genetic algorithms, and support vector machines, were employed by analysts, as new tools for solving chemical problems.

1.1.1.2 Content of Chemometrics

Development of chemometrics has provided many new ideas, new approaches, and new concepts for solving problems in all chemical branches such as analytical chemistry, food chemistry, environmental chemistry, medicinal chemistry, organic chemistry, and chemical engineering. Its research content almost covers the entire process of chemical measurement (Fig. 1.1), mainly including the following parts [9, 10].

(1) Sampling theory and method

Sampling is the first step of analysis. The reliability of analytical results is directly related to whether the sampling is correct or rational. The purpose of analysis or

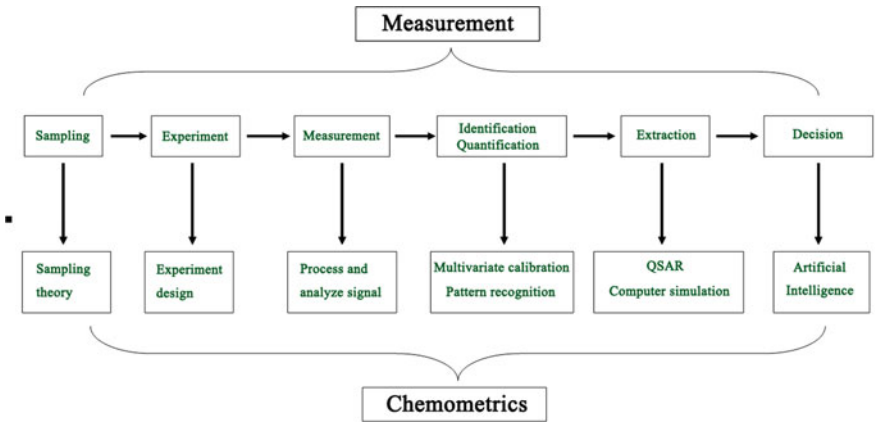

Fig. 1.1 Correspondence between the chemometrics and chemical measurement

testing is to obtain the unbiased information about the entire object based on the data measured from a sectional sample. Sampling refers to the mathematical theory of how to collect samples. Commonly used sampling methods involve heterogeneous solid materials, dynamic processes, and quality inspections.

(2) Experimental design and optimization

Experimental design and optimization need to design and arrange experiments and optimize measurement conditions so as to improve work efficiency. Orthogonal design and simplex optimization method are still the main strategy for experimental optimization. Its purpose is to obtain as much information as possible about the relationship between the target and the factors with the fewest number of trials. Besides, some global optimizations, such as simulated annealing algorithm, genetic algorithm, and particle swarm algorithm, are also being practiced.

(3) Signal processing

Interference signal and noise are often mixed in the analysis signal. By use of signal smoothing, filtering, transformation, peak splitting, curve fitting, derivation, and integration techniques, analysis signals can be reliably distinguished and detected from interference signals, and the signal-to-noise ratio can be improved.

(4) Resolution and calibration

Multivariate resolution and calibration are the core content and also the most distinctive part of chemometrics. Calibration is a mathematical process that extracts useful information from the instrument signal. Its purpose is to establish the relationship between the analysis signal and the concentration for the quantification of the analyte. Multivariate calibration is a method used to improve the selectivity and reliability of analysis that is suitable for a variety of instrument signals, such as spectrum, mass spectrum, and chromatographic data. It correlates the independent variable (measurement information) of the training set with the dependent variable (the property of interest, such as the concentration of an analyte in a complex system or other physical and chemical properties) so as to establish multivariate calibration models. For unknown samples, when the measurement information is obtained, the concentration or property parameters, that used to be measured by laborious, time-consuming, and costly standard methods, can be predicted according to the established model.

Multivariate resolution can extract various response curves of pure substances (spectral curve, pH curve, time curve, elution curve and concentration curve, etc.) from the analysis data of various evolution processes of unknown mixtures without need to know the type and composition of unknown samples in advance. Common multivariate resolution includes self-mode curve resolution (SMCR), evolving factor analysis (EFA), window factor analysis (WFA), heuristic evolving latent projections (HELP), projection rotation factor analysis (PRFA), generalized rank annihilation method (GRAM), Tucker3, parallel factor analysis (PARAFAC), alternating trilinear decomposition (ALTD), and so on. It can solve problems that trouble traditional analytical chemistry, such as the analysis of complex multi-component equilibrium and kinetic systems, the detection of peak purity of complex systems in

chromatography and its hyphenated methods, and the resolution of overlapping peaks.

(5) Pattern recognition

Chemical pattern recognition is to select the characteristics of samples, find the rules of classification, and then classify and identify unknown sample sets according to the rules of classification. If sample is known, then classified; if unknown, the classification depends entirely on the natural characteristics of the sample. Chemical pattern recognition can be used to interpret spectral data, study structure-activity relationships, classify drugs, determine pollution sources, diagnose early stage of cancer, and identify authentic products, etc. It provides very useful information for decision-making and process optimization.

(6) Computer simulation

Simulation is an important means of using computer to study chemical reactions, measuring methods, and analyzing data. Monte Carlo simulation is one of the most commonly used simulation methods.

(7) Quantitative structure-activity relationships

Quantitative structure-activity relationship (QSAR) uses multivariate calibration and pattern recognition methods to find out the quantitative relationship between structure, properties, and biological activity from a series of compounds with the already known activities, then predict the activity of new compounds, and guide the design of new compounds.

(8) Chemical database and library searching

With the daily increase of spectrum data, various databases appeared, such as compound structure databases, various spectrum databases, physical property databases, etc. The rapid retrieval and effective use of data have become an important research content of computer processing information.

(9) Artificial intelligence and chemical expert system

The chemical expert system is an intelligent computer program system that applies chemical knowledge and logical reasoning to solve chemical problems. It covers molecular structure analysis, selection of the best measurement, and separation conditions for various instruments (chromatography, spectroscopy, etc.), etc.

Almost all of the above chemometrics contents are involved in the spectroscopic analysis, but actually, they have their own key points and particularities. In addition, there are also new focus on calibration transfer, outlier sample identification, and model evaluation methods. The chemometric methods applied to modern spectroscopic analysis mainly include the following five aspects [11–13].

(a) Spectral preprocessing and variable selection methods, such as derivative, Fourier transform, wavelet transform, genetic algorithm, etc., weaken or eliminate the influence of various non-target factors on the spectrum, remove irrelevant information variables as possible, improve resolution and sensitivity, and enhance the predictive ability and robustness of the calibration model.

(b) Multivariate calibration methods for establishing quantitative models, such as multiple linear regression (MLR), principal component regression (PCR), partial least squares (PLS), artificial neural network (ANN), and support vector machine regression (SVR), etc. The purpose is to build analytical model for predicting the physical properties or chemical compositions of unknown samples.
(c) Pattern recognition methods and outlier detection methods, such as minimum distance discrimination method, SIMCA and KNN method for recognition, as well as spectral residual root mean square method and nearest neighbor distance method for outlier detection, etc. The purpose is to cluster or identify different types of samples, and to determine whether the sample to be tested is within the coverage of the quantitative model, and to ensure the accuracy of the prediction results.
(d) For signals obtained by hyphenated analysis methods (excitation-emission three-dimensional fluorescence spectroscopy) or spectral imaging (near-infrared, infrared and Raman imaging, etc.), multidimensional resolution, and calibration methods, such as Tucker3, PARAFAC, ATLD, and multi-way PLS methods, can distinguish the response signals of multiple analytes with similar properties at the same time, and directly quantitatively determine the analyte components of interest in the presence of unknown interferences.
(e) calibration transfer methods, such as direct standardization (DS), piecewise direct standardization (PDS), and Shenk's algorithm, etc., reliably transfer the qualitative or quantitative calibration model established on one instrument to other identical or similar instruments, or use the model established under a certain condition for the spectra collected by the same instrument under another conditions, thereby reducing the time and cost required for calibration.

1.1.1.3 Necessity of Chemometrics

Application of chemometrics to the quantitative and qualitative analyses of spectroscopy in many cases makes the analysis result a significant level-up. Its functions can be summarized into the following aspects.

(1) Multivariate calibration, as shown in Fig. 1.2, can improve the accuracy and precision of analysis. Factor analysis methods such as principal component regression and partial least squares can not only make use of the full spectrum but also significantly reduce the interference of coexisting components and background. The concentration of multiple components can be directly determined without chemical separation.

The basis of spectral quantification is the Lambert-Beer law. The linear relationship is based on the assumption of monochromatic light and dilute solution, without considering the interaction between light-absorbing molecules and the neighboring molecules. In practice, the relationship between absorbance and concentration of actual samples, especially natural complex (agro-products, petroleum, etc.) is usually

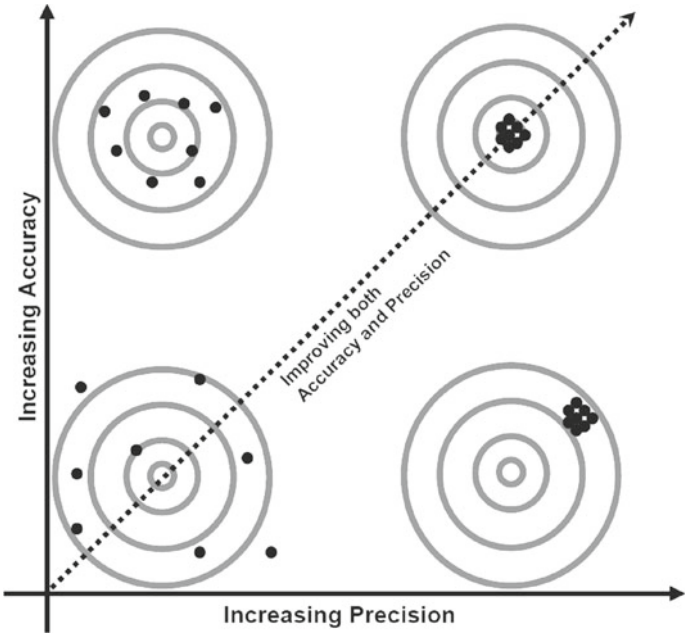

Fig. 1.2 Scheme to improve the accuracy and repeatability of analytical testing

not a simple linear relationship. The traditional single-wavelength calibration curve method can no longer generate satisfactory result. Take determination of fat content in meat using near-infrared spectroscopy, for example, only the absorbance at 940 nm (the characteristic absorption band of methylene third overtone) cannot establish an accurate calibration curve (as shown in Fig. 1.3), with the correlation coefficient R of only 0.23. Instead, the short-wave near-infrared spectrum (850–1050 nm) combined with PLS is used to establish a multivariate calibration model, a far more accurate prediction results can be obtained (as shown in Fig. 1.4), with R of 0.97 at the same concentration range [14].

(2) Signal processing technology can improve the S/N ratio of the instrument, increase sensitivity, eliminate interference, extract useful information hidden in the spectrum, separate overlapping peaks, and improve resolution of the spectrum. For example, methods as Fourier and wavelet transform can smooth, de-noise, and compress the spectrum, reliably distinguish and detect useful signals from the interferences, providing high-quality characteristic variables for multivariate calibration.

Figure 1.5 shows the Raman spectra of the same mineral from different origins in the international RRUFF mineral database. Due to the interference of fluorescence, the spectra vary in great difference. However, after the baseline correction by the asymmetric least squares, the Raman spectra of the same mineral have good similarity (Fig. 1.6) [15].

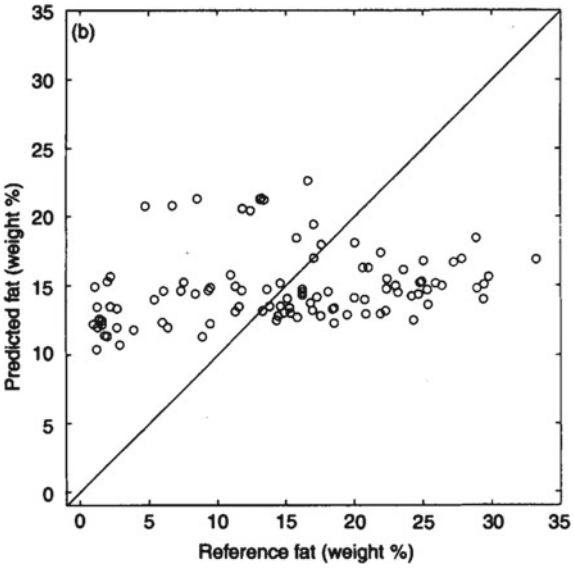

Fig. 1.3 Unary linear regression results of absorbance at 940 nm

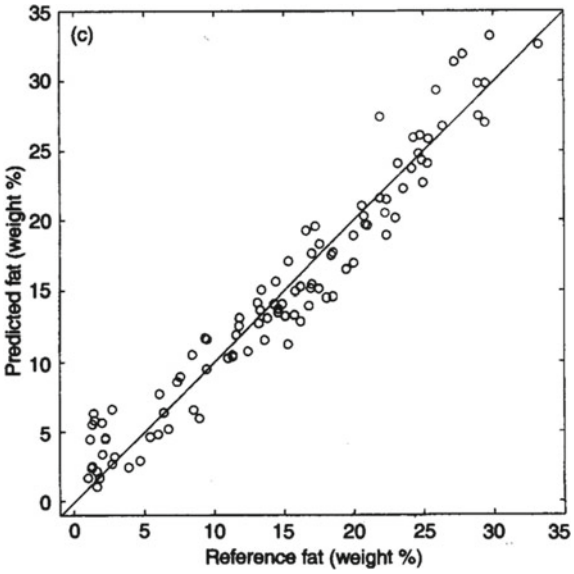

Fig. 1.4 The calibration result of the shortwave NIR full spectrum-PLS method

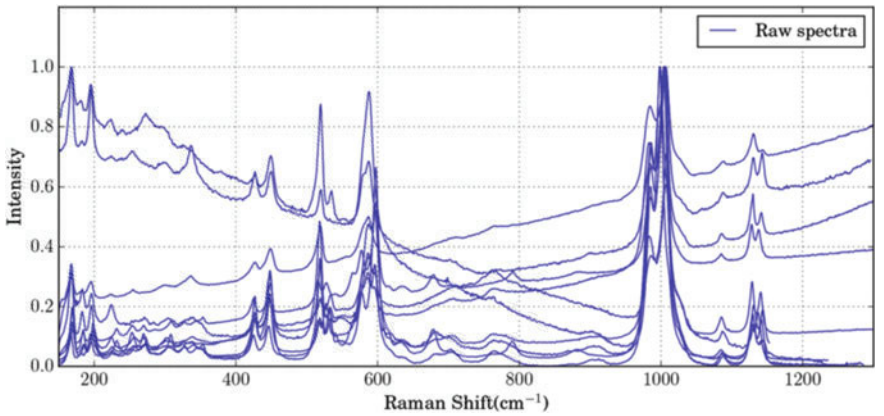

Fig. 1.5 Ten original Raman spectra of the same mineral from different origins in the RRUFF mineral database

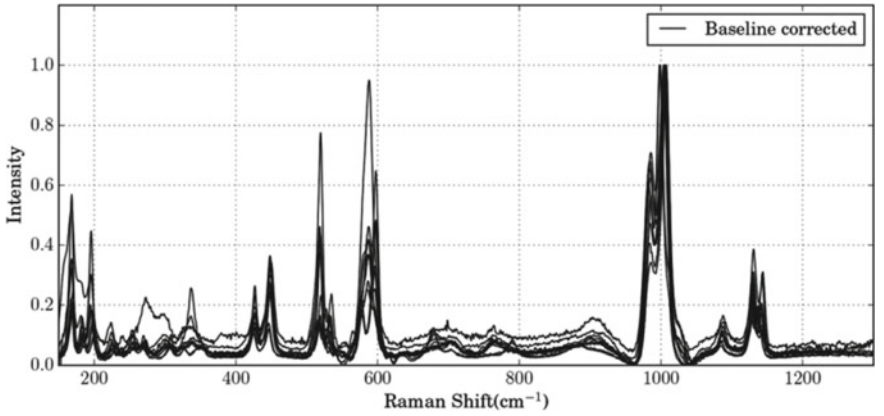

Fig. 1.6 Spectra of Fig. 1.5 after baseline correction

Figure 1.7 is the original spectrum of NIR diffuse reflectance spectra of flour. Affected by particle size and sample heterogeneity, the baseline drift is serious, making the spectral change not related to its composition concentration linearly. After the second derivative preprocessing, it can be seen that not only the baseline drift has been corrected, but also many characteristic peaks have been extracted in Fig. 1.8.

Figure 1.9a is the original spectra of the detection point in the landing area acquired by the China Yutu 2 patrol rover reaching the surface of the moon's back. Figure 1.9b is the spectra after processing by the continuous removal method (envelope removal method). It can be seen that this method effectively enhances the reflection characteristics of the spectral curve and provides the possibility for further analysis of the chemical composition of the lunar mantle [16].

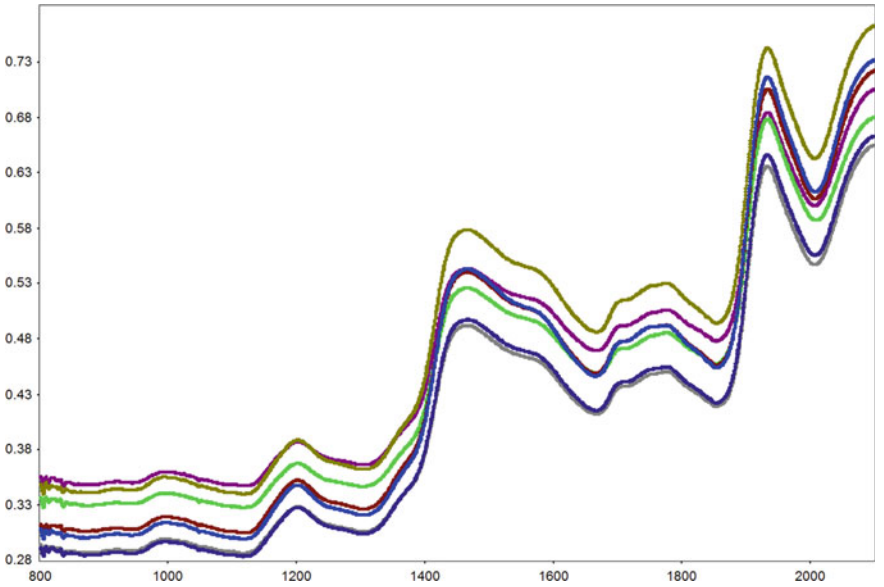

Fig. 1.7 Diffuse reflectance NIRS of different flour samples

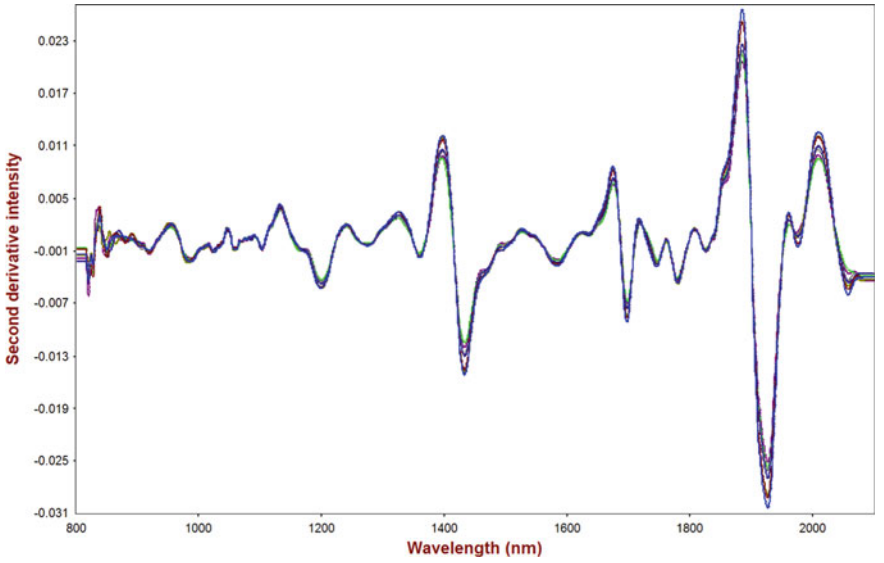

Fig. 1.8 Spectra of Fig. 1.7 after the second derivative processing

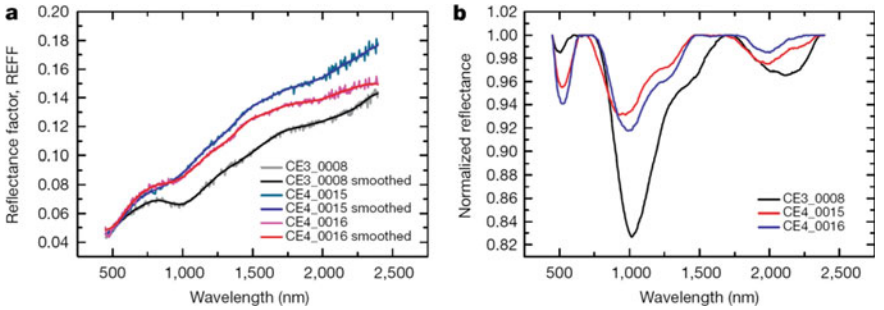

Fig. 1.9 a Diffuse reflectance near-infrared spectra of minerals on the lunar surface; Fig. 1.9 b Spectra processed by continuum removal method

(3) Pattern recognition can make spectral analysis no longer a mere provider of analytical data, but a provider of chemical information as well as a direct participant and solver of chemical issues. For example, spectra with pattern recognition methods can accurately identify authentic products such as drugs, food, and cosmetics, as well, can diagnose early stage of cancer, identify sources of oil spills.

Figure 1.10 is the MIR spectra of the root-end substances of bird feathers from different genders, in which spectra of males and females cannot be identified by the traditional characteristic peak method, because they all reflect the functional groups in proteins, nucleic acids, phospholipids, carbohydrates, and ribose. But, after extracting the scores of the first and third principal components (Fig. 1.11), the gender of the bird can be clearly distinguished by PCA.

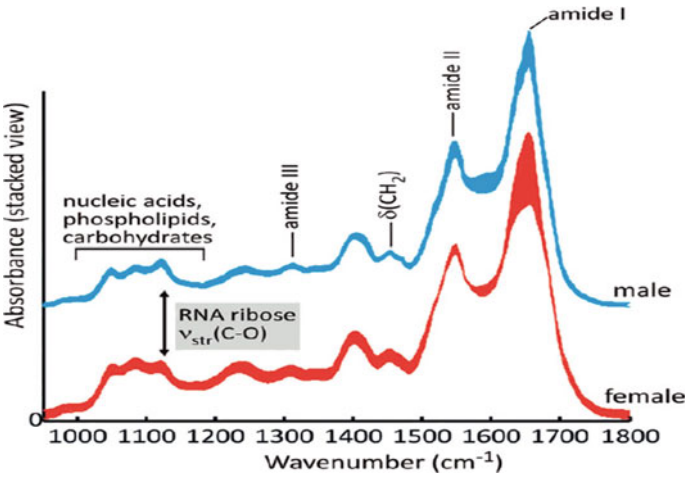

Fig. 1.10 Mid-infrared spectra of the root-tip material of different gender bird feathers

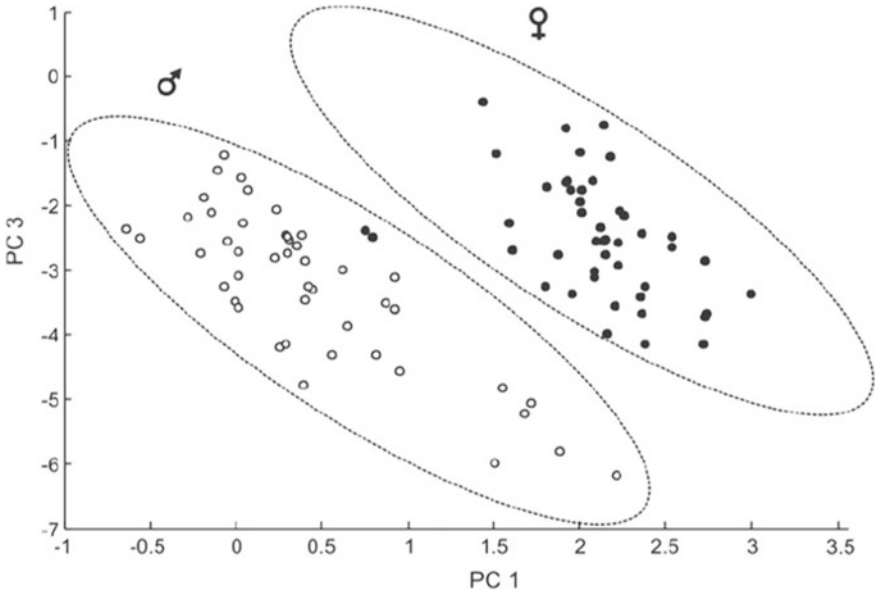

Fig. 1.11 The first and third principal component diagrams after principal component analysis

In the industrial process, some control variables are often related to each other. Separate statistics on these variables often lead to situations where abnormal conditions are not easily confirmed. As shown in Fig. 1.12, the individual temperature and pH variables in each production process are both within the controllable range, but it is easy to identify abnormal points by multivariate statistical methods.

1.1.1.4 Attention in Application

Chemometrics is the application of statistics, mathematics, and computer technology in chemistry. Namely, chemistry is the basis of all the applications, and any those out of chemistry is unreliable. When using chemometrics, a deep understanding and mastery of the field involved in the problem or relevant chemical background should be possessed first. For instance, to use NIRS in the analysis of petrochemical products, it is so much necessary to master certain conventional analytical techniques of petrochemical products and the basic principles of NIRS, then it is possible to establish a reasonable model by chemometrics. Otherwise, a very dangerous result would inevitably arise.

Therefore, modern process analytical technology with chemometrics and spectroscopy is considered to be a highly intersecting comprehensive discipline and also a complete system integrating cutting-edge science and novel technology. It includes engineering technology disciplines with analytical instruments, optics, and electronic

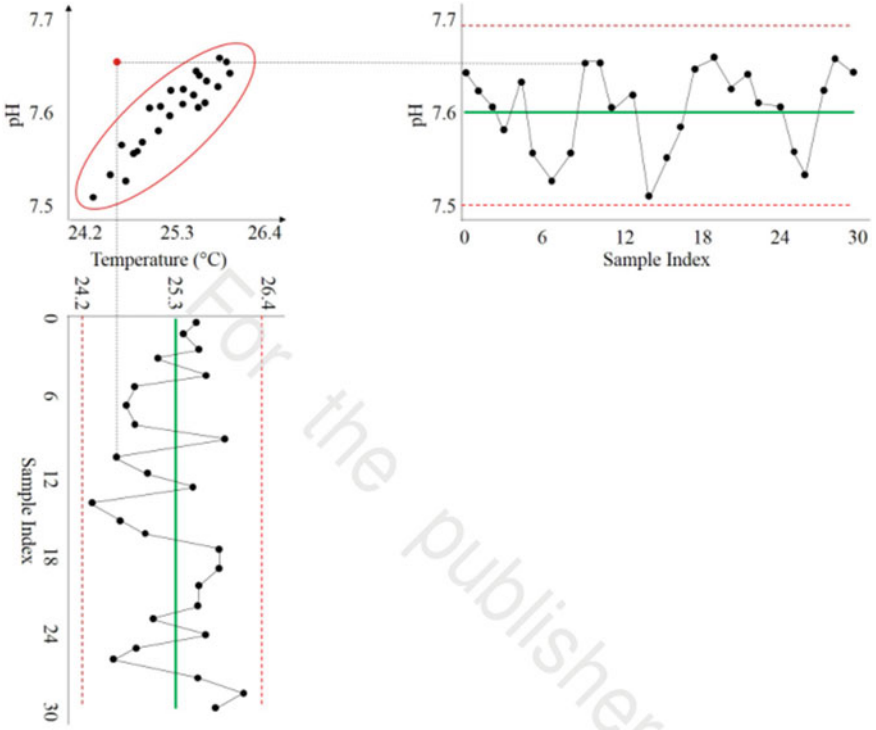

Fig. 1.12 Single and multi-variable control chart for judging abnormal points in the production process

engineering, and also applied basic disciplines with petrochemistry, food chemistry, medicinal chemistry, and soil chemistry, etc.

When dealing with practical problems, it is necessary to choose the appropriate chemometric method according to the specific case, instead of using the latest or the most complicated method. In fact, some basic chemometric concepts can address many application problems [17]. Using the simplest method to obtain satisfactory results is an important principle need to follow when choosing chemometric methods. Of course, this requires proficiency in some basic concepts and algorithm principles of chemometrics.

1.1.2 Analysis of Spectroscopy Combined with Chemometrics

1.1.2.1 Establishment of Calibration Model

In recent years, with the continuous improvement of instrument performance and measurement accessories, the analytical technology of molecular spectroscopy combined with chemometrics is being applied in many fields at an astonishing speed.

As shown in Fig. 1.13, spectroscopy combined with chemometrics methods for analysis mostly use the same mode, that is, a calibration model is established based on a set of known samples, which is called calibration samples or training samples. Based on the spectra of these samples and their corresponding reference data, a calibration or recognition model is established. For the sample to be tested, only its spectrum needs to be measured, and the quantitative or qualitative results based on the established model will be obtained.

The basic steps for building a quantitative calibration model are as follows:

(1) Collection of calibration samples

There are two requirements for calibration samples. One is that the sample should be representative. Its composition should include all the chemical components contained in the sample to be predicted in the future, and its variation range should be greater than that of the corresponding property of the sample to be predicted. Specifically, the variation range is usually greater than five times the reproducibility of the reference method, and it is evenly distributed throughout the range. For example, if the reproducibility of the gasoline octane number determined by the standard method is 0.7 units, then variation range of the calibration sample is at least 3.5 units. The second requirement is that the number should be adequate enough to effectively extract the mathematical relationship between the spectra and the components to be predicted. For a simple test system, at least 60 representative samples are required. For a complex system, at least over one hundred of representative samples are required.

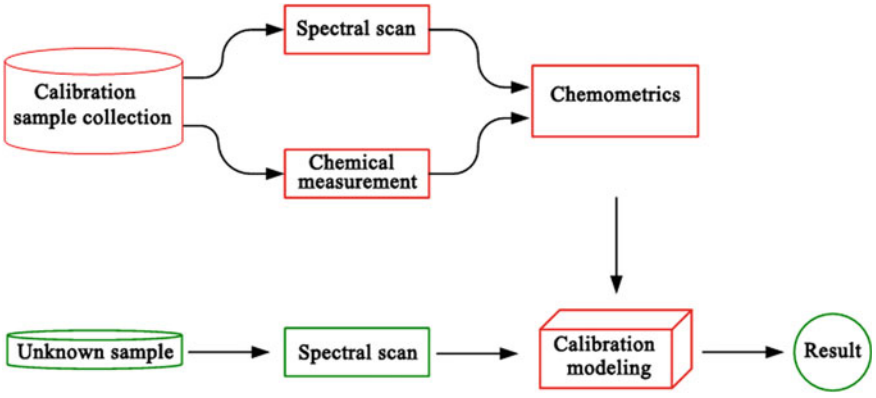

Fig. 1.13 Process of establishment of calibration model and prediction of unknown sample

For collection of natural samples, a variety of influencing factors should be considered. For example, when collecting crop samples, it should include samples of different climates, growing conditions, varieties, textures, and harvest seasons, etc. Online chemical testing should include samples under various process conditions, such as raw materials, temperature, pressure, and catalysts, etc.

(2) Acquisition of spectra

For near-infrared spectroscopy (NIRS), modes of transmission, diffuse reflection, and diffuse transmission can be selected according to the different objects. Even the same diffuse reflection method, there are different measurement accessories like integrating spheres, diffuse reflection probes, etc. Thus, the optimal selection of acquisition conditions and standardized measurement are the core content of spectra collection. The spectra acquisition to be optimized mainly include temperature, optical path, resolution, number of spectral accumulations, and wavelength range, as well as, sample pretreatments such as milling of solid samples, extraction of liquid samples, or fruit slices, etc. In most cases, the samples used for NIRS measurement do not require any pretreatment.

To obtain uniformly measured spectra, standardized collection of spectra is very important, that is, spectral measurement conditions of all samples in the same calibration set should be as consistent as possible. Plus, sampling (such as sample inhomogeneity issues) and loading (such as the density of solid particles, the direction of liquid cuvettes, the orientation of single grains or fruits, etc.) should also be standardized.

(3) Selection of calibration sample

Samples that are analyzed in the laboratory usually have thousands of inspections in a few months, but it is possible that more than 80% of these samples are duplicate samples. So, it is necessary to select the representative samples to establish a calibration model. It can not only increase the speed of modeling but also reduce the storage space of the library. Furthermore, when encountering samples outside the model boundaries, fewer samples can increase the range of application of the model and facilitate model update and maintenance. Plus, the cost will be huge.

PCA is usually performed on the spectra of all calibration samples, and then a certain number of representative samples are selected according to their distribution in the principal component space (PCs), such as the commonly used K-S method. When selecting calibration samples, attention should be paid to the outlier samples. In the spatial distribution of PCs, these outliers are significantly different from others, which may contain other components or the extreme concentrations.

(4) Measurement of reference method

The accuracy of the reference data has a greater impact on the prediction of the quantitative model. Therefore, most of the reference data used in modeling are measured by standard methods or conventional analytical methods. If necessary, the accuracy and repeatability of these conventional methods should to be evaluated. To obtain the high-accuracy reference data, sometimes it is necessary to take the average value